Power and Performance Analysis of Scalable Photonic Networks for Exascale Architecture

Avinash K. Kodi
Electrical Engineering and Computer Science
Ohio University
Athens, Ohio 45701
Email: kodi@ohio.edu

Brian Neel
AMD Research
7171 Southwest Parkway
Austin, Texas, 78735
Email: brian.neel@amd.com

William C. Brantley
AMD Research
7171 Southwest Parkway
Austin, Texas, 78735
E-mail: bill.brantley@amd.com

***Abstract*—As bandwidth demands from high-performance computing (HPC) applications have increased, scalable topology and power-efficiency of network technology are becoming critical parameters for exascale design. Dragonfly topology provides low diameter for exascale networks however, fewer global links reduce the bisection bandwidth while high-radix routers increase the router complexity and area overhead. Further, performance/watt delivered by metallic interconnects significantly increases the power consumed by the network. In this paper, we propose multiple-level topologies that utilize scalable HPC topologies such as k-ary n-cube, flattened butterfly and dragonfly for intra-cabinet and inter-cabinet levels that can lead to higher bisection, manageable radix, and reduced link costs, although at higher packet latency due to increased diameter. As photonic technology offers higher bandwidth density, better power efficiency, and higher performance/watt, our proposed exascale network is designed with photonic transceivers such as vertical-cavity surface emitting lasers (VCSELs) and photodiodes, and complementary-metal-oxide semiconductor (CMOS) routers. Our analytical and simulation studies show that multiple-level topologies can achieve 10-15% more throughput while consuming 40% less power when compared to single-level topologies, and VCSEL-based transceivers can deliver the injection bandwidth of 256 GB/sec/direction while consuming 2-3 MW in an exascale systems.**

I. Introduction

Technology scaling combined with increased demands from high-performance computing (HPC) applications are accelerating the growth and performance of future supercomputers. The next frontier in supercomputers is exascale machines that can deliver (10^{18}) or exaflop computational capability. Exascale machines can be built by combining hundreds of thousands individual nodes in which each node, or accelerated processing unit (APU) combines central processing units (CPUs), and graphical processing units (GPUs), and stacked DRAM modules for higher computational capability. A critical component that connects all the nodes is the interconnection network, which should deliver tera-bits per second (Tb/s) of inter-node communication bandwidth within the allocated power budget. For current petaflop machines, the network, which includes the channel and the router, consumes 10-12% of the total power budget [1]. The current Cray XC supercomputer series, with Aries router can scale to 92,544 nodes, and has an injection bandwidth of 4.7-5.25 GB/sec/direction using both electrical and optical links [1]. Therefore, in this paper, we propose the design of exascale networks with three specific design challenges: (i) scale the topology to 100,000 nodes[1], (ii) deliver an injection bandwidth of 256 GB/sec/direction, and (iii) minimize the network power (links, routers) to 10-12% of overall power budget (capped at 20 MW).

Network topology has been evaluated extensively for HPC systems starting with direct networks such as the k-ary, n-cube, flattened butterfly [2] and dragonfly [3] topologies and indirect networks such as the folded Clos or fat-tree networks [4]. Cray XC was built on dragonfly topology instead of folded-Clos because it avoids the need to add network stages as the system size increases [1]. While a dragonfly network has low diameter for exascale networks, fewer global links reduce the bisection bandwidth and require adaptive routing to prevent hot-spots due to congestion. Moreover, the number of ports in a high-radix router affects the router cost when implemented with alternate technologies. In this paper, we advocate multiple-level network topologies similar to two-tier hierarchy (local and global) of dragonfly networks, but that are designed with different or similar scalable HPC topologies can (i) minimize the radix, (ii) increase the bisection bandwidth, and (iii) reduce the global link costs, particularly when designed with optical technology.

The proposed injection bandwidth of 256 GB/sec is many orders of magnitude higher than current petaflop machines. The only known solution to achieve such high data rates while minimizing power consumption is optical technologies [5], [6], [7], [8], [9], [10], [11], [12]. Hybrid integration by direct modulation using on-chip package vertical-cavity surface emitting lasers (VCSELs) and photodiode (PIN) arrays have been demonstrated to deliver up to 25 Gb/sec of data rate per laser at 1-2 pJ/bit [13] [14], [15]. Hybrid integration is a near-term solution because technology is mature, wafer-level testing is available, easier assembly (80,000 per three inch wafers) is possible and high reliability and high-speed modulation are available. Silicon photonics is the alternate technology solution that can deliver high bandwidth-density at higher energy-efficiency and area-efficiency [7], [8], [9]. It is expected that silicon photonic wavelength-division multiplexed (WDM) devices will achieve energy requirements of ∼0.1-0.2

This research was partially funded under Government Contract No. DE-AC52-8MA27344 and subcontract B600716.

[1]Complementary exascale research conducted on the processing side has determined that proposed APUs can deliver 10^{13} flops; therefore, in this work we assume that we will need 100,000 nodes for an exascale machine.

978-1-5090-0172-9/15/$31.00 ©2015 IEEE

pJ/bit at distances less than 1 m. However, as the technology is immature due to thermal sensitivity, we evaluate the proposed architecture using VCSELs. In [23], exascale topologies were evaluated using simplistic models using photonic technology. However, topology analysis was limited to small scale networks (1000 nodes) without layout and intra- and inter-cabinet connections taken into account. In this work, we make the following novel contributions: (i) We explore the design-space of building exascale networks by analyzing a multiple-level approach to local and global networks. We analyze different direct topologies such as the k-ary n-cube, flattened butterfly and dragonfly topologies and indirect networks such as the folded-Clos topologies. Our analysis indicates that multiple-level topologies can reduce the radix, increase bisection, reduce the average hop count, and reduce the number of long optical cables at a slight increase in diameter. (ii) We analyze the design and feasible implementation of optical interconnects by evaluating a VCSEL-based solution by taking into account waveguide layout and cabling costs. Our results indicate that multiple hierarchical topologies can provide both power and area-efficiency for exascale systems. (iii) Our simulation results on a cycle-accurate network simulator shows that multiple-level topologies can achieve 10-15% more throughput than single-level topologies while consuming 20-40% less power for upto 16K nodes.

II. Network Topology

A. Single-level Topology

In this section, we optimize the network parameters of different network topologies for 100,000 nodes. We consider the network characteristics of the four scalable topologies: k-ary n-cube, k-ary n-tree, k-ary n-fly and dragonfly [16]. Each of these topologies have been implemented in prior supercomputers [17], [18], [19], [1]. For instance, k-ary n-cube is edge-symmetric, exploits locality for near-neighbor communication, and offers path diversity and load-balancing capabilities. Similarly, k-ary n-tree or folded-Clos (fat-tree) offers multiple roots that increase path diversity and provide scalable bisection with increasing number of nodes. The k-ary n-fly or flattened butterfly offers one-hop connections within a dimension, reducing both packet latency and power. Dragonfly topologies offer reduced diameter without the excessive link cost of flattened butterfly by directly connecting all nodes within a group using local links and all the groups using global links. Table I shows the overall analysis of single-level topologies.

B. Multi-level Topology

The motivation of designing exascale networks using a multiple-level approach is based on the fact that nodes generally are first combined to reduce the number of switches (via concentration) and then grouped to form a cabinet. Several of these cabinets then are connected to form the entire system. From a scalability and serviceability viewpoint, networks must be modular at the cabinet-level so that they can be serviced in case of faults and added to with new cabinets when workloads demand more computing (scale-out). This naturally splits the system design into two levels; intra-cabinet network and inter-cabinet network. At the intra-cabinet level, all the system super-nodes (after concentration) can be connected via the backplane using optical waveguides. Each of these super-nodes can be connected directly to other cabinets at the inter-cabinet level using optical fibers. To reach 100,000 nodes, each cabinet would contain approximately 500 nodes, and we would need 200 cabinets in total.

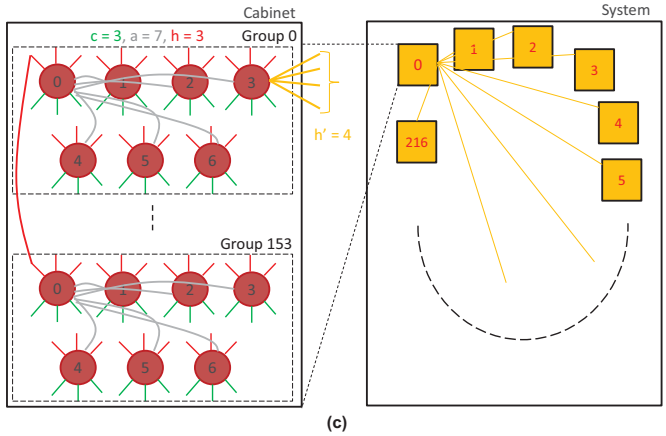

Fig. 1. Multi-level dragonfly topology design (a=7, c=3, h=3 and h=4).

As an example, Figure 1 shows two levels of dragonfly network. Unlike other topologies, there are several combinations for designing multiple-level dragonfly networks (e.g., the first-level topology yields the total number of routers within a cabinet). To connect all routers (inter-cabinet), either the global links can be connected with a 1:1 ratio or they can be overprovisioned (1:2 or 1:4). This increases the global bisection width as there are more links at the inter-cabinet level when these are overprovisioned. To illustrate the overprovisioned example, consider the first level (intra-cabinet) with a=7, c=h=3 as shown on Figure 1. The total number of nodes at the intra-cabinet level, $N = ac(ah+1) = 462$. The total number of routers within a cabinet is $R = a(ah+1) = 154$. To reach 100,000 nodes, there must be at least 216 cabinets. This implies that with a 1:1 global link connection per router, a single cabinet can only connect to 154 cabinets. However, if we overprovision 1:2, then we can connect all cabinets and are left with 92 extra global links. While overprovisioned global links increase the bisection width, they also increase the link cost. In this example, we overprovision 1:4 per router (h'=4) which results in an approximate overprovisioning of 1:3 at the cabinet level connections. We have analyzed several configurations of multi-level dragonfly designs.

Table I shows the overall analysis of single and multiple-level topologies - the evaluation shows the network parameters, degree/radix, diameter, bisection width, average hop count, total number of links (and percentage of long cables), and total number of switches. The first four rows in Table I are from our analysis on single-level optimization. Higher-dimension k-ary n-cube networks (k=7, n=6) have a low degree but higher diameter and link cost. Flattened butterfly networks (c=4, k=4, n=8) reduce the switch cost (due to concentration) and the average hop count, but have both higher link cost and larger degree requirements. Fat-tree topology (k=10, n=4) shows a high bisection width (50% of the total number of nodes), low diameter but high switch counts. In a fat-tree topology, the number of levels dictates the total cost; more levels implies a lower degree or smaller radix but a higher number of links;

Hierarchical Design Topology1 + Topology2	Para--meters	R	D	B_c	H_{avg}	L (% of long links)	S
7-ary 6-cube (kn)	k=7,n=6	13	21	67K	12	1300K (46%)	100K
Flattened Butterfly (FBfly)	c=4,k=4,n=8	32	8	65K	4	800K (38%)	25K
Fat-tree (FT)	k=10,n=4	20	8	50K	7.8	600K (40%)	50K
Dragonfly (Dfly)	a=26,c=12, h=12	50	3-5	26K	2	430K (25%)	8.7K
3DTorus+ Fbfly (kn+Fbfly)	c=4,k=5,n=3 k=6,n=3	13	10.5	40K	$\sim$5	670K (60%)	27K
3DTorus+ Dfly (kn+Dfly)	c=4,k=5,n=3 a=6,c=4,h=4	14	8.5	25K	$\sim$4	350K (48%)	27K
Bfly+ Dfly	c=4,k=5,n=3 a=6,c=4,h=4	20	4	25K	$\sim$2	500K (20%)	27K
Dfly+ Dfly (a=7)	a=7,c=3,h=3 h'=4	16	7	35K	$\sim$3	530K (25%)	33K
Dfly+ Dfly (a=8)	a=8,c=2,h=2 h'=3	14	7	67K	$\sim$3	660K (21%)	47K
Dfly+ Dfly (a=5)	a=7,c=3,h=3 h'=5	15	7	43K	$\sim$3	500K (33%)	33K
Dfly+ FT	a=7,c=3,h=3 k=16,n=2	16	8	67K	$\sim$4	550K (27%)	41K

TABLE I. PERFORMANCE EVALUATION OF SINGLE- AND MULTIPLE-LEVEL TOPOLOGIES WHERE R IS THE RADIX, D IS THE DIAMETER, B_c IS THE BISECTION WIDTH, H_{avg} IS THE AVERAGE HOP DISTANCE, L IS THE TOTAL NUMBER OF LINKS (WITH PERCENTAGE OF LONG LINKS), AND S IS THE TOTAL NUMBER OF SWITCHES.

fewer levels implies a higher degree but a lower number of links. The dragonfly topology (a=26, c=12, h=12) shows low diameter, average hop count, and low number of switches. However, the dragonfly topology offers lower bisection width and the radix of the switch is very high because it has to support the local radii, concentration, and the global links. As the bisection is determined only by the global links, fewer global links implies less bisection ($\sim$25% of the total number of nodes).

From Table I, for multiple-level designs shown, the total radix is reduced to less than 20, the diameter is less than 10 and the average hop count is less than 5. For ($kn + Fbfly$) and ($Fbfly + Fbfly$) topologies, the link costs dominate due to the flattened butterfly topology. For ($kn + Dfly$) and ($Bfly + Dfly$) topologies, because the network topologies are not overprovisioned, the bisection is 25% of the injection bandwidth. Although there are several combinations of ($Dfly + Dfly$), we show three combinations that provide higher bisection bandwidth with reduced radix. These topologies also indicate higher link costs due to overprovisioning of global channels.

III. PHOTONIC INTERCONNECTS

We consider VCSEL-based optical interconnects to satisfy bandwidth and energy-efficiency requirements for exascale topologies. Figure 2 shows a typical optical interconnect link consisting of the serializer, VCSEL drivers, waveguide, photodiodes, transimpedance amplifier (TIA), limiting/sense amplifier (LA/SA) and deserializer. To achieve a link bandwidth of 256 GB/sec or 2 Tb/s, we need 64 VCSELs each running at 32 Gb/s. Each VCSEL array will contain four VCSELs and occupy 1 mm $\times$ 0.25 mm. This is similar to the area required for the photodiodes (PD) to receive the photonic signals. As these VCSEL and PD arrays would be attached to the bottom of the organic carrier of a router, substrate pads need to be eliminated to fit the transceivers. The size of the package pad is 1 mm $\times$ 0.6 mm, and pitch (separation between pads) of these pads is 1 mm. Therefore, to pack four VCSEL arrays (16 total VCSELs) and four PD arrays, we need to eliminate 8 mm of package pads on the substrate edge. Moreover, polymer waveguides are needed to route signals from the VCSELs to the PDs. Polymer waveguides have a 60-micron pitch. Therefore, 16 waveguides need 1 mm and 64 waveguides need 4 mm.

To determine the area overhead, the analysis investigated three VCSEL/PD layouts with waveguides as shown in Figure 3(a) through (c). In Figure 3(a), the analysis extracted/inserted signals via waveguides on the sides of the VCSEL/PD package. The area required for this organization was 16 mm $\times$ 4 mm = 64 mm^2. This design ensures that there are no waveguide crossings, signals are extracted from both sides, and the transmitter and receiver waveguide sets are separated by more than the minimum pitch. This design requires the optical transceivers to be mounted at the edge of the chip. Figure 3(b) shows an alternate design in which the optical transceivers are not at the edge of the pitch, which may simplify the design because the electrical drivers can be located closer to logic and not on the edge of the chip. This design will require additional waveguides to drive to and receive from the edge of the chip. Assuming that the distances are nominal (1-2 mm), the total area overhead of this design is 8 mm $\times$ 6 mm = 48 mm^2. However, because of the minimum package pad pitch, the actual length is 10 mm and not 8 mm, so the total area is 60 mm^2. In Figure 3(c), the waveguides are designed such that the outgoing/incoming directions are orthogonal to the VCSEL/PD layout. This design consumes the most area, 10 mm $\times$ 12 mm = 120 mm^2. This design will be required if package layout requires edge connectors that cannot be mounted on the edge where the VCSEL/PD array is located. Of the three proposed designs shown in Figure 3, (a) restricts the mounting to the edge of the chip while consuming the least area, (b) permits flexibility in transceiver and logic placement, and (c) may be required if there is a mismatch between edge connectors and the transceiver location.

A. Router Microarchitecture

Figure 4(a) shows the router microarchitecture and Figure 4(b) shows the router pipeline. The proposed high-radix switch has either waveguides (from on-board and backplane) or fibers (from inter-cabinet) as input/output ports. Each port has 64 $\times$ 32 Gb/s of photonic signals arriving from or departing to the router. Each of the photonic links is converted to electrical signals via the optical-to-electrical (O/E) blocks shown. This corresponds to the first router pipeline stage. This is followed by converting parallel links into serial using a 8:1 parallel-to-serial (de-serdes) to generate 256 bits of data. The head flit is

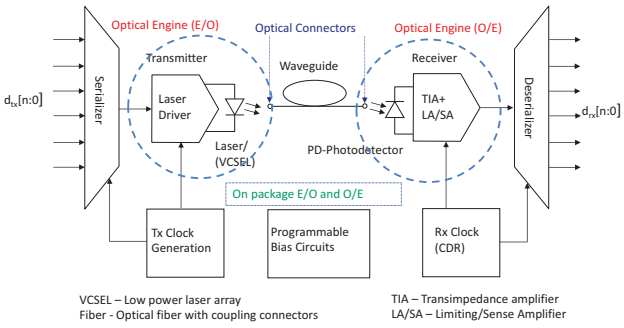

Fig. 2. Design of a photonic link consisting of serializer, laser driver, waveguide, receiver (TIA, LA/SA), and deserializer.

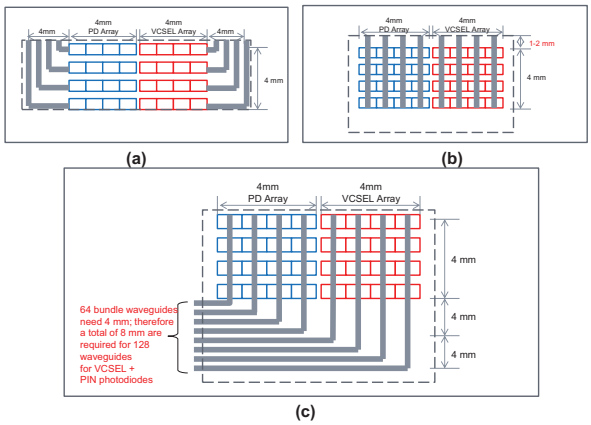

Fig. 3. (a-c) Different VCSEL-PD layouts.

extracted for routing information; we propose source routing, which has the output port selection embedded in the header flit. This process simplifies route computation and reduces the routing logic to a simple lookup. The packet already has the virtual channel (VC) or buffer information embedded to simplify the destination SRAM buffer selection (via a demux). We also embed flow-control information within the flit. We implement credit-based flow control for our proposed architecture; therefore, every incoming flit maintains credit information that is transferred to VC allocation. The second pipeline stage is buffer write combined with route computation (RC) look-up. For the header flit, the information then is written to the VC state table maintained at the router. This enables payload flits to follow the same route to the destination. The third stage is VC allocation (VA) stage, when a buffer at the downstream router is allocated. The credit information is critical to indicate if the flit can propagate to the downstream router. The VC also will create credit information to be sent to the downstream router. The fourth pipeline stage is switch allocation (SA), in which all the flits contend at the switch stage. We propose a non-speculative SA that follows the VC allocation. The fifth router pipeline stage is switch traversal (ST). Here, the serial data is converted back to parallel data using a 1:8 serial-to-parallel (serdes) converter and then transmitted using the VCSELs. There are a total of six router pipeline stages (O/E, buffer write, VA, SA, ST and E/O) in our proposed architecture.

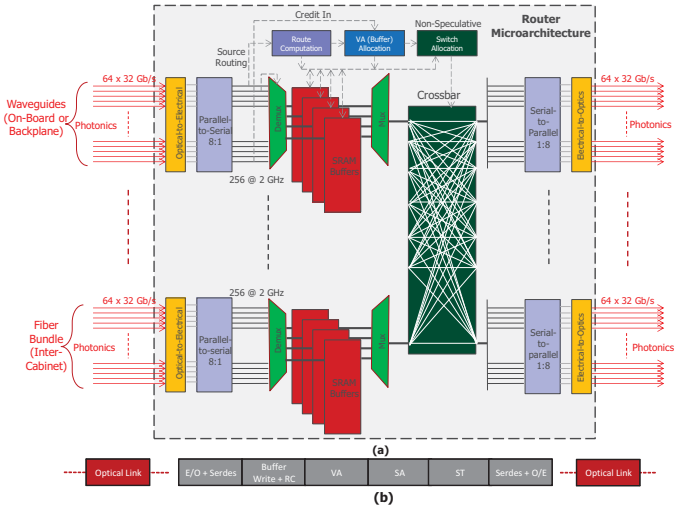

Fig. 4. (a) Router microarchitecture and (b) router pipeline stages: (i) optical-to-electrical conversion + Parallel-to-Serial (O/E + De-serdes), (ii) buffer write + route computation (RC), (iii) virtual channel allocation (VA), (iv) switch allocation (SA), (v) switch traversal (ST) and (vi) serial-to-parallel + optical-to-electrical (serdes + E/O).

B. System Design

Figure 5 shows the overall system design of the proposed exascale system. Figure 5(a) shows the proposed system board, which consists of several nodes concentrated and connected to a router. Each node is connected to the router using polymer waveguides. The proposed router connects to other nodes, the backplane (intra-cabinet links), and via fibers to other cabinets. We propose connecting all system boards via a backplane that consists of polymer waveguides. For inter-cabinet connections, we propose to use fibers; therefore, waveguides are coupled to fibers via Ferrule MT (multi-terminal) connectors. Figure 5(b) and (c) show the proposed layout of the exascale system in which we adopted the parameters from the current Titan machine. With a cabinet width of 0.5 m, depth of 1.5 m, height of 2 m, tray size of 1 m, number of cabinets/row of 25, total number of rows of 8, and aisle width of 2 m, we estimate the row-to-row distance to be 12.5 m and maximum column distance to be 26 m. Therefore, we estimate the maximum manhattan cable length to be 44.5 m for global communication whereas the maximum backplane length for local communication is 2 m. This is used to estimate accurately the round-trip latency and to estimate the buffers required at the router as follows: buffers = distance (meters) $\times$ (speed of light over fiber (nanoseconds/meter)) $\times$ (1 buffer slot/packet) $\times$ data rate (packets/nanoseconds) $\times$ 2 (roundtrip). This estimates the maximum global buffers at 890 slots and local buffers at 40. We rounded the number of buffer slots to 900 and global and local buffer slots to 48, with 6 VCs to prevent deadlocks and head-of-line (HoL) blocking.

IV. Performance Evaluation

A. Power Analysis

In this evaluation, we estimate the router and link power and area for different topologies. For CMOS router power, DSENT0.9 is used [20]. DSENT models D-Flip Flop (DFF) for buffers and multiplexer-based crossbar. For a complete evaluation, we also consider switch allocator and clocking. We

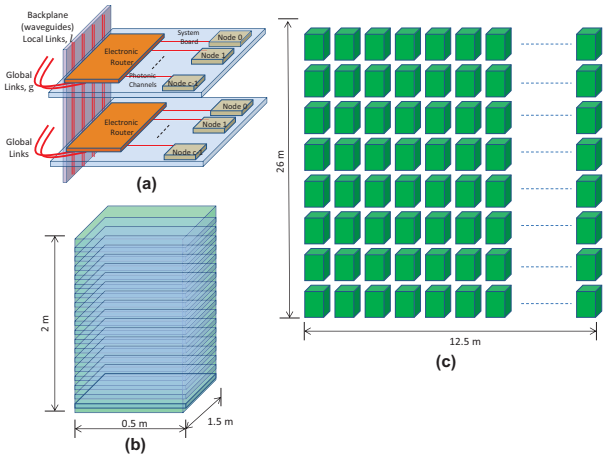

Fig. 5. (a) Backplane and intra-board optical connections, (b) typical cabinet dimensions (adopted from Titan), and (c) layout of cabinets (25×8).

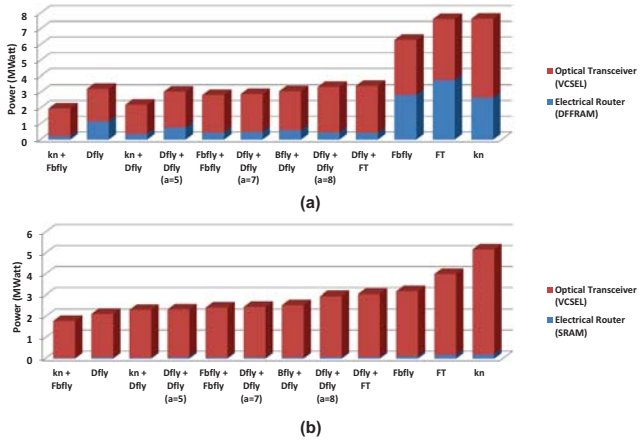

Fig. 6. Network power estimation for 100,000-node exascale system with (a) DFF-RAM buffers and (b) SRAM buffers.

model a two-stage switch allocation. The first stage is used to arbitrate between VCs in the same input port, and the second stage is used to arbitrate between input ports. Broadcast-based H-tree is modeled for clocking. We consider 32-nm bulk CMOS technology node for modeling CMOS components because power and area models are available; however, we consider 32-nm bulk CMOS a very conservative technology choice in the exascale time frame. For on-chip communication, DFF is the choice for buffers. However, for exascale topology, the roundtrip latency for global channels can be very high due to time of flight over long links. Therefore, SRAM is a better technology choice because it is denser and more power-efficient than DFF. CACTI 6.0 was used to model and simulate SRAM buffers.

For the optical transmitter, we project the total laser driver power at 30 mW at 3V supply, clock generation and distribution for 16 lanes at 40 mW (2.5 mW/lane), serializer at 10 mW with 16 inputs at 2 Gb/sec which gives the total transmitter power to be 42.5 mW/lane or 1.33 pJ/bit at 32 Gb/sec. For the optical receiver, our evaluation assumes a three-stage TIA at 1.5V, 7 mA and -15 dBm sensitivity, clock and data recovery (CDR), and deserializer for a total receiver power of 0.53 pJ/bit. Therefore, the total energy efficiency achieved at 32 Gb/s is 1.86 pJ/bit or 60 mW. This implies that if we design 64 VCSEL transmitters at 60 mW, then the total power is 3.84 watts per 256 GB/sec or 7.68 watts per bidirectional port. The power numbers are obtained by a combination of technology scaling and published numbers from literature.

Figure 6(a) shows the network power for 100,000 nodes using DFF-RAM and Figure 6(b) shows the power using SRAM buffers. As seen, SRAM reduces the impact on the overall power consumption compared to DFF due to the large number of buffers needed to overcome the round-trip time. This is especially true for long global channels connecting cabinets that span distances of 12.5 meters $\times$ 26 meters (from Titan). Except for dragonfly, all single-level topologies consume the most power for both SRAM and DFF-RAM buffers. Single tier dragonfly reduces the total power due to the smaller number of switches. The other scalable topology is k-ary n-cube, which also consumes less power due to higher number of local connections. Most of the combined topologies evaluated (Dfly+Dfly, Bfly+Dfly, Dfly+FT) consume more power than k-ary n-cube networks and single-level networks. An exascale system is expected to consume 20 MW of power including the processor, memory and network. Current petaflop machines use 10-12% of the system power for the network, which implies that most of our multiple-level topologies (2-3 MW) are viable candidates. Figure 7 shows power estimation if the bandwidth of the optical transceiver is reduced to (a) 50% (128 GB/sec/direction) and (b) 25% (64 GB/sec/direction) of the injection bandwidth using only SRAM buffers. Because CMOS router components (buffers/crossbars) have lower power dissipation than optical transceivers, reducing link bandwidth to 50% of injection bandwidth directly benefits the overall link power consumption. As observed, the power dissipation decreases by 50% when link bandwidth is reduced to 50% of injection bandwidth. However, when the link bandwidth is reduced to 25% (64 GB/sec/direction), then CMOS routers start affecting the overall power consumption. In this case, the power dissipation decreases to 30-35% rather than 25% because CMOS components will consume more power. With DFF-RAM components, the power reduction is 25-30% when the link bandwidth is reduced to 50% of injection bandwidth as DFF-RAM buffers consume a substantial portion of the total network power. Clearly, SRAM-based buffers are the better choice due to lower power dissipation. For link rates greater than 4 GB/sec, optical transceivers will dominate the overall power consumption.

B. Area Analysis

Figure 8 shows the overall area for the optical and CMOS portions of the router required for different topologies. The area evaluation here is restricted to CMOS portions (buffers, crossbars) and optical transceivers of the router die. The area parameters were extracted from DSENT for DFF-RAM-based buffers, crossbar, clocking, and allocators. CACTI provided the area overhead for SRAM-based buffers. Single-level dragonfly consumes the most area simply because of the high radix, which is almost 3 times greater than other combined topologies. We assume an area overhead of 60 mm^2 (Figure 3(b)) for the optical transceiver layout. Most combined topologies have an area of 1000 mm^2 compared to Aries router from Cray

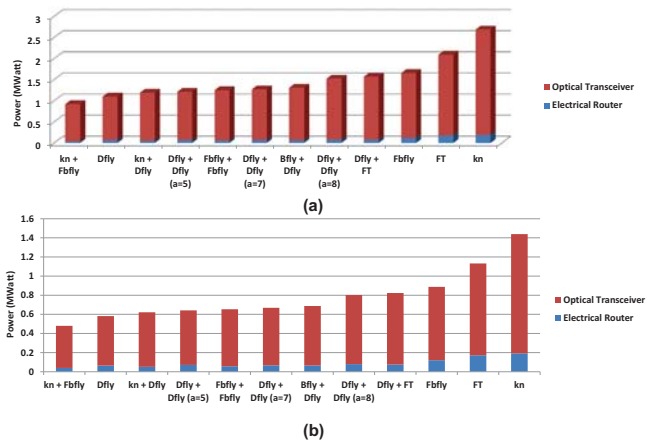

Fig. 7. Network power estimation for 100,000-node exascale system with SRAM buffers with (a) 50% (128 GB/sec/direction) and (b) 25% (64 GB/sec/direction) of injection bandwidth.

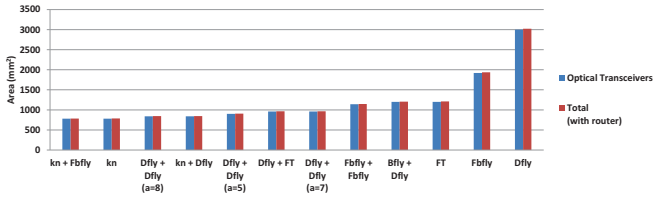

Fig. 8. Area estimation for optical transceiver and total router area including buffers, crossbars, and allocators.

XC, which has an area of 16.6 $mm \times 18.9\ mm$ (or 313.74 mm^2), our assumed area is only 3 times larger. Dragonfly topology has a significant area overhead due to high-radix design; advanced technology solutions such as 3D stacking could address such the area problems. The area overhead evaluated here is for only the router components (i.e. CMOS router and optical transceivers); optical waveguides at the board and backplane levels are excluded.

C. Throughput, Latency and Power: 1,024 Nodes

To gain insight into the network performance, we modeled and tested different single-level and multiple-level topologies using a cycle-accurate network simulator. The proposed network evaluation merges single-level networks such as mesh, flattened butterfly and dragonfly and multi-level networks such as mesh$_{local}$-dragonfly$_{global}$ (G$_{dragon}$-L$_{mesh}$) flattened butterfly$_{local}$-dragonfly$_{global}$ (G$_{dragon}$-L$_{fb}$), and dragonfly$_{local}$-dragonfly$_{global}$ (G$_{dragon}$-L$_{dragon}$). Fat tree was not considered in the performance measurement because all networks under consideration are direct networks. All designs were tested with different synthetic traffic patterns such as uniform random, bit reversal, butterfly, matrix transpose, complement, perfect shuffle and worst case traffic pattern for the network under test.

Figures 9(a) through (d) show the network latency for uniform random, bit reversal, butterfly, and matrix transpose traffic patterns for 1,024 nodes. For uniform traffic, both dragonfly and flattened butterfly show the best latency, which is almost 10% better when compared to multi-level designs. This is expected as the hop count in multi-level topologies is greater

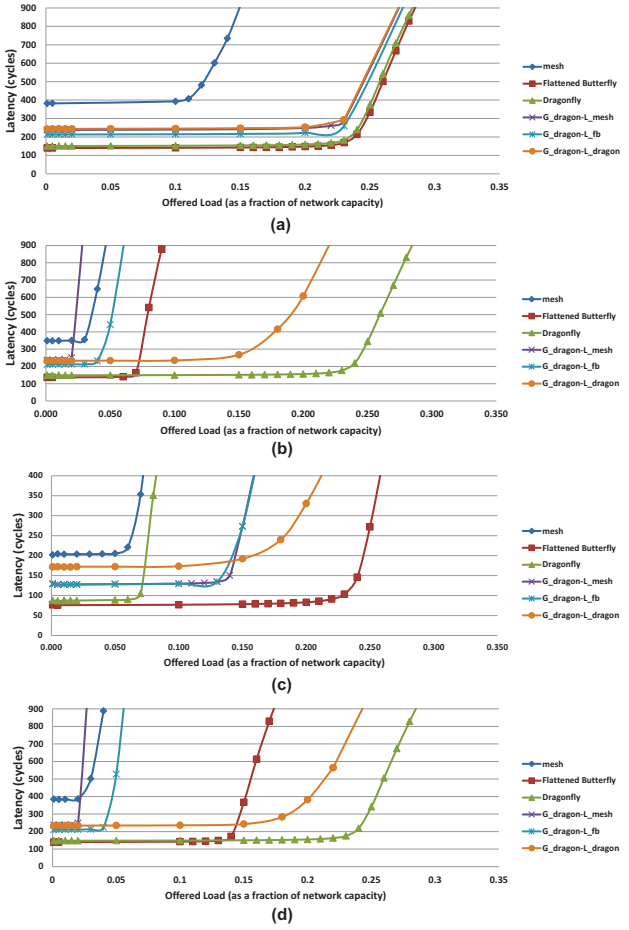

Fig. 9. Network latency measured for various topologies for 1,024 nodes for (a) uniform random, (b) bit reversal, (c) butterfly and (d) matrix transpose traffic pattern.

than in dragonfly and flattened butterfly topologies. The combined topologies G$_{dragon}$-L$_{dragon}$, G$_{dragon}$-L$_{fb}$ and G$_{dragon}$-L$_{mesh}$ perform reasonably well given that these topologies increase the diameter of the network. For bit reversal traffic, dragonfly topology is best due to the communication pattern. Dragonfly saturates at 25% network load whereas the next closest topology of G$_{dragon}$-L$_{dragon}$ saturates as 18%. The remaining topologies saturate much earlier; the traffic creates almost a worst-case traffic scenario for mesh and flattened butterfly networks.

For butterfly, single-level flattened butterfly provides the best performance. The traffic created in both networks forces packets to travel to the furthest nodes in different dimensions. As flattened butterfly connects all nodes within a single dimension, the hop count decreases, and both complement and butterfly traffic patterns favor flattened butterfly topology. In both scenarios, flattened butterfly saturates at 25% of network load. For butterfly traffic, multi-level dragonfly topology provides the next best performance. G$_{dragon}$-L$_{dragon}$ saturates at 20% of network load, and G$_{dragon}$-L$_{fb}$ and G$_{dragon}$-L$_{mesh}$ saturate at 15% of network load, but all multi-level topologies have higher zero-load latency. Because multi-level topologies increase the diameter and hop count, the zero-load latency (when there are no other packets in the network) is higher. For complement traffic, all multi-level topologies saturate at

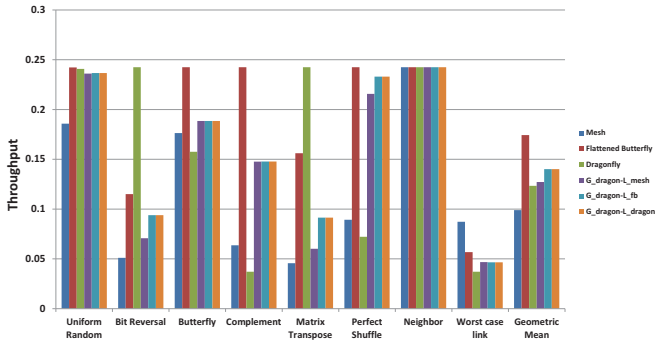

Fig. 10. Saturation throughput for 1,024 nodes for various traffic patterns.

15% of the network load. Multi-level topologies provide a design trade-off between radix and performance. Flattened butterfly requires very high radix, whereas multi-level topologies balance the radix while delivering reasonable performance. For matrix transpose, dragonfly topology provides the best performance by saturating at 25% of network load. G_{dragon}-L_{dragon} performs the next best by saturating at 20% of network load.

Figure 10 shows the saturation throughput for various traffic patterns. The last column shows the geometric mean for various traffic patterns. The results follow the earlier network saturation points; if a network topology saturates at higher network load, then the throughput is also higher. For instance, for uniform random, flattened butterfly and dragonfly provide the highest throughput, which is marginally higher when compared to the multi-level topologies. The geometric mean of the averages provide some insights into how these diverse traffic will affect performance; flattened butterfly provides the best overall performance for the majority of traffic patterns; however its increased radix and router complexity render it infeasible for an exascale network. Although easier to implement, mesh topology has extremely high diameter, which again makes it uncompetitive. When dragonfly and all the remaining multiple-level topologies that implement dragonfly at the global level are considered, multiple-level topologies perform 10-15% better than a single-level topology. One reason for the improvement is the increased path diversity in multiple-level topology. Another reason is that the routing algorithm implemented in our models is mostly shortest path, which could lead to increased congestion. Multiple-level topologies provide improved performance without having to resort to Valiant's algorithm [22] for distributing the hot-spots. Moreover, the same performance is delivered at much reduced radix, thereby reducing the router complexity and cost.

Figure 11(a) shows the power dissipated for 1,024 nodes with DFF-RAM buffers, and Figure 11(b) shows the power dissipated with SRAM-based buffers. The multiple-level topologies consume 40% less power than single-level dragonfly and 60% less power than flattened butterfly for both scenarios. The higher radix of the router contributes to increased router power as well as the optical links. However, as the size of the network increases, higher radix topologies will have a power advantage because they can connect more routers directly than multiple-level topologies.

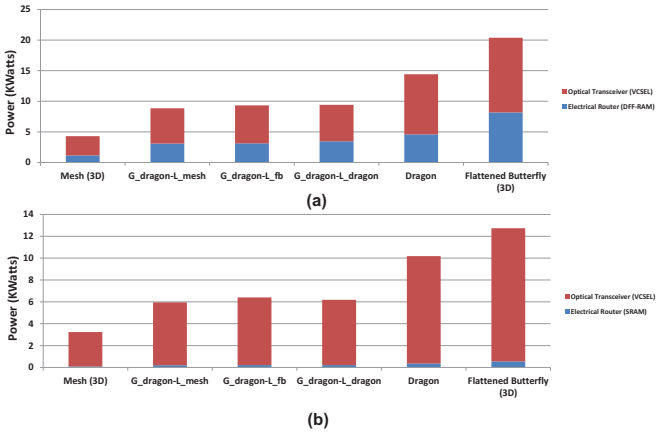

Fig. 11. Power dissipated for 1K nodes for different topologies with (a) DFF-RAM and (b) SRAM buffers for uniform random traffic.

D. Sensitivity Studies: Network Size

We increased the size of the network to 16,384 nodes and tested variations of dragonfly networks (i.e. G_{dragon}-L_{dragon}, G_{dragon}-L_{fb} and G_{dragon}-L_{mesh}) with a single-level dragonfly topology. Figure 12(a) and (b) show the network latency for uniform random and bit reversal traffic patterns. For uniform (Figure 12(a)), the multiple-level topologies show an increase in steady-state latency compared to a single-level dragonfly topology. All networks saturate between 23% and 25% indicating that multiple-level designs show similar performance for uniform random. For bit reversal (Figure 12(b)), G_{dragon}-L_{fb} and G_{dragon}-L_{mesh} topologies show early saturation (less than 5%) due to local network saturating. For dragonfly topology, this is not the worst-case traffic pattern; therefore, both G_{dragon}-L_{dragon} and single-level dragonfly saturate beyond 20% of network load.

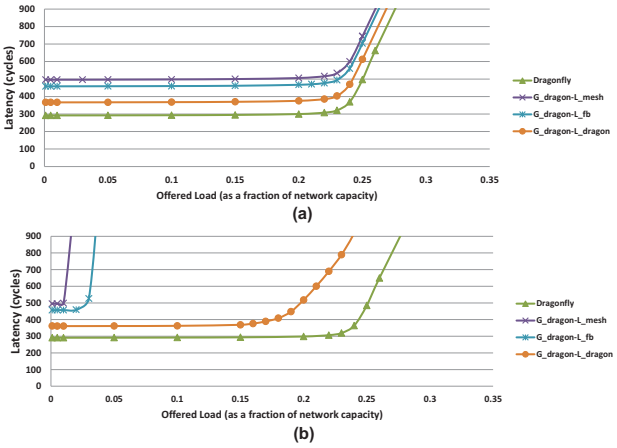

Fig. 12. Network latency measured for various topologies for 16,384 nodes for (a) uniform random and (b) bit reversal traffic pattern.

Figure 13 shows the network saturation throughput for all traffic scenarios. The last column shows the geometric mean of all the traffic patterns. As observed, multiple-levels of dragonfly topologies such as G_{dragon}-L_{dragon}, G_{dragon}-L_{fb} and G_{dragon}-L_{mesh} perform almost 50% better than single-level dragonfly topology. Both complement and perfect shuffle patterns affect the throughput of single-level dragonfly

topology due to congestion in the network. Although multiple-levels increase the latency, additional routes allow for path diversity, which reduces the impact on network congestion. Therefore, multiple-levels of dragonfly design shows improved throughput with increase in network size. We expect similar results when the network size increases to more than 16,384 nodes.

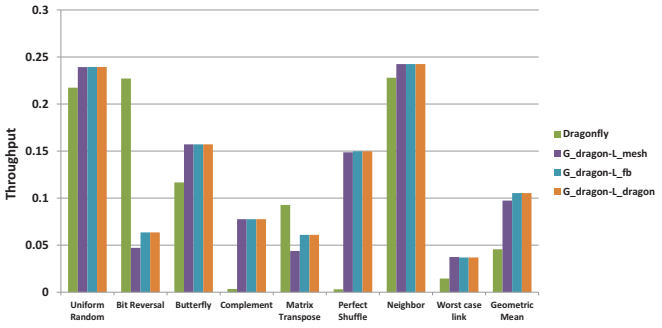

Fig. 13. Saturation throughput for 16,384 nodes for various traffic patterns.

V. Conclusions

In this paper, we identify the three challenges associated with design of the interconnect architecture for exascale: scaling nodes to 100,000 nodes, delivering injection bandwidth of 256 GB/sec/direction, and minimizing network power to 2-3 MW. One approach to address the shortcomings of a single-level topology (lower bisection, higher radix) is to develop multi-level topologies by splitting the exascale network into intra-cabinet and inter-cabinet networks. The proposed multi-level topologies combine other well-known topologies such as k-ary n-cube, flattened butterfly and dragonfly topologies to improve network characteristics such as reduced radix, higher bisection and reduced cable costs. Photonic technology can provide high performance/watt, higher bandwidth density and improved scalability compared to metallic interconnects. Photonic interconnects using arrays of VCSELs and photodiodes were considered to achieve 256 GB/sec/direction of injection bandwidth. Our evaluation showed that the network power is between 2-3 MW with SRAM buffers for the multiple-level topologies we evaluated. In addition, the minimum power of the VCSELs at 32 Gb/s required to achieve a BER of 10^{-15} with a minimum receiver sensitivity of -15 dBm was evaluated by considering the optical losses due to packaging at the board and system levels. Our simulation results indicate that multiple-level topologies can deliver 10-15% higher throughput while reducing the power consumed by 40% although at a higher packet latency when compared to a single-level topology.

VI. Acknowledgment

We thank Petre Popescu for providing the analysis for VCSEL power estimation and thank Steve Reinhardt for valuable discussions involving the design of the architecture.

References

[1] B. Alverson, E. Froese, L. Kaplan, and D. Roweth, "Cray xc series network," CRAY, White Paper WP-Aries01-1112, 2012.

[2] J. Kim, W. J. Dally, and D. Abts, "Flattened butterfly: Cost-efficient topology for high-radix networks," in *Proceedings of 34th Annual International Symposium on Computer Architecture(ISCA)*, June 2007, pp. 126 – 137.

[3] J. Kim, W. Dally, S. Scott, and D. Abts, "Technology-driven, highly-scalable dragonfly topology," in *Proceedings of International Symposium on Computer Architecture (ISCA)*, June 2008, pp. 77 – 88.

[4] F. Petrini, A. Hoisie, S. Coll, and E. Frachtenberg, "The quadrics high performance clustering technology," *IEEE Micro*, vol. 22, no. 1, pp. 46–57, 2002.

[5] D. A. B. Miller, "Device requirements for optical interconnects to silicon chips," *in Proceedings of the IEEE, Special Issue on Silicon Photonics*, vol. 97, pp. 1166–1185, 2009.

[6] R. G. Beausoleil, P. J. Kuekes, G. S. Snider, S.-Y. Wang, and R. S. Williams, "Nanoelectronic and nanophotonic interconnect," *Proceedings of the IEEE*, vol. 96, no. 2, pp. 230–247, February 2008.

[7] A. V. Krishnamoorthy, R. Ho, X. Zheng, H. Schwetman, J. Lexau, P. Koka, G. Li, I. Shubin, and J. E. Cunningham, "Computer systems based on silicon photonic interconnects," in *Proceedings of the IEEE*, vol. 97, no. 7, June 2009, pp. 1337–1361.

[8] A. Biberman, K. Preston, G. Hendry, N. Sherwood-Droz, J. Chan, J. S. Levy, M. Lipson, and K. Bergman, "Photonic network-on-chip architectures using multilayer deposited silicon materials for high-performance chip multiprocessors," *ACM Journal on Emerging Technologies in Computing Systems*, vol. 7, pp. 1–25, July 2011.

[9] M. Georgas, J. Leu, B. Moss, C. Sun, and V. Stojanovic, "Addressing link-level design tradeoffs for integrated photonic interconnects," in *CICC*, 2011, pp. 1–8.

[10] A. K. Kodi and A. Louri, "Energy-efficient and bandwidth reconfigurable photonic networks for hpc systems," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 17, pp. 384–395, April 2011.

[11] R. Morris, A. Kodi, and A. Louri, "Reconfiguration of 3d photonic on-chip interconnects for maximizing performance and improving fault tolerance," in *45th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO-45)*, 2012.

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

[13] M. A. Taubenblatt, "Optical interconnects for high-performance computing," *Journal of Lightwave Technology*, vol. 30, no. 4, pp. 448–457, 2012.

[14] http://www.hpcwire.com/topic/networks/finisar-intros-150-gbps-parallel-active-optical-cable-48806432.html.

[15] http://www.intel.com/pressroom/kits/hpc/index.htm.

[16] W. Dally and B. Towles, *Principles and Practices of Interconnection Networks*. Morgan Kaufmann, 2003.

[17] N. R. Adiga et al., "Blue gene/l torus interconnection network," *IBM Journal on Research and Development*, vol. 49, no. 2/3, 2005.

[18] A. Dhodapkar et al., "Seamicro sm 10000-64 server: Building datacenter servers using cell phone chips," in *Hot Chips 23*, August 2011.

[19] B. Arimilli et al., "The percs high-performance interconnect," in *Proceedings of the 18th Symposium on High-Performance Interconnects (HotI10)*, August 2010, pp. 75–82.

[20] C. Sun et al., "Dsent a tool connecting emerging photonics with electronics for opto-electronic networks-on-chip modeling," in *6th ACM/IEEE International Symposium on Networks-on-Chip*, May 2012.

[21] R. Dangel et al., "Polymer-waveguide-based board-level optical interconnect technology for datacom applications," *IEEE Transactions on Advanced Packaging*, vol. 31, no. 4, pp. 759–767, November 2008.

[22] L. G. Valiant and G. J. Brebner, "Universal schemes for parallel communication," in *Proc. of the ACM Symposium of the Theory of Computing*, 1981, pp. 263–277.

[23] A. Kodi, B. Neel, and W. Brantley, "Photonic Interconnects for Exascale and Datacenter Architectures," in *IEEE Micro Magazine*, 2014, pp. 18–30.