SAND2012-3524 C

SAND2012-3524C

Multi-Layer Silicon Nanophotonic Network Chips for High-Performance Computing

***Abstract*—One of the key facets of high performance scientific application performance is communication, which involves the amount of data movement required and the parallel scalability of an application. Historically, the communication subsystem has consisted of MPI running on dense compute nodes interconnected by specialized packet-switched networks. Silicon photonics has presented a promising solution to some of the performance and power challenges in HPC, and has become a likely candidate for replacing or augmenting current network designs. In this work, we use multi-layer silicon photonics to implement circuit-switches which enable high bandwidth communication for very large scale machines by using the existing MPI rendezvous protocol. We explore a variety of network topologies with a detailed physical-layer analysis of the design of the photonic subsystem as well as simulations of large parallel scientific applications.**

I. Introduction

High performance computing is plagued by interrelated problems in scaling to exascale capabilities, including expressing application parallelism, data management, power, cost, and fault tolerance. It is generally believed that revolutionary innovation is required in a number of these areas to overcome the challenges in reaching an affordable exascale solution. New emerging technologies are expected to play a large role in alleviating the constraints in hardware, while enabling game-changing redesign of software.

As machine size, compute density, compute capability, machine power, and machine investment continue to increase, optical communication has made its way into data centers to alleviate some of the problems of scalability. Optical active cabling is used today simply to provide links which are able to connect switches across a machine room (tens of meters) because optics suffers much less from distance-dependent performance. Optics is also employed for its density, using a combination of multi-core fibers and/or wavelength division multiplexing (WDM) to reduce the weight and size of cables. However, novel use of optics is expected to play a larger role in not just replacing current electrical implementations, but enabling new architectures with different performance characteristics.

Optical packet-switches are promising for latency-sensitive applications, and can be implemented with silicon optical amplifiers (SOAs) [1], [2] or arrayed waveguide gratings (AWGs) [3]. However, they can be difficult to implement and scale, expensive, and can be limited in bandwidth by the modulation datarate. Circuit-switching is typically easier to implement, and has been achieved with technologies such as 3D MEMS [4] and beam steering with piezoelectrics [5], and has been proposed before for use in data centers [6].

Architectural studies have shown the benefits of optical circuit-switch architectures for some classes of applications [7], [8], however circuit-setup time for these technologies is on the order of milliseconds, which limits their use for many high-performance applications that have dynamic communication behavior, despite proposed optimization techniques [9], [10]. Silicon photonics is a potentially revolutionary development in that silicon photonic broadband switches offer switching times on the order of nanoseconds [11], which can yield fast reconfiguration of optical circuit paths. Moreover, using multi-layer deposited photonics, it is now possible to consider building high-radix silicon photonic switch chips that can connect large data centers [12].

The idea of using photonic circuit-switches actually lines up perfectly with today's MPI-based implementations of scientific applications. MPI typically uses a *rendezvous* handshake protocol to establish buffer space at receiving nodes before sending large chunks of data in order to avoid multiple memory copy operations. A size threshold is set which dictates that large messages must first send a small request message, while smaller messages may be directly sent. Also, many specialized high-performance implementations perform this handshake in the network interface cards (NICs) to avoid having to involve the MPI software layer.

We propose the use of the MPI rendezvous protocol that is used today for implementing the path-setup protocol that is required for photonic circuit-switches to establish an end-to-end path before data can be sent. In effect, we use the small message that must be sent anyway in establishing receiver buffer space to also reserve circuit-path resources in a photonic network chip. This network chip connects a group of compute nodes to augment current high-performance networks, as shown in Figure 1. Cheap, low-speed electrical links are connected to the chip to provide a control plane for establishing circuit paths. Optical fiber is laterally coupled to the photonic switch, which sits on top of the control plane. Through the use of multi-layer deposited photonics, this chip can be fabricated and packaged in a single process run, and no flip-chip bonding or chip stacking is needed.

In this work, we design our photonic network chips using indirect network topologies like the butterfly, Clos, and fat-tree, which were originally designed for use in circuit-switched networks. Other work has used indirect networks [13], [14] with silicon photonics, but only in the context of network-on-chip where tiled access points are required, and only with single-layer crystalline silicon photonics. We assume the electronic control plane is a low-speed mirror network that

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

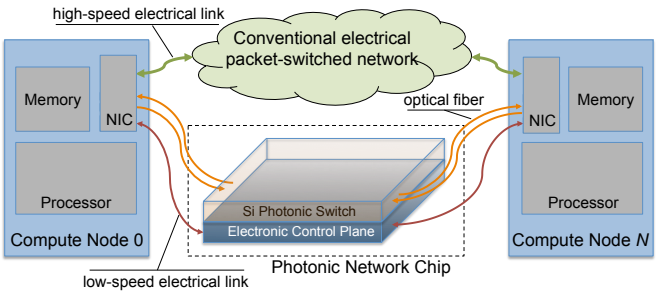


Figure 1: Proposed network architecture.

controls the photonic switching elements, a technique that has been previously explored in the context of networks-on-chip [15].

Finally, we evaluate our designs with simulations that capture every aspect of the communication characteristics of key scientific proxy applications. This work is the first to combine the use of existing MPI protocol implementations and multi-layer photonic network design to benefit high-performance scientific applications.

II. Multi-Layer Photonics

Silicon photonic devices have typically been fabricated using pure crystalline silicon with an SiO_2 cladding, which offers good electrical and optical properties in the telecom C-band (around $\lambda = 1.55\mu m$). Silicon photonic waveguides, modulators, detectors, filters, and broadband switches have all been fabricated and shown to work well. The ring resonator has become a common workhorse for many of these functions because of the flexibility in its design to target areas of the wavelength domain in WDM systems, which is invaluable for achieving high-density communication. Though the ring resonator has experimentally been highly sensitive to fabrication variations and temperature fluctuations, efforts to combat these effects are underway such as ring heating, using narrow waveguides and polymer cladding [16], using slotted waveguides [17], and using Mach-Zehnder interference correction [18]. However, crystalline silicon is limited to a single layer. Waveguide crossings can be engineered to allow the layout of complex interconnect designs [19], but can often become a major contributing factor to network insertion loss [20].

Recently, multi-layer fabrication using a silicon nitride and poly-silicon material system has gained attention because of the flexibility gained in layout by using multiple vertical optical layers [21]. Silicon nitride actually exhibits lower loss than pure crystalline silicon over the telecom frequency range and can be used for transport and filtering [22], but is not optically active. Poly-silicon, though much higher loss, can be used for modulation [23], switching, and detection [24], while vertically coupling to nitride for transport. This scheme, shown in Figure 2, can avoid waveguide crossings by using two layers of nitride for transport, a single layer of poly-silicon for switching, and broadband vertical couplers [25] for connecting additional layers of nitride. Additionally, nitride and poly can be deposited directly on top of CMOS electronics, avoiding

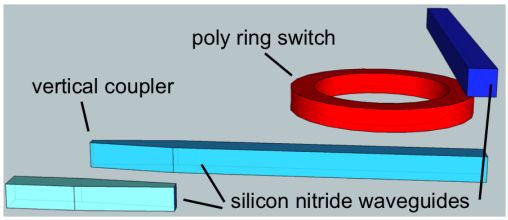


Figure 2: An example of multi-layer devices using silicon nitride and polycrystalline silicon.

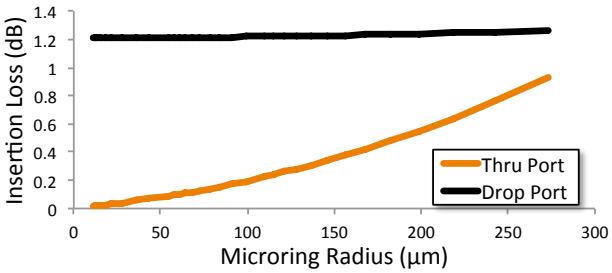


Figure 3: Microring insertion loss versus ring radius.

the need for costly packaging such as flip-chip bonding or chip stacking. Using the nitride-poly material system, the idea of creating a multi-layer silicon photonic interconnect has been proposed [26] for NoC applications, which we use as a foundation for the devices and analysis in this work.

Some of our analysis relies on the fact that the loss characteristics of the ring change as the radius increases, shown in Figure 3. We assume broadband switches which are designed such that they are ON-resonance when *off*, and can be pushed OFF-resonance by injecting carriers into the ring. This type of switch design is easier to implement because the effect of the carrier injection itself does not have to be corrected for over time to align to a specific resonance.

Table I lists the optical power budget parameters we assume in our design analysis. These parameters dictate how much optical power the network can sustain, which happens to be the difference between the switch maximum power (17.8 dBm) and the detector sensitivity (-17 dBm), or 34.8 dB. Table II lists the insertion loss parameters for each device we use in our analysis.

III. Architectural Analysis

The network topologies we consider in this work are indirect networks which have been designed to be amenable to circuit-switching. Though previous work has attempted to use them for network-on-chip applications, they are typically not considered because layout is awkward for tiled cores which require that access points generally be distributed regularly around a chip. Because we are applying them to large data center network chips whose ports are on the edges of the chip and because we are employing multi-layer photonics to avoid crossings, these topologies are among the right choices for this application. Their design, layout, and analysis is a novel contribution to the field.

Table I: Optical Power Budget Parameters

Parameter	Symbol	Value
Waveguide Maximum Injected	P_{wg}	30 dBm (1000 mW)
Switch Maximum Injected (On)	P_{switch}	17.8 dBm (60 mW)
Photodetector Sensitivity	η_{det}	–17 dBm (0.02 mW)

Table II: Optical Device Insertion Loss Parameters

Parameter	Symbol	Value
Waveguide Propagation (Nitride)	ζ_{ni}	0.1 dB/cm
Waveguide Propagation (Poly)	ζ_{poly}	5 dB/cm
Waveguide Bend	ζ_{bend}	0.005 dB/90°
Modulation	ζ_{mod}	1.2 dB
Photodetector Filter Drop Port	ζ_{det}	0.5 dB
Photodetector Filter Through Port	ζ_{filt}	0.05 dB
Vertical Coupler (Poly to Nitride)	ζ_{vert}	0.1 dB
Chip couplers	ζ_{chip}	0.5 dB

A. Insertion Loss Analysis

One of the most important analyses in the design of a silicon photonic network is an insertion loss analysis, determines how many wavelengths can be used in WDM for circuit-switched networks, if any. The design space being explored by this analysis is an optimization problem that can be summarized by the following descriptions which correspond to Figure 4:

1) Only a range of the optical spectrum is useful, constrained by either commodity lasers and modulators, or the free spectral range (FSR) of modulators. This useful area we call FSR_{use}.
2) Broadband ring switches can increase their wavelength density by increasing the ring radius, but at the cost of larger footprint and higher through port loss.
3) Higher insertion loss means that each wavelength requires more power to overcome it, meaning that the optical power budget determined by the nonlinear threshold and detector sensitivity can be divided into a smaller number of wavelengths.

In effect, choosing the right broadband switch ring radius is key to maximizing the number of wavelengths that can fit into the FSR while keeping the insertion loss low enough for that number of wavelengths to be reliably injected such that they overcome the worst case network loss. Previous work has outlined good practices in insertion loss analysis [20], as well as work specifically targeting multi-layer silicon photonics [26], and we follow a similar approach for this work by constructing hierarchical analytical models of the photonic subsystems to determine the number of wavelengths that can be used for WDM transmission.

B. Radix-N Switch Design

One basic building block of the network topologies that we will consider makes use of a regularly-structured ring matrix which creates a crossbar switch. Figure 5 shows this design for a 4×4 switch using 2 layers of nitride and a single active poly layer in the middle. This switch also removes the bottom right ring which is not needed, and replaces it with a bended vertical coupler. Though other 4×4 switch designs have been proposed which are more compact and have less loss [27],

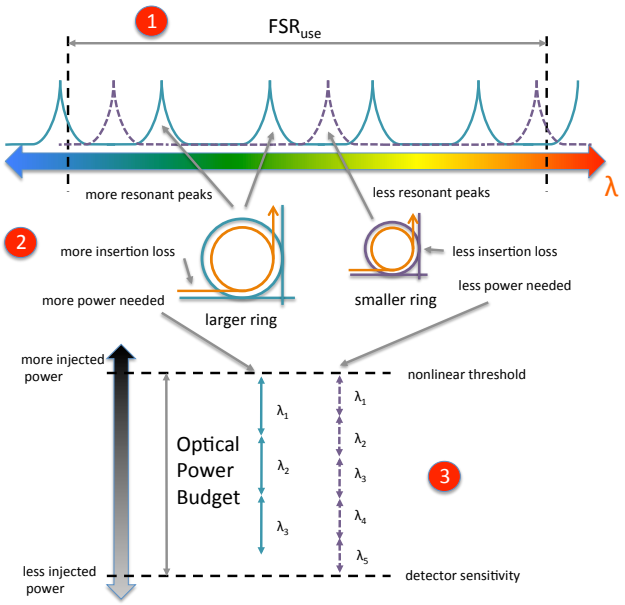


Figure 4: The design space optimization of an optical power budget for a broadband-switch based photonic network.

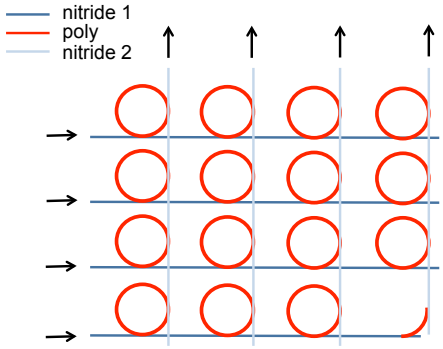


Figure 5: Multi-layer ring switch matrix, here a 4×4 crossbar.

they are not easily scalable beyond four ports and they don't offer U-turns which are required for staged indirect networks.

Assuming that the individual ring drop loss (ϕ_{drop}) is larger than ring through loss (ϕ_{thru}), we can model the worst case loss of this switch with the equation:

$$\Omega_{matrix} = \phi_{drop} + \phi_{thru} \times (2R - 3) + 2L_{sw}\zeta_{ni} \quad (1)$$

where R is the switch radix, ϕ_{drop} and ϕ_{thru} are the radius-dependent ring losses from Figure 3, and $L_{sw} = 2(r \times +50\mu m)R$ with r being the ring radius.

C. Topology 1: Butterfly

The first topology we will consider is the unidirectional butterfly, a staged indirect network which can be found in Figure 6. Note that the colors indicate which layer (nitride1, poly, nitride2) each connection is implemented on. By using poly for the short straight connections, and each nitride layer for the crossing connections, we can completely avoid waveguide crossings. However, using our switch matrix design, vertical couplers are required for all poly straight connections at each switch, and (R-2) of all middle-stage switches. From Figure

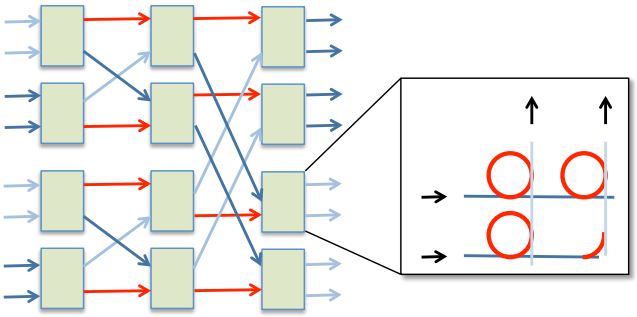


Figure 6: The unidirectional butterfly, here a 3-stage 2-ary.

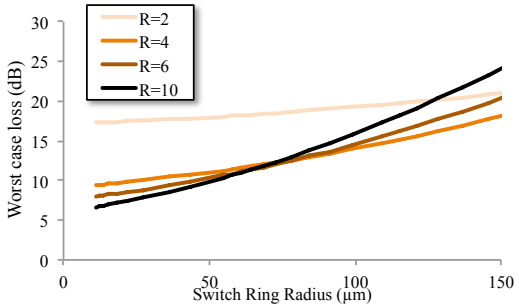


Figure 7: Worst-case insertion loss of 1000-port butterfly using different radix switches.

6, it should be clear that the switch is scalable to an arbitrary number of ports by adding more stages, and that crossings can always be avoided by using R layers of nitride, where R is the bidirectional switch radix.

We model the worst-case insertion loss of the butterfly as follows:

$$\Omega_{bfly} = S\Omega_{matrix} + 2\zeta_{chip} + \Omega_{prop} + \Omega_{couplers} \quad (2)$$

where S is the number of stages, Ω_{matrix} is the loss from passing through a single switch from Section III-B, $\Omega_{couplers}$ is the total of all vertical couplers needed, and Ω_{prop} is the loss from propagation between switches. The two missing terms Ω_{prop} and $\Omega_{couplers}$ can be defined as follows:

$$\Omega_{prop} = \sqrt{2}L_{chip}\zeta_{ni} + 3S\zeta_{bend} \quad (3)$$

$$\Omega_{couplers} = (S-2)(2\zeta_{vert}) \quad (4)$$

assuming the worst-case path is using cross-paths on nitride, and using a simplification of the propagation distance as the diagonal of the length of the squared chip edge (L_{chip}) plus three bends per stage.

Using these equations, we can plot the worst-case loss for a switch chip with at least 1000 ports using 2, 4, 6, and 10-port radix switches (and 10, 5, 4, and 3 stages, respectively), found in Figure 7. At small ring switch radii, there is an advantage to using higher-radix switches because the reduced number of stages saves on the loss incurred with dropping through a ring at each stage. However, as ring radius increases the 4-ary butterfly quickly wins.

Taking the loss into consideration, we can also plot the number of wavelengths that would fit into the optical power budget

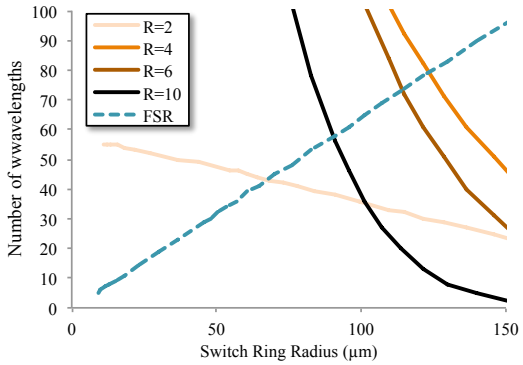


Figure 8: Number of wavelengths allowed of 1000-port butterfly.

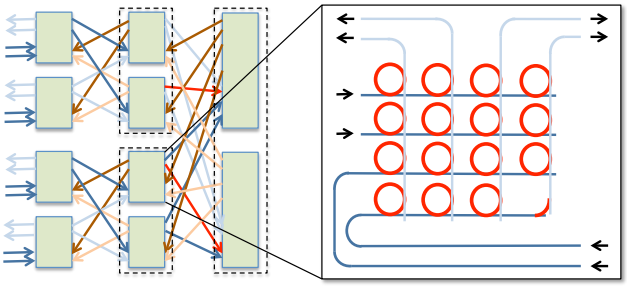


Figure 9: The Fat Tree network, with 3-stages and 2×2 bi-directional switches.

of 34.8dB, found in Figure 8. Also plotted is the number of wavelengths limited by the FSR_{use}, which increases with switch ring radius. The optimal design point is the intersection of these two lines, with the 4-ary 5-stage butterfly achieving 71 wavelengths.

D. Topology 2: Fat-Tree

A fat tree [28] can be implemented similar to a butterfly in its structure and connections, but it uses either unfolding or bi-directional ports to provide path diversity with predictable bandwidth characteristics. A fat tree implemented with homogenous switches (that is, port count of switches does not increase with level as it should logically) is rearrangeably nonblocking, though this is not implemented to simplify routing logic. As depicted in Figure 9, the matrix switch layout must be modified slightly to provide logically bi-directional ports. Also, in order to implement the crossing connections between stages, it is necessary to use more than two layers of nitride which can gradually vertically couple into the poly layer that resides in the middle of the stack. In general, the fat tree will require $2R$ layers of nitride, where R is the bi-directional radix of the switches. In Figure 9, up-links are shown using the two nitride layers closest to the poly, and down-links use the layers further away in the vertical stack.

Our modeling of the worst-case insertion loss in the fat tree is similar to the butterfly:

$$\Omega_{ftree} = (2S-1)\Omega_{matrix'} + 2\zeta_{chip} + \Omega_{prop} + \Omega_{couplers} \quad (5)$$

where S is the number of stages, $\Omega_{matrix'}$ is the loss from passing through a single modified matrix switch, $\Omega_{couplers}$ is

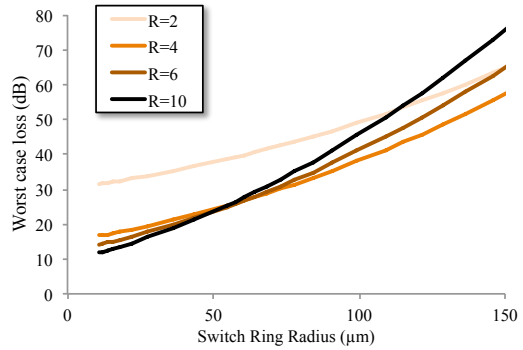


Figure 10: Worst-case insertion loss of 1000-port fat tree network with a range of values for switch radix.

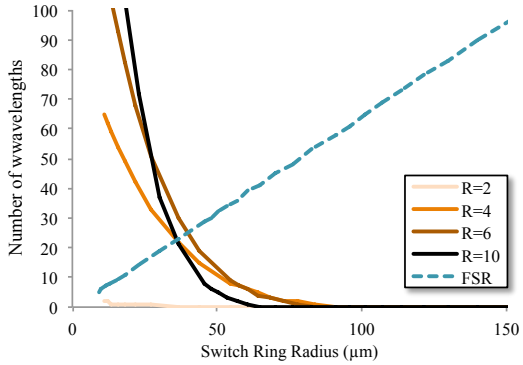


Figure 11: Number of wavelengths allowed of 1000-port fat tree network.

the total of all vertical couplers needed, and Ω_{prop} is the loss from propagation between switches. Again, the missing terms are:

$$\Omega_{prop} = \sqrt{2}L_{chip}\zeta_{ni} + 3S\zeta_{bend} \quad (6)$$

$$\Omega_{couplers} = 2\zeta_{vert}(S-1) + 4\zeta_{vert}(S-1) \quad (7)$$

$$\Omega_{matrix'} = \Omega_{matrix} + 2\zeta_{bend} + 1.5L_{sw}\zeta_{ni} \quad (8)$$

where L_{chip} is the length of the chip edge and L_{sw} is the length of the switch determined by ring radius.

Like the butterfly, we can plot the worst-case insertion loss of a 1000-port fat tree for different values of the bidirectional switch radix R, seen in Figure 13. Again, using the 4-port switch yields the lowest insertion loss in the area of interest. However, plotting the wavelengths in Figure 10, we see that the 6-port implementation actually yields more wavelengths at 23 when considering the FSR. Predictably, this is less bandwidth than the butterfly, but the fat tree gains path diversity which reduces the contention in the switch chip.

E. Topology 3: Clos

The Clos network topology [29] is a 3-staged indirect network that can be designed such that it is nonblocking. Referring to Figure 12, a configuration which has m switches with n input ports and k output ports in the first stage, and k switches with m input and output ports in the second stage, the network is nonblocking if $k \geq 2n-1$. The Clos network

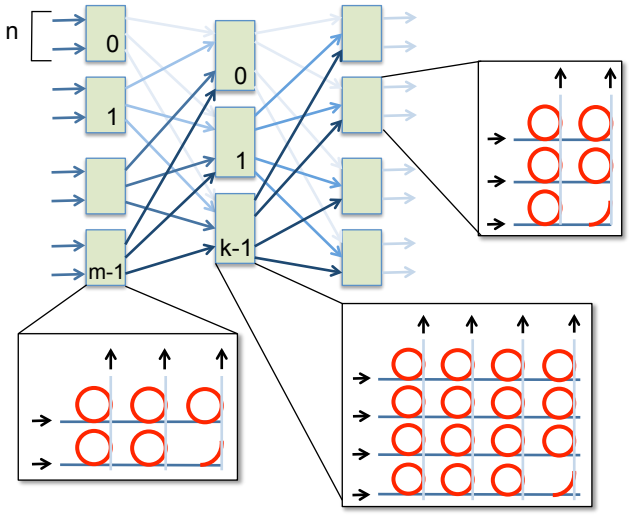


Figure 12: The unidirectional Clos network, here a 3-stage with n=2, k=3, and m=4.

also requires m layers of nitride to completely avoid crossings in the layout, as shown in Figure 12.

To model the worst-case loss through the clos, we can use the following equation:

$$\Omega_{clos} = \Omega_1 + \Omega_2 + \Omega_3 + \Omega_{vert} + \Omega_{prop} + 2\zeta_{chip} \quad (9)$$

where Ω_{1-3} are the losses for the switches at each stage, Ω_{vert} is the vertical coupling loss for the top-most layer of nitride to the bottom layer of nitride that feeds into the switch, and Ω_{prop} is the propagation loss between stages, all of which can be defined as follows:

$$\Omega_1 = (n-2+k-1)\phi_{thru} + \phi_{drop} \quad (10)$$

$$\Omega_2 = (2m-3)\phi_{thru} + \phi_{drop} \quad (11)$$

$$\Omega_3 = \Omega_1 \quad (12)$$

$$\Omega_{vert} = 4(\frac{m}{2}+1)\zeta_{vert} \quad (13)$$

If we assume a 1000-port switch chip, with $k = 2n-1$ and $m = 1000/n$ then we can plot the worst-case insertion loss of the Clos for different values of n, seen in Figure 13. As it turns out, a value of 32 for n seems to be the most optimal of those considered, offering the best tradeoff of second stage switch size (increasing the number of ζ_{thru} added up) and the number of vertical couplers needed for additional nitride layers (increasing with larger m).

Taking the loss into consideration, we can again plot the number of wavelengths that would fit into the optical power budget of 34.8dB, found in Figure 14 along with the FSR-limited wavelengths. The $n = 32$ line maximizes the number of wavelengths at a ring switch radius of 27μm with 19 wavelengths. Though this is less than the butterfly and fattree, the Clos is designed as nonblocking.

F. Analysis Summary

After detailed analytical modeling of the physical characteristics of switch chips implemented with different topologies,

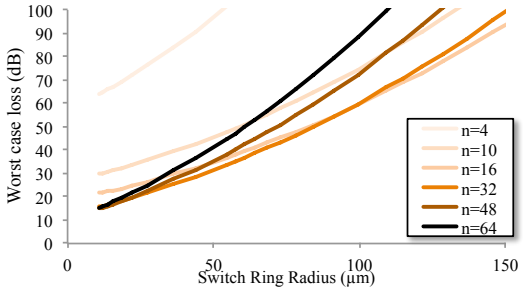


Figure 13: Worst-case insertion loss of 1000-port Clos network with a range of values for n.

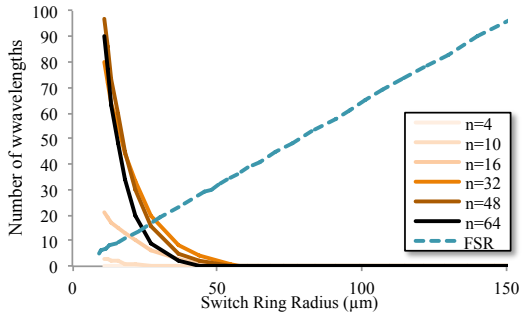


Figure 14: Number of wavelengths allowed of 1000-port Clos network.

we find different bandwidth results for each which is summarized in Table III.

However, each network has different blocking and therefore contention characteristics, which can only be determined through simulation. The next section properly evaluates this space.

IV. Simulation Results

In this section, we evaluate the effectiveness of using circuit-switched silicon photonic network chips for high performance computing by simulating the execution of scientific applications with SST/macro, a coarse-grained simulator aimed at evaluating the communication characteristics of large parallel applications at scale [30], [31]. SST/macro contains a fully implemented model of the MPI interface and protocols, as well as performance models of processors, memory, NICs, and network switches. Besides determining the value in using fast circuit-switching for MPI rendezvous messages, the network chips that we designed in Section III have different blocking characteristics which can only be effectively evaluated in simulation. In this work, we use four scientific application *skeletons* as representative workloads for HPC.

A. Application Skeletons

A useful vehicle for evaluating a large complex design space is the proxy application, which is a piece of code which is meant to represent the characteristics of a specific full application, but is smaller, likely simpler, and easier (and quicker) to run than its parent code.

SST/macro is meant to run *skeleton* applications, or applications that retain the communication and control information

Table III: Analysis Summary

Topology	# Wavelengths
Butterfly	71
Fat Tree	23
Clos	19

of the original code but abstract away any computation that is used to produce a real numerical result. This allows us to model an entire application at scale while looking at the features we are interested in, namely communication and its dynamic run-time behavior, without requiring massive computing resources to do so. Often, a model of computation will be constructed that describes the processor utilization (*e.g.* flops) and memory use (*e.g.* bytes accessed) so that undeterministic behavior (such as when using MPI_Waitsome) is close to the real application, and so that it can be validated as a whole. In this work, we use *communication skeletons* for evaluation which have no computation model so that we can isolate the communication-dependent aspect of the original applications.

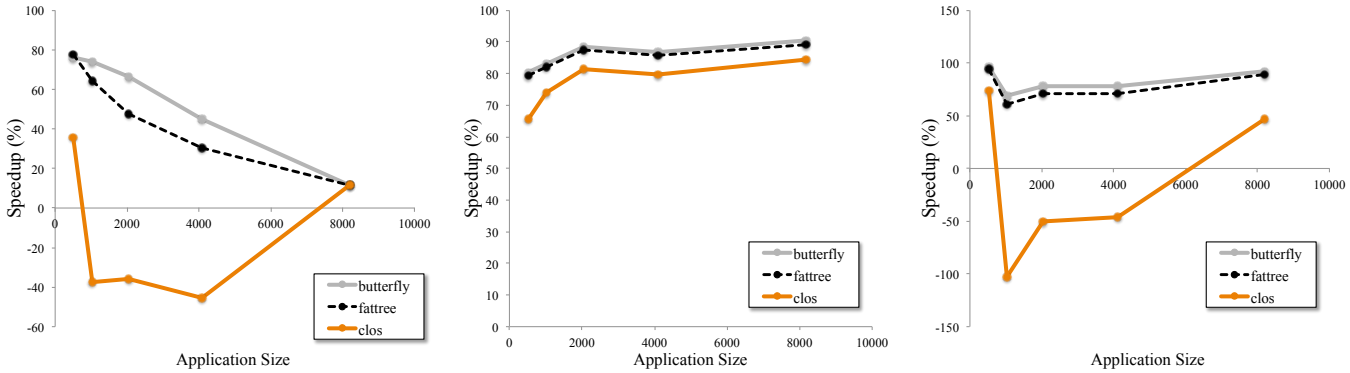
For our experiments, we consider three scientific skeleton applications included in the SST/macro distribution, listed below:

- **miniMD**: MiniMD is a molecular dynamics micro-application from the Mantevo project [32]. MiniMD was created to investigate improving spatial-decomposition particle simulations as a simpler, but more accessible and easily built and executed version of LAMMPS [33]. In this work we use weak scaling starting from a 256×256×256 problem with 100 time steps.
- **miniGhost**: miniGhost [34] is a proxy application for CTH [35], which is a multi-material, large deformation, strong shock wave, solid mechanics code. In this study, we model a 3-dimensional shaped-charge problem of 80×192×80 with 40 variables.
- **LU**: LU [36] is an application-level benchmarking code supplied as part of the NAS Parallel Benchmark suite (NPB). The algorithm solves a synthetic system of non-linear PDEs using a symmetric successive over-relaxation (SSOR) kernel employing a two-phase wavefront sweep through the 3-dimensional data domain. This paper focuses on weak scaling starting from a 512×1024×256 problem, retaining 32 grid points per rank.

To illustrate how the performance changes as the applications scale, we investigate them from 512 to 8k ranks. We assign 8 ranks to each multi-core processor, which results in a mapping that fits entirely into the nodes that are connected by a single photonic network chip. Future work will explore optimal network design and application mapping with node counts that exceed the switch port count.

B. Simulation Setup

SST/macro contain a number of parameterized models for use in network modeling. In our proposed architecture, a conventional high-speed packet-switched network is used for eager-send MPI messages, and a separate low-speed electronic network is used to arbitrate circuit-path setup for the high-



(a) miniMD (b) LU (c) miniGhost

Figure 16: Speedup using proposed photonic network chip.

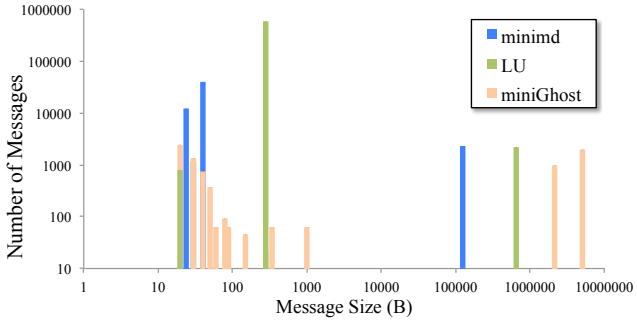


Figure 15: Message size characteristics of applications.

Table IV: Simulation Network Parameters

Parameter	Value
Electronic Data Network	
Topology	3D Torus
Switch Input Buffer Size	64kB
Switch Output Buffer Size	128kB
Link Bandwidth	10 Gbps
Link Latency	30ns
Switch Crossbar Bandwidth	20 Gbps
Switch Crossbar Latency	$2ns$
Switch Arbitration Latency	$2ns$
Packet MTU	8kB
Virtual Channels	4
Electronic Circuit Setup Network	
Switch Input Buffer Size	8kB
Switch Output Buffer Size	16kB
Link Bandwidth	1 Gbps
Link Latency	100ns
Switch Crossbar Bandwidth	2 Gbps
Switch Crossbar Latency	$2ns$
Switch Arbitration Latency	$2ns$
Packet MTU	256B
Virtual Channels	2
Circuit Setup Timeout	$10\mu s$
Photonic	
Circuit Propagation Delay	$100ns$
Single Wavelength Datarate	10 Gbps
Protocol	
MPI Rendezvous Threshold	8kB
Circuit Setup Timeout	$15\mu s$

bandwidth photonic circuit-switches. Table IV lists the parameters for the different parts of the networks that were used.

For circuit-path setup, network interface cards (NICs) on nodes sending data initiate a circuit configuration by sending out a *SETUP* message that attempts to reserve photonic resources. Setup messages that make it to the destination are turned into a *PATH-ACK* and sent back to the source, who can then transmit the data through the photonic network. After data transmission is complete, a *TEARDOWN* message is sent to release photonic resources.

Often in circuit-switching, a path-blocked protocol is used for congestion management which returns path-setup messages which cannot reserve photonic resources back to the source to release resources it has reserved along the way. Because of the Clos and butterfly topologies we are considering here, using a path-blocked protocol is not easy to use because a message cannot be sent backwards to the switch from which it came. Therefore, we use a path-timeout protocol which automatically sends out a path-teardown message and a new path-setup message from the source after a set time if it has not received a path-ack. An incremental randomized backoff wait time is used between the timeout (sending a teardown) and sending out the new setup.

C. Simulation Results

First, we characterize the applications in terms of their message size distribution, which is the determining factor in how many messages will clear the MPI rendezvous threshold and transmit on the photonic network. Figure 15 shows a message-size histogram for the applications we consider. All of them display a variety of message sizes, with a larger number of small messages and a small number of very large messages, which is typical of many scientific applications which cannot avoid exchanging large arrays of data. Since the threshold for the MPI rendezvous protocol is typically on the order of 8kB, this provides a clear separation between messages that will pass through the photonic circuit switched network and those that will take the electronic packet-switched one.

Figure 16 shows the results of the speedup attained for each application using the photonic network chip over a purely electrical baseline. In all cases, the butterfly achieves the best speedup despite it being a blocking network, on average 70-90%. The applications studied here make better use of

more bandwidth (enabled by the lower insertion loss of the design, and therefore more wavelengths) than a network with less blocking because these applications, like many others, exhibit a smaller number of very large messages that can highly stress a packet-switched network. Weak scaling in miniMD and miniGhost was not perfectly achieved because of the way the problem is decomposed among MPI ranks, which led to variations in performance especially for the Clos network. Overall, the circuit-switching network chip and the extremely high bandwidth provided by WDM multi-layer photonics can alleviate bandwidth-limited phases of application communication while relying on low-latency conventional packet-switching for collectives and other small messages.

V. Conclusion

Silicon photonics offers a promising solution to circuit-switch setup time which has the potential to benefit high-performance computing, scientific or otherwise. Multi-layer deposited materials enables the design of more complex, larger scale interconnects that are necessary in next-generation machines. The extremely high bandwidth offered by this technology, coupled with the use of existing MPI protocol implementations achieved 70-90% speedup of the communication for a range of proxy applications of varying sizes.

References

[1] H. Wang, K. Bergman, C. Gray, and D. Keezer, "Demonstration of end-to-end bit-parallel memory transactions across the ultra-low latency data vortex optical packet switch," in *Conference on Optical Fiber Communication (OFC)*, march 2010, pp. 1 –3.

[2] C. Lai, D. Brunina, and K. Bergman, "Demonstration of 8x40-gb/s wavelength-striped packet switching in a multi-terabit capacity optical network test-bed," in *IEEE Photonics Society, 2010 23rd Annual Meeting of the*, nov. 2010, pp. 688 –689.

[3] R. Proietti *et al.*, "40 gb/s 8x8 low-latency optical switch for data centers," in *Optical Fiber Communication Conference*. Optical Society of America, 2011, p. OMV4.

[4] J. Bowers, "Low power 3d mems optical switches," in *IEEE/LEOS International Conference on Optical MEMS and Nanophotonics*, aug. 2009, pp. 152 –153.

[5] D. Scrymgeour and others., "Hybrid electrooptic and piezoelectric laser beam steering in two dimensions," *Lightwave Technology, Journal of*, vol. 23, no. 9, pp. 2772 – 2777, sept. 2005.

[6] G. Wang *et al.*, "c-through: part-time optics in data centers," *SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 4, Aug. 2010.

[7] L. Schares *et al.*, "A reconfigurable interconnect fabric with optical circuit switch and software optimizer for stream computing systems," in *Conference on Optical Fiber Communication*, march 2009, pp. 1 –3.

[8] N. Farrington *et al.*, "Hardware requirements for optical circuit switched data center networks," in *Optical Fiber Communication Conference*. Optical Society of America, 2011, p. OTuH3.

[9] S. Takizawa, T. Endo, and S. Matsuoka, "Locality aware mpi communication on a commodity opto-electronic hybrid network," in *IEEE International Symposium on Parallel and Dist. Processing.*, april 2008, pp. 1 –8.

[10] S. Shao, A. Jones, and R. Melhem, "Compiler techniques for efficient communications in circuit switched networks for multiprocessor systems," *IEEE Trans. on Parallel and Dist. Systems*, vol. 20, no. 3, pp. 331 –345, march 2009.

[11] A. Biberman *et al.*, "Broadband silicon photonic electrooptic switch for photonic interconnection networks," *Photonics Technology Letters, IEEE*, vol. 23, no. 8, pp. 504 –506, april15, 2011.

[12] A. Biberman *et al.*, "CMOS-compatible scalable photonic switch architecture using 3D-integrated deposited silicon materials for high-performance data center networks," in *Optical Fiber Communication Conference (OFC)*, march 2011, pp. 1 –3.

[13] H. Gu, J. Xu, and W. Zhang, "A low-power fat tree-based optical network-on-chip for multiprocessor system-on-chip," in *Design, Automation Test in Europe Conference Exhibition*, april 2009, pp. 3 –8.

[14] A. Joshi *et al.*, "Silicon-photonic clos networks for global on-chip communication," in *Proceedings of the 3rd ACM/IEEE International Symposium on Networks-on-Chip.*, may 2009, pp. 124 –133.

[15] G. Hendry *et al.*, "Analysis of photonic networks for a chip multiprocessor using scientific applications," in *Proceedings of the 3rd ACM/IEEE International Symposium on Networks-on-Chip*, 2009, pp. 104–113.

[16] J. Teng *et al.*, "Athermal silicon-on-insulator ring resonators byoverlaying a polymer cladding on narrowedwaveguides," *Opt. Express*, vol. 17, no. 17, pp. 14 627–14 633, Aug 2009.

[17] L. Zhou *et al.*, "Towards athermal optically-interconnected computing system using slotted silicon microring resonators and RF-photonic comb generation," *Applied Physics A: Materials Science & Processing*, vol. 95, no. 4, pp. 1101–1109, Jun. 2009.

[18] B. Guha, B. B. C. Kyotoku, and M. Lipson, "CMOS-compatible athermal silicon microring resonators," *Opt. Express*, vol. 18, no. 4, pp. 3487–3493, Feb 2010.

[19] T. Fukazawa, T. Hirano, F. Ohno, and T. Baba, "Low loss intersection of si photonic wire waveguides," *Japanese Journal of Applied Physics*, vol. 43, no. 2, pp. 646–647, 2004.

[20] J. Chan, G. Hendry, A. Biberman, and K. Bergman, "Architectural exploration of chip-scale photonic interconnection network designs using physical-layer analysis," *Journal of Lightwave Technology*, vol. 28, no. 9, pp. 1305 –1315, may. 2010.

[21] K. Preston, B. Schmidt, and M. Lipson, "Polysilicon photonic resonators for large-scale 3d integration of optical networks," *Opt. Express*, vol. 15, no. 25, pp. 17 283–17 290, Dec 2007.

[22] A. Gondarenko, J. S. Levy, and M. Lipson, "High confinement micron-scale silicon nitride high q ring resonator," *Opt. Express*, vol. 17, no. 14, pp. 11 366–11 370, Jul 2009.

[23] K. Preston, P. Dong, B. Schmidt, and M. Lipson, "High-speed all-optical modulation using polycrystalline silicon microring resonators." *Applied Physics Letters*, vol. 92, no. 15, 2008.

[24] K. Preston, Y. H. Lee, M. Zhang, and M. Lipson, "Waveguide-integrated telecom-wavelength photodiode in deposited silicon," *Opt. Lett.*, vol. 36, no. 1, pp. 52–54, Jan. 2011.

[25] R. Sun *et al.*, "Impedance matching vertical optical waveguide couplers for dense high index contrast circuits," *Opt. Express*, vol. 16, no. 16, pp. 11 682–11 690, 2008.

[26] A. Biberman *et al.*, "Photonic network-on-chip architectures using multi-layer deposited silicon materials for high-performance chip multiprocessors," *ACM J. on Emerging Tech. in Computing Sys.*, vol. 7, no. 2, 2011.

[27] J. Chan, A. Biberman, B. G. Lee, and K. Bergman, "Insertion loss analysis in a photonic interconnection network for on-chip and off-chip communications," *Nature*, no. I, pp. 300–301, 2008.

[28] C. E. Leiserson, "Fat-trees: universal networks for hardware-efficient supercomputing," *IEEE Trans. Comput.*, vol. 34, no. 10, pp. 892–901, Oct. 1985.

[29] C. Clos, "A Study of Non-blocking switching networks," *Bell Syst. Tech. J.*, vol. 32, pp. 406–424, 1953.

[30] C. L. Janssen *et al.*, "A simulator for large-scale parallel computer architectures." *IJDST*, vol. 1, no. 2, pp. 57–73, 2010.

[31] C. L. Janssen, H. Adalsteinsson, and J. P. Kenny, "Using simulation to design extremescale applications and architectures: programming model exploration," *SIGMETRICS Perform. Eval. Rev.*, vol. 38, pp. 4–8, March 2011.

[32] M. A. Heroux *et al.*, "Improving performance via mini-applications," Sandia National Labs, Tech. Rep. SAND2009-5574, September 2009. [Online]. Available: https://software.sandia.gov/mantevo

[33] "LAMMPS molecular dynamics simulator," 2009. [Online]. Available: http://lammps.sandia.gov/index.html

[34] R. F. Barrett *et al.*, "Poster: mini-applications: vehicles for co-design," in *Proc. of Supercomputing Companion*, 2011, pp. 1–2.

[35] E. S. Hertel, Jr. *et al.*, "Cth: A software family for multi-dimensional shock physics analysis," in *in Proceedings of the 19th International Symposium on Shock Waves*, 1993, pp. 377–382.

[36] M. Yarrow and R. D. Wijngaart, "Communication improvement for the lu nas parallel benchmark: A model for efficient parallel relaxation schemes." NASA Ames Research Center, Tech. Rep. NAS- 97-032, November 1997.