Design Space Exploration of Photonic Interconnects

by

Chen Sun

B.S., University of California, Berkeley (2009)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2011

Author ....
Department of Electrical Engineering and Computer Science
September 2, 2011

Certified by ....
Vladimir Stojanović
Associate Professor
Thesis Supervisor

Accepted by ....
Leslie A. Kolodziejski
Chair, Department Committee on Graduate Students

Design Space Exploration of Photonic Interconnects

by

Chen Sun

Submitted to the Department of Electrical Engineering and Computer Science
on September 2, 2011, in partial fulfillment of the
requirements for the degree of
Master of Science in Electrical Engineering and Computer Science

Abstract

As processors scale deep into the multi-core and many-core regimes, bandwidth and energy-efficiency of the on-die interconnect network have become paramount design issues. Recognizing potential limits of electrical interconnects, emerging nanophotonic integration has been recently proposed as a potential technology option for both on-chip and chip-to-chip applications. As optical links avoid the capacitive, resistive and signal integrity limits imposed upon electrical interconnects, the introduction of integrated photonics allows for efficient realization of physical connectivity that are costly to accomplish electrically. While many recent works have since cited the potential benefits of optics, inherent design tradeoffs of photonic datapath and backend components remain relatively unknown at the system-level.

This thesis develops insights regarding the behavior of electrical and hybrid opto-electrical networks and systems. We present power and area models that capture the behavior of electrical interface circuits and their interactions with optical devices. To animate these models in the context of a full system, we contribute DSENT, a novel physical modeling framework capable of estimating the costs of generalized digital electronics, mixed-signal interface circuitry, and optical links. With DSENT, we enable fast power and area evaluation of entire networks to connect the dynamics of an underlying photonics interconnect to that of an otherwise electrical system. Using our methodolody, we perform a technology-driven design space exploration of intra-chip networks and highlight the importance of thermal tuning and parasitic receiver capacitances in network power consumption. We show that the performance gains enabled by photonics-inspired architectures can enable savings in total system energy even if the network is more costly. Finally, we propose a photonically interconnected DRAM system as a solution to the core-to-DRAM bandwidth bottleneck. By attacking energy consumption at the DRAM channel, chip, and bank level with integrated photoncis, we cut the power consumption of the DRAM system by 10× while remaining area neutral when compared to a projected electrical baseline.

Thesis Supervisor: Vladimir Stojanović
Title: Associate Professor

Acknowledgments

I would first like to acknowledge my thesis supervisor, Professor Vladimir Stojanović, for his mentorship. His guidance and persistence have helped me accomplish much more than what I thought I was originally capable of.

I have had the pleasure to work with many collaborators in my research. I thank Michael Georgas, Jonathan Leu, Benjamin Moss, Jason Orcutt, and Oguzhan Uyar for their help with chip tapeouts as well as their guidance in the photonics project. I gratefully acknowledge the contributions of Chia-Hsin Owen Chen, George Kurian, Jason Miller, Professor Anant Agarwal and Professor Li-Shiuan Peh of MIT CSAIL for their assistance in the co-development of architectural modeling tools. I acknowledge Scott Beamer, Yong-Jin Kwon, Professor Chris Batten, and Professor Krste Asanovic, for their collaborative efforts and for introducing me to the photonics project while I was still at Berkeley. I would also like to thank Professor Ajay Joshi, whose insights helped me greatly when I had just arrived at MIT.

I thank the current and past members of ISG who have provided friendship and helpful discussion. They are, in alphabetical order, Wei An, Fred Chen, Hossein Fariborzi, Michael Georgas, Ajay Joshi, Byungsub Kim, Jonathan Leu, Yan Li, Zhipeng Li, Fabian Lim, Benjamin Moss, Sanquan Song, Ranko Sredojevic, and Oguzhan Uyar. I offer a special thanks to our summer intern, Ana Klimovic, for helping with my research while I worked on this thesis. I acknowledge all my friends both new and old for all the fun we've had.

Finally, I would like to thank my parents and family for everything they have done to make this work possible.

Support

The work in this thesis is funded in part by Intel, DARPA, APIC, and IFC.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

As CMOS technology continues to scale into the deep sub-100 nm regime, improvements in transistor density are mainly being utilized to achieve larger core counts. The clock frequency wars of the late 1990s and early 2000s are no longer sustainable, as the end of Dennard (constant-field) transistor scaling prevents aggressive single-core performance scaling given power density constraints. Improvements in processor performance have since been achieved through processor parallelism. As a result, we have witnessed in recent years a massive explosion in the number of processor cores (Figure 1-1), bringing about the rise of the many-core era.

Increasing core counts, however, place an ever-increasing burden on the on-die interconnection network to support the additional levels of coordination needed between parallel processor cores, memories, and specialized IP blocks both on and off-chip. The primary challenge involves the trade-off between network performance and implementation cost. Modern chip multiprocessors (CMPs) are typically power-constrained, with the interconnect consuming a significant fraction of the system area and power budget [20, 59]. As core counts continue to scale, more complex interconnect topologies are necessary to support future bandwidth and latency requirements, without disproportionately impacting power and area costs.

Currently, the state of electrical interconnects (Figure 1-2) presents a grim prospect. The tradeoff between I/O bandwidth density and energy efficiency, as well as poor scaling of on-chip wires, threatens to push projected processor pin counts and inter-

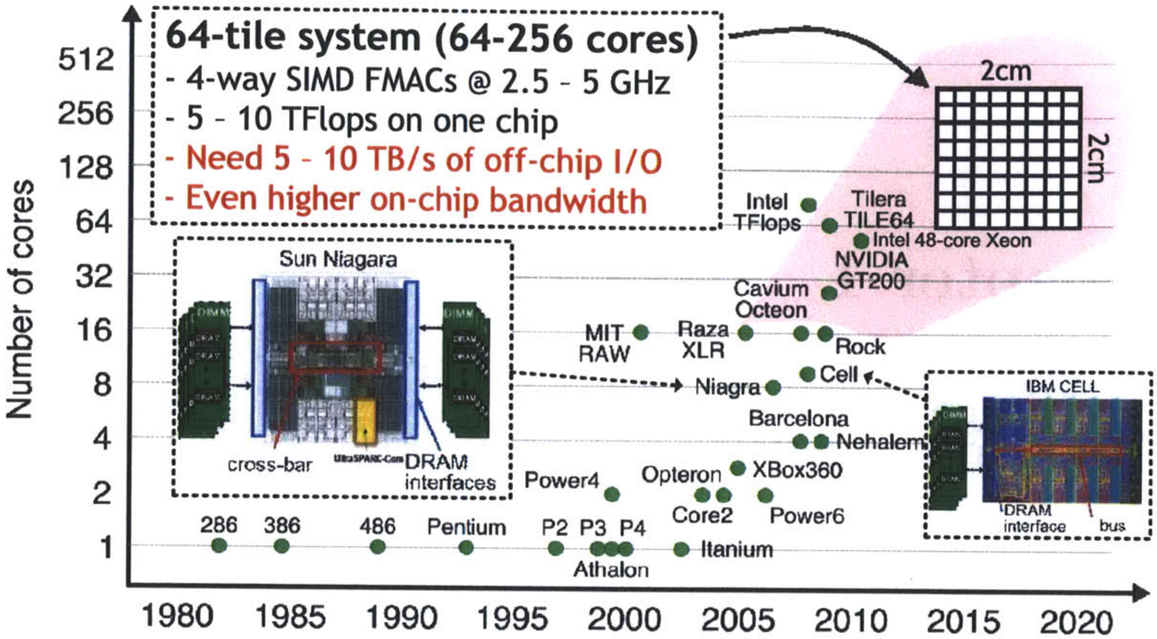
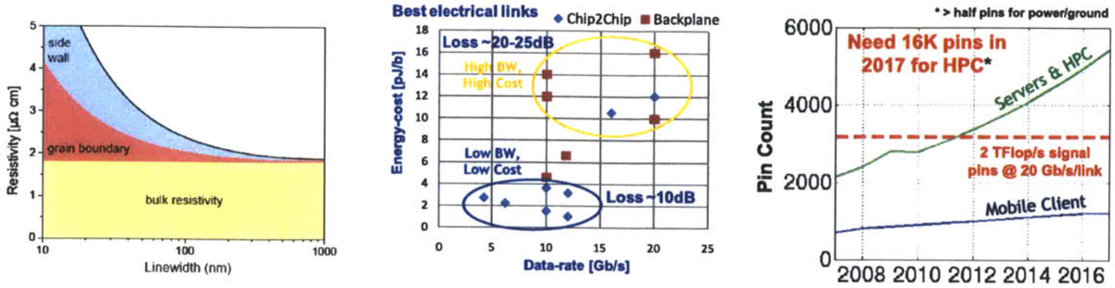


Figure 1-1: Processor core count scaling into the many-core era

connect power consumption past practical limits (Figure 1-3). Without a sufficiently scalable method of improving both bandwidth density and link power consumtpion, many-core processors are doomed to bandwidth starvation.

Silicon photonics is a new disruptive technology that promises to overcome the traditional barriers of electrical interconnects. With direct integration of optical devices with CMOS electronics, the benefits of optical links can be enjoyed at chip-scale granularity. Of its many benefits, unparalleled channel capacity enabled by wavelength-division multiplexing and low energy cost per bit over long distances stand out as key enablers for a denser and more energy efficient communication fabric. As demonstrated in existing works [3, 25, 36, 65, 70], photonics-augmented architectures are projected to satisfy the demands of many-core systems for both on-chip and off-chip I/O. Photonics is currently immature, however, and there is a clear need to capture the inherent tradeoffs of the technology to steer further device development and drive optimization.

We organize the rest of this thesis as follows. In Chapter 2 we provide an overview of photonics technology and explain in detail the challenges faced by on-chip and off-



(a) On-chip wire resistivity (b) I/O efficiency vs data-rate (c) Package pin count scaling

Figure 1-2: The current state of interconnects

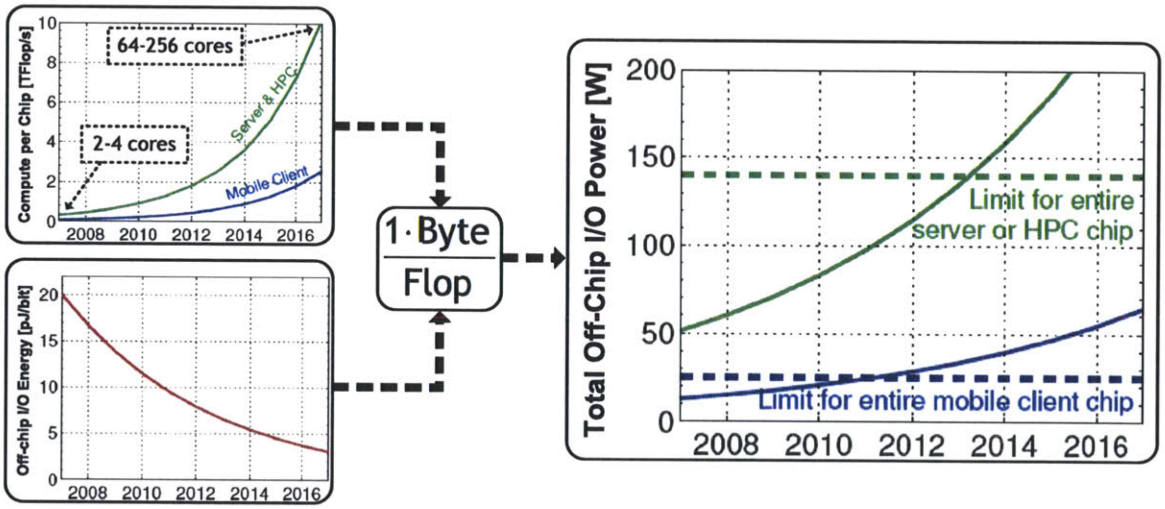


Figure 1-3: I/O power scaling

chip interconnection networks. Chapter 3 discusses our models of an optical link, focusing on the active data-path elements as well as thermal tuning. In Chapter 4, we propose DSENT, a new architectural modeling tool that integrates both photonic models and a generalized electrical modeling framework to enable cost evaluations of full opto-electrical networks. We use our framework in Chapter 5 to perform evaluations of intra-chip networks and compare photonics with electronics. Finally, in Chapter 6, we propose a new photonically-interconnected DRAM memory system to address the looming memory bandwidth bottleneck.

Chapter 2

Background

In this chapter, we present a brief background concerning both optical interconnects and electrical networks. We introduce the basic optical devices and components that enable nanophotonic integration and wavelength-division multiplexed (WDM) optical links. We next outline the challenges faced by networks-on-chip (NoC) and chip-to-chip links, focusing on the challenges faced by present-day electrical solutions and recent work on how integrated optics may be used to remedy the situation.

2.1 Photonic Devices and Interconnects

The concept of using light as a replacement for electrical signaling has been in practice for several decades. Demonstrating superior channel capacities, low channel loss, and immunity to electro-magnetic interference, fiber-optic communication systems have gradually replaced their electrical counterparts beginning with the first fiber-optic Transatlantic cable in 1988 [60]. Technological advances has since enabled optical signaling to gradually find its way into smaller-scale computer systems where its high bandwidth and energy-efficient properties were leveraged in server backplanes and super-computers [61]. Continued technology development and miniaturization, as well as advances in silicon micro-fabrication, have recently enabled the integration of active optical components and nano-scale electronics in the form of silicon photonics, which employs silicon as the primary optical medium [2, 33, 53]. As a result, direct integration of CMOS electronics and silicon photonics is now a distinct possibility. We explore in this section the basics of the technology and the opportunities that arise in the interconnect domain.

2.1.1 Photonic Building Blocks

A key challenge of photonic integration is identifying the set of building blocks that a mostly electronic system can use. To allow the production of these blocks on a commercial scale, they must be compatible with CMOS chip fabrication flows in terms of form factor and performance as well as material selection and device yields.

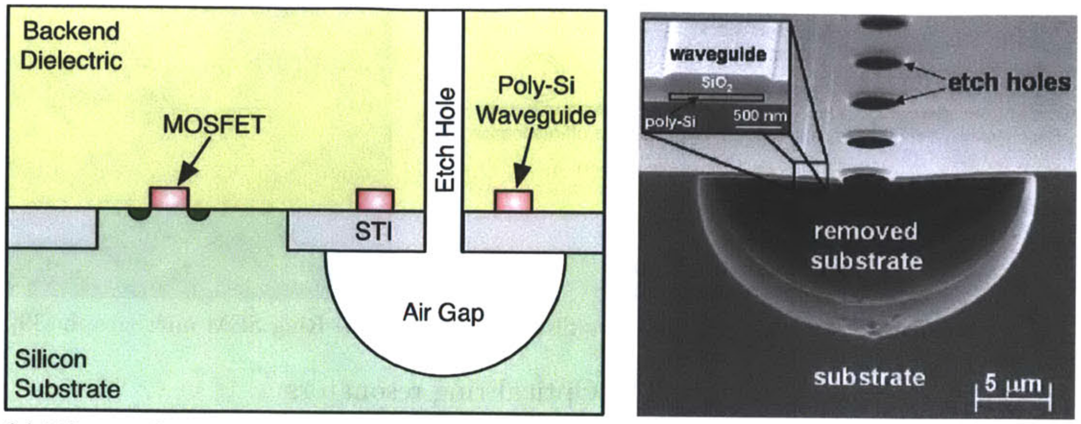
Waveguides and Couplers

Waveguides are the primary means of routing light within the confines of a chip. This is accomplished by surrounding a silicon core, which has a high index of refraction, with cladding material of a lower index of refraction SiO_2 from buried oxides or backend dielectrics [32], deposited polymers [57], and even air gaps [53]. High index contrast between the core and cladding allows for bend radii on the order of a few microns to be achieved with minimal loss.

Vertical grating couplers [63] allow light to be directed both into and out-of the plane of the chip. Light traveling in waveguides on the chip can be coupled (at an angle) into optical fibers to be brought off-chip and off-chip light can be coupled into on-chip waveguides.

Laser

Both on-chip and off-chip lasers are possible options as sources of light for on-chip optics. Leveraging the existing fiber-optic communication infrastructure, off-chip continuous wave lasers are commercially available at reasonably high efficiencies. Integrated on-chip laser sources can also be be found in the form of Ge-based [41] or hybrid silicon/indium phosphide [22] lasers. While on-chip lasers have markedly lower efficiencies due to process constraints and device immaturity, we note that close



(a) Waveguide using air gap for strong confinement

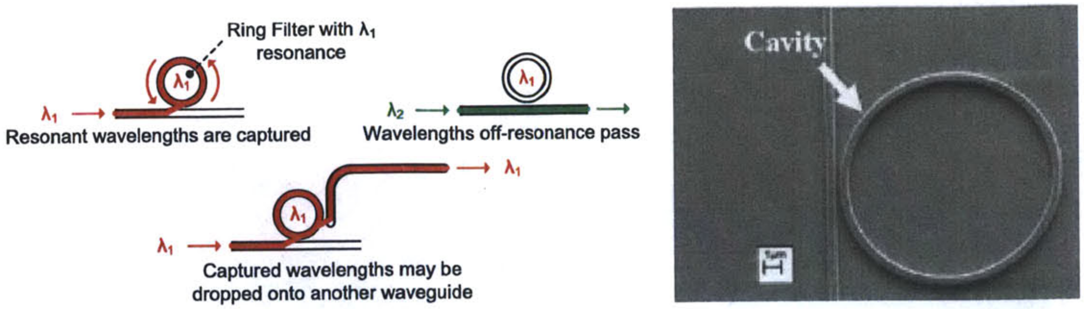
(b) Waveguide SEM Micrograph

Figure 2-1: A waveguide fabricated in a bulk CMOS process using an undercut air-gap to increase cladding thickness [53].

proximity to electronics enables lower laser distribution losses (as they do not have to couple onto the chip in the case of a off-chip laser) and creates opportunities for finer-grained power throttling of the laser.

Ring Resonator

The primary active component is the optical ring resonator. As opposed to the more traditional and cumbersome Mach-Zehnder structures [11, 46], rings rely on a resonant structure to increase the optical length. As a result, they are much smaller than their Mach-Zehnder counterparts [40] – ring radii are typically less than 10 μm– allowing for hundreds of thousands of ring resonators to fit on a single die. When coupled to a waveguide, the resonant properties of the ring allow it to perform as a notch filter; wavelengths not at the resonant wavelengths pass by unaffected while resonant wavelengths are trapped in the ring and be potentially *dropped* onto another waveguide. The resonant wavelength of each ring can be controlled by adjusting the device geometry or the index of refraction. These properties enable rings to perform wavelength-selective filtering, allowing actions to be performed for a particular wavelength without affecting other wavelengths present on the same waveguide.


Ring Filter with λ1 resonance
λ1
λ1
Resonant wavelengths are captured
λ1
λ2
Wavelengths off-resonance pass
λ1
λ1
λ1
Captured wavelengths may be dropped onto another waveguide
Cavity

(a) Rings as wavelength selective filters (b) Ring SEM micrograph [39]

Figure 2-2: Optical ring resonators

Modulators

Electrical modulation of the resonant wavelength is achieved by varying the index of refraction through either *carrier-injection* or *carrier-depletion* mechanisms. Both utilize the free carrier plasma dispersion effect [55] to move a ring's resonance in and out of the laser frequency, achieving on-off keying of that wavelength.

Detectors and Receivers

A photodetector is responsible for converting optical power into electrical current, which can then be sensed by a receiver circuit [15] and resolved to electrical ones and zeros. Both germanium and SiGe alloys (which are already used in modern processes for strain engineering) are potential photodetector materials [32, 53]. As the absorption spectrum of the detector depends on the material itself, however, photodetectors standalone are wideband. A wavelength-selective photodetector requires a ring resonator to drop light of a single wavelength onto the photodetector, while letting others pass.

2.1.2 Photonic Links

An example of an on-chip WDM photonic link is shown in Figure 2-3. Two pairs of senders and receivers, A and B, are able to transmit two independent bitstreams simultaneously using only a single on-chip waveguide. Two wavelengths, λ_1 and λ_2, are provided by an external laser source and output onto a single-mode optical fiber.

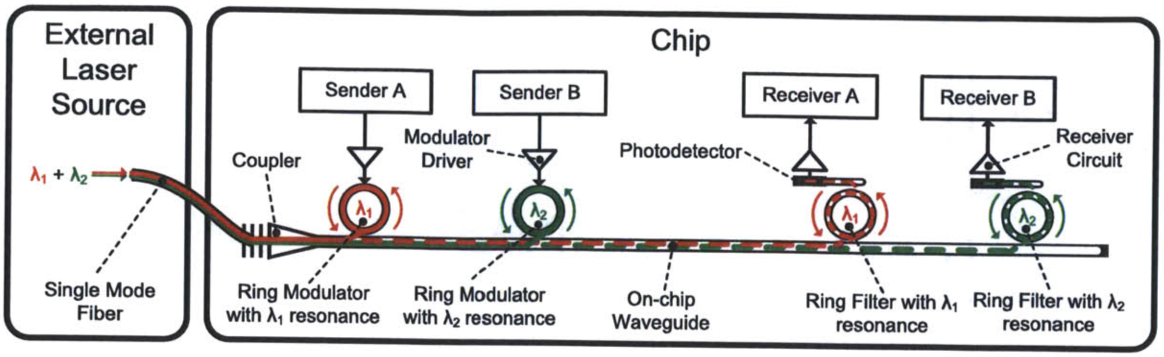


Figure 2-3: A wavelength division multiplexed intra-chip photonic link. *Sender A* modulates data on λ_1 which is received by *Receiver A* while *Sender B* communicates with *Receiver B* via λ_2.

The two wavelengths are then coupled and directed into the plane of the chip through a vertical coupler. The first ring modulator, which is tuned to λ_1, allows *Sender A* to modulate λ_1, encoding its datastream upon it. λ_2 is untouched by this modulator, as it is not at the resonant wavelength. The second modulator, tuned to λ_2, allows *Sender B* to modulate λ_2, without affecting λ_1. On the receive side, λ_1 is caught by a resonant ring filter and dropped onto the photodiode and receiver circuit of *Receiver A*. Likewise, λ_2 is caught and given to *Receiver B*. As such, each wavelength can be treated individually, allowing for many independent bitstreams to share the same optical waveguide in a WDM link.

Note that both endpoints of a photonic link do not necessarily have to lie on the same chip. Light can be coupled seamlessly on and off each chip using vertical couplers, enabling links to span across several chips without the need to rebuffer. However, because light propagates in one direction, one would require a minimum of two wavelengths, one for each direction, to build a bi-directional link. An example of a bidirectional inter-chip link is shown in Figure 2-4. λ_1 travels from left to right and carries information from *Chip A* to *Chip B* while λ_2 travels from right to left and carries information from *Chip B* to *Chip A*.

In general, optical links demonstrate the following physical advantages over their electrical equivalents:

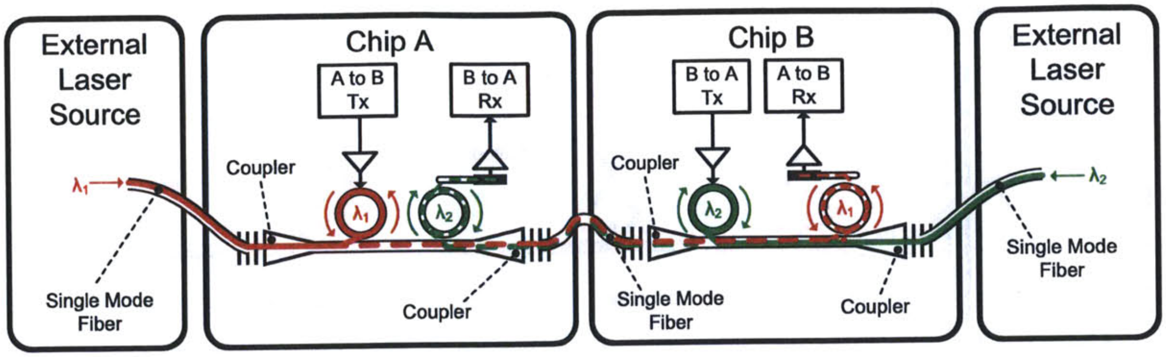


Figure 2-4: A bi-directional wavelength division multiplexed chip-to-chip photonic link. *Chip A* modulates data on λ_1 which is received by *Chip B*. *Chip B* modulates data on λ_2 which is received by *Chip A*.

1. Seamless off-chip links. No need for additional off-chip drivers or buffers, saving area and power.

2. Energy-efficient long-distance links. Eliminates energy spent charging and discharging parasitic wiring capacitances that dominate energy consumption in long on-chip electrical wires. Improved signal integrity when compared to board-level traces such that simpler transceiver circuits may be used.

3. Speed of light signal propagation. Optical fibers and waveguides do not suffer resistive and capacitative delays of on-chip electrical wires.

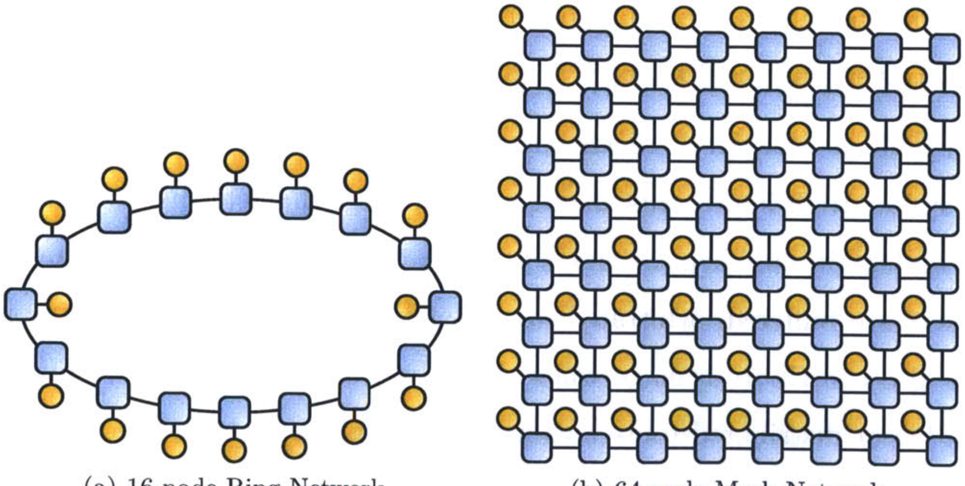
4. High capacity links. Wavelength division multiplexing allows many signals to share the same waveguide or optical fiber, offering superior I/O bandwidth.

2.2 Interconnect Networks

Interconnection networks span both intra-chip and inter-chip domains. Achievable bandwidth, latency, and energy cost per bit send across the network are the key performance and energy efficiency metrics. We examine the challenges faced by modern networks and recently proposed solutions.

2.2.1 Intra-Chip Networks

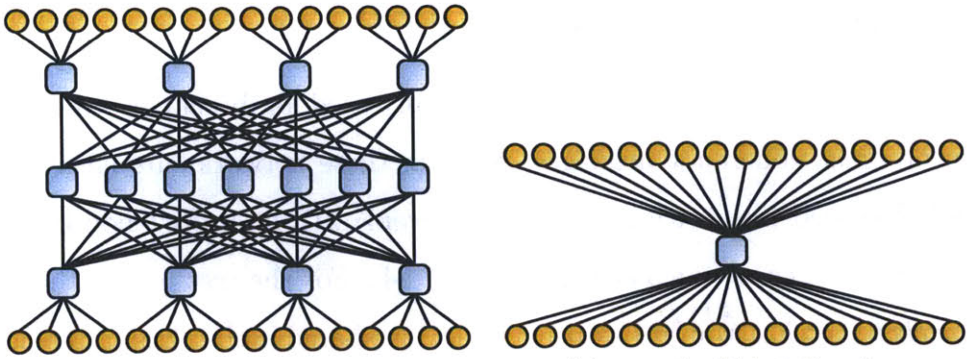
An intra-chip interconnect network is typically used for direct core-to-core communication or cache coherency messages. Data in these networks are either broadcast on a global shared bus, such as the IBM Power4 [21], or routed hop-by-hop using on-chip routers, such as the Tilera processors [64]. For the rest of this thesis, we examine *packet-switched* router-based networks, as shared global bus implementations are not scalable beyond only a handful of cores. In a packet-switched network, data is grouped into packets that are queued, buffered, and routed at each router in the network, resulting in traffic-dependent latency of each packet. This is in contrast to a *circuit-switched* network, which attempts to set up a dedicated link from source to destination before data transmission can begin, incurring a setup/teardown time overhead. As network traffic in multi-core NoC architectures consist mostly of short control messages or cache lines (generally 32–64 Bytes), setup/teardown overheads become significant. As a result, packet-switching has become the general norm for modern NoCs [13].


(a) 16-node Ring Network

(b) 64-node Mesh Network

Figure 2-5: Examples of high-diameter low-radix networks on chip. Squares represent routers, circles represent network endpoints, and lines represent links.

The goal of the intra-chip network is to achieve an all-to-all connectivity between network endpoints, be it cores, memory controllers, or caches. We separate these networks into two broad categories. *High-diameter low-radix* networks, such as the



(a) 16-node Clos Network

(b) 16-node Global Crossbar

Figure 2-6: Examples of low-diameter high-radix networks on chip. Squares represent routers, circles represent network endpoints, and lines represent links.

ring and mesh networks shown in Figure 2-5, require a large average number of hops between endpoints (high-diameter) but with a small degree of switching at each router (low-radix). Alternatively, global crossbars, and clos networks (Figure 2-6) are considered *low-diameter high-radix networks*, characterized by a smaller average number of hops and a larger degree of switching at each router. This inverse relationship between diameter and radix is inherent, as the ability to switch between a larger number of different paths at each router implies a smaller necessary number of router and link traversals (hops) to reach a given destination.

The tradeoff between diameter and radix carries several implications in the NoC design space. High hop counts characteristic of high-diameter networks are undesirable as each additional hop through a router carries both a latency and power penalty as packets are buffered, arbitrated, and switched at the router of each hop. Extra circuits also contribute leakage when the network is inactive. High radix networks contain big or distributed routers as well as long global links to reach the next hop, both of which are difficult to layout and costly to traverse, incurring a larger per-hop energy penalty. To address some of these issues, router bypassing and "fast-lane" techniques [35, 67] have been proposed as ways improve latencies and power consumption of routers. Low-swing [9] and equalized [24] links have also been identified as techniques for reducing link energy costs and delays.

Optical interconnects provide many interesting opportunities in the intra-chip domain. Despite a higher fixed cost at the modulator and receiver endpoints, power consumption of optical links is relatively insensitive to distance. The photonic clos network proposed by Joshi et al. [25] uses this property to efficiently implement the long global links characteristic of clos based topologies. Distributed optical crossbar topologies such as ATAC [36] and Firefly [70] leverage both the inherent delay advantage of photonics and efficiency of global optical links to greatly improve network latencies at a marginal cost overhead. High-diameter optical torus networks, such as Phastlane [10], use a combination of low optical delays and aggressive switch bypassing to eliminate the drawbacks of a high-diameter network.

Photonics technology itself, however, remains immature and there is still yet a great deal of uncertainty on its capabilities. Evaluations of photonic architectures have not yet evolved past the use of fixed energy costs and losses [4, 25, 65, 70], which also vary significantly from study to study. There is a clear need to revisit the initial assumptions of the technology and to develop more detailed photonic device and circuit models to capture the inherent costs and trade-offs of photonic technology.

2.2.2 Inter-Chip Interconnects

Inter-chip or chip-to-chip interconnects are used primarily for communication from processor to off-chip DRAM and other board-based components. In current generation processors, processor-side package pin count constraints force a high data-rate per I/O pin in order to meet bandwidth demands. As core counts increase, off-chip bandwidth must also grow accordingly to support the larger throughput and avoid performance penalties due to bandwidth starvation. Poor projected growth of the number of package pins due to limits on external I/O pitch [23] will cause I/O bandwidth to become a significant bottleneck in the near future.

The I/O bandwidth problem can be best illustrated by the bus-based link between on-die memory controllers and DRAM modules. To overcome package pin limitations, current generation memory interfaces Rambus XDR/XDR2 [58] and buffered memory [14, 44] push I/O data-rates to the electrical extremes in order to achieve high I/O bandwidths given a small number of package pins. Extreme data-rates are not only sub-optimal in terms of energy, but also constrain the capacity of memory channel; each added DRAM socket introduces a transmission line discontinuity in the board-level trace, effectively degrading signal integrity, lowering the achievable data-rate, and producing a conflict between bigger memory capacities and high bandwidths. As such, current electrical techniques do not allow memory channel bandwidth and capacity to grow simultaneuously.

The concept of building chip-stacked processor and DRAM systems connected using arrays of through-silicon-vias (TSVs) [31] has gained momentum in recent years. The idea is to allow for efficient communication between stacked chips using dense arrays of vias running at the core data-rate (the "wide-but-slow" approach). Thermally, however, chip-stacking DRAM and processor is undesirable. The increased density of electronics creates creates high temperatures, accelerating leakage and forcing aggressive DRAM refresh rates. Furthermore, to enable parallel communication between the processor die and a large number of stacked DRAM, DRAM chips could potentially be skewered by hundreds of vias, greatly impacting area efficiency. Yield considerations and TSV aspect ratios also put an inherent limit on the stack height. While still a promising solution to curb off-chip bandwidth requirements, methods to move bits efficiently off-package (or off-stack) are still necessary.

Integrated optical technology presents an opportunity to side-step many of the issues faced by electronics. A single I/O fiber implementing wavelength-division multiplexing can support dozens of independent parallel links, providing per-fiber bandwidth equivalent to that of many electrical pins and achieving far denser inter-chip connectivity. As package pin limitations are avoided, optical links can run at a slower, more energy-efficient data-rate like the aforementioned TSV approach. Batten et al. [3] proposed a global photonics crossbar to connect processors directly to memory, using monolithically integrated photonics to overcome pin bandwidth limitations and seamlessly move bits both on and off-chip at a lower a lower energy/bit than an electrical network. The *Corona* [65] architecture uses TSVs to connect a processor die with a dedicated optical die. The optical die implements both an intra-chip network for on-die communication and an inter-chip network to optically connect to off-chip

DRAM. While photonics I/O can achieve a gain over electrical I/O, further gains are realizable if the DRAM system is redesigned with photonics in mind.

2.3 Contributions of this Thesis

Modeling of Optical Links

While optical losses provide a simple first-order approximation for passive photonic devices, interactions between interface circuitry and active photonic devices are far more complex. In chapter 3, we develop models for nanophotonic links and demonstrate important tradeoffs between datapath and backend elements. We form a ring tuning methology that combats process and temperature variations on resonant ring devices, a signicant hurdle in nanophotonic integration. Finally, we evaluate a complete optical link and observe the tradeoffs leading to an optimal energy point.

Parts of the chapter are to appear in:
[16] M. Georgas, J. Leu, B. Moss, C. Sun, and V. Stojanović. Addressing Link-Level Design Tradeoffs For Integrated Photonic Interconnects. *Custom Integrated Circuits Conference.* Sep 2011.

DSENT – A Design Space Exploration of Networks Tool

Cross-hierarchical power and area estimation is paramount for evaluating the impact of architecture-, circuit- and technology-level changes on a full system. In chapter 4, we present DSENT, a Design Space Exploration of Networks Tool. DSENT provides a framework that is applicable to a wide-range of designs commonly found in modern NoCs, serving as a platform for implemention of both electrical and photonic models. We illustrate the salient features of our approach and validate our results using SPICE-level simulation.

DSENT is developed in collaboration with Chia-Hsin Owen Chen of Prof. Li-Shiuan Peh's group in MIT CSAIL.

Electronics vs. Optics – An Intra-Chip Network Comparison

To motivate the adoption of integrated photonics, optical-electrical systems must demonstrate a clear and consistent advantage over electrical alternatives. Chapter 5 presents technology-driven system evaluations across a representative sample of intra-chip core-to-core networks. We compare photonics with scaled-electrical technology and revisit the ATAC [36] architecture at an advanced process node. We show that to remain competitive with scaled-electrical interconnects, optical interconnects must similarly scale.

We perform the ATAC study in collaboration with George Kurian of Prof. Anant Agarwal's group in MIT CSAIL.

Photonically Interconnected DRAM

Integrated photonics can potentially overcome the package-limited pin bandwidth constraints for a processor to DRAM interconnect. This, however, requires reorganization of the current bus-based approach as well as a redesign of the DRAM part itself to support photonics. In Chapter 6, we explore the idea of a photonically interconnected DRAM system and form recommendations for the optimal level of photonic penetration to achieve high energy efficiency. We improve upon the DRAM models of *CACTI* [7] and use this framework to evaluate our architecture. We show that while optics can net an immediate win through simple one-to-one replacement of electrical I/O transceivers, more significant gains can be achieved by co-designing the DRAM chip architecture to better take advantage of this technology.

Parts of the chapter appears in:
[4] S. Beamer, C. Sun, Y.-J. Kwon, A. Joshi, C. Batten, V. Stojanović, and K. Asanović. Re-architecting DRAM Memory Systems With Monolithically Integrated Silicon Photonics. *Int'l Symp. on Computer Architecture.* Jun 2010.

Chapter 3

Modeling of Optical Links

On-chip nanophotonic integration has been recently recognized as a potential technology option for both on-chip and off-chip interconnects. Though many works have cited the numerous potential advantages of photonic interconnects, such as superior bandwidth density and energy-efficiency [4, 25], inherent design tradeoffs of photonic link and backend components remain relatively unexplored, yet they are critical for bringing out the true energy-efficiency of photonics. In particular, the extent of the costs associated with ring resonance tuning have not been sufficiently quantified. In this chapter, we develop component models linking device, process and circuit parameters to interconnect bandwidth and power consumption. We outline several potential strategies for ring tuning and perform a design exploration of a wavelength-division multiplexed point-to-point link under various aggregate throughput scenarios.

3.1 Data-Path Components of a WDM Link

The two interface circuits responsible for electrical-to-optical and optical-to-electrical conversion are the modulator drivers and receivers. The properties of these circuits affect not only their individual power consumption, but also impact the characteristics of the optical devices they control and hence the laser power. This section illustrates the dynamics of these components and the role they plan in the overall link model.

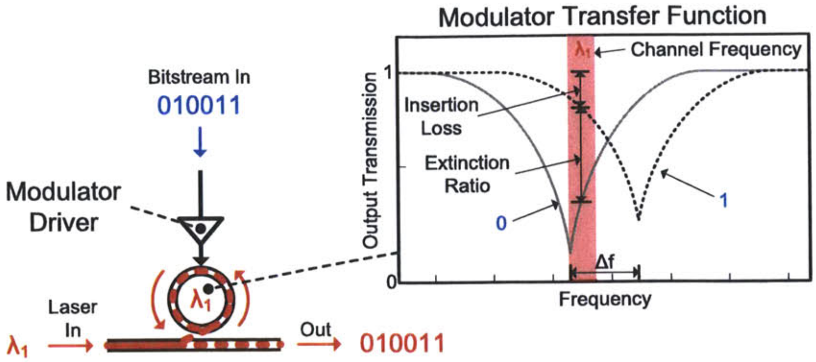


Figure 3-1: Modulation of the channel frequency λ_1 using a carrier depletion modulator

3.1.1 Ring Modulators

Optical ring modulators come in primarily two different flavors: *carrier-injection* and *carrier-depletion*. For both types of modulators, carriers are moved around to influence a material's index of refraction through the free carrier plasma dispersion effect [55]. Typically, carrier-injection modulators suffer from high power consumption and rampant self-heating effects [49] due to carrier recombination in the forward-biased junction. As such, we focus on a reverse-biased carrier-depletion topology, as it is the more promising design for integrated optical links.

A ring modulator works by on-off keying a wavelength of light that nominally matches the ring's resonance. The two key parameters for a modulator are its insertion loss and extinction ratio. The insertion loss is the ratio of input to output light intensity when modulator is outputting an optical *one* and the extinction ratio is the ratio of light intensity between an optical *one* and a *zero*. To construct the modulator model, we extend the device models in [55] to find the necessary amount of charge the modulator driver must move in order to hit the desired insertion loss and extinction ratio (IL and ER):

$$\Delta Q = Q_1 - Q_0 \tag{3.1a}$$

$$= Q_{HWHM} \cdot \left(\sqrt{\frac{1 - T_n \cdot IL}{IL - 1}} - \sqrt{\frac{1 - T_n \cdot ER \cdot IL}{ER \cdot IL - 1}} \right) \tag{3.1b}$$

In this equation, T_n is the transmisivity of the ring at the resonant wavelength (the bottom of the notch). Q_{HWHM} is the amount of charge needed to move a ring's resonance to its half-width half-maximum point and is inversely proportional to the quality factor of the ring. When a high modulator data-rate is required, the ring's quality factor must be sacrificed to support the higher bandwidth. As such, Q_{HWHM} and ΔQ will increase with data-rate. Using equations for a reverse-biased junction, ΔQ can be mapped to a corresponding modulator drive voltage, V_D, required to deplete the necessary amount of charge:

$$\Delta Q = Q(V_D) - Q(0) = \int_0^{V_D} \frac{C_{j0}}{\sqrt{1+\frac{V}{V_{bi}}}} \cdot dV \tag{3.2a}$$

$$V_D = V_{bi} \cdot \left(\left(\frac{\Delta Q}{2 \cdot V_{bi} \cdot C_{j0}} + 1 \right)^2 - 1 \right) \tag{3.2b}$$

where V_{bi} is the built-in junction potential and C_{j0} is the junction capacitance under no external bias. An effective junction capacitance can then be deduced using the charge and the drive voltage, $C_{eff} = \Delta Q(f)/V_D$. A driver circuit can then be built and sized appropriately to drive V_D across C_{eff} at the desired data-rate. The overall energy for the modulator can be expressed as:

$$E_{mod} = E_{driver} + E_{pre-driver} \tag{3.3a}$$

$$= \frac{1}{\gamma} \cdot C_{eff} \cdot V_{DD} \cdot V_D + E_{pre-driver}(f) \tag{3.3b}$$

$$\tag{3.3c}$$

where γ is the efficiency of generating a supply voltage of V_D and $E_{pre-driver}(f)$ is the energy consumed by the pre-driver stages, which is dependent on both the data rate, f, and transistor performance. We note that E_{mod} will grow quickly with the data-rate, as high speeds requires not only a more aggressive modulator design, but also a larger Q_{HWHM} to achieve a bigger resonance shift.

3.1.2 Receivers

An optical receiver is reponsible for sensing the presence-of or lack-of light at a photodetector to convert the optical signal into digital ones and zeros. There are three primary types of receivers: resistive, transimpededance amplifier, and integrating. We focus mainly on an integrating receiver, which consists of a photodetector connected across the input terminals of a sense-amplifier, as it is regarded as the most energy-efficient topology under the conditions present in integrated photonics [15]. A key figure of merit for the receiver is the sensitivity (the laser power required at the connected photodetector) needed reliable resolve each bit. We start with a simplified expression for the required voltage buildup necessary at input terminals of the sense-amp:

$$V_d = v_s + v_{os} + \Phi(BER) \cdot \sqrt{\sum \sigma_n^2} \tag{3.4}$$

which is the sum of the sense-amp minimum latching input swing (v_s), the sense amplifier offset mismatch (v_{os}), and all gaussian noise sources multiplied by the number of standard deviations corresponding to the receiver bit error rate. The required input can then be mapped to the receiver sensitivty, P_{sense}:

$$P_{sense} = \frac{1}{R_{pd}} \cdot \frac{ER}{ER-1} \cdot V_d \cdot C_{par} \cdot f_{data} \tag{3.5}$$

where R_{pd} is the photodetector responsivity (in terms of Amps/Watt), ER is the extinction ratio provided by the modulator, C_{par} is the total parasitic capacitance present at the sense amplifier input node, and f_{data} is the data rate of the receiver. The total laser power required at the laser source is simply the sum of the power needed by each receiver receiving light from the laser, after applying gain needed to overcome optical path losses:

$$P_{laser} = \sum P_{sense,i} \cdot 10^{loss_i/10} \tag{3.6}$$

Where $P_{sense,i}$ is the laser power required at photodetector i and $loss_i$ is the loss to that photodetector, given in dB.

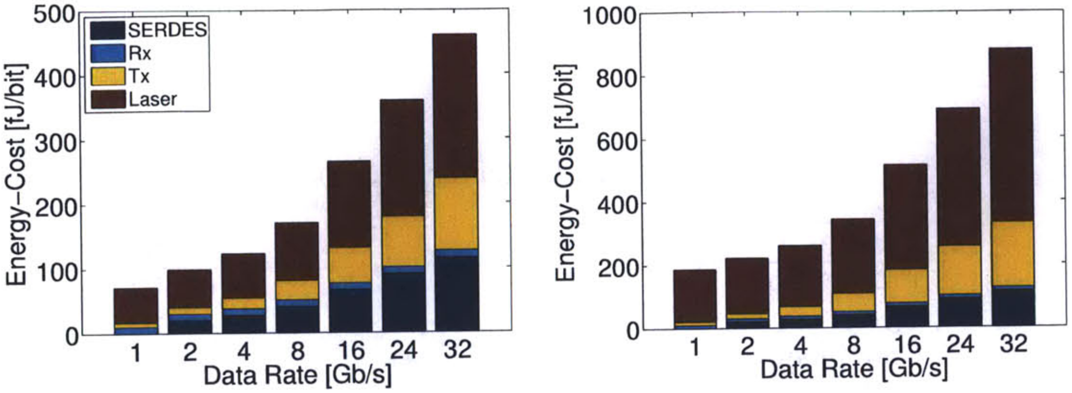
3.1.3 Data-Path Evaluation

Table 3.1: Link Evaluation Parameters

Parameter	Value
Process Node	45 nm SOI
V_{DD}	1.0 V
Core Frequency	1 GHz
SERDES Topology	Mux/Demux Tree
Bit Error Rate (BER)	10^{-15}
Device to Circuit Parasitic Cap C_P	20 fF
Wavelength Band λ_0	1300 nm
Photodiode Responsivity	1.1 A/W
Wall-plug Laser Efficiency P_{laser}/P_{elec}	0.3
Channel Loss	10-15 dB
Insertion Loss IL_{dB} (Optimized)	0.05-5.0 dB
Extinction Ratio ER_{dB} (Optimized)	0.01-10 dB

To illustrate the interactions between the modulator and receiver and the impact on wall-plug laser power, we perform a power optimization across modulator insertion loss, extinction ratio, and receiver topologies for different link data-rates. We note that we assume a core frequency of 1 GHz and thus higher data-rates must pay an associated serialization/deserialization (SERDES) cost. Figure 3-2 shows the energy-per-bit breakdowns for two optical loss scenarios.

In all plots, the laser power is the dominant energy consumer, increasing quickly with data rate as aggressive data-rates force a relaxation of modulator insertion loss and extinction ratio, shown in Figure 3-3. Switching from a 5 dB optical path loss to a 10 dB amplifies not only the laser component, but the modulator component as well. This is a result of the optimal insertion loss and extinction ratio of the modulator adjusting to reflect the greater laser power, as seen when comparing Figure 3-3a with Figures 3-3b. We note that when considering only data-path elements, it is strictly favorable to run at low data-rates. This trend does not hold true, however, once ring tuning is considered.



(a) Loss=5 dB, C_{par}=20 fF

(b) Loss=10 dB, C_{par}=20 fF

Figure 3-2: Energy per bit breakdowns as data-rate changes for the link data-path elements. Losses of 5 dB and 10 dB represent the intra-chip and inter-chip link scenarios, respectively.

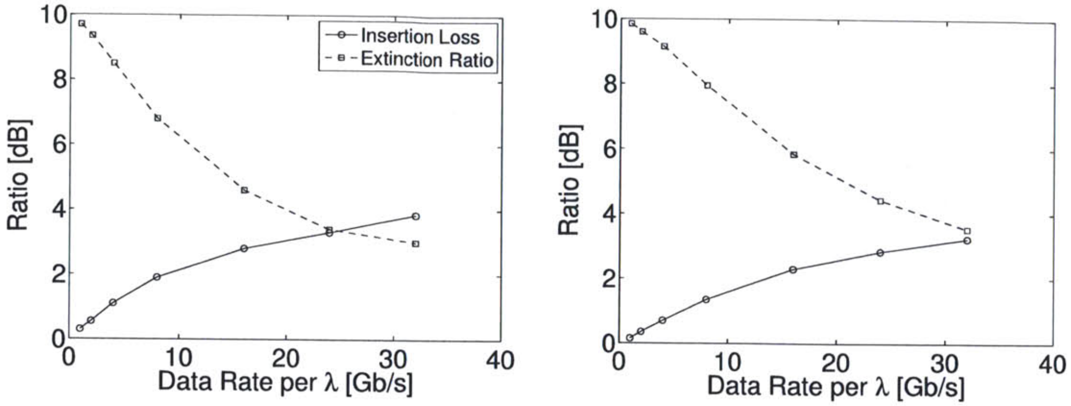
3.2 Ring Resonance Tuning

An integrated WDM link relies primarily upon optical ring resonators to perform channel selection using the ring's resonant frequency. Sensitivity of a ring's resonant frequency to the ring dimensions and the index of refraction leaves them particularly vulnerable to process- and temperature-induced resonant frequency mismatches [53, 62].

3.2.1 Process and Temperature Mismatches

For rings built with gate poly-silicon on commercial CMOS bulk processes, dimensional process variations can result in resonance mismatches of up to 90 GHz and absolute die-to-die mismatches of 600 GHz or more [53]. Similarly, mismatches with standard deviations in the range of 20-70 GHz for same-die and 150-220 GHz for die-to-die have been observed for rings built with SOI [62]. As local- and systematic-level process mismatches differ greatly in magnitude and tuning implications, we model them using σ_{rL} and σ_{rS}, corresponding to the standard deviations characteristic of local ring-to-ring and global systematic mismatches, respectively.

A strong thermal dependence in the index of refraction of silicon causes ring



(a) Loss=5 dB, C_{par}=20 fF

(b) Loss=10 dB, C_{par}=20 fF

Figure 3-3: Optimal modulator insertion loss and extinction ratio across a range of data-rates. At higher data-rates, the optimal insertion loss increases and the extinction ratio decreases as it becomes progressively more costly maintain favorable ratios.

resonances to drift with temperature. $\frac{\Delta f}{\Delta T}$ in the range of -10 GHz/K have been observed [53] [47], implying that a shift of several hundred GHz can be expected in a hostile thermal environment, such as that of a high-performance processor. Unlike static process variations, however, thermal fluctuations are time-dependent, requiring active tuning to stabilize ring resonances. We note, however, the temperature of rings in the same filter bank looks relatively uniform, and as such, contributes to the systematic mismatch.

3.2.2 Methods for Resonance Tuning

To mitigate mismatch introduced by process and temperature, we present several strategies for ring tuning. As an example, we consider the problem of tuning a set of receive-side rings to a set of WDM channel frequencies placed at fixed frequency intervals (Figure 3-4). Given that ring resonances repeat (with separation between peaks defined as the ring's free spectral range, or FSR), we require that all channel frequencies fit within one FSR. The effect of local and systematic variations on this setup is shown in Figure 3-5; For sub-channel wavelength alignment, resistive heaters are fabricated alongside each ring for thermal control, and are driven by a relatively low-bandwidth, receive-data driven control loop.

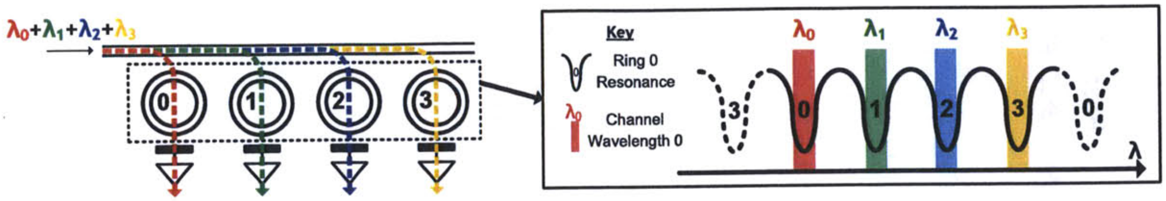
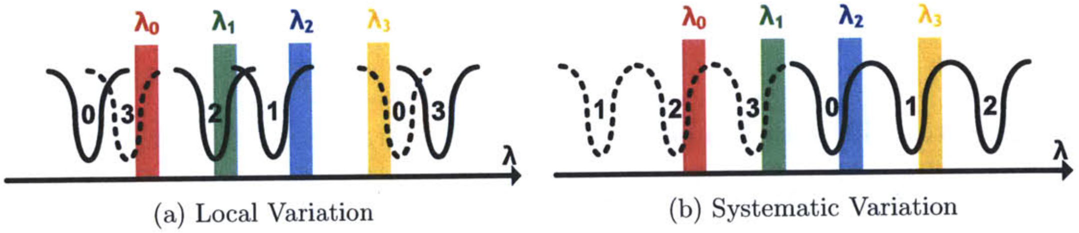


Figure 3-4: A tuning diagram representing a set of 4 perfectly aligned rings. Different WDM Channels are depicted by color and constitute the 4 vertical bars. There are only 4 rings present in this picture and each ring's resonance repeats every FSR. In the tuning diagram, these extra resonances are depicted as dashed curves.



(a) Local Variation

(b) Systematic Variation

Figure 3-5: Tuning diagrams illustrating the effect of local and systematic variations on ring resonances and channel wavelength alignment

We classify ring tuning strategies into four broad categories. The first is *full thermal* tuning, where each ring is assigned a fixed channel wavelength and thus must always tune to that absolute wavelength. Though large resonance shifts are achievable, this straightforward approach comes at a steep power cost. In particular, the inability to cool rings (to blue-shift resonances) below the ambient chip temperature implies that a large blue-shift bias must be applied to the rings at fabrication time such that the resonance can always be red-shifted to the desired channel wavelength across all operating temperatures and process corners. Given the high uncertainties in absolute resonances due to process variations, this bias may need to be as large as 1 THz (6 nm wavelength bias at 1300 nm), requiring temperature heatups of 100 K or more. Though post-process steps such as undercut [53] can imrpove thermal isolation (increasing heating efficiency), additional problems such as thermal cross-talk amongst nearby rings make such large heatups impractical and prohibitively expensive.

The second category is *full athermalization* [57]. In the context of no process variations, full athermalization requires no active tuning at all and thus represents an ideal tuning scenario. With process variations, however, post-process UV trimming on a ring-by-ring basis is an almost absolute requirement, as athermal rings have lost the means with which process variations can be tuned out. Nevertheless, a trimmed design requires no tuning power while in operation.

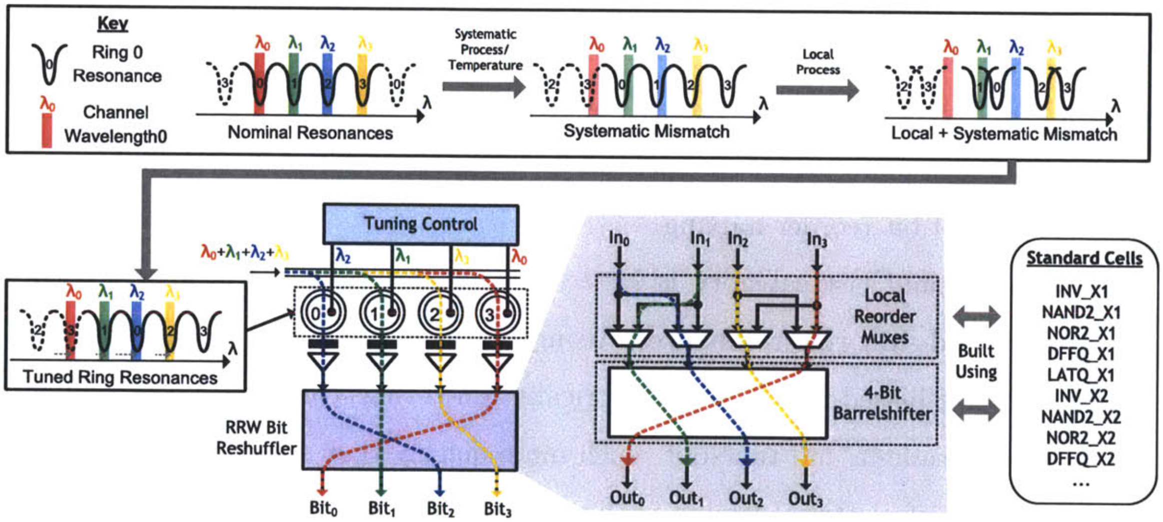


Figure 3-6: Electrical bit reshuffling backend implementing a revolving ring window for a set of 4 receive-side rings. The backend allows for a degree of flexibility in ring channel frequency assignment, such that ring heating power can be reduced. Though not shown, the same technique can be used for transmit-side rings.

The third category is *windowed* tuning, where rings are afforded a "window" of channel frequencies that they are allowed to tune to. Strategies that fall under this category rely upon an electrical backend to reposition bits back to their correct positions. Extending upon the *sliding* ring window method used by Nitta et al. [49], which only works across a limited range of temperatures, we exploit ring resonance repetition to form a *revolving* ring window (RRW), relying instead on the next resonance of the ring to avoid the need to blue-shift 3-6. This strategy comes coupled with a two-stage electrical bit-reshuffling backend. The first stage consists of a set of muxes that perform bit-reordering in the event that local variations shift a ring's resonance beyond that of its neighbors with larger local variations implying a larger degree of muxing. The second stage is a barrel-shifter, which compensates for systematic process and temperature shifts common to all rings in the filter bank by applying a global shift in bit positions. When desired channel frequencies are spaced evenly within one resonance repetition, the necessary tuning distance of each ring becomes proportional to the channel separation, decoupling tuning power from the severity of systematic temperature or process variations. The electrical backend also allows the option to place extra rings/receivers to create a degree of freedom for which rings to actually use. With all rings spaced evenly across the FSR, extra rings reduce the mean tuning distance and allow for further tuning power reduction. This must be balanced with area costs and a more expensive backend as it requires a larger barrel-shifter and a higher degree of bit-reorder muxing.

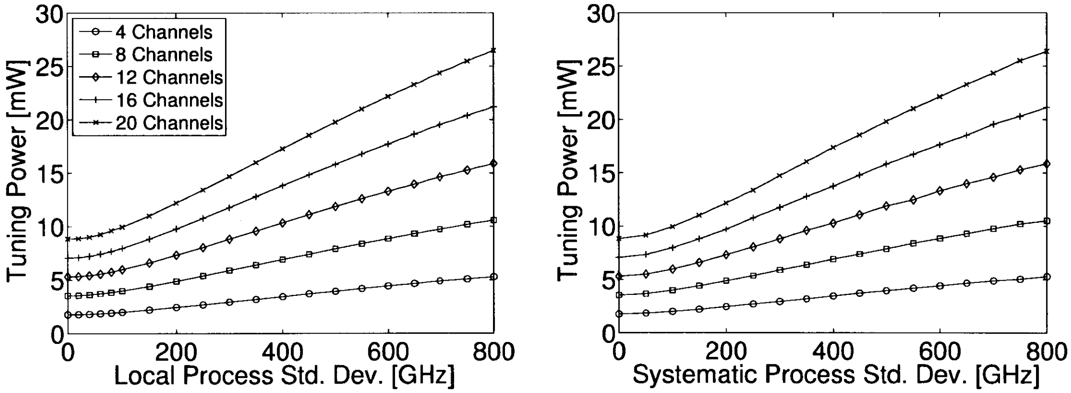
The fourth and last category is *electrically-assisted* tuning, which uses carrier-depletion to red-shift ring resonances. Using the same resonance detuning principle used for reverse-biased modulators, electrically tuned rings consumes no static power and is able to tune-in and tune-out much more quickly than through thermals. The chief drawback of electrical tuning is the far inferior tuning range, typically able to achieve no more 100 GHz. By itself, the small tunable range of fails to make any significant headway in combating process or temperature variations. To be effective, this method must be coupled with ring windowing strategies, such as RRW, that decouple the tuning range from large systematic variations. Should electrical shifts be just short of the required distance, heaters may still be fabricated to bridge any remaining distances. We note that electrical tuning can be used with athermalized rings, should process variations remain small enough.

3.2.3 Evaluation of Resonance Tuning Methods

To evaluate the costs of each tuning strategy, we develop a *Monte Carlo* tuning model. For each tuning scenario, a set of rings in a ring filter bank with some desired resonances is fabricated. To simulate the effects of local process variations, we randomize the resonance of each ring using σ_{rL} and apply global systematic variations using σ_{rS}. The model then attempts to tune the set of rings across a range of temperatures. If successful, the tuning power cost is reported. The experiment is performed 1000 times for each parameter combination (fabrication bias, degree of local muxing, etc.) to find the optimum tuning strategy for a given yield target.

Table 3.2: Tuning Model Evaluation Parameters

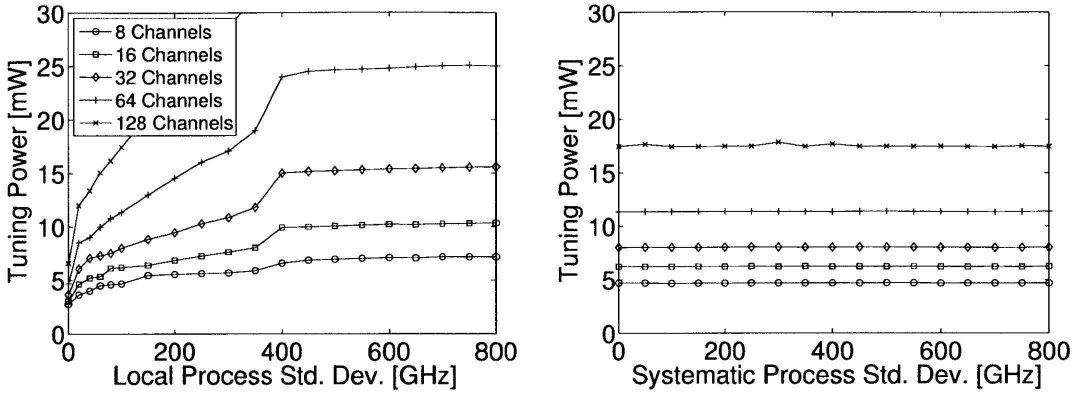
Parameter	**Value**
Process	45 nm SOI
Temperature Range	300-360 K
Aggregate Link Throughput	128 Gb/s
Free Spectral Range (FSR) R=3um ring	4 THz
Heating Efficiency	0.1 K/μW
Tuning Efficiency ($\frac{\Delta f}{\Delta T}$)	10 GHz/K
Local Process Variation (σ_{rL})	varies
Systematic Process Variation (σ_{rS})	varies
Tuner Controller Power	10 μW/Ring
Electrical Tuning Limit	50 GHz
Yield Target	99 %



(a) Tuning power vs σ_{rL}, (σ_{rS}=100 GHz) (b) Tuning power vs σ_{rS}, (σ_{rL}=100 GHz)

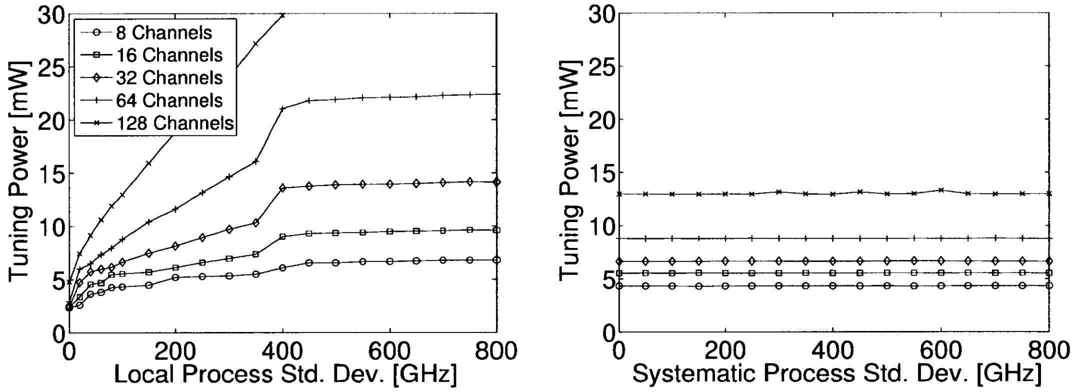
Figure 3-7: Tuning power vs process variation at various channelizations for the full thermal tuning scenario.

The power needed to perform full thermal tuning is shown in Figure 3-7 across a range of process variations (σ_{rS}, σ_{rL}) and channelizations. Across process variations, the power cost of full thermal tuning approximately tracks $\sqrt{\sigma_{rS}^2 + \sigma_{rL}^2}$, stemming



(a) Power vs σ_{rL}, (σ_{rS}=100 GHz) (b) Power vs σ_{rS}, (σ_{rL}=100 GHz)

Figure 3-8: Tuning power vs process variation at various channelizations with a revolving ring window tuning strategy.



(a) Power vs σ_{rL}, (σ_{rS}=100 GHz) (b) Power vs σ_{rS}, (σ_{rL}=100 GHz)

Figure 3-9: Tuning power vs process variation at various channelizations with a tuning stratgy employing both a revolving ring window and electrical assistance.

from the increase in fabrication red-shift bias needed in order to maintain the same yield given higher process variations. The increase in tuning power is also linear with the number of channels, proportional to the number of rings that require tuning.

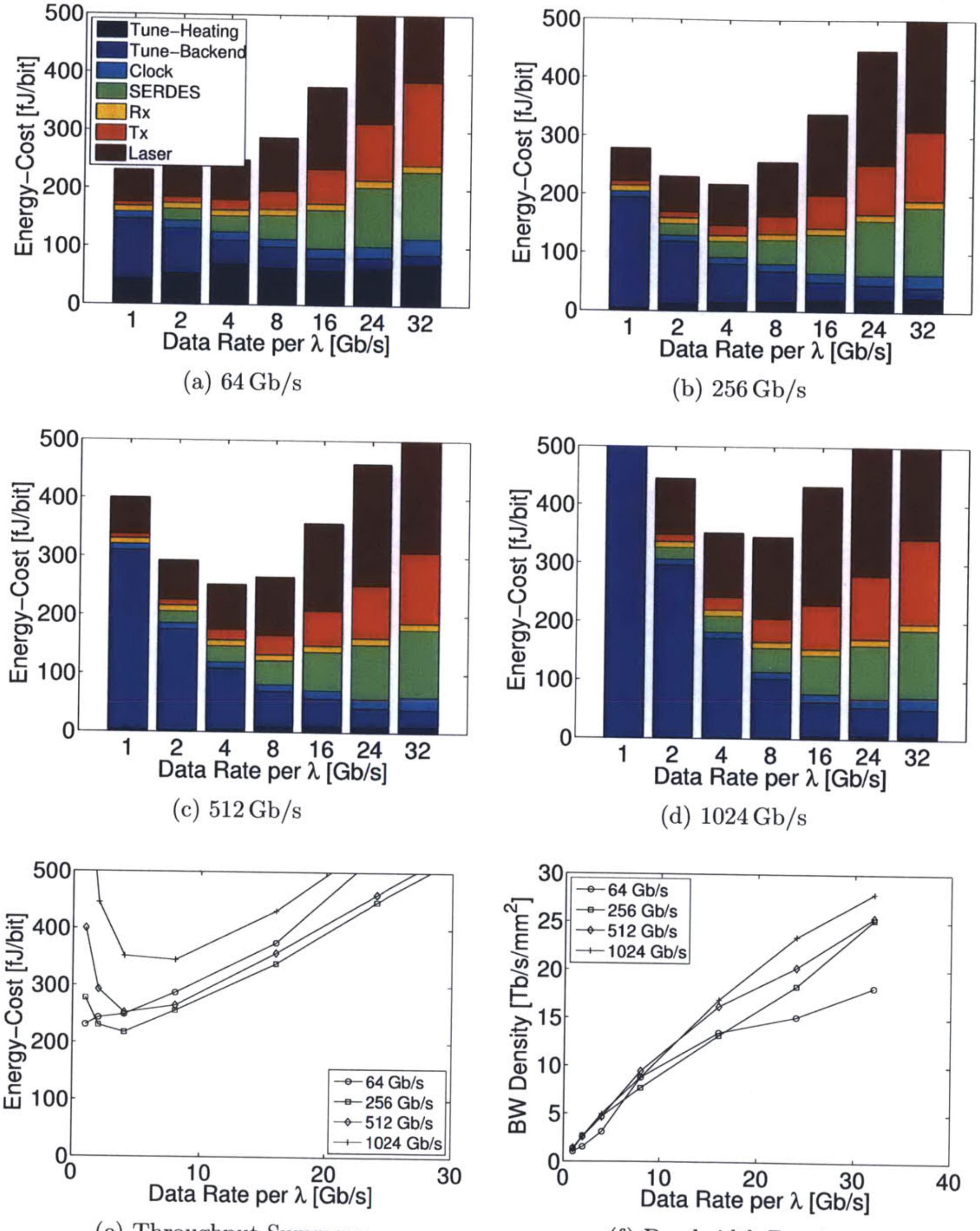
Implementing a revolving ring window, we show that tuning power can be successfully decoupled from σ_{rS} (Figure 3-8). Local variations (σ_{rL}) still affect the tuning power, as a larger σ_{rL} requires a larger degree of bit-reorder multiplexing. The tuning power scales gracefully with the number of channels, owing to the decrease in channel-to-channel separation (and tuning distance) of each ring. Revolving ring windows with electrically assisted tuning allows for even further reductions in tuning power. As shown in Figure 3-9, cases with high numbers of channels benefit most as the channel separation is small enough to be covered electrically, without using heaters. Using this backend, we demonstrate a 5-10X tuning power reduction at dense WDM channelizations while maintaining tuning robustness across a range of process variations.

3.3 Full Evaluation of a WDM Link

Combining models for both data-path and tuning, we perform a full link-level evaluation of a WDM link to quantify power consumption tradeoffs. In our evaluation, we explore links with 4 different aggregate throughput design points, 64 Gb/s, 256 Gb/s, 512 Gb/s, 1024 Gb/s, corresponding to minimum, medium, high, and maximum bandwidth scenarios.

Figure 3-10 shows that tuning power dominates at lower data-rates (since there are more channels given fixed throughput) and decreases with data-rate. Modulator, laser, SERDES, and receiver energies increase with data-rate and dominate at high rate-rates. At all throughput scenarios, an optimal energy balance is achieved at around 4-8 Gb/s. An overall energy-optimal point occurs at around 200 fJ/bit for a link with 256 Gb/s of aggregate throughput and 4 Gb/s data-rate.

At this energy optimal point, we see that the energy consumption is roughly an even 3-way split between tuning, laser, and mod/rx/SERDES. As tuning power is now mostly dominated by the backend electrical components, this energy will scale favorably with technology and can be optimized using custom design. A full electrical tuning backend is also unnecessary on both modulate- and receive-side barrel-shifts and bit-reordering only need to be performed once meaning backend power can be cut by another 50 %. Refinement of photodetector responsivity and parasitic capacitances as well as lower-loss optical devices with improved electrical laser efficiencies can bring about further reductions in wall-plug laser power. It can be expected that energy/bit will drop to sub-100 fJ with device development, process scaling and overall link component refinement. We investigate the scaling of optical links in greater detail in the intra-chip network comparison presented in Chapter 5. Photonic links can provide around 10 Tb/s/mm^2 of bandwidth density at a data-rate of 8 Gb/s, a significant advantage over bump-pitch limited (1 Tb/s/mm^2) and package pin limited (0.05 Tb/s/mm^2) electrical links.



(a) 64 Gb/s

(b) 256 Gb/s

(c) 512 Gb/s

(d) 1024 Gb/s

(e) Throughput Summary

(f) Bandwidth Density

Figure 3-10: Optimized link energy per bit vs. data-rate for different aggregate link throughputs for an optical loss=5 dB, C_{par}=10 fF. For tuning, we assume a bit-reshuffler backend and electrically-assisted tuning with local variation σ_{rL}=40 GHz and systematic variation σ_{rS}=200 GHz. Other parameters used are found in Table 3.1 and able 3.2. Note that the number of WDM channels changes with data rate to maintain a constant through in each plot (Number of Channels = Throughput / Data-Rate).

Chapter 4

DSENT – A Design Space Exploration of Networks Tool

To provide a physical modeling framework to implement our models, we introduce DSENT, a Design Space Exploration of Networks Tool. DSENT enables cross-hierarchical area and power evaluation of on-die interconnects, quantifying the impact of technology-, circuit-, and architecture-level changes on the overall system. DSENT provides electrical and optical interconnect models, integrated on a flexible framework that is applicable to a wide range of NoC designs. In this chapter, we illustrate the most salient features of our approach.

4.1 Motivation

DSENT was originally envisioned as a tool that focused primarily on the modeling of photonic link components in an NoC system. As such, the first iteration of the tool did not contain the infrastructure for electrical modeling and was always used in conjunction with *Orion* [26], a tool which estimates power and area consumption for on-chip routers given electrical technology parameters. While this initial toolflow worked well for the receiver and modulator components of Chapter 3 and photonic devices, we realized, however, that we could not model the backend digital electronics as *Orion* contained only parametrized models of routers. To address these shortcomings, DSENT was redeveloped as a full network physical modeling tool integrating both electrical and optical models in a generalized framework, allowing evaluation of hybrid electro-optical networks and direct comparisons between optics and electronics.

4.2 Related Work

To understand the importance of a flexible electrical-optical tool framework, we examine the methodologies used by several other widely-used architectural modeling tools.

DSENT's electrical network component models are closely related to those in *Orion* [26], which estimates power and area consumption for on-chip routers and links given some input set of technology parameters. While *Orion* is a valuable tool for modeling and optimization, it is not without its drawbacks. *Orion* lacks a delay model for router components, allowing for router clock frequencies to be set abitrarily without an impact on energy/cycle or area. *Orion* also suffers from a somewhat irregular choice of technology parameters. Low-level technology parameters (drain side-wall capacitance, for example) used by *Orion* are non-trivial to obtain for predictive technologies, even when SPICE models are available. Furthermore, analytical calculations made using these low-level parameters accumulate a significant amount of error by the time router and network level estimates are reported.

CACTI is another physical modeling and optimization tool focusing on SRAM and caches [7], with recently added models for DRAM. *McPAT* [38] and *Wattch* [6] are tools derived from the original *CACTI* framework to model area, power, and cycle-time of processors. As opposed to detailed modeling of individual components, *McPAT* applies instead a series of technology trends to curve-fit predictions with published processor results. However, all three lack the freedom needed for design space exploration, as technology parameters and architectural assumptions are embedded in the code.

PhoenixSim [8] is the result of recent work in photonics modeling, bringing about improved architectural visibility concerning the tradeoffs of photonic networks. However, *PhoenixSim* lacks electrical models, including electro-optical interface circuits and link backends, and relies upon *Orion* for all electrical links and routers. Energy estimations for electrical interface circuitry – modulator drivers, receivers, and thermal tuning – use fixed numbers, losing many of the interesting dynamics when transistor technology, data-rate, and tuning scenarios vary. PhoenixSim, as a result, could not capture any of the trade-offs we illustrated in Chapter 3 and cannot be used for area- or power-optimal link design.

To address shortcomings of these existing tools, DSENT aims to provide a consistent and generalizable electrical framework for predictive modeling and design space exploration of electrical and hybridized optical-electrical NoCs. Though current models built using the framework target primarily interconnects, our flexible framework can be used to form accurate power and area estimates of many other components, including caches and processors.

4.3 Framework Overview

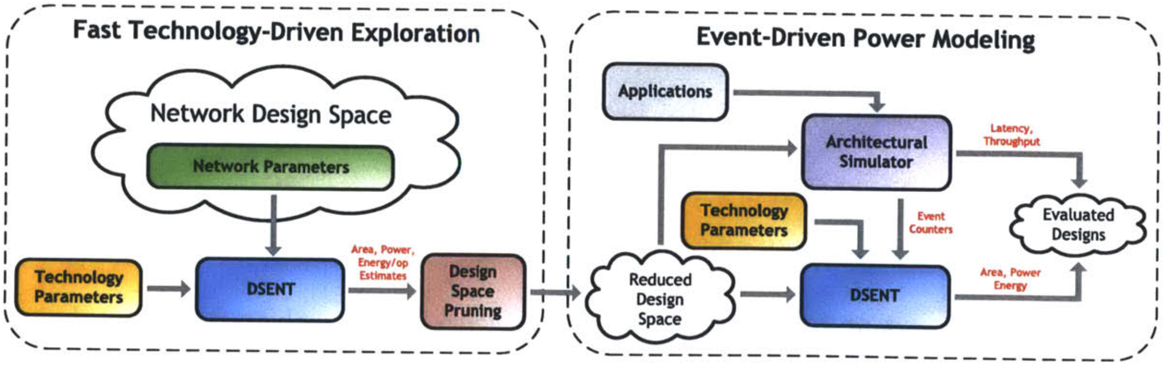


Figure 4-1: Using DSENT for fast design space exploration and integration with an architectural simulator for event-driven power modeling.

DSENT is written in C++ and designed for two primary usage modes (Figure 4-1). When used standalone, DSENT functions as a fast design space exploration tool capable of rapid power/area evaluation of hundreds of different network configurations, allowing for impractical or inefficient networks to be quickly identified and pruned before detailed cycle-accurate evaluation. When integrated with an architectural simulator [1, 45], DSENT can be used to generate traffic dependent power traces and area estimations for the network. We design DSENT to be easily extensible to allow user input of additional models.

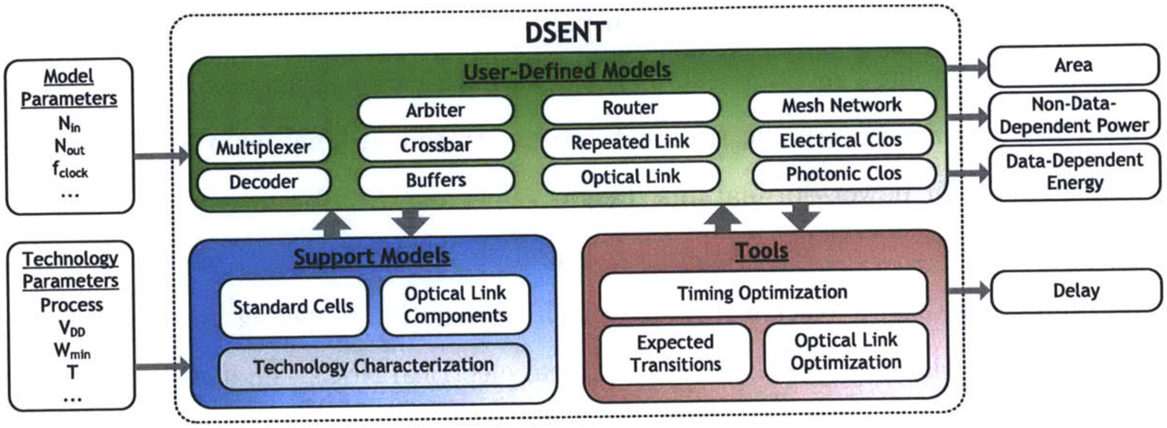


Figure 4-2: The DSENT framework with examples of network-related user-defined models.

Today's modeling tools either provide very good accuracy over a small set of calibrated topologies, or allow flexible design space exploration at the expense of accuracy. DSENT overcomes these trade-offs to perform fast and accurate physical modeling by using circuit- and logic-level techniques to simplify model input specification without sacrificing modeling accuracy. Our approach improves upon the methodology used by currently-available architectural modeling tools by demonstrating the following concepts:

- A flexible area, energy and power interface between sets of models
- State-dependent analytical leakage model with leakage calculations that depend on temperature, technology parameters, and logic states.
- Standard cell model generation during model evaluation using specified technology parameters.
- Generalized delay calculation and energy-optimal timing optimization, applicable to arbitrary pieces of digital logic, incorporating wire resistances and capacitances.

- Expected transition probability model, allowing for calculation of average energy per operation and leakage (using the state-dependent model) to vary based upon signal probabilities.

The DSENT framework, shown in Figure 4-2, can be separated into three distinct parts: user-defined models, support models, and tools. To ease development of user-defined models, much of the inherent modeling complexity is offloaded onto support models and tools. As such, most user-defined models involve simple instantiation of support models, relying upon tools to perform analysis and optimization. Like an actual electrical system, model modularity is emphasized to allow for model reuse and to reduce the amount of required user input. Furthermore, models can leverage instancing and multiplicity to reduce the amount of repetitive evaluation and reduce total run-time. For currently available models, full network evaluation time varies from under a second to a few seconds based upon model size and complexity. DSENT leaves open the option to allow, for example, all one thousand tiles of a thousand core system to be evaluated individually, at the cost of evaluation time.

4.4 Technology and Supporting Models

The gap between technology parameters and a higher-level set of building- blocks is bridged by DSENT's support model infrastructure. For construction of digital electronics, DSENT generates a full standard cell library from which all other digital models are derived. For photonics, DSENT creates a set of optical components from which photonic links and networks can be derived. This section highlights a few key aspects of support models as the hierarchical composition of models.

4.4.1 Area, Power and Energy – A Universal Model Interface

To allow for hierarchical composition and reuse of models, DSENT requires each model to output only three primary results: area, *non-data-dependent* power, and *data-dependent* energy. The DSENT framework does not enforce any requirements as to how each model obtains values for these results. As such, model complexity and accuracy is a choice the user can make based upon what is known regarding the properties and behavior of the model.

The required results form the set from which all other physical cost metrics can be derived for a generic system. The area field is just the area estimate of a design. For NoC components, further distinction can be made between active silicon area, wiring area, and photonic device area (for optical components). For power estimation, DSENT distinguishes between data-dependent and non-data-dependent sources of power consumption. Non-data-dependent power is power that is consumed by a component regardless of whether the circuit is being used or sitting idle, such as leakage and ungated clock power. A data-dependent energy is defined for each event or transaction that the model makes, and refers to the energy consumed by the transaction. Crossbar traversal, buffer read and buffer write are examples of events in the context of a router with individual energy costs. The total power consumption of a model is thus defined as

$$P_{total} = P_{NDD} + \sum E_i \cdot f_i \tag{4.1}$$

where P_{NDD} is the total non-data-dependent power of the model and E_i, f_i are the energy cost of event and the frequency of such events, respectively.

Though both area and the non-data-dependent component of power may be estimated statically, data-dependent power calculation requires knowledge of the overall system behavior and activities. An architectural simulator can be used to supply the event counts at the network or router-level, such as router or link traversals. Events at the gate- and transistor-level, however, are too low-level to be kept track of by these means, motivating our expected transition probability approach, to be discussed in Section 4.5.2. We note that while the calculation of exact event counts is not the main focus of the framework, we can form analytical approximations of network event counts for fixed traffic patterns to obtain rough network energy/bit metrics.

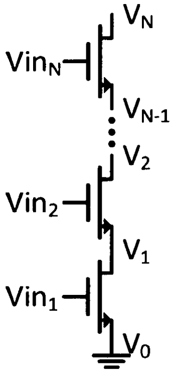


Figure 4-3: N-stack of NMOS transistors. Leakage through this stack can be calculated analytically for each input combination.

It is this principle that allows DSENT, standalone, to perform fast parameter space sweeps to quickly identify and prune bad or impractical designs.

4.4.2 State-Dependent Leakage Model

Instead of leakage tables, which require excessive characterization for every single stack-size, input-state, and transistor size, DSENT adopts an analytical leakage model. Transistor leakage of an NMOS transistor can be described by:

$$I_{leak} = W \cdot I_o \cdot 10^{\frac{V_{gs}}{s_1} + \frac{V_{ds}-1}{s_2}} \tag{4.2}$$

W is the transistor width, V_{gs} is the voltage across gate and source, V_{ds} is the voltage across drain and source, I_o is the off current measured at $V_{gs} = 0$ and $V_{ds} = V_{DD}$, and s_1 and s_2 are subthreshold swing and subthreshold DIBL swing, respectively. We equate the leakage current of all transistors for an arbitrary N-stack of NMOS transistors (Figure 4-3), formulating the following set of linear equations for node voltages in the stack:

$$V_i = \frac{1}{2 \cdot s_1 + s_2} \Big(s_1 \cdot V_{i+1} + (s_1 + s_2) \cdot V_{i-1} \\ + s_1 \cdot s_2 \cdot \log_{10}\left(\frac{W_{i+1}}{W_i}\right) + s_2 \cdot (Vin_{i+1} - Vin_i) \Big) \tag{4.3a}$$

$$V_0 = 0 \tag{4.3b}$$

$$V_N = V_{DD} \tag{4.3c}$$

V_i, Vin_i and W_i correspond to the drain voltage, input gate voltage, and the width, respectively, of the i-th transistor in the stack. Using $V_0 = 0$ and $V_N = V_{DD}$ as bounds, we obtain a leakage current by solving for all other node voltages and plugging the result back to get leakage. Transistors found to be on (in the triode region) in the stack are assumed to have a V_{ds} of 0, as on-resistance is minimal compared to off-resistance. Though not shown, leakage for a PMOS pull-up stack follows the same set of equations but with flipped polarities. As the dependence of I_o on temperature is generally exponential, temperature effects may be fitted using a third swing term s_3. Temperature effects on s_1 and s_2, are also analytically calculated.

4.4.3 Standard Cells

As the usage of standard cells is practically universal in modern digital design flows, detailed timing, leakage, and energy/op characterization at the standard cell level can enable a high degree of modeling accuracy. The DSENT framework applies standard cell design heuristics extrapolated from open-source [48] libraries and calibrated with commercial standard cells to generate a standard cell models portable across technology and constructed during run-time(Figure 4-4). We strive to rely on only a minimalistic set of technology parameters, shown in Table 4.1, that best capture the major characteristics of deep sub-100 nm technologies without diving into transistor modeling. Both interconnect and transistor properties are paramount at these processes, as interconnect parasitics play an ever larger role due to poor scaling trends [54]. These parameters can all be obtained from ITRS roadmap projection

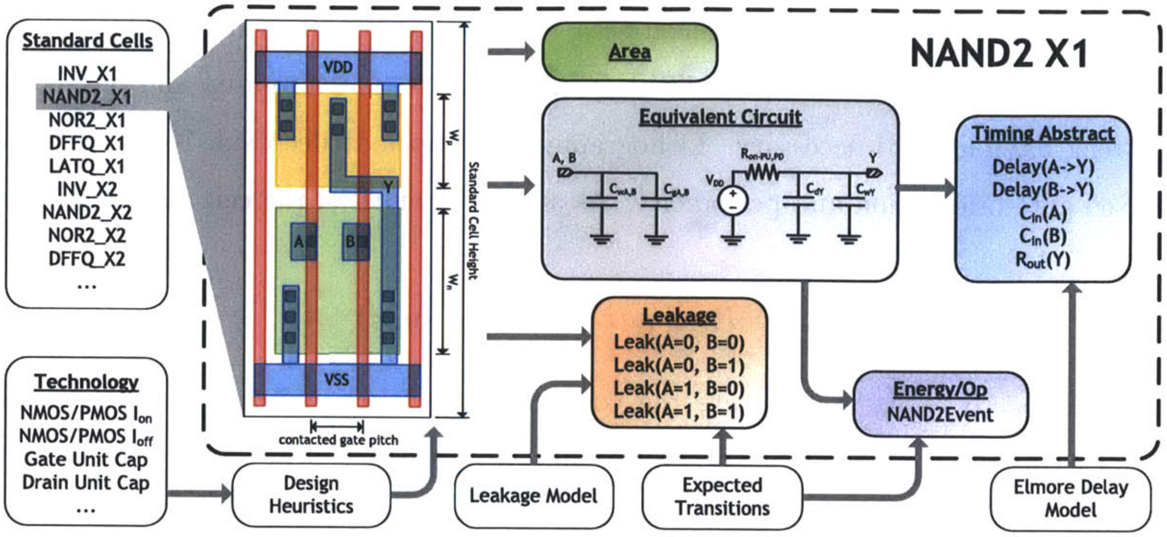


Figure 4-4: Standard cell model generation and characterization. In this example, a NAND2 standard cell is shown and characterized.

tables for predictive technologies or characterized from SPICE and process design documentation when available.

Table 4.1: DSENT Electrical Technology Parameters

Process Parameters	**Interconnect Parameters**
Process Supply Voltage (V_{DD})	Wiring Layers (Local, Global, etc.)
Minimum Gate Width	Minimum Wire Width
Contacted Gate Pitch	Minimum Wire Spacing
Gate Capacitance / Width	Resistivity
Drain Capacitance / Width	Wire Thickness
Effective On Current / Width (P/NMOS)	Dielectric Thickness
Single-transistor Off Current (P/NMOS)	Dielectric Constant
Subthreshold Swing (P/NMOS)	
DIBL Swing (P/NMOS)	

To generate a standard cell library, DSENT begins by picking a global standard cell height:

$$H_{cell} = H_{ex} + \alpha \cdot (1 + \beta) \cdot W_{min} \tag{4.4}$$

β represents the P-to-N ratio, W_{min} is the minimum transistor width, and H_{ex} is the extra height needed to fit in supply rails and diffusion region separation. α is heuristically picked such that large (high driving strength) standard cells do not require an excessive number of transistor folds and small (low driving strength) cells do not waste too much active silicon area. Given a desired drive strength and function for each standard cell, DSENT sizes transistors to match pull-up and pull-down strengths, folding if necessary. Lithography limitations at deep sub-100 nm force a fixed gate orientation and periodicity. As such, the width of the cell can be written as:

$$W_{cell} = P_{cg} \cdot [max\,(N_{NMOS}, N_{PMOS}) + 1] \tag{4.5}$$

where P_{cg}, N_{NMOS}, N_{PMOS} are the contacted gate pitch, number of NMOS pull down transistor, and number of PMOS pull up transistors, respectively. The extra gate pitch in the width is allocated for the separation between standard cells.

With transistor sizes and cell dimensions known, each cell is analytically evaluated to extract relevant characteristics. Leakages are generated using the state-dependent leakage model. Cell dimensions are used to generate standard-cell-level wiring parasitics. Next, DSENT uses an equivalent circuit for each cell to obtain timing information used by DSENT's timing model (Section 4.5.1) and per-node capacitances used to generate data-dependent energy costs. Finally, all results are cached by the framework to form the library for this technology. From here, standard cells can be instantiated to form bigger electrical building blocks. The tool can also output Liberty [51] format libraries the industry standard used by commercial hardware tools for logic synthesis.

4.4.4 Optical Components and Links

DSENT implements the full set of optical device and component-based models presented in Chapter 3. For passive optical devices, simple calculations using optical path losses are sufficient in providing first-order approximation for the behavior of optical links. Active devices and circuits, such as modulators, receivers, and thermal tuning backends utilize both analytical models to describe the analog behavior of the circuit as well as standard cells for digital components. Integration of both electrical and optical components allows DSENT's models to capture the tradeoff between electrical and optical power. Optical link- and network-level models are composed through the specification of connectivity between various optical components. Upon model evaluation, DSENT traces optical paths to build tables consisting of the lasers, modulators, and detectors/receivers that are active for particular wavelengths. Optical path losses to each receiver in the table can be used to calculate the required laser power of an arbitrary optical network.

4.5 Tools and Optimizations

DSENT features a simple set of tools for analysis and optimization of models. For electrical designs, a built-in delay calculation and timing optimization allows for models to adapt to different performance requirements. DSENT also supports a simple probability propagation scheme to estimate gate-level switching power. This section describes these tools as well as some of their limitations.

4.5.1 Delay Calculation and Timing Optimization

To allow models to scale with transistor performance and clock frequency targets, we apply a first-order delay estimation and timing optimization. Using timing information in the standard cell models, chains of logic are mapped to resistance-capacitance (RC) trees using gate drive strengths and capacitances, shown in Figure 4-5. An Elmore delay estimate [18, 56] between two points i and k can be formed by summing the product of each resistance and the total downstream capacitance it sees:

$$t_{d,i-k} = ln(2) \cdot \sum_{n=i}^{k} \sum_{m=n}^{k} R_n \cdot C_m \tag{4.6}$$

We note that any specified resistances or capacitances for wiring parasitics is automatically factored along the way.

If a register-to-register delay constraint, such as one imposed by the clock period, is not satisfied, timing optimization is required to meet the delay target. To this end, we employ a greedy incremental timing optimization algorithm, shown in Figure 4-6. We start with the identification of a critical path. Next, we find a node to optimize

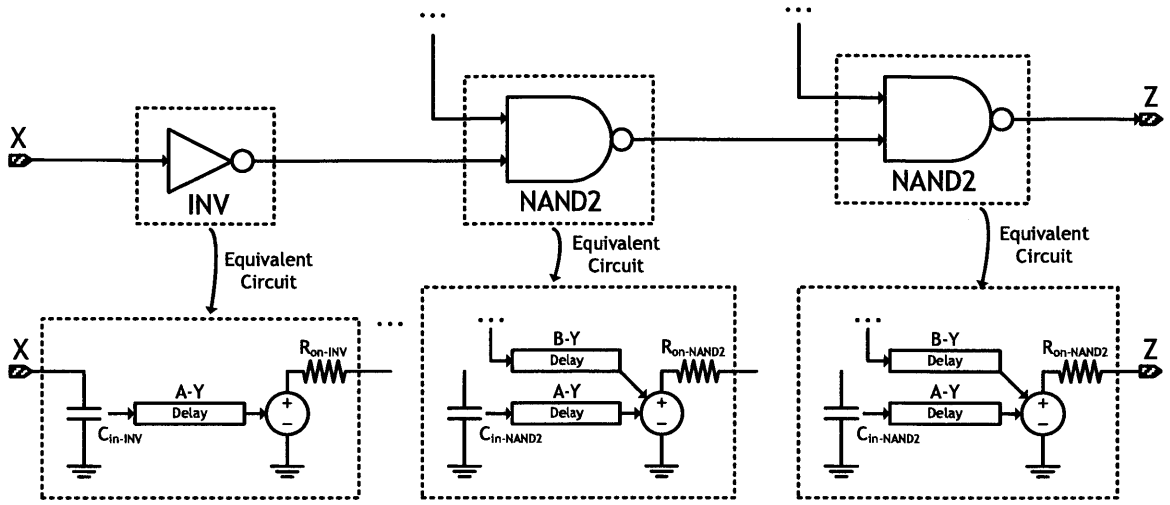


Figure 4-5: DSENT's delay calculation framework. Delay is estimated by mapping standard cells to sets of input capacitances and output drive resistances.

to improve the delay on the path, namely, a small gate driving a large output load. Finally, we size up that node and repeat these three steps until the delay constraint is met or if we realize that it is not possible and give up. Our method optimizes for minimum energy given a delay requirement, as opposed to logical-effort based approaches employed by existing tools [7, 38], which optimize for minimum delay without regards for energy. Though lacking the rigorousness of timing optimization algorithms used by commercial hardware synthesis tools, our approach runs fast and performs well given its simplicity.

4.5.2 Expected Transitions

The primary source of data-dependent energy consumption in CMOS devices comes through the charging and discharging of transistor gate and wiring capacitances. For every transition of a node with capacitance C to voltage V, we dissipate an energy of $E = \frac{1}{2}C \cdot V^2$. To calculate data-dependent power usage, we sum the energy dissipation of all such transitions multiplied by their frequency of occurence, $P_{DD} = \sum C_i \cdot V_i^2 \cdot f_i$. Node capacitance C_i can be calculated for each model and, for digital logic, V_i is the supply voltage. The frequency of occurence, f_i, however, is much more difficult to estimate accurately as it depends on the exact pattern of bits

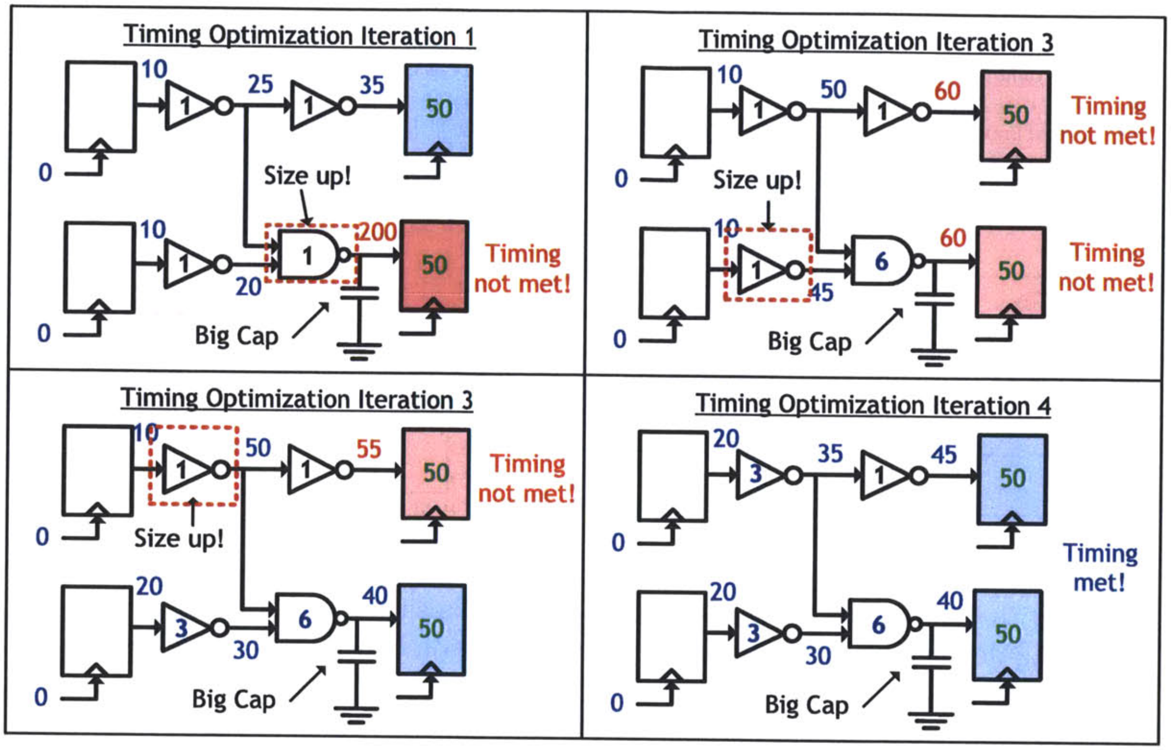


Figure 4-6: An example of DSENT's timing optimization model. Using delay calculations, timing optimization may incrementally size-up cells until all delay constraints are met. In this example, a 50 time unit delay constraint is imposed on all register to register paths.

flowing through the logic. As event counts and signal information at the logic gate level are generally not available except through simulation of structural netlists, our framework uses a simplified expected transition probability model [42] to estimate the average frequency of occurance for switching events.

As shown in Figure 4-7, each node is assigned four probabilities — P_{00}, P_{01}, P_{10}, P_{11} — at each input node of a model, corresponding to the chances of a 0-to-0, 0-to-1, 1-to-0, and 1-to-1 transition occuring for a given cycle, respectively. Each model uses probabilities to calculate the average frequency of occurence of each switching event and the average leakage power (using the state-dependent leakage). Probabilties will change based on how the model is used, producing different power and leakage numbers. The impact of clock gating, for example, can be estimated by simply setting P_{01} and P_{10} transition probablities of the clock to be 0 for the idle usage state. As digital logic is composed of standard cells, transition probability calculation and

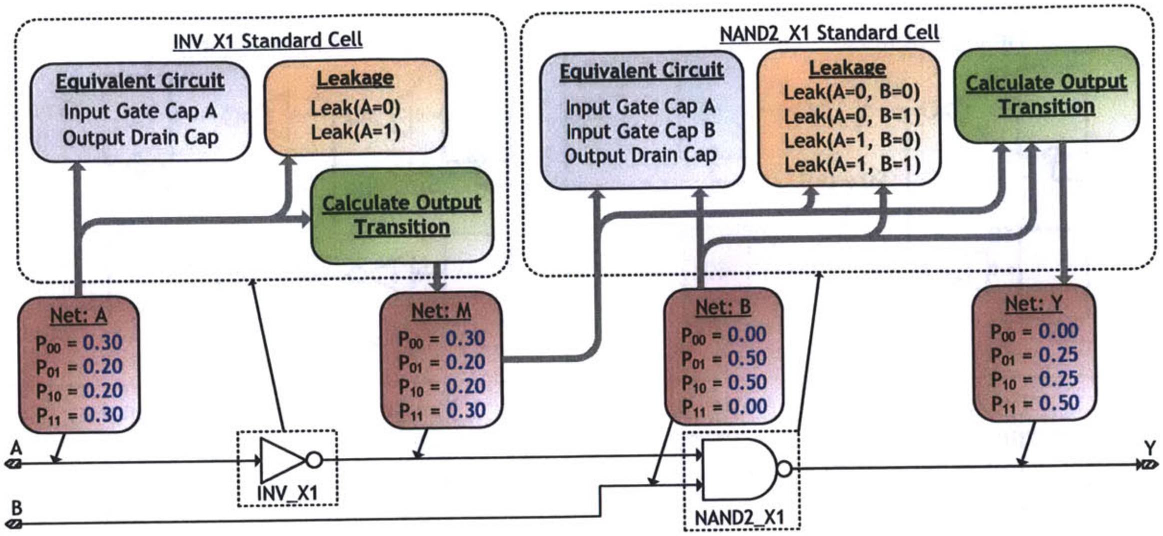


Figure 4-7: An example illustrating the propagation of transition probabilities in DSENT. In the context of digital components, transition probabilities are used for both data-dependent gate switching power estimation and leakage power estimation.

propagation can be performed at the cell level and automatically propagated from one cell to the next, hiding much of the complexity of this approach. Probabilities derived using this model are also used with state-dependent leakage in the standard cells to form accurate leakage calculations. Interestingly, when input transition probabilities are set to represent fixed 0s and 1s (i.e. probability of 0-to-0 transitions = 1.0), applying probability propagation results in logic evaluation.

4.5.3 Known Tool Limitations

Complex control logic Though the DSENT framework aims to simplify modeling, it is not without its limitations. One such limitation is the lack of logic synthesis capability built into the tool, complicating declaration of complex control logic. Likewise, specification of wiring capacitances and resistances is difficult without model placement information. For structured data-path elements, such as crossbars and buffers, wiring parasitics can be calculated analytically on a case-by-case basis by assuming a particular floorplan. The effect of wiring parasitics for unstructured control logic is much more difficult to derive.

Correlated transition probabilities To allow probability calculations remain local, each block assumes that its inputs are independent from one another. As such, modeling inaccuracy can occur when signals are highly correlated. This problem can be mitigated by providing conditional probabilities between pairs of inputs, though manual user intervention is required for these cases as it is not handled automatically by the framework. Fortunately, this problem does not significantly impact data-path elements, as data bits flowing through them are generally weakly correlated.

4.6 User-Defined Models and Validation

4.6.1 Available Models

While the DSENT modeling framework can be applied to any generic electrical system, our model library currently emphasizes components for NoCs. In addition to just routers and links, the DSENT framework allows for modeling of much broader set of network-level components. A representative subset of electrical building blocks currently modeled by DSENT is shown in Table 4.2. Where applicable, we adopt the textbook circuit structures found in [12] for NoC components.

Table 4.2: Subset of Available DSENT Models

12 Standard Cells, 10 sizes each	Separable Allocator	Mux-Tree Serializer
Latch Based RAM	Ripple Adder	Demux-Tree Deserializer
D-Flip-Flop Based RAM	Router	Barrel-Shifter
Multiplexer	Repeated Link	Electrical Mesh
Crossbar	Decoder	Electrical Clos

4.6.2 Validation of an NoC Router

To validate our approach, we model an NoC router on a 45 nm SOI process using DSENT and compare against SPICE simulation using the transistor models available for that process. The parameters of the router are shown in Table 4.3 and the results of our comparison are summarized in Table 4.4. Overall, DSENT's total power estimate is within 10% of SPICE. For most router subcomponents, DSENT is within 5–15%. We note that the high discrepency between DSENT and SPICE for router control power is due to the presence of highly correlated signals present in control logic, which DSENT assumes are independent (a framework limitation discussed in Section 4.5.3). However, control power represents only a small fraction of the total router power allowing DSENT to maintain a respectable accuracy for the full router. Interestingly, estimates produced by Orion 2.0 [26] are way off. We note that the 500-cycle SPICE simulation of the router took approximately 20 hours to complete. DSENT's total run-time was under a second.

Table 4.3: Configuration of the Validated Router

Parameters	**Values**
Clock Frequency	1 GHz
Number Ports	6
Flit width	64 bit
Total Number VCs per port	8
Flit Injection Rate	0.16 flits/cycle/port
Total Buffers per Port	16
Buffer type	D-Flip-Flop based RAM
Arbiter type	Matrix arbister
Crossbar type	Multiplexer crossbar

Table 4.4: Comparison of DSENT Results with SPICE Simulations and Orion

	SPICE	**Orion2.0**	**DSENT**
Buffer	6.93 mW	0.84 mW (-88%)	7.55 mW (8.9%)
Crossbar	2.14 mW	9.75 mW (355%)	2.06 mW (-3.7%)
Control	0.75 mW	0.3 mW (-60%)	0.93 mW (24%)
Clock	0.74 mW	3.86 mW (436%)	0.63 mW (-14.9%)
Total	10.6 mW	14.8 mW (40%)	11.2 mW (5.7%)
Total Area	0.070 mm^2	0.154 mm^2 (120%)	0.062 mm^2 (-11.0%)

4.7 Future Work on DSENT

Additional Models Using the DSENT framework, we plan to evaluate and compare an ever broader range of networks and network components. In the near future, we plan to integrate models for SRAM, which are needed for models of SRAM-based buffers in routers and caches. Models for several electrical link techniques, such as low-swing and equalization, are also planned to better represent the capabilities of electrical links. As silicon photonics is currently a rapidly evolving field, optical models will change to reflect new designs and components. Validation of optical models with their respective components may also be possible as they are fabricated and made available.

Framework Improvements Addressing some of the tool's framework limitations concerning the specification of random control logic, we first plan to improve the calculation of transition probabilities for more accurate power estimation. Next, as a long-term goal, we plan to use structural verilog netlists for model specification and functionality. This creates the possibility for DSENT to read post-synthesis or post-place-and-route netlists to ease modeling of complex logic. Auto-generation of Liberty timing file descriptions of standard cells will also allow the use of hardware synthesis tools to aid model specification. In lieu of knowing actual placement information, we plan to derive some analytical heuristics to automatically factor wiring parasitics, as they play ever larger roles for advanced processes. The ultimate goal of these improvements is a flow that can potentially take in verilog and output power and area, portable across any technology node while still allowing custom (non-verilog) models to be integrated into the flow.

Chapter 5

Electronics vs. Optics – An Intra-Chip Network Comparison

As significant process development is required to introduce photonics into advanced CMOS fabrication processes, photonics-inspired architectures must show clear advantages over their electrical counterparts. Furthermore, optical interconnect technologies must provide a roadmap for further improvement or risk falling behind to aggressive electrical scaling. In this chapter, we use the modeling infrastructure developed in Chapters 3 and 4 to evaluate the costs and benefits of photonics as compared to their electrical counterparts. We identify areas in which photonic device designers must improve upon to remain ahead in intra-chip networks domain.

5.1 Technology Scaling for Intra-Chip Networks

While the introduction of optics as an interconnect technology potentially carries architectural and power advantages, it is unclear whether these advantages will continue to hold as electrical technology matures. In this section, we compare several different configurations of 64-core and 256-core network topologies using an architectural simulator as our performance model. Using DSENT, we examine the effects of technology scaling on network energy cost per bit. Based on our analysis, we identify parasitic capacitance incurred during photonics integration and electrically-assisted tuning as promising areas for further photonics device improvement.

5.1.1 Experiment Setup

We choose three representative network topologies to evaluate: the electrical mesh, representing a low-radix high-diameter electrical network, electrical clos, a high-radix low-diameter electrical network, and the photonic clos [25], a hybrid optical-electrical network. We compare these topologies at the 45 nm SOI, 22 nm SOI, and 11 nm Tri-Gate transistor nodes, representing the present, near-future, and far-future technology scaling scenarios, respectively. To obtain technology parameters needed by DSENT, models for 22 nm and 11 nm were derived using the virtual-source transport [30] and parasitic capacitance model [68] using projections on what is likely to be technologically feasible during those time frames [29]. For 45 nm, we characterize and extract the technology parameters needed by DSENT using SPICE models available for this technology.

Table 5.1: Parameters for Intra-Chip Network Evaluation

Parameters	**Values**
Process Nodes	45 nm SOI, 22 nm SOI, 11 nm Tri-Gate
Process Supply Voltages	1.0, 0.8, 0.6 V
Temperature	340 K
Network Clock Frequency	1 GHz
Photonic Link Configuration	64 λ at 1 Gb/s per λ
Ring Tuning Strategy (for Photonic Clos)	RRW w/ Electrical Assist
Photonic Device Parameters	Same as Figure 3-10
Number of Cores	64, 256
Total Area (Divided evenly among cores)	600 mm^2
Router Configuration	3 Pipeline Stages
Mesh Link Delay	1 cycle
Clos Link Delay	2 cycles
Clos Config (Num. Ing x Mid x Eg Routers)	8 x M x 8 (64), 16 x M x 16 (256)
Network Flit Width	64 bits
Traffic Pattern	Uniform Random

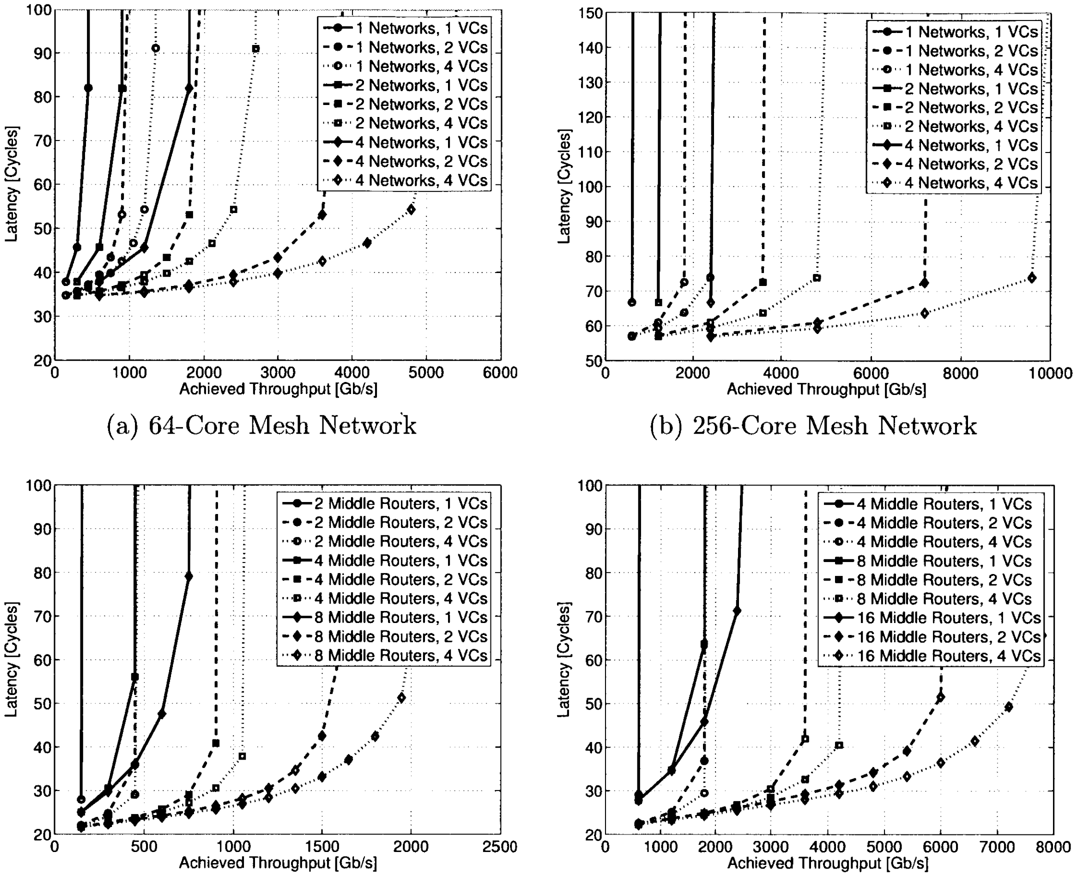
To obtain network-level event counts with which to animate DSENT models, we use the GEM5 [5] architectural simulator, which employs Garnet [1] as the underlying network simulator. For the studies presented in this section, we inject packets into the network in a uniform random traffic pattern. Table 5.1 shows the technology and performance parameters used in our experiment.

5.1.2 Network Performance Evaluation

To put the performance of each network into context, we use GEM5 to gauge the average network latencies vs. network injection rates across a range of network configurations. For the mesh, we vary the number of virtual channels (VCs) and the number of multiple mesh networks. For the clos, we sweep the number of VCs and the number of middle-stage routers while keeping the number of ingress and egress routers fixed at 8 for the 64-core clos and 16 for the 256-core clos. All networks are assumed to run at 1 GHz.

Figure 5-1 shows the result of our performance experiment. We note that these results are technology independent (and also irrespective of either photonic or electrical implementations), as they represent performance given only the logical network structure and connectivity. For both mesh and clos, the number of VCs has a profound impact on network performance. In all cases, going from a single VC (solid line) to 2 VCs (dashed line) results in a doubling or more of network saturation throughput while moving from 2 VCs to 4 VCs provides only an extra 30 50 %. Saturation throughput is directly proportional to the number of multiple networks for the mesh and the number of middle routers for the clos. Comparing between the two topologies, clos carries a significant zero-load latency advantage over the mesh, as is expected of a low-diameter topology.

We note that because we use a uniform random traffic pattern, we thus made the assumption that network latency and network injection rates are independent. In the context of real applications, however, injection rates are correlated with latencies through the application performance. Though not considered for the studies in this section, we describe this effect in more detail when we revisit the ATAC architecture in Section 5.2.



(a) 64-Core Mesh Network

(b) 256-Core Mesh Network

(c) 64-Core Clos Network

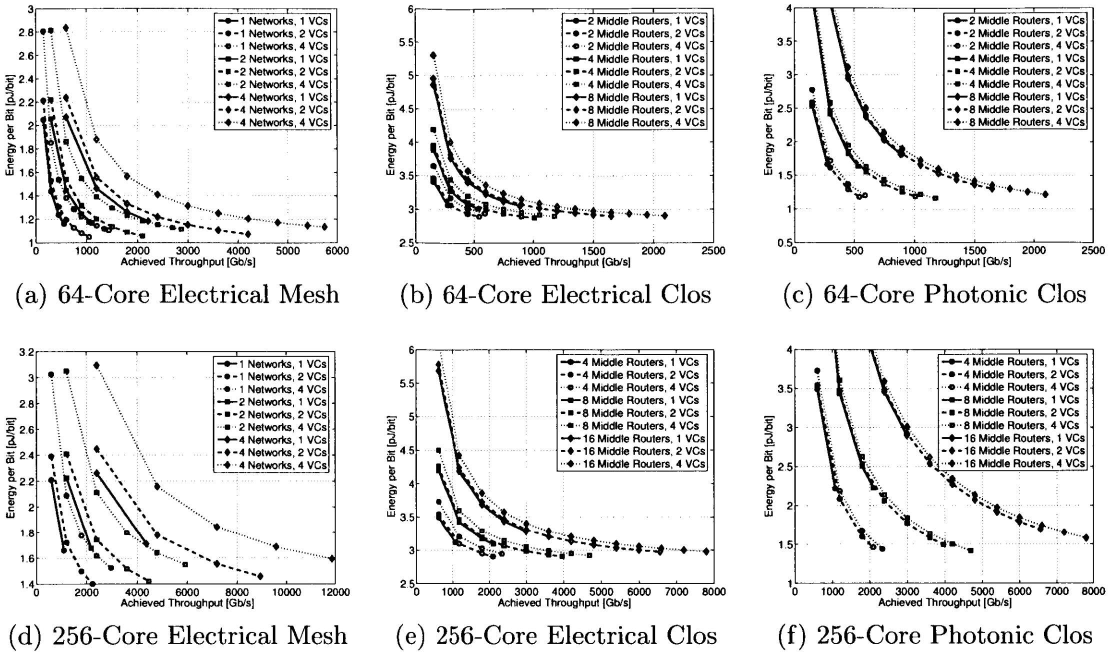
(d) 256-Core Clos Network

Figure 5-1: Average latency comparison between the 64-core and 256-core mesh and clos topologies. Note that the latencies and throughputs in these plots are technology independent.

5.1.3 Network Efficiencies at the 45 nm Technology Node

Using event counters generated by GEM5 and physical models of DSENT, we compare the energy per bit cost of each network configuration at the baseline 45 nm technology node, shown in Figure 5-2. Photonic links run at 1 Gb/s per λ, which also happens to be optimal for the given flit width and network clock frequency.

From Figure 5-2, we make several interesting observations. For all networks, the energy per bit rises sharply at low utilizations due to non-data-dependent power, which adds more to energy per bit when fewer bits are delivered by the network. This trend causes all networks designed for aggressive throughputs to perform less



(a) 64-Core Electrical Mesh (b) 64-Core Electrical Clos (c) 64-Core Photonic Clos

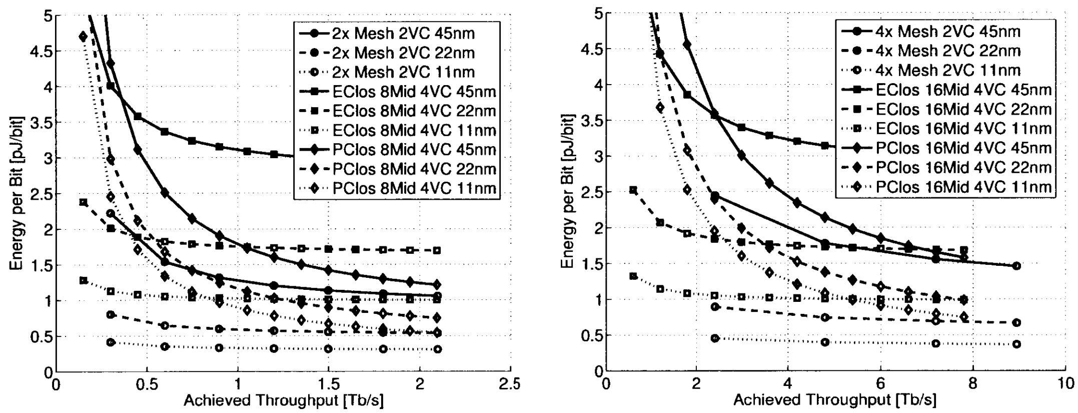
(d) 256-Core Electrical Mesh (e) 256-Core Electrical Clos (f) 256-Core Photonic Clos

Figure 5-2: Energy/bit of various configurations of the three network topologies versus network injection rates for 64-core systems (top) and 256-core systems (bottom). Each line ends at the throughput the network saturates, which we define as roughly where the average network latency reaches 2X the zero-load latency.

efficiently at lower throughputs than those sized appropriately, as more network resources implies more idle components and non-data-dependent power consumption (more leakage, clocking, laser, etc.). When multiple networks are considered for a mesh, the choice of two virtual channels appears to be optimal. When moving from 1 to 2 VCs, network capacity is improved by roughly 2X, while energy/bit increases marginally. Moving from 2 VCs to 4 VCs, however, results in a smaller performance gain but carries a much steeper overhead in energy/bit. As a result, 2 parallel networks with 2 VCs has a lower energy per bit than a single network with 4 VCs. For the electrical and photonic clos topologies, the overhead of adding additional VCs is lower relative to the total, as packets traverse through only three routers.

5.1.4 Scaling From 45 nm to 22 nm and 11 nm

We select the optimal configuration of each network at a throughput corresponding to roughly 0.5 flits/cycle/core (an achieved throughput of 2 Tb/s for a 64-core network and 8 Tb/s for a 256-core network). For 64-core networks, these are the mesh with 2 networks and 2 VCs, the electrical clos with 8 middle routers and 4 VCs, and the photonic clos with 8 middle routers and 4 VCs. For a 256-core network, these are the mesh with 4 networks and 2 VCs, the electrical clos with 16 middle routers and 4 VCs, and the photonic clos with 16 middle routers and 4 VCs. We plot the effects of technology scaling for these networks for both the 64-core and 256-core scenarios, shown in Figure 5-3.



(a) Technology Comparison – 64 cores

(b) Technology Comparison – 256 cores

Figure 5-3: Comparison of the three networks across technology nodes for both the 64-core and 256-core scenarios.

This plot highlights the slower scaling of the photonic clos with technology as compared to the fully electrical topologies. For the 64-core scenario at half network utilization (a throughput of roughly 1 Tb/s), the photonic clos scales from 1.8 pJ/bit at 45 nm to 0.9 pJ/bit at 11 nm. The electrical topologies scale much better, scaling an impressive 1.3 0.3 pJ/bit for the mesh and 3.1 1.0 pJ/bit for the clos. The scaling rate disparity is more apparent for the 256-core configuration and the 11 nm photonic clos loses to the 11 nm electrical clos at sub-5 Tb/s throughputs.

The lackluster efficiency of the photonic clos network at the 11 nm can be attributed to two main reasons. The first reason is that while electrical portions of the photonic clos network scale with technology (routers and ring tuning backend) many of the photonic components do not (ring heating power, receiver parasitic capacitance, optical losses, etc.). The second reason is due to the higher proportion of non-data-dependent power. The photonic clos achieves improved energy efficiency primarily by trading data-dependent energy of electrical link traversals with non-data-dependent (always-on) ring-tuning and laser. As electrical links scale while laser and ring-tuning remain stagnant, non-data-dependent power becomes an increasingly significant fraction of the total. As non-data-dependent power sources are unaffected by network utilization, the energy per bit of the photonic clos rises sharply at lower utilizations, allowing it to be overtaken by the data-dependent energy-dominated electrical clos at lower utilizations.

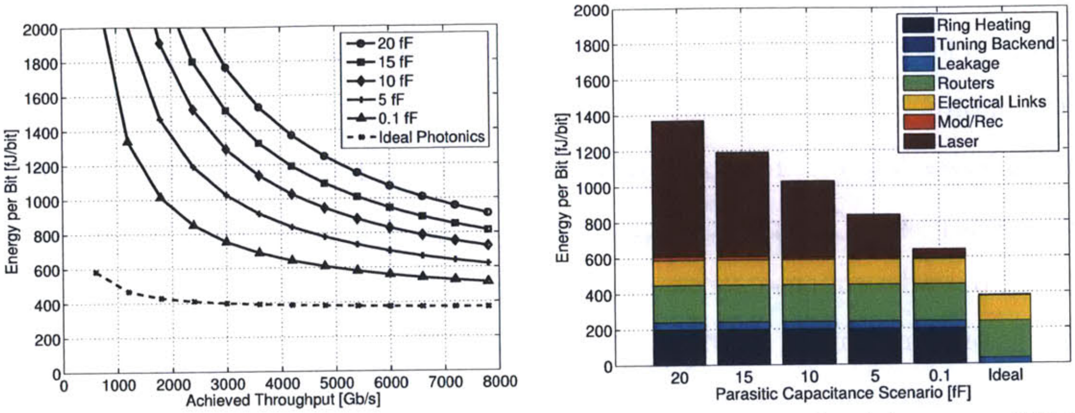
5.1.5 Scaling Photonic Interconnects

For photonics to remain competitive with electrical alternatives at the 11 nm node and beyond, it must similarly scale. Here, we identify potential strategies for reducing laser and ring heating costs, the two primary sources of non-data-dependent power consumption in an optical link. Though we use the photonic clos network as the basis of our evaluation, these techniques are applicable to a far more general class of photonic networks.

Laser Power Reduction

As shown in Equation 3.6, there are two primary ways to reduce laser power consumption of the network. The straightforward way is to reduce the optical path loss from the laser to receiver. This requirement can be achieved through overall photonic device refinement and should remain a paramount goal for optics device designers. The more subtle and circuit-oriented approach involves the improvement of receiver sensitivity which reduces the amount of light required at the photo-detector.

As the receiver itself is a circuit, its sensitivity will improve slightly with technology



(a) Energy per bit vs C_{par}

(b) Energy per bit breakdown at 4 Tb/s throughput

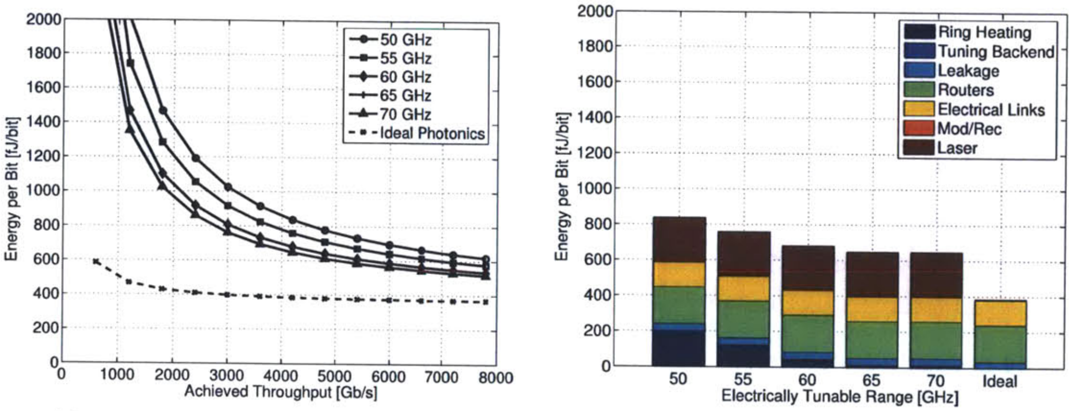
Figure 5-4: Effect of reducing the parasitic interconnect capacitance between interface circuitry and photonic devices for a 256-core photonic clos. An ideal photonics case (no losses, no parasitics, 100 % laser efficiency, perfect rings) is also shown for comparison. All results use the 11 nm technology node.

as transistors become faster and smaller. However, the factor dominating the receiver sensitivity is not the speed of the circuit itself, but rather the capacitances (C_{par}) present at the input terminals of the receiver. A reduction in C_{par} will result in a proportional improvement in receiver sensitivity and laser power, shown in Figure 5-4. This is consistent with the trend predicted by Equation 3.5.

We note the parasitic capacitance, C_{par} is affected mainly by the distance between photonic devices and electronics. Heterogeneous photonics integration, where electronics sit on one die and connect to a dedicated photonics die using a TSV, will inevitably present a large C_{par} to the receiver due to interconnect capacitances of the TSV [31]. Monolithic integration [3, 53], where electronics and photonics are put on the same die, can enable a close proximity between electronics and photonics. This method allows for C_{par} to be reduced to that of on-chip wires, resulting in a desirable decrease in required laser power. While monolithic integration may result in increased losses due to the added difficulty of having both photonics and electronics in the same process, this drawback must be weighed against the potential savings brought about by decreased parasitics.

Ring Tuning Improvement

With laser power reduced, ring heating power represents the dominant source of non-data-dependent power. While our approach of using electrical assistance and a revolving ring window allowed for a reduction in ring heating power, ring heating was still required due to the limited tuning range achievable by the depletion-mode modulator. As shown in Figure 5-5, even a slight increase in the electrically tunable range yields huge savings in heating power, falling to a negligible level when the range is greater than the channel separation (63 GHz for this configuration).



(a) Energy per bit vs tunable range

(b) Energy per bit breakdown at 4 Tb/s

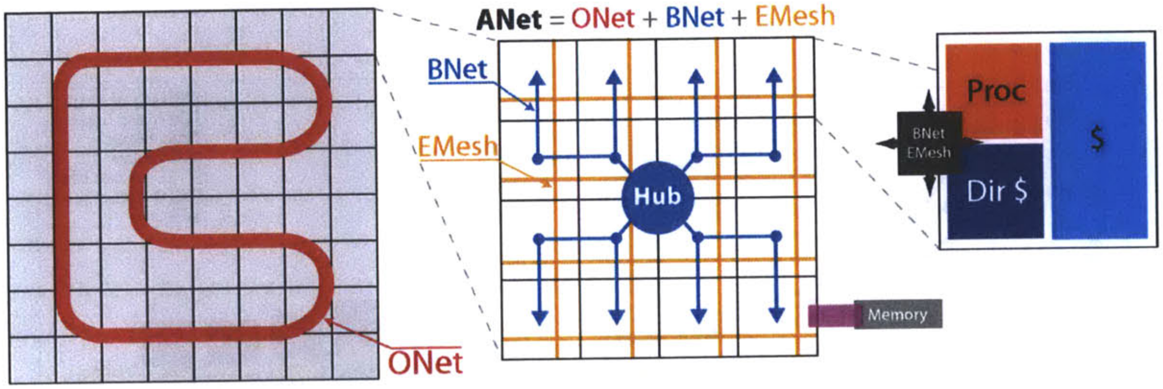
Figure 5-5: The effect of improving the electrically tunable range limit to reduce power used to heat up rings for a 256-core photonic clos. A 5 fF C_{par} is assumed. Ideal photonics scenario is shown for comparison. All results use the 11 nm technology node.

With both a reduction in C_{par} and an improved electrically tunable range, energy per bit in the photonic clos improves by roughly 2X from the original design. Future device loss improvements will likely drive down the laser power component further, making electrical components the dominant source of energy consumption.

5.2 Revisiting the ATAC Architecture

In the previous section, we injected random packets into a network and evaluated the energy it cost to deliver them to their destination. While this method of synthetic traffic generation provides a straightforward approach in evaluating networks under various loads, it does not necessarily capture the effect of network latencies in the context of an application. In this section, we revisit the 1024-core ATAC architecture proposed by Kurian et al. [36] from an application-driven full-system power and performance perspective.

5.2.1 Overview of the ATAC Architecture



(a) 64 optically-connected clusters

(b) Electrical cluster network connecting 16 cores

(c) Tile architecture

Figure 5-6: The ATAC Network – Image from [36]

The 1024-core ATAC network, shown in Figure 5-6, consists of an optical crossbar (the *ONet*) overlayed on top of a packet-switched electrical mesh (*EMesh*). The overall system is split into 64 individual groups of cores called *clusters*, connected together optically by the optical crossbar. Each cluster has a dedicated set of wavelengths or waveguides with which it can broadcast to all other clusters at any time. This is similar to the single-writer multiple-reader crossbar topology discussed in [70]. A cluster consists of 16 cores and contains a centralized *hub* that connects to the global optical crossbar. The hub handles all electrical-to-optical and optical-to-electrical conversions necessary to send and receive packets on the ONet. For a packet to gain access to the ONet, it must first traverse the EMesh to reach the hub of a cluster. For a core to receive a packet from the optical crossbar, it is sent from the hub to cluster-level electrical broadcast trees, the *BNet*, so that it is seen by all the cores in the cluster or sent through the EMesh to reach the destination core.

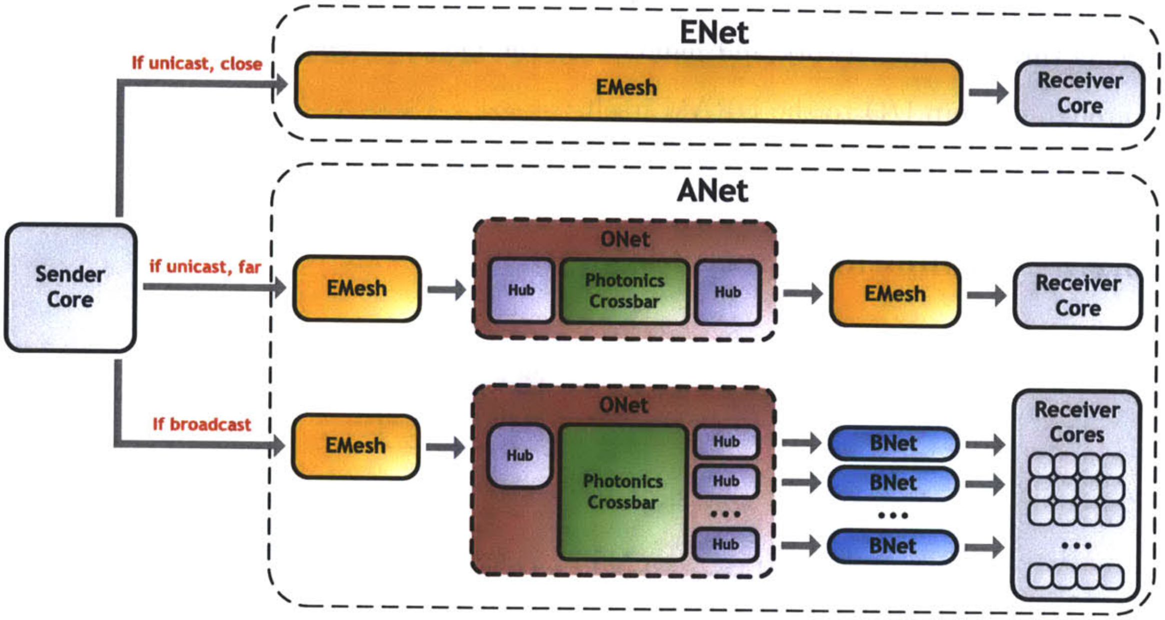


Figure 5-7: Packet Traversal Paths in the ATAC Network

When a packet is sent by a core, there are several different paths (Figure 5-7) it can take to reach its destination. If the packet is a unicast packet, it has the choice of using the *ENet*, the path that uses only the undering EMesh, or the ANet, the path that uses the photonics crossbar. This choice is made on a per-packet basis, with the architecture preferring the ENet for short-distance communcation and the ANet for long-distance. If a packet is broadcast packet, it is send using the ANet, using the inherent broadcast capabilities of both the ONet and BNet to quickly fan out and reach all recipients.

As discussed in [36], the ATAC architecture supports *ACKwise*, a modified limited directory-based cache coherence protocol leveraging the broadcast capabilities of the architecture. *ACKwise* implements a directory distributed across all cores with each core being the *home* to a set of addresses. Each entry in the directory is responsible for keeping track of up to k sharers or set a global bit if the number of sharers exceeds k. If an exclusive access request is made to for entry where the global bit is set, a broadcast is sent across the network to invalidate all sharers. For accesses to off-chip DRAM, the ATAC architecture assumes a dedicated memory controller per cluster, for a total of 64 memory controllers in the 1024-core case. DRAM access requests are made through the network and handled by the memory controller before being sent through off-chip I/O to the DRAM itself.

5.2.2 Evaluation Methodology

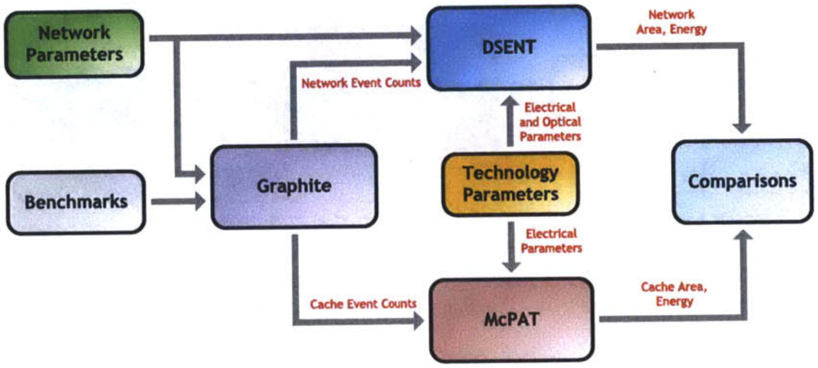


Figure 5-8: ATAC Network Evaluation Methodology

The evaluation methodology, shown in Figure 5-8, is similar to that of the previous section, albeit with a few differences. Instead of synthetic random traffic, we use five different applications from the SPLASH-2 [69] benchmark suite. We also use the distributed multi-core Graphite [45] simulator, as it is a lot faster than GEM5 at running 1024-core simulations. DSENT is used for modeling the network and McPAT/CACTI [38] for the caches. While executing each application, Graphite outputs event counts for the network and caches which are fed into DSENT and McPAT to produce power estimates for their respective components. We assume the 11 nm technology node, as it is the projected node in which 1024 reasonably powerful cores can possibly fit on a single die. The full set of electrical and optical technology parameters is shown in Table 5.2. Performance-related parameters are shown in Table 5.3.

Table 5.2: Technology Parameters for ATAC Network Evaluation Refer to Section 5.1.1 for source of the 11 nm models.

Parameters	**Values**
Process Node	11 nm Tri-Gate
Process Supply Voltages	0.6 V
Temperature	340 K
Laser Efficiency (P_{laser}/P_{elec})	30 %
Wavelength Band	1300 nm
Ring Free Spectral Range	4 THz
Waveguide Pitch	4 µm
Waveguide Loss	0.2 dB/cm
Waveguide Non-linearity Limit	30 mW
Ring Through Loss	0.0001 dB
Ring Drop Loss	1.0 dB
Ring Area	200 μm^2
Ring Heating Efficiency	0.1 K/µW
Ring Tuning Efficiency	10 GHz/K
Ring Electrically Tunable Range	50 GHz
Photodetector Responsivity	1.1 A/W
Photonic Link Configuration	64 λ at 1 Gb/s per λ

5.2.3 Performance Comparison of Network and Coherency Protocols

For the network, we compare ATAC against two separate electrical baselines. The *mesh-pure* network is just a simple electrical mesh network with no native hardware support for broadcasts. The *mesh-bcast* network implements flit multi-casting at each router and can thus broadcast packets much more quickly. For the protocol, we compare *ACKwise* against a traditional directory-based protocol (*Dir-NB*) and a directory-based protocol that uses broadcasts once the capacity of the sharer list is exceeded (*Dir-B*). The main difference between *Dir-B* and *ACKwise* is that *ACKwise* requires acknowledgements from only the actual sharers whereas *Dir-B* needs acknowledgements from all the cores in the system as it does not keep track of the number of sharers once the sharer list reaches capacity.

Figure 5-9 shows the run-time of each benchmark application across the different combinations of networks and coherency protocols. In all benchmarks, the *Dir-NB* protocol fares poorly, as it must restrict the total number of sharers to the capacity of

Table 5.3: Performance Parameters for ATAC Network Evaluation

Common Parameters	**Values**
Clock Frequency	1 GHz
Core Type	In-Order
L1 Cache Size per Core	32 KB
L2 Cache Size per Core	256 KB
Flit Size	64 bits
Baseline Mesh Parameters	**Values**
Router Delay	1 Cycle
Link Delay	1 Cycle
ATAC Parameters	**Values**
EMesh Router Delay	1 Cycle
EMesh Link Delay	1 Cycle
ONet Delay	3 Cycles
BNet Delay	1 Cycle

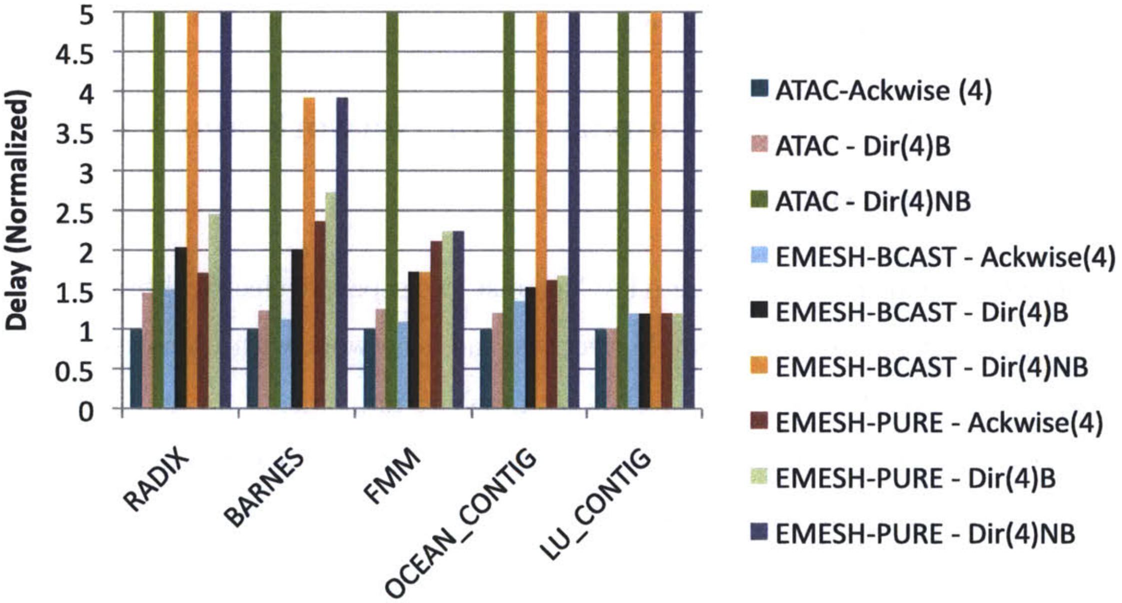
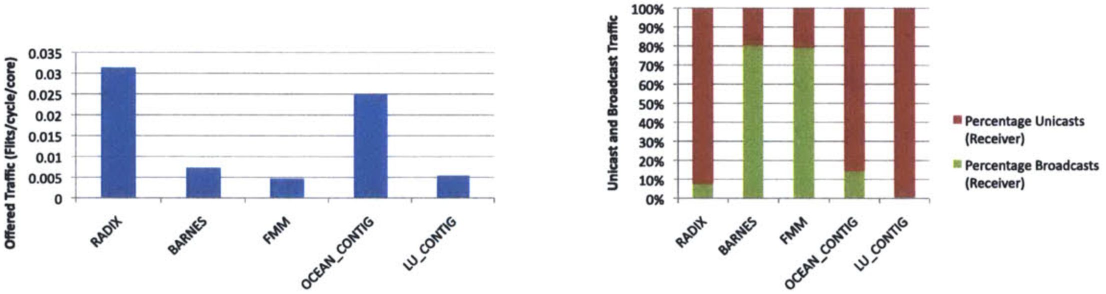


Figure 5-9: Runtime Comparison of Cache Coherency Protocols – We fix the number of sharers for each protocol at 4.

the sharer list. The *Dir-B* and *ACKWise* protocols perform far better, with *ACKWise* taking a significant lead as it does not have to collect unicast acknowledgements from all 1024 cores. Overall, the ATAC network performs noticeably better than the two mesh baselines using the *Dir-NB* or *ACKWise* protocols due to the low-latency broadcast capabilities of the ANet.

5.2.4 Application-Driven Energy Comparison

While ATAC demonstrates clear performance benefits over the baseline mesh networks, it is not necessarily more energy efficient. In Figure 5-10, we see that all five SPLASH-2 benchmarks push very few flits through the network. From the insights we developed in Section 5.1, non-data-dependent laser and ring tuning power can quickly overwhelm the data-dependent energy advantage of photonics at low utilizations. For this reason, we consider two additional photonics technology options – a power-gated laser and fully athermalized rings with trimming – as means to mitigate non-data-dependent power.

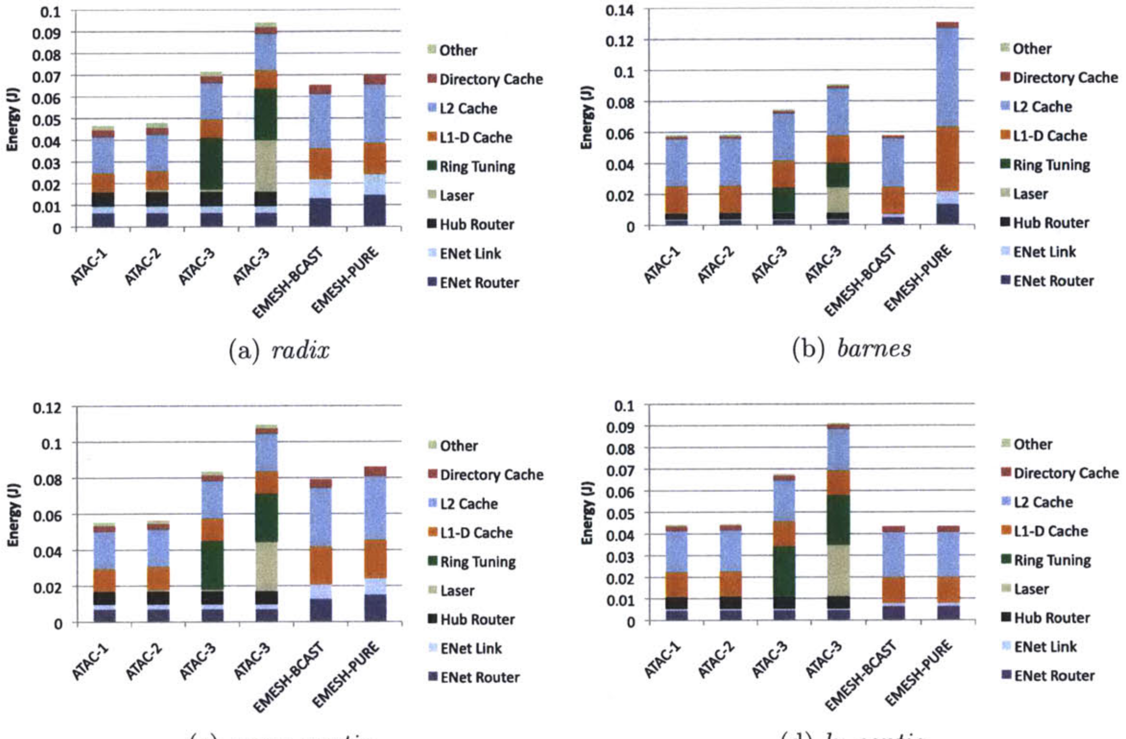


(a) Average flit injection rate

(b) Percentage of bits received from broadcasts

Figure 5-10: Benchmark Network Utilization Statistics for ACKWise – The percentage of bits received from broadcasts is shown in (a) for each benchmark. The average flit injection rate of each benchmark is shown in (b). All results are shown for the *ACKWise* protocol with 4 sharers. Note that the second (right) *ATAC-3* column is mistakenly labeled, and should instead be *ATAC-4*.

Using the additional technology options, we split the ATAC architectures into four different variants. The *ATAC-4* variant represents the vanilla ATAC architecture, with an always-on laser and the electrically-assisted RRW tuning strategy outlined in Section 3.2. The *ATAC-3* design assumes a laser that is power-gated on a packet-by-packet basis, turning the laser into a data-dependent power component. *ATAC-2* assumes both a power-gated laser and athermalized ring resonators with post-process trimming, such that tuning power is eliminated. We include *ATAC-1*, which uses idealized photonics, as another comparison point.



(a) *radix*

(b) *barnes*

(c) *ocean contig*

(d) *lu contig*

Figure 5-11: Benchmark Energy Consumption Breakdowns – This shows total energy amount of energy consumed by the cache/network to run the benchmark. Each field includes both data-dependent and non-data-dependent sources of energy consumption. Not shown is the *fmm* benchmark which has a similar profile as *barnes*. All results are shown for the *ACKWise* protocol with 4 sharers.

Figure 5-11 shows the results of our evaluation. Comparing the *ATAC-3,4* topologies with *ATAC-1,2*, we can see that, as expected, the non-data-dependent components of power dominate in these benchmarks. The large reduction in energy when laser power-gating is applied (*ATAC-3* vs. *ATAC-2*) is largely attributed to the prolonged idle periods in which laser can be turned off to save power. The miniscule difference in energy between idealized photonics (*ATAC-1*) and a laser power-gated athermal design (*ATAC-2*) further highlights the underutilization of the laser.

We would like to point out that *none* of the ATAC network variants are favored over the mesh baselines from a pure network energy/bit perspective. It is thus quite interesting to see that *ATAC-1,2* can hold a sizable advantage in total application energy across several benchmarks. The cause of this apparent pheonomenon is the fact that application energy usage is correlated with run-time through the presence non-data-dependent power; the longer the application takes to run, the greater the idle-time of its various components. This effect is most clearly seen in the *radix* benchmark, where ATAC holds a notable run-time lead over the electrical baselines. Though the number of cache accesses (which is data-dependent energy) is identical across all designs, the caches in *mesh-pure* and *mesh-bcast* spend longer time idling. As a result, significantly more leakage and clocking energy is burnt, giving *ATAC-1,2* a clear total energy lead. Without a significant lead in run-time, such as the case with *lu contig*, ATAC's total application energy trails that of the baselines.

The main takeaway of this study is not necessarily to present or showcase the ATAC architecture as an optimal design which it certainly is not but to demonstrate that photonics does not necessarily have to win in raw network energy/bit to remain beneficial. *Total* system power and performance are the real figures of merit that photonics-inspired designs improve upon.

5.3 Network Comparison Summary

Using the models and evaluation methodologies developed in Chapters 3 and 4, we quantified the costs and benefits of various intra-chip networks. Our results show that, from a network energy/bit perspective, optical interconnects must scale aggressively to remain competitive with their scaled-electrical counterparts and highlight parasitic capacitance reduction and improved electrically-assisted tuning as key areas of improvement. Next, we use our evaluation framework to perform an application-driven network comparison featuring the previously proposed ATAC architecture. We show that in the context of a full system, a photonics-inspired architecture can lose in interconnect energy per bit yet still come out ahead overall in total system energy.

Chapter 6

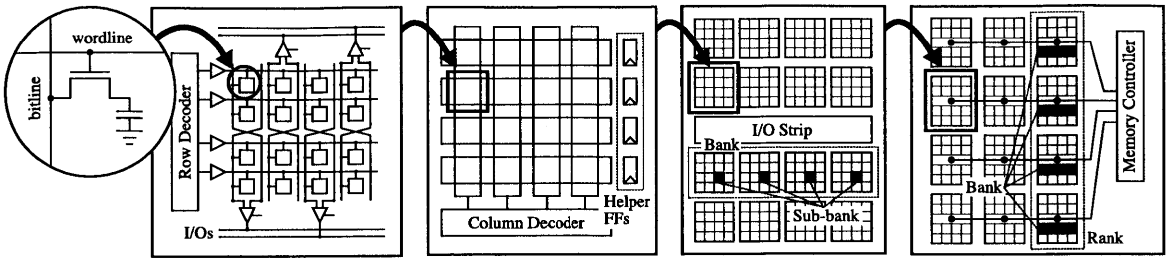
Photonically Interconnected DRAM

In the previous chapter, our evaluations and comparisons encompassed that of the on-chip core-to-core interconnect. As raw throughput of manycore machines continues to scale with Moore's Law, off-chip memory bandwidth must also scale proportionally to prevent a bottleneck. In this chapter, we extend the exploration of photonic interconnects into the chip-to-chip domain through the design of a photonically interconnected DRAM (PIDRAM) system. Using the modern DRAM architecture as a baseline, we redesign both the off-chip and cross-chip interconnect as well as improve bank access efficiency using capabilities afforded by monolithic photonics.

6.1 DRAM Technology

In order to provide fast and energy-efficient access to billions of storage cells while remaining tileable and expandable, modern DRAM systems employ multiple levels of hierarchy. Figure 6-1 shows the structure for contemporary DRAM architectures.

As DRAM chips are considered commodity parts, chip-level organization of a DRAM chip is driven primarily by cost. As such, DRAMs typically use a minimal number of metal wiring layers to reduce wafer and mask costs, contains cell array redundancy to improve yields, and packs as many bits as possible in a given die area.



(a) Array Core (b) Array Block (c) Chip (d) Channel

Figure 6-1: Hierarchical View of a DRAM Memory System

At the bottom of the DRAM hierarchy lies the DRAM *cell*, consisting of a simple access transistor and a storage capacitor that holds a stored bit. As opposed to SRAM, capacitative storage is dynamic and contains no mechanism for regenerating the stored value. Data retention during prolonged idle periods is achieved through periodic refreshes, where cells are read out and rewritten to manually regenerate the stored charge. As such, the access transistor is typically a super high threshold voltage device designed to minimize storage cell leakage during idle periods.

DRAM cells are arrayed in two-dimensional grids and combined with peripheral access circuitry to form an *array core* (Figure 6-1(a)). Each row of cells in an array core shares a wordline driven by peripheral wordline drivers. Wordline voltages are typically boosted above the nominal chip supply due voltage to achieve strong turn-on of the high threshold access transistor. Each column of cells shares a bitline leading to a sense-amplifier found at the edge of the array core. As the bitline capacitance is larger than the capacitance of a storage cell, the sense amplifiers are needed to regenerate the low-swing signal on the bitline during cell reads. On activation of an array core, every cell in the activated row is read by the sense amplifiers. However, only a few bits will be transferred over the array core I/O lines during a column access. Before a different row can be activated, all actived bits are written back into their respective storage cells and the bitlines are precharged to ready them for the next activation. As such most of bits read from a row are never accessed before a different row is activated, except under specific workloads. Due to intrinsic capacitances, array cores are also limited to a modest size that grows very slowly with respect to

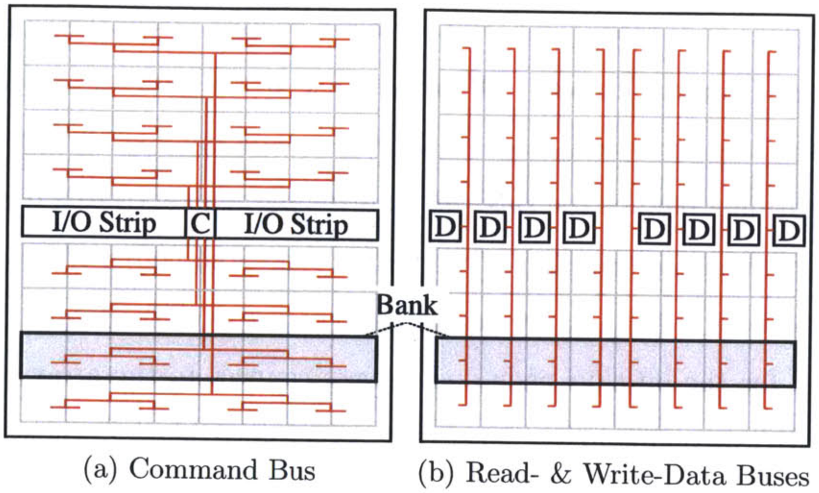


Figure 6-2: Organization of a Modern DRAM Chip – The command bus (a) broadcasts control bits from the command I/O pins to all array blocks on the chip. The read- and write-data (b) buses and array blocks are bit-sliced across the chip to match the locations of the I/O pins. (C = command I/O pins, D = off-chip data I/O pins, on-chip electrical buses shown in red). The example shown has eight banks with eight array blocks each.

technology scaling. For this chapter, we assume a a folded bitline [28, 66] array-core architecture consisting of 512 wordlines and 1024 bitlines, with an individual DRAM cell area of $8F^2$ where F is the feature size. However, our general assumptions are also valid for the open bitline ($6F^2$) and vertical access transistor ($4F^2$) architectures, which are expected to become prevalent once improved DRAM process control is demonstrated [66].

An *array block* is a group of array cores that share circuitry such that only one of the array cores is active at a time (Figure 6-1(b)). Each array core shares its sense-amplifiers and I/O lines with the array cores physically located above and below it, and the array block provides its cores with a global predecoder and shared helper flip-flops for latching data signals entering or leaving the array block. As only one array core can be enabled at a time within an array block, the access width of an array block is equivalent to the number of I/O lines from a single array core.

A *bank* is an independently controllable unit that is made up of several array blocks working together in lockstep (Figure 6-1(c)). The number of array blocks per bank sets the bank's access width. Array blocks from the same bank do not need to be placed near each other, and they are often striped across the chip pitch match data I/O pins. On a bank access, all array blocks of the bank are activated, each of which activates one array core, each of which activates one row. The set of activated array cores within a bank is the *sub-bank* and the set of all activated rows is the *page*.

A *chip* contains multiple banks that time-share the chip's I/O pins to reduce overheads and help hide bank latencies times (Figure 6-1(c)). Figure 6-2 shows how the I/O strip for the off-chip pads and drivers connect to the array blocks of each bank. The DRAM command bus must reach every array block in the chip, so a gated hierarchical H-tree is used to distribute control and address information from the centralized command pins in the middle of the I/O strip (Figure 6-2(a)). The read- and write-data buses are striped across the chip such that each array blocks in each column connects to the same data bus pin in the I/O strip (Figure 6-2(b)).

A *channel* consists of a collection of banks distributed across one or more DRAM chips (Figure 6-1(d)). A memory controller manages a channel, which consists of three logical buses: the command bus, the read-data bus, and the write-data bus. Often, the read-data and write-data buses share the same set of physical wires. To increase bandwidth, multiple DRAM chips are often ganged in parallel as a *rank*, with a slice of each bank present on each chip. To further scale bandwidth, the system can have multiple memory channels. To increase capacity, multiple ranks can be placed on the same channel, but with only one accessed at a time.

6.2 PIDRAM Memory System

Figure 6-3 illustrates our proposed PIDRAM memory system. The system implements multiple independent PIDRAM memory channels. Each channel is managed by a PIDRAM memory controller on the processor-side and connects with a set of PIDRAM chips found on a PIDRAM DIMM. PIDRAM memory controllers are distributed across the processor chip and connect to cores through an on-chip network, which may also be photonic [25, 36, 65, 70]. In the rest of this chapter, we will focus on the implementation of a single PIDRAM memory channel.

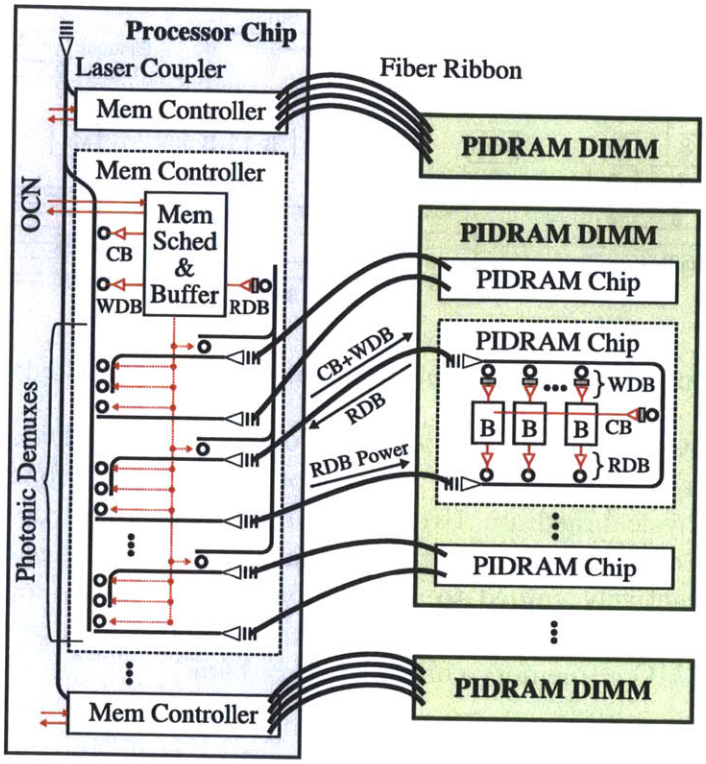
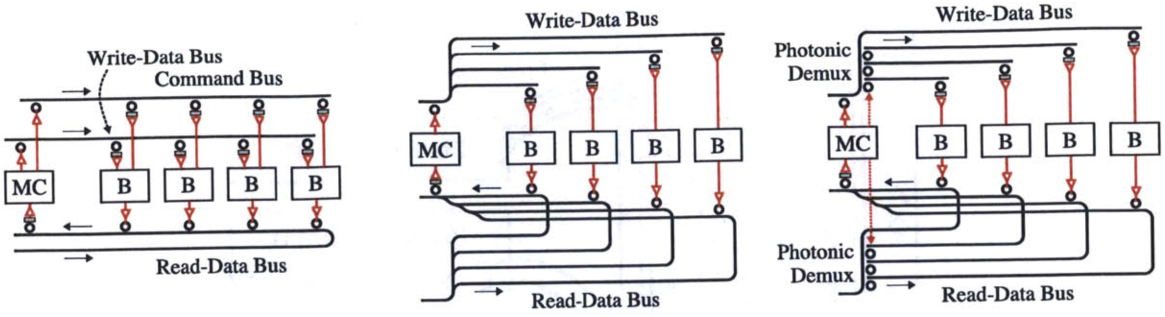


Figure 6-3: PIDRAM Memory System – Each PIDRAM memory channel connects to a PIDRAM DIMM via a fiber ribbon. The memory controller manages the command bus (CB), write-data bus (WDB), and read-data bus (RDB), which are wavelength division multiplexed onto the same fiber. Photonic demuxes guide power to only the active PIDRAM chip. (OCN = on-chip network, B = PIDRAM bank, each ring represents multiple rings for multi-wavelength buses, on-chip electrical wiring shown in red)

6.2.1 PIDRAM Channel Organization

Figure 6-4 illustrates three ways to implement the three logical buses of a memory channel – command bus, write-data bus, and read-data bus – using the available photonic building blocks. For now we assume that a PIDRAM bank never needs to be distributed across multiple chips, an assumption which we will revisit later.

Figure 6-4(a) shows a *shared photonic bus*. This is logically identical to a standard electrical bus and similar to existing photonic DRAM proposals [19, 65]. In this implementation, the memory controller first broadcasts a command to all banks and each bank determines if it is the target bank for the command. On a PIDRAM write command, the target bank will tune-in its photonic receiver on the write-data



(a) Shared Photonic Buses (b) Split Photonic Buses (c) Guided Photonic Buses

Figure 6-4: Photonic Implementations of Command, Write-Data, and Read-Data Buses – (a) *shared photonic buses* where optical power is broadcast to all banks along a shared physical medium, (b) *split photonic buses* where optical power is split between multiple direct connections to each bank, and (c) *guided photonic buses* where optical power is actively guided to a single bank. For clarity, command bus is not shown in (b,c), but it can be implemented in a similar fashion as the corresponding write-data bus. (MC = memory controller, B = bank)

bus while all other banks tune out. The memory controller places write data on the write-data bus, to be received and written by only the target bank. The interaction of the command and write-data bus resembles the single-writer multiple-reader buses described found in [34, 70]. For a PIDRAM read command, just the target bank will perform the read operation and then use its modulator on the read-data bus to send the data back to the memory controller. The read-data bus resembles the multiple-writer single-reader buses of [65]. However, the memory controller schedules the read-data bus to avoid any need for global arbitration.

At first glance, the shared photonic bus seems attractive since, when the bus is active, all of the optical laser power is fully utilized. Unfortunately, the losses quickly add up and make the optical laser power an *exponential* function of the number of banks. If all of the banks are on the same PIDRAM chip, then the losses can be manageable. However, to scale to larger capacities, we will need to "daisy-chain" the shared photonic bus through multiple PIDRAM chips [65]. Each PIDRAM chip adds two coupler losses, waveguide losses, and extra ring losses, making the shared photonic bus feasible only for connecting banks within a PIDRAM chip as opposed to connecting banks across PIDRAM chips.

Figure 6-4(b) shows an alternative implementation that we call a *split photonic bus*, which divides the long shared bus into multiple branches. In the command and write-data bus, modulated laser power is still sent to all receivers, and in the read-data bus, laser power is still sent to all modulators. The split nature of the bus, however, means that the total laser power is roughly a *linear* function of the number of banks. If each bank was on its own PIDRAM chip, then we would use a couple of fibers per chip (one for modulated data and one for laser power) to connect the memory controller to each of the PIDRAM chips. Each optical path in the write-data bus would only traverse one optical coupler to leave the processor chip and one optical coupler to enter the PIDRAM chip regardless of the total number of banks. This implementation breaks the exponential relation between laser power and the number of banks at the cost of increasing the number of fibers. In this example with four banks per bus, we will need to use four waveguides/fibers where we only needed one in the shared photonic bus. The significant bandwidth density of our photonic technology can potentially make this a reasonable tradeoff.

To further reduce the required optical power, we introduce the the concept of a *guided photonic bus*, shown in Figure 6-4(c), which uses optical power guiding in the form of photonic demultiplexers to actively direct power to just the target bank. Each photonic demultiplexer uses an array of either ring or comb filters, and these filters are actively tuned by the memory controller to guide light down the desired branch, leaving the unused branches dark. For the command and write-data bus, the photonic demultiplexer is placed after the modulator to direct the modulated light to the target bank. For the read-data bus, the photonic demultiplexer is placed before the modulators to allow the memory controller to manage when to guide the light to the target bank for modulation. Since the optical power is always guided down a single branch, the total laser power is roughly *constant* and independent of the number of banks. The optical loss overhead from the photonic demultiplexers and the reduced bandwidth density from the branching make the guided bus approach most attractive when working with large per-bank optical losses.

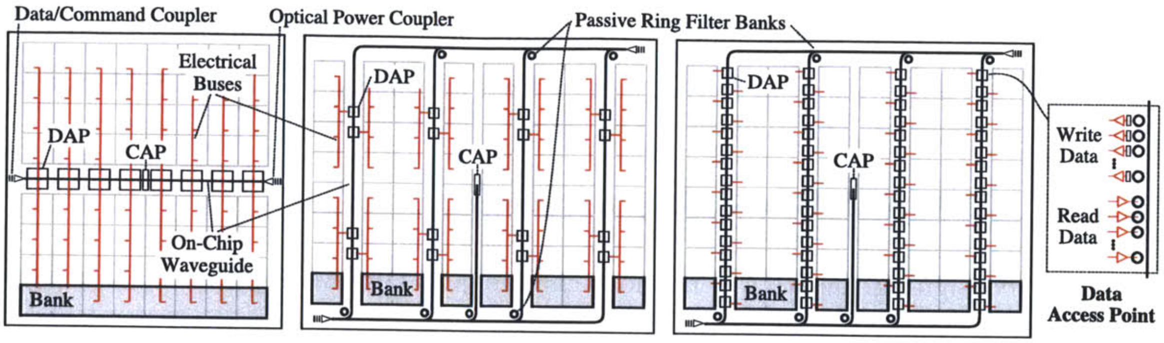
We use a hybrid approach to implement each of the three logical buses. The memory scheduler within the memory controller orchestrates access to each bus to avoid conflicts. The command bus is implemented with a single wavelength on a guided photonic bus. The command wavelength is actively guided to the PIDRAM chip containing the target bank. Once on the PIDRAM chip, a single receiver converts the command into the electrical domain and broadcasts the command to all banks on the chip. Very high bandwidth channels supporting many banks may require additional command wavelengths to ensure sufficient command bandwidth. Both the write-data and read-data buses implement the guided photonic bus to actively guide optical power to a single PIDRAM chip of a channel and the shared photonic bus to distribute the data once within the chip. For economic and packaging reasons, we assume each PIDRAM chip will only have two fibers: one for the three buses and one for the unmodulated read-data optical power. Figure 6-3 illustrates how the command, write-data, and read-data buses are wavelength division multiplexed onto the same fiber. Routing the read-data optical power through the processor chip is necessary for a guided photonic bus implementation so that the each photonic demultiplexer can be positioned within its appropriate memory controller.

6.2.2 PIDRAM Chip Organization

Previously, we motivated guided photonic buses to implement the inter-chip portion of the command, write-data, and read-data buses while using shared photonic buses for the intra-chip portion. There is an important design trade-off in terms of how much of the on-chip portion of these buses should be implemented optically versus electrically. This design choice is primarily driven by trade-offs in area and power.

Figure 6-5(a) illustrates the approach labeled *P1*, where the electrical I/O strip in Figure 6-2 is replaced with a horizontal waveguide and multiple *photonic access points*. Each photonic access point converts the corresponding bus between the optical and electrical domains. The on-chip electrical H-tree command bus and vertical electrical data buses remain as in traditional electrical DRAM shown in Figure 6-2.

Figures 6-5(b) and 6-5(c) illustrate our approach for bring photonics deeper into



(a) P1 Floorplan (b) P2 Floorplan (c) P8 Floorplan

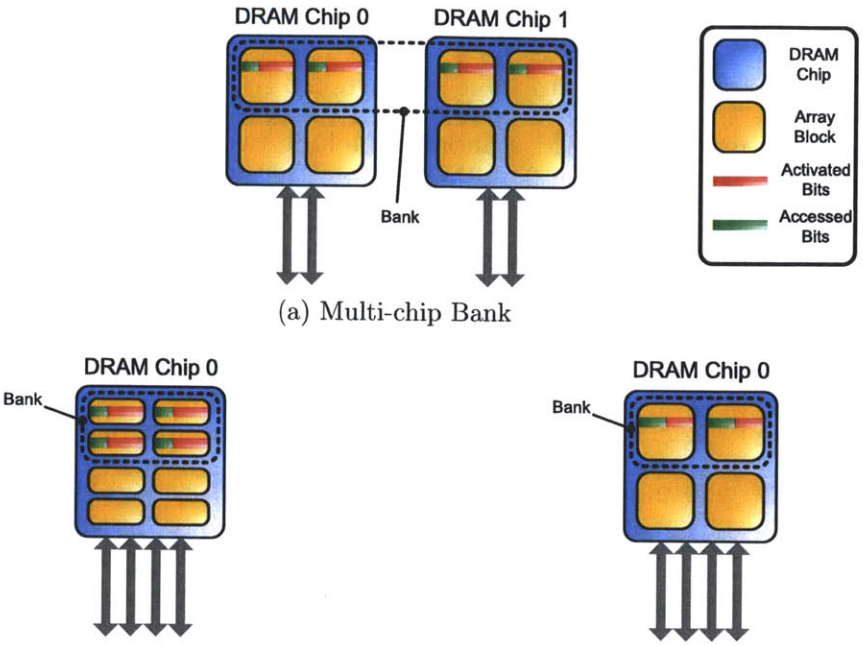
Figure 6-5: PIDRAM Chip Floorplans – Three floorplans are shown for an example PIDRAM chip with eight banks and eight array blocks per bank. For all floorplans, the photonic command bus ends at the command access point (CAP), and an electrical H-tree broadcasts control bits from the command access point to all array blocks. For clarity, the on-chip electrical command bus is not shown, but it is similar to that shown in Figure 6-2(a). The data buses shown in the floorplans gradually extend the photonics deeper into the PIDRAM chip: (a) *P1* uses photonic chip I/O for the data buses but fully electrical on-chip data bus implementations, (b) *P2* uses seamless on-chip/off-chip photonics to distribute the data bus to a group of four banks, and (c) *P8* uses photonics all the way to each bank. (CAP = command access point, DAP = data access point, on-chip electrical buses shown in red)

the chip to improve intra-chip energy-efficiency. We use a *waterfall floorplan*, where the waveguides are distributed across the chip. The horizontal waveguides contain all of the wavelengths and the optically passive ring filter banks at the top and bottom of the waterfall direct different subsets of wavelengths into each vertical waveguide. Each of these vertical waveguides is analogous to the electrical vertical buses in *P1*, so a bank can still be striped across the chip horizontally to allow easy access to the on-chip photonic interconnect. Various waterfall floorplans are possible that correspond to more or less photonic access points. For a *Pn* floorplan, *n* indicates the number of partitions along each vertical electrical data bus. All of the photonic circuits have to be replicated at each data access point for each bus partition. This increases the fixed link power due to link transceiver circuits and ring heaters. It can also potentially lead to higher optical losses, due to the increased number of rings on the optical path. In Section 6.3 we further evaluate these trade-offs. Our photonic floorplans all use the same on-chip command bus implementation as traditional electrical DRAM: a command access point is positioned in the middle of the chip and an electrical H-tree command bus broadcasts the control and address information to all array blocks.

6.2.3 PIDRAM Bank Organization

After redesigning the DRAM memory channel and DRAM chip to use an energy-efficient photonic interconnect, the next limiting factor is the power consumed activating bits in the banks themselves. In traditional electrical DRAM memory systems, banks are almost always distributed across multiple chips (Figure 6-6a). This is because they are pin-bandwidth limited to a few bits per bus clock cycle. For example, to obtain all 512 bits needed for a 64-byte cache line, a bank might be striped across eight different chips and each access activates all chips in parallel. Unfortunately, this leads to large page sizes and wasted energy as most of the activated page is unused. Alternatively, electrical DRAM memory systems could activate just a single chip and wait for the entire cache line to stream out. However, the limited pin bandwidth per chip will significantly increase serialization latency. A PIDRAM memory channel supports a much higher chip I/O bandwidth. A single data fiber or waveguide can provide 80 GB/s in each direction, enabling an entire cache line to be fetched from a single chip without incurring significant serialization latency. As such, we propose to localize a bank (and an entire cache-line access) to a single PIDRAM chip. As only a single chip is activated per access, this can reduce the page size by a factor of eight or more, resulting in significant energy savings as fewer total bits are activated per access.

For a single bank within a chip to support an access width of an entire cache line, we must either increase the number of array blocks so that more array cores can be accessed in parallel (Figure 6-6b), or increase the number of bits accessed in each array core (Figure 6-6c). Increasing the number of array blocks can signficantly degrade area efficiency due to less amortization of shared circuitry area overheads. Furthermore, the number of activated bits (page size) increases if more array cores are accessed in parallel, which negatively impacts energy efficiency. This can be mitigated by reducing the row size of each array core, which will again cost area. Increasing the



(a) Multi-chip Bank

(b) Single-chip Bank w/ More Array Blocks (c) Single-chip Bank w/ More Array Core I/Os

Figure 6-6: DRAM Bank Organizations - Each DRAM bank can be striped across multiple chips (a) or contained entirely within a single-chip (b, c). To preserve the access width when going to a single chip, we must either increase the number of array blocks (b) or increase the I/O width of each array core (c).

number of bits accessed per array core is the more preferable approach. Though it also requires area to fit wiring pitches for additional array core I/O lines, it reduces the number of array blocks that need to be accessed in parallel, improving the ratio of accessed bits to activated bits.

We note that these same strategies can be applied to existing electrically-interconnected DRAMs. However, there is currently little incentive to make this change because the energy savings within the bank are dwarfed by the electrical inter-chip and intra-chip interconnect energy. Even if the energy savings are worth the cost in area, current designs are pin-bandwidth limited and would see no benefit in supporting wide bank access widths due to the increase in serialization latency. The improvements in bandwidth density and energy efficiency realized through photonics are key enablers for single-chip access and reducing page activation energy.

By interleaving accesses to multiple parallel banks, more banks per chip can hide bank busy times and match the bandwidth of the DRAM core with the I/O. However, rapid bank interleaving puts strain on the power delivery network of a DRAM chip as page activations draw high instantaneous current. To reduce the cost of the power delivery network, modern DRAM standards include two timing constraints, t_{RRD} and t_{FAW}, which mandate minimum intervals between activate commands [43]. These constraints are the result of aggressive cost-cutting, and could potentially handicap the benefits of additional banks by limiting the number of activates per unit time. Recent industry focus on improving DRAM core efficiency have yielded designs that reduce t_{RRD} and t_{FAW} significantly [50]. Furthermore, implementing inter-chip communication with photonics frees up a number of pins, which can be used as power pins to improve power delivery. Power drawn per activate is also reduced if the number of array core I/Os is increased, since fewer bits need to be activated.

6.3 Evaluation of a Single PIDRAM Chip

In this section, we compare various PIDRAM configurations and floorplans to a baseline electrical DRAM implementation with the same capacity. This baseline design is labeled *E1* and is similar to that described in Section 6.1. We limit our study to a memory channel consisting of a single PIDRAM chip, and we ignore pin bandwidth-density constraints for the electrical baseline. In the next section, we will explore scaling to multi-chip PIDRAM memory channels.

6.3.1 Evaluation Methodology

To evaluate the energy efficiency and area tradeoffs of proposed PIDRAM architectures, we use heavily modified DRAM models of the *CACTI* modeling tool [7]. Though we were able to use some of *CACTI*'s original models for details such as decoder sizing, gate area calculations and technology parameter scaling, the design space we explored required a complete overhaul of *CACTI*'s assumed DRAM organization and hierarchy. To this end, we built our own architectural models for the DRAM core, from circuit-level changes at the array core level and array block level, to the higher bank organization as shown in Figure 6-1. To validate our electrical models, we tested them against known points for a range of processes and configurations. In addition to covering the baseline electrical DRAM design, we accounted for the overhead of each relevant photonic design in our models and developed a comprehensive methodology for calculating the power and area overheads of off-chip I/O for both the electrical and photonic cases of interest. We note that since we model our architecture in *CACTI* as opposed to DSENT, we move back to fixed energy costs and instead bound our design space with both aggressive and conservative projections for photonic devices. All energy and area calculations presented are for a 32 nm DRAM process.

Table 6.1: Photonic Device Parameters for PIDRAM Evaluation – Based on coupler designs in [63], waveguide losses from [17], filter designs in [52] as well as our preliminary photonic transceiver test chips and ongoing device work.

Photonic Device Parameters	**Aggressive Value**	**Conservative Value**
Coupler loss	0.5 dB	1 dB
Splitter loss	0.2 dB	0.2 dB
Non-linearity loss at 30 mW	1 dB	1 dB
Modulator insertion loss	1 dB	1 dB
Waveguide loss	2 dB/cm	4 dB/cm
Waveguide crossing loss	0.05 dB	0.05 dB
Filter through loss	1e-4 dB	1e-3 dB
Filter drop loss	1 dB	1 dB
Photodetector loss	1 dB	1 dB
Laser efficiency	30 %	50 %
Receiver sensitivity at 10 Gb/s	-20 dBm	-20 dBm

To quantify the performance of each DRAM design, we use a detailed cycle-level microarchitectural C++ simulator with synthetic traffic patterns to issue loads and stores at a rate capped by the number of in-flight messages. The memory controller converts requests into DRAM commands which are issued based on a round-robin arbitration scheme and various timing constraints based on contemporary timing parameters found in the Micron DDR3-SDRAM data sheet [43]. These timing parameters are also in agreement with our modified *CACTI*. We simulate a range of different designs by varying: floorplan, number of I/Os per array core, number of banks, and the channel bandwidth. We use the events and statistics from the simu-

Table 6.2: Electrical and Photonic I/O Energies fj/bt = average energy per bit-time assuming 50energy includes clock and leakage, thermal tuning energy assumes 20 K temperature range. Electrical I/O projected from an 8 pJ/bt at 16 Gb/s design in a 40 nm DRAM process [37], to a 5 pJ/bt at 20 Gb/s design in a 32 nm DRAM process. Photonic I/O runs at 10 Gb/s/wavelength. Photonic projections based on our own preliminary test chips and ongoing circuit designs.

Data-Dependent I/O Energy	**Aggressive Value**	**Conservative Value**
Electrical Transceiver	1050 fJ/bt	1050 fJ/bit
Optical Receiver	20 fJ/bt	50 fJ/bt
Optical Modulator	40 fJ/bt	100 fJ/bt
Non-Data-Dependent I/O Energy	**Aggressive Value**	**Conservative Value**
Electrical Transceiver	1450 fJ/bt	1450 fJ/bit
Optical Receiver	10 fJ/bt	30 fJ/bt
Optical Modulator	5 fJ/bt	20 fJ/bt
Thermal Tuning Cost Per Ring	16 fJ/bt	32 fJ/bt

lator to animate our DRAM and photonic device models to compute the energy per accessed bit.

We find that for random traffic, a bank with a 512-bit access width has a bi-directional data bandwidth of approximately 10 Gb/s independent of system size, which matches our analytical model. Since each wavelength (λ) has a uni-directional bandwidth of 10 Gb/s, this translates to an equivalent bandwidth of 1/2 λ in each direction under balanced read and write workloads. Accordingly, we find the knee in the curve of sustained random bandwidth versus number of banks occurs when the number of λ per direction is half the number of banks.

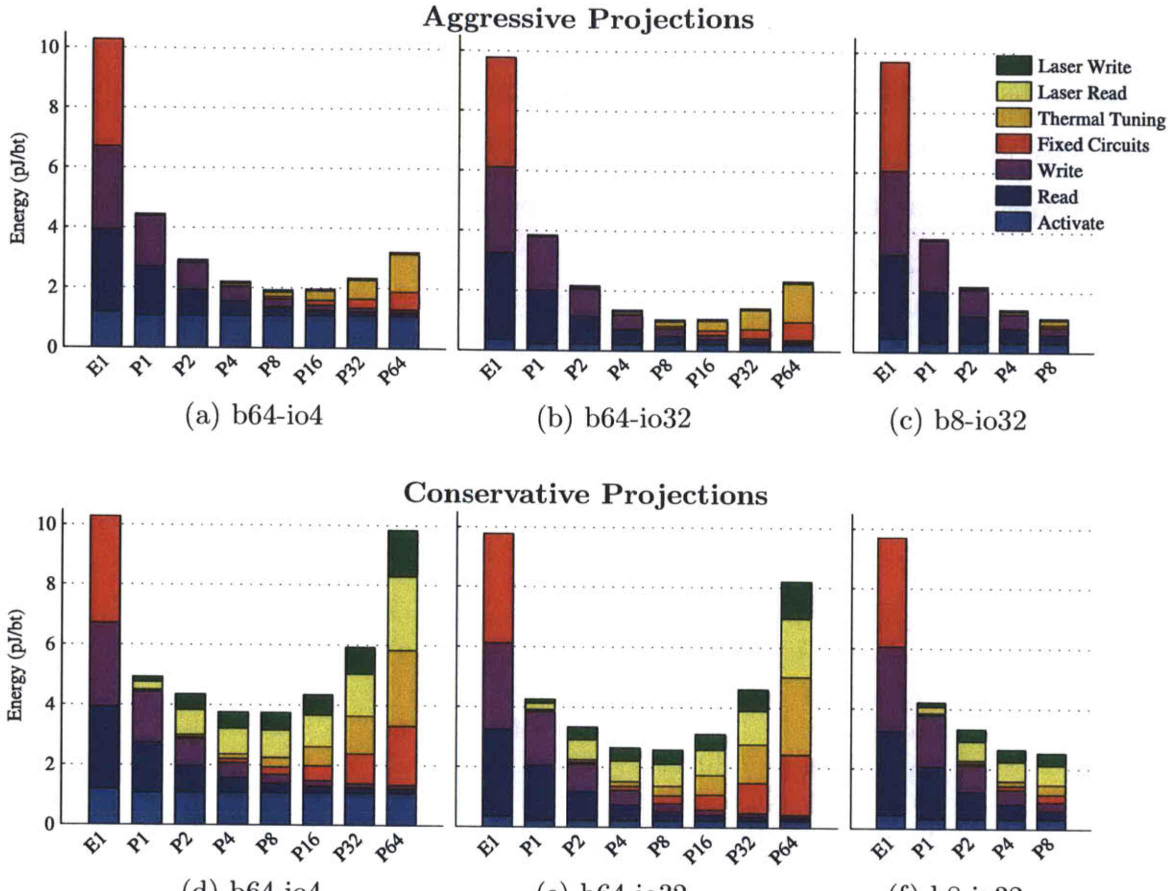
For streaming traffic the effective bank bandwidth is higher, however, we believe random traffic is more representative of expected system traffic in future systems. In the manycore era, even if every core has locality in its access stream, there will be so many of them, that from the point of view of any memory controller, accesses will appear random. An intelligent memory controller could reorder the accesses to re-extract some of the locality, but this is unlikely to scale to many cores. Consequently, we perform most of our design and analysis assuming random traffic.

Latency is not an important figure of merit for this work because we do not expect PIDRAM to affect it significantly. We do not change the array core internals, which sets many of the inherent latencies for accessing DRAM. Moreover, our bank bandwidths are sufficiently sized such that the serialization latency is not significant, especially for random traffic, when compared to the inherent DRAM latencies. As to be expected, as the channel approaches peak utilization, the latency does rise dramatically due to queuing delays.

Table 6.3: Representative Configurations We explored designs consisting of 8, 16, 32, and 64 banks, each with 4, 8, 16, 32 and 64 λ/dir, and 4, 8, 16, and 32 I/Os per array core. All configurations were evaluated for all floorplans possible with that configuration.

Parameter	b64-io4	b64-io32	b8-io32
Access Width	512 bits	512 bits	512 bits
Capacity	8 Gb	8 Gb	8 Gb
Banks	64	64	8
Bandwidth (λ / direction)	32	32	4
I/Os per Array Core	4	32	32
Array Blocks Per Bank	128	16	16
Page Size (Kb)	128	16	16
Floorplans	E1–P64	E1–P64	E1–P8

Although we evaluate hundreds of design points with our methodology, we will limit the rest of this section to the three configurations shown in Table 6.3. These configurations are either optimal or are representative for their given parameters. The *b64-io4* configuration and the *b64-io32* configuration represent high-bandwidth PIDRAM chips and the *b8-io32* configuration represents low-bandwidth PIDRAM chips. The *b64-io4* and *b64-io32* configurations differ in the number of array blocks and array core I/Os and are both included to demonstrate the tradeoff of the number of array core I/Os and number of array blocks for a fixed access width. All of our configurations are for a capacity of 8 Gb, which yields a reasonably sized chip given the 32 nm DRAM process technology. The DRAM-chip access width (bits per request) is 512 bits, which is scaled up from the 64 bits in contemporary DRAM. This is to enable the transfer of a 64-byte cache line from a single chip with a single request.



(a) b64-io4

(b) b64-io32

(c) b8-io32

(d) b64-io4

(e) b64-io32

(f) b8-io32

Figure 6-7: Energy Breakdown for Various Floorplans – The top row (a–c) and bottom row (d–f) assume the aggressive and conservative sets of photonic device parameters found in Table 6.1, respectively. Results for (a,b,d,e) are at a peak bandwidth of ≈500 Gb/s and (c,f) are at a peak bandwidth of ≈60 Gb/s with random traffic. Thermal tuning energy assumes 20 K temperature range, and non-data-dependent circuits energy includes clock and leakage. Read energy includes chip I/O read, cross-chip read, and bank read energy. Write energy includes chip I/O write, cross-chip write, and bank write energy. Activate energy includes chip I/O command, cross-chip row address energy, and bank activate energy.

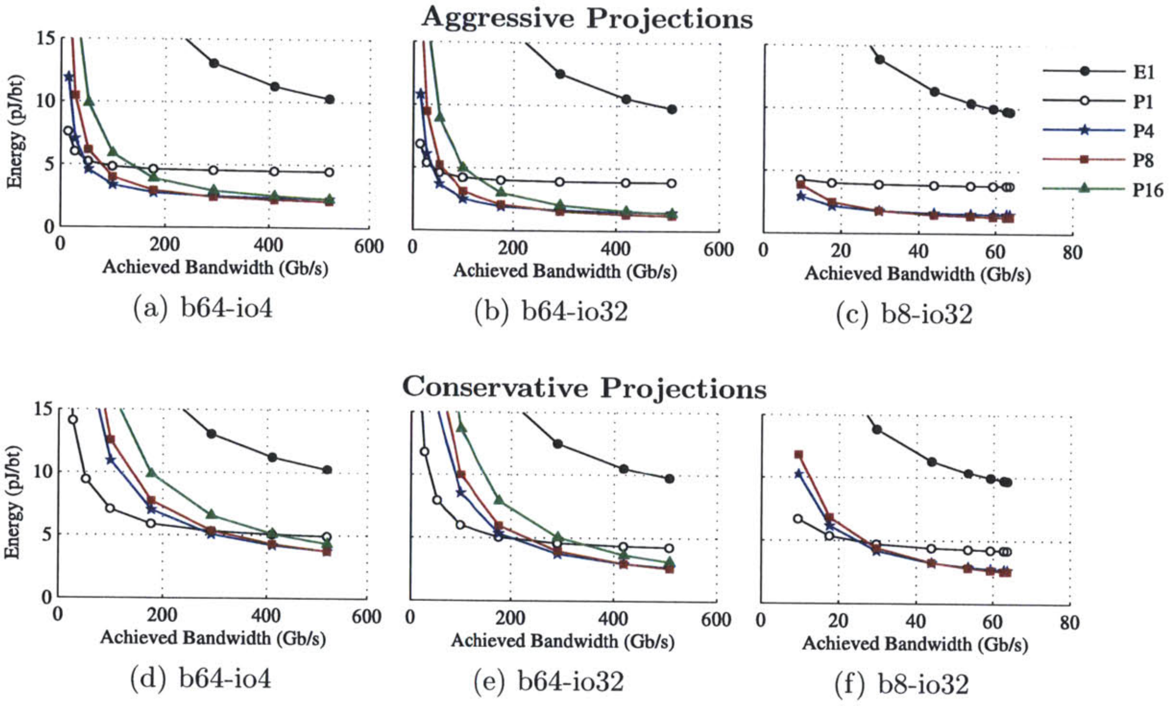
6.3.2 Energy Breakdown

Figure 6-7 shows the energy-efficiency breakdown for various floorplans implementing our three representative PIDRAM configurations. Each design is subjected to a random traffic pattern at peak utilization and the results are shown for the aggressive and conservative photonic technology projections. Across all designs it is clear that replacing the off-chip links with photonics is advantageous, as *E1* towers above the rest of the designs. How far photonics is taken on chip, however, is a much richer design space. To achieve the optimal energy efficiency requires balancing both the data-dependent and non-data-dependent components of the overall energy. For Figure 6-7, the non-data-dependent energy includes: electrical laser power for the write bus, electrical laser power for the read bus, non-data-dependent circuit energy including clock and leakage, and thermal tuning energy. As shown in Figure 6-7(a), *P1* spends the majority of the energy on intra-chip communication (write and read energy) because the data must traverse long global wires to get to each bank. Taking photonics all the way to each array block with *P64* minimizes the cross-chip energy, but requires a large number of photonic access points (since the photonic access points in *P1* are replicated 64 times in the case of *P64*), contributing to the large non-data-dependent component of the total energy. This is due to the non-data-dependent energy cost of photonic transceiver circuits and the energy spent on ring thermal tuning. By sharing the photonic access points across eight banks, the optimal design is *P8*. This design balances the data-dependent savings of using intra-chip photonics with its non-data-dependent overheads.

Once the off-chip and cross-chip energies have been reduced (as in the *P8* floorplan for the *b64-io4* configuration), the activation energy becomes dominant. The benefits of moving to the *b64-io32* configuration, which increases the number of bits we read or write from each array core to 32, can be seen in Figure 6-7(b). The much smaller page size further reduces the activate energy cost, and overall this optimized design is 10× more energy efficient than the baseline electrical design. Figure 6-7(c) shows similar tradeoffs for the low-bandwidth *b8-io32* chip.

Figure 6-7(d f) shows the same designs as Figure 6-7(a c), but for conservative silicon-photonic technology assumptions. Replacing the off-chip links with silicon photonics still helps significantly, but bringing photonics across the chip closer to the array blocks is less of an improvement. This is a consequence of the lossier components which require more laser power. The optimal floorplan still appears to be *P8*, but it has a smaller margin over *P1*. Changing the number of I/Os per array core still proves to be beneficial, but this improvement is diluted.

6.3.3 Energy vs. Utilization



(a) b64-io4 (b) b64-io32 (c) b8-io32

(d) b64-io4 (e) b64-io32 (f) b8-io32

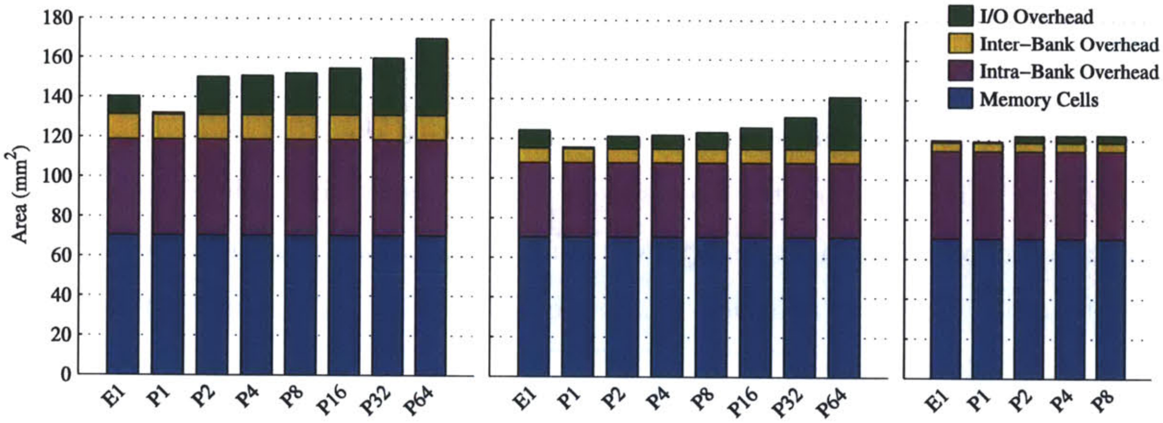
Figure 6-8: Energy vs. Utilization – The top row (a–c) and bottom row (d–f) assume the aggressive and conservative sets of photonic device parameters found in Table 6.1, respectively. To reduce clutter, we plot only the three most energy efficient waterfall floorplans (*P4*, *P8*, *P16*). *P16* is not shown for (c,f) since *b8-io32* only has eight banks.

Although low power at peak throughput is important, a system designer is often just as concerned about energy efficiency at low utilization. For a given design, we scale back the utilization by reducing the number of messages that can be in flight, and the results are shown in Figure 6-8(a–c). As expected, the energy per bit increases as utilization decreases due to the non-data-dependent power components. The non-data-dependent circuit overheads in the transceivers of the electrical baseline are significant enough to result in poor energy-efficiency regardless of utilization.

Designs with a larger fraction of non-data-dependent power will have a steeper slope, and this tradeoff can clearly be seen when comparing *P8* and *P16* to *P1*. The higher numbered *Pn* floorplans do better at high-utilization cases, as the global electrical wires connecting the array blocks to the photonic access points are shorter. However, they do worse than the *P1* floorplan at low utilization because the non-data-dependent ring tuning and idle photonic circuits adds up. Overall, non-data-dependent power punishes systems with under-utilized resources. As such, desired system throughput and utilization will also affect the choice of the optimal design.

Figure 6-8(d–f) shows the effects of less capable photonic devices, which result in a relatively large penalty for low utilization of high-bandwidth systems. This most notably affects the *P4*, *P8*, and *P16* floorplans.

6.3.4 Area



(a) b64-io4 (b) b64-io32 (c) b8-io32

Figure 6-9: Area Breakdown for Various DRAM Designs with Aggressive Photonic Projections – The I/O overhead includes the area costs of circuits that form the access points as well as the area overhead of the waveguides. The inter-bank overhead is the area cost of wires and buffers used to bring bits from the banks to the photonic access points. The intra-bank overhead refers to the area taken up by the peripheral circuitry of each bank (e.g., decoders and sense-amplifiers). The memory cell area is the area taken up by the actual DRAM cells.

Figure 6-9 shows the total area breakdown of each design. Replacing the off-chip links with photonics results in significant area savings (*E1* vs. *P1*) due to the large size of bump-pitch limited electrical off-chip I/Os. Taking photonics deeper on-chip results in the jump in I/O overhead area between the *P1* and *P2* floorplans which can be explained by the move from the single photonic strip in *P1* to the waterfall in *P2* and above. Since we assume a very conservative 50 µm width for each photonic trench in the area calculations, the vertical trenches needed by the waterfall present a noticeable area increase. Interestingly, the losses in the conservative case require the laser power per wavelength to be so high that only 8 wavelengths can be supported per waveguide to stay within the 30 mW nonlinearity limit. This requires 16 waveguides instead of one in *P1* and two per column instead of one for the waterfall of *P2* and above. However, it results in an additional area overhead of less than 1 mm^2 due to the compact waveguide pitch in each trench. A design with a higher number of I/Os per array core (*b64-io32*) yields a more area efficient design than one with few I/Os per array core and many array blocks (*b64-io4*), consistent with our insights in Section 6.2.3.

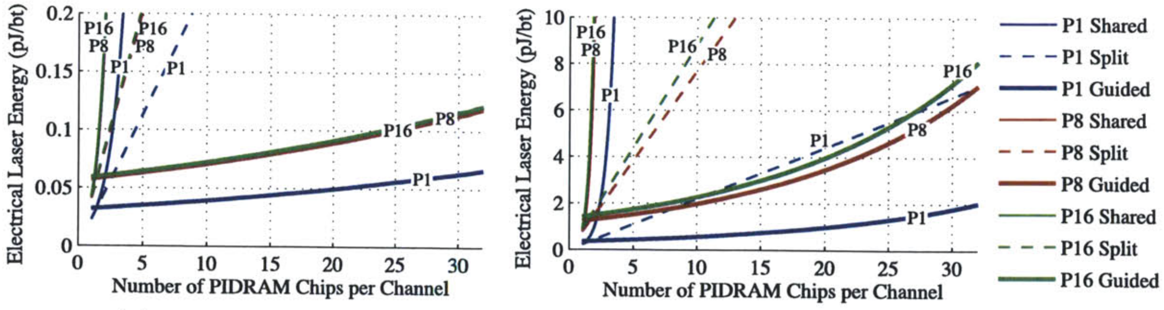
Overall, the most energy efficient design (*b64-io32-P8*) has slightly smaller area than the electrical baseline (*b64-io32-E1*). We note that the I/O area of the electrical baseline is bump-pitch limited and unlikely to scale much with technology, setting a lower bound on I/O area usage given the number of I/O pins. Photonic access points, on the other hand, are relatively small and will continue shrinking with scaling of electrical back-end circuits.

6.4 Scaling PIDRAM Memory Channels

When the number of PIDRAM chips per channel is scaled to increase capacity, the primary concern is the amount of laser power needed to overcome the extra losses that result from the overhead of adding more chips. In this section, we first quantitatively examine the laser power trade-offs between the shared, split, and guided photonic bus approaches, before qualitatively discussing 3D integration as a complementary technique for further capacity scaling.

6.4.1 Optical Power Guiding

For our -20 dBm receiver sensitivity and 30% laser efficiency, an optical path loss in the range of roughly 15 25 dB is needed to keep the background laser power below the link energy cost. With a daisy-chained shared bus approach, the optical loss



(a) b64-io32, Aggressive

(b) b64-io32, Conservative

Figure 6-10: Electrical Laser Power Scaling for Multi-Chip PIDRAM Memory Channels – Laser power increases more slowly with the proposed guided photonic bus implementation versus either the shared or the split photonic bus implementations.

grows exponentially by the loss through a PIDRAM chip (3.5–7 dB aggressive or 7–13 dB conservative, depending on the floorplan) for each additional chip on the channel. With 5.5 dB (aggressive) to 10 dB (conservative) already lost in the memory controller waveguides, couplers, and rings, this approach becomes impractical beyond one or two chips. With 32 *b64-io32-P8* chips sitting on a channel implemented as a shared bus, the optical loss grows to 213 dB and 407 dB for the aggressive and conservative projections, respectively.

The split bus approach fares significantly better than shared bus as the required laser power grows roughly linearly with the number of chips per channel. For a single *b64-io32-P8* chip channel, the optical loss is 12 dB aggressive and 22 dB conservative, and grows to 27 dB and 37 dB when 32 *b64-io32-P8* chips are attached the channel for the aggressive and conservative projections, respectively.

With a guided bus, the laser power is sent only to the necessary chip. The fixed loss in the memory controller increases by 2–3 dB due to the extra power guiding ring and the need to also couple the read path laser in and out of the memory controller. A second increase in the memory controller loss results from the power guiding rings added to the memory controller with each additional chip. More rings along the path means more ring through loss and longer waveguides within the memory controller, amounting to an extra 0.1 dB to 0.3 dB loss for each additional chip. A guided bus channel with 32 *b64-io32-P8* chips has an optical loss of 17 dB and 33 dB for the aggressive and conservative projections, respectively.

Figure 6-10 shows how much the laser power contributes to the overall energy/bit for several floorplans of the *b64-io32* configuration. We can see that the guided bus designs have much more room to scale, as the shared and split bus approaches quickly become infeasible after only a few chips. As expected, designs that do not go as far into the PIDRAM chip consume less power, which makes sense since the PIDRAM chips themselves contribute less loss to the optical critical path. Interestingly, with conservative components, the split bus in Figure 6-10(b) can outperform the guided bus for smaller number of chips per channel, because the loss-overhead of guiding on the memory controller side is bigger than the linear increase in power required for the split bus.

6.4.2 3D Integration

Three dimensional stacking is a complementary technology to photonics. This technology can be used to increase the capacity of PIDRAM memory channels without additional fiber wiring and packaging overhead. We introduce the concept of a PIDRAM cube, which is a collection of stacked PIDRAM dies (e.g., as in [27]), connected electrically by through-silicon vias and optically by vertical coupling in through-silicon via holes.

Stacking can be especially useful for high-capacity systems, where a significant fraction of the fibers would be unused with optical power guiding. By stacking these chips in a PIDRAM cube and adding a second stage of power guiding within the stack, we can reduce the number of packages and fibers in the system while maintaining the same capacity and bandwidth. The first stage of power guiding determines which PIDRAM cube gets the channel, while the second stage determines which die in the cube gets the channel.

With our photonic design, all of the dies in the stack after the base die will be the same, which greatly reduces the manufacturing costs. For example, for a stack of eight die, the generic die needs a total of 16 couplers. Only one in each direction will be active in any given die, and the others will be drilled-out by the TSV holes. Although stacking DRAM chips on top of the CPU die may increase the DRAM chip temperature (and hence refresh power), stacking PIDRAM chips in a cube away from the CPU should have minimal temperature effects on PIDRAM leakage and photonic components as overall PIDRAM chip power dissipation is relatively small.

6.5 Summary

In this chapter, we developed a new DRAM achitecture that leveraged the emerging photonics technology to overcome pin-bandwidth density and energy efficiency limitations of traditional DRAM systems. Simple one-to-one replacement of electrical I/O with photonic I/O resulted in an immediate gain, while the balance between data-dependent and non-data-dependent power components determined the optimal penetration of photonic components for intra-chip links. We showed that the superior bandwidth density of photonics can be used to further enhance the architectural efficiency of DRAM banks, enabling fast single-chip access and a drastically reduced DRAM page size to minimize the number of wasted activated bits. These techniques yielded an approximate 10× improvement in energy per accessed bit over projected electrical-only DRAM chips with matching area footprints. When expanded to multichip solutions using optical power guiding, our proposed PIDRAM system can scale gracefully to meet a wide range of future system capacity and bandwidth demands.

Chapter 7

Conclusion

Integrated photonics is an emerging technology offering many opportunities for relieving the interconnect bottleneck of high-performance processors and architectures. With photonics, the potential for orders-of-magnitude improvement in bandwidth density and energy efficiency creates many technology-driven architectural opportunities. Like any other pieces of new technology, however, its arrival must be greeted with cautious optimism, as its exact costs and benefits are not sufficiently understood or quantified.

To address the lack of visibility into the architectural tradeoffs that govern integrated photonics, this thesis develops a set of models designed to illustrate the interactions between link components and capture the interactions between photonic devices and electrical interface circuits. Identifying thermal tuning as a significant source of power consumption, we outline strategies for a revolving-ring-window tuning scheme and motivate electrically-assisted tuning, allowing tuning power to scale gracefully to a large number of WDM wavelengths. We show that the the three dominant sources of energy consumption in a photonic link   laser, modulator, and tuning   can be optimized by balancing link data-rates, modulator insertion losses, and extinction ratios. The links themselves, however, are only a small component of the overall interconnection network. To capture the dynamics of an entire network under variable technology scenarios, this work presents DSENT, which integrates our link models as part of a full network evaluation framework. Our methodology allows for a portable set of models that scale across a range of technology scenarios and usage cases.

Using our models, we evaluate and compare several case-studies. Our results show that while photonics must continue improving key technology and device parameters to compete with scaled electronics in intra-chip network energy efficiency, photonics-inspired architectures can achieve lower total system energy by improving application run-times, even if the network component becomes more expensive. With our insights, we propose a photonically interconnected DRAM system that tackles the pin bandwidth limitations of existing systems to meet future bandwidth and capacity demands.

The results of this thesis will be used to drive future work at the device-, circuit-, and architecture-level. With a greater understanding of the basic building blocks, the system-level costs and benefits, and the architectural design goals, we take the next step into making integrated electro-optical systems a reality.

Bibliography

[1] N. Agarwal et al. GARNET: A detailed on-chip network model inside a full-system simulator. In *ISPASS*, pages 33 42, Apr 2009.

[2] T. Barwicz et al. Silicon photonics for compact, energy-efficient interconnects. *Journal of Optical Networking*, 6(1):63 73, Jan 2007.

[3] Chris Batten et al. Building manycore processor-to-dram networks with monolithic silicon photonics. *Int'l Symp. on High-Performance Interconnects*, Aug 2008.

[4] Scott Beamer, Chen Sun, Yong-Jin Kwon, Ajay Joshi, Christopher Batten, Vladimir Stojanović, and Krste Asanović. Re-architecting DRAM memory systems with monolithically integrated silicon photonics. In *Int'l Symp. on Computer Architecture*, pages 129 140, New York, NY, USA, 2010. ACM.

[5] Nathan L. Binkert et al. The M5 simulator: Modeling networked systems. *IEEE Micro*, 26:52 60, 2006. ISSN 0272-1732.

[6] David Brooks, Vivek Tiwari, and Margaret Martonosi. Wattch: a framework for architectural-level power analysis and optimizations. In *Proceedings of the 27th annual international symposium on Computer architecture*, Int'l Symp. on Computer Architecture, pages 83 94, 2000.

[7] CACTI6.5. CACTI6.5. Online Website,
`http://www.hpl.hp.com/research/cacti`.

[8] Johnnie Chan et al. PhoenixSim: a simulator for physical-layer analysis of chip-scale photonic interconnection networks. Apr 2010.

[9] Chia-Hsin Owen Chen et al. Physical vs. virtual express topologies with low-swing links for future many-core nocs. *Networks-on-Chip, International Symposium on*, 0:173 180, 2010.

[10] Mark J. Cianchetti et al. Phastlane: a rapid transit optical routing network. In *Int'l Symp. on Computer Architecture*, pages 441 450, 2009.

[11] P. Dainesi et al. CMOS compatible fully integrated machŰzehnder interferometer in SOI technology. *IEEE Photonics Technology Letters*, 12:660 662, June 2000.

[12] William J. Dally and Brian Towles. *Principles and Practices of Interconnection Networks.* Morgan Kaufmann Publishers, 2004.

[13] W.J. Dally and B. Towles. Route packets, not wires: on-chip interconnection networks. In *Design Automation Conference, 2001. Proceedings*, pages 684 689, 2001.

[14] B. Ganesh et al. Fully-buffered DIMM memory architectures: Understanding mechanisms, overheads and scaling. In *High Performance Computer Architecture, 2007. HPCA 2007. IEEE 13th International Symposium on*, pages 109 120, Feb 2007.

[15] M. Georgas et al. A monolithically-integrated optical receiver in standard 45-nm soi. *European Solid-State Circuits Conference*, Sep 2011.

[16] Michael Georgas, Jonathan Leu, Benjamin Moss, Chen Sun, and Vladimir Stojanović. Addressing link-level design tradeoffs for integrated photonic interconnects. *Custom Integrated Circuits Conference*, Sep 2011.

[17] C. Gunn. CMOS photonics for high-speed interconnects. *IEEE Micro*, 26(2): 58 66, Mar/Apr 2006.

[18] R. Gupta, B. Tutuianu, and L.T. Pileggi. The elmore delay as a bound for rc trees with generalized input signals. *IEEE TCAD*, 16(1):95 104, Jan 1997.

[19] A. Hadke et al. OCDIMM: Scaling the DRAM memory wall using WDM based optical interconnects. *Int'l Symp. on High-Performance Interconnects*, Aug 2008.

[20] J. Howard *et al.* A 48-core IA-32 message-passing processor with DVFS in 45nm CMOS. In *ISSCC*, pages 108 109, Feb 2010.

[21] IBM. IBM Power4 processor. Website, `http://www.research.ibm.com/power4`.

[22] Intel. Intel hybrid silicon laser. Website, `http://techresearch.intel.com/ProjectDetails.aspx?Id=149`.

[23] ITRS. ITRS 2010 update, assembly and packaging chapter. Website, `http://www.itrs.net/Links/2010ITRS/Home2010.htm`.

[24] Ajay Joshi et al. Designing energy-efficient low-diameter on-chip networks with equalized interconnects. *Int'l Symp. on High-Performance Interconnects*, Aug 2009.

[25] Ajay Joshi et al. Silicon-photonic clos networks for global on-chip communication. *Int'l Symp. on Networks-on-Chip*, May 2009.

[26] A. Kahng et al. ORION 2.0: A fast and accurate NoC power and area model for early-stage design space exploration. *Design Automation and Test in Europe*, Apr 2009.

[27] U. Kang et al. 8 Gb 3D DDR3 DRAM using through-silicon-via technology. *Int'l Solid-State Circuits Conf.*, Feb 2009.

[28] Brent Keeth et al. *DRAM Circuit Design: Fundamental and High-Speed Topics.* Wiley-IEEE Press, 2008.

[29] A. Khakifirooz and D.A. Antoniadis. MOSFET performance scaling - part II: Future directions. *Electron Devices, IEEE Transactions on*, 55(6):1401 1408, 2008.

[30] A. Khakifirooz, O.M. Nayfeh, and D. Antoniadis. A simple semiempirical short-channel MOSFET current-voltage model continuous across all regions of operation and employing only physical parameters. *Electron Devices, IEEE Transactions on*, 56(8):1674 1680, Aug 2009.

[31] Jung-Sik Kim et al. A 1.2V 12.8GB/s 2Gb mobile Wide-I/O DRAM with 4x 128 I/Os using TSV-based stacking. In *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2011 IEEE International*, pages 496 498, Feb 2011.

[32] Lionel Kimerling. Silicon microphotonics. *Applied Surface Science*, 159:8 13, June 2000.

[33] Lionel Kimerling et al. Electronic-photonic integrated circuits on the CMOS platform. *Proceedings of the SPIE*, 6125, Mar 2006.

[34] N. Kırman et al. Leveraging optical technology in future bus-based chip multiprocessors. *MICRO*, Dec 2006.

[35] Amit Kumar et al. Express virtual channels: Towards the ideal interconnection fabric. *Int'l Symp. on Computer Architecture*, June 2007.

[36] George Kurian et al. ATAC: A 1000-core cache-coherent processor with on-chip optical network. *Parallel Architectures and Compilation Techniques*, Sep 2010.

[37] Haechang Lee et al. A 16 Gb/s/link, 64 GB/s bidirectional asymmetric memory interface. *IEEE Journal of Solid-State Circuits*, 44(4):1235 1247, Apr 2009.

[38] Sheng Li et al. McPAT: An integrated power, area, and timing modeling framework for multicore and manycore architectures. *MICRO*, Dec 2009.

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

[40] B. E. Little et al. Ultra-compact $Si - SiO_2$ micro ring resonator optical channel dropping filters. *IEEE Photonics Technology Letters*, 10:549 551, Apr 1998.

[41] Jifeng Liu et al. Ge-on-si laser operating at room temperature. *Opt. Lett.*, 35 (5):679 681, Mar 2010.

[42] R. Marculescu, D. Marculescu, and M. Pedram. Probabilistic modeling of dependencies during switching activity analysis. *IEEE TCAD*, 17(2):73 83, Feb 1998.

[43] Micron. Micron DDR SDRAM products. Online Datasheet, `http://www.micron.com/products/dram/ddr3`, .

[44] Micron. Micron LRDIMM memory. Website, `http://www.micron.com/products/dram_modules/lrdimm.html`, .

[45] Jason E. Miller et al. Graphite: A distributed parallel simulator for multicores. Jan 2010.

[46] A. Narasimha et al. A fully-integrated 4x10 Gb/s DWDM optoelectronic transceiver in astandard 0.13 µm CMOS SOI. *Int'l Solid-State Circuits Conf.*, Feb 2007.

[47] Magdalena S. Nawrocka et al. Tunable silicon microring resonator with wide free spectral range. *Applied Physics Letters*, 89(7), 2006.

[48] NCSU FreePDK45. NCSU FreePDK45. Online Website, `http://www.eda.ncsu.edu/wiki/FreePDK`.

[49] C. Nitta, M. Farrens, and V. Akella. Addressing system-level trimming issues in on-chip nanophotonic networks. *HPCA '11*, Feb 2011.

[50] Tae-Young Oh et al. A 7 Gb/s/pin GDDR5 SDRAM with 2.5 ns bank-to-bank active time and no bank-group restriction. *Int'l Solid-State Circuits Conf.*, Feb 2010.

[51] Open Source Liberty. Open source liberty. Website, `http://www.opensourceliberty.org/`.

[52] Jason Orcutt et al. Demonstration of an electronic photonic integrated circuit in a commercial scaled bulk CMOS process. *Conf. on Lasers and Electro-Optics*, May 2008.

[53] Jason S. Orcutt et al. Nanophotonic integration in state-of-the-art CMOS foundries. *Optics Express*, 19(3):2335 2346, Jan 2011.

[54] Dac C. Pham et al. Overview of the architecture, circuit design, and physical implementation of a first-generation cell processor. *IEEE Journal of Solid-State Circuits*, 41(1):179 196, Jan 2006.

[55] Clifford Pollock and Michal Lipson. *Integrated Optics.* Springer, 2003.

[56] Jan M. Rabaey, Anantha Chandrakasan, and Borivoje Nikolic. *Digital Integrated Circuits: A Design Perspective, second edition.* Prentice Hall, 2003.

[57] Vivek Raghunathan et al. Athermal silicon ring resonators. In *Integrated Photonics Research, Silicon and Nanophotonics*, page IMC5. Optical Society of America, 2010.

[58] Rambus. Rambus XDR2 memory. Website,
`http://www.rambus.com/in/technology/solutions/xdr2/index.html`.

[59] Reid J. Riedlinger et al. A 32nm 3.1 billion transistor 12-wide-issue Itanium processor for mission-critical servers. *Int'l Solid-State Circuits Conf.*, Feb 2011.

[60] P. K. Runge. Undersea lightwave systems. *Optics and Photonics News*, 1(11): 9 12, 1990.

[61] J. R. Sauer et al. Photonic interconnects for gigabit multicomputer communications. *LTS, IEEE*, 3(3):12 19, Aug 1992.

[62] S. Selvaraja et al. Fabrication of uniform photonic devices using 193nm optical lithography in silicon-on-insulator. *ECIO '08*, 2008.

[63] Dirk Taillaert, Peter Bienstman, and Roel Baets. Compact efficient broadband grating coupler for silicon-on-insulator waveguides. *Optics Letters*, 29(23):2749 2751, Dec 2004.

[64] Tilera. Tilera Gx 8000. Website,
`http://www.tilera.com/products/processors/TILE-Gx-8000`.

[65] D. Vantrease et al. Corona: System implications of emerging nano-photonic technology. *Int'l Symp. on Computer Architecture*, June 2008.

[66] Thomas Vogelsang. Understanding the energy consumption of dynamic random access memories. *MICRO*, Dec 2010.

[67] Hangsheng Wang et al. Power-driven design of router microarchitectures in on-chip networks. In *Proceedings of the 36th annual IEEE/ACM International Symposium on Microarchitecture*, MICRO 36, pages 105 , Washington, DC, USA, 2003. IEEE Computer Society. ISBN 0-7695-2043-X.

[68] Lan Wei, F. Boeuf, T. Skotnicki, and H.-S.P. Wong. Parasitic capacitances: Analytical models and impact on circuit-level performance. *Electron Devices, IEEE Transactions on*, 58(5):1361 1370, May 2011.

[69] S. C. Woo et al. The SPLASH-2 programs: Characterization and methodological considerations. In *Int'l Symp. on Computer Architecture*, June 1995.

[70] P. Yan et al. Firefly: Illuminating on-chip networks with nanophotonics. *Int'l Symp. on Computer Architecture*, June 2009.