Photonic Interconnects for Exascale and Datacenter Architectures

Multitier network topologies combine scalable topologies for local and global interconnects to improve bisection, minimize radix, and reduce link costs, albeit at higher packet latency. The authors envision an entire exascale network comprising photonic links for communication and CMOS routers for switching. Compared to the single-level dragonfly topology, multitier topologies provided similar power and latency, higher bisection, and reduced area overhead.

Avinash Karanth Kodi
Ohio University

Brian Neel
William C. Brantley
Advanced Micro Devices

Technology scaling and increased demands from high-performance computing applications are accelerating the growth and performance of future supercomputers and large datacenters. The next frontier in supercomputers is exascale machines that can deliver exaflop (10^{18}) computational capability. Exascale machines can be built by combining hundreds of thousands of individual nodes that combine CPUs and GPUs with stacked DRAM modules.

The interconnection network, which connects all the exascale system's nodes, should deliver high communication bandwidth within the allocated power budget.[1] The interconnection network topology determines critical network characteristics, such as the switch degree or radix, diameter (maximum hop between two nodes), bisection width, and number of links. These variables directly relate to the exascale computing system's power, performance, and area.

Researchers have studied network topology extensively for high-performance computing 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 topologies.[4] The Cray XC[5] was designed on the dragonfly topology instead of folded-Clos because dragonfly avoids the need to add network stages as the system size increases. The dragonfly topology comprises groups of nodes; at the intragroup (local) level, the routers are fully connected, and at the intergroup (global) level, all groups are fully connected. Although the dragonfly network has low diameter for exascale networks, it has fewer global links, which reduces the available bisection bandwidth, and its minimal path diversity could cause congestion, requiring adaptive routing. Moreover, the number of ports in a high-radix router affects the router area and design complexity.

Published by the IEEE Computer Society

Table 1. Network characteristics of scalable topologies for exascale systems.

Topology	Node (N)	Degree or radix (d)	Diameter (D)	Bisection width (Bc)	Avg. hop count (H_{avg})	No. of routers	No. of links
k-ary, n-cube	k^n	$2n+1$	$nk/2$	$4k^{n-1}$	$nk/4 \rightarrow$ even k; $n(k/4-1/4k) \rightarrow$ odd k	k^n	$N(2n+1)$
Fat tree (k-ary, n-tree)	k^n	$2k$	$2n = 2\log_k N$	$N/2$	$2[n-(1/(k-1))]$	$N/k(\log_k N)$	$N(\log_k N+1)$
Flattened butterfly (k-ary, n-fly, c = concentration)	ck^n	$c+n(k-1)$	n	$(k/2)^2k^{n-1}$	$n/2$	k^n	$(N/c)[c+n(k-1)]$
Dragonfly (a = local routers, h = global links, c = concentration)	$ac(ah+1)$	$c+a-1+h$	3 (2 local, 1 global)	$(ah/2)^2$	~ 2	$a(ah+1)$	$a(ah+1) \times (c+a+h-1)$

In this article, we advocate multitier network topologies with a two-tier hierarchy similar to a dragonfly network (local and global), but designed with topologies that minimize radix and router complexity, increase bisection bandwidth, and provide similar throughput with slightly higher latency. We target 100,000 nodes for exascale networks, comparable to the Cray XC series with 92,544 nodes.

Because power consumed by metallic interconnects increases with distance, researchers are investigating emerging optical technologies for implementing interconnection networks in large-scale networks. Hybrid integration by direct modulation using an on-chip package of vertical-cavity surface-emitting lasers (VCSELs) can deliver data rates up to 25 Gbits/second (Gbps) at 1 picojoule (pJ) per bit of power consumption.[6] Commercial VCSELs are commodity components, and most of the optics available for data communication today are based on VCSEL technology. Hybrid integration is a near-term solution because the technology is mature, and wafer-level testing, easy assembly (80,000 per 3-inch wafer), high reliability, and high-speed modulation are available.

Silicon photonics is the alternate technology solution; it can deliver high-density bandwidth at higher energy and area efficiencies.[7] Both academia and industry are actively pursuing fabrication of silicon photonic devices. Silicon photonic-enabled, wavelength-division multiplexed devices are expected to achieve energy requirements of 1 pJ/bit at short distances (less than 10 cm). However, the technology is immature, and thermal sensitivity and device capability are still being researched. In this article, we analyze the impact of VCSELs for implementing exascale systems with an injection bandwidth of 256 GBps or 2 terabits/second (Tbps).

Single-tier and multitier network topology

We develop network topologies using single-tier and multitier approaches employing four scalable topologies (k-ary n-cube, flattened butterfly, dragonfly, and fat tree), because these topologies have been implemented in prior supercomputers.[8-12] Table 1 shows the network characteristics of these scalable topologies. To determine the ideal parameters using each of these topologies, we ran analytical models by varying one of the parameters used for designing the exascale network with a single-tier topology. For example, with k-ary n-cube networks, we asked, what ideal values of k and n will yield 100,000 nodes such that network degree and diameter are minimized while providing high

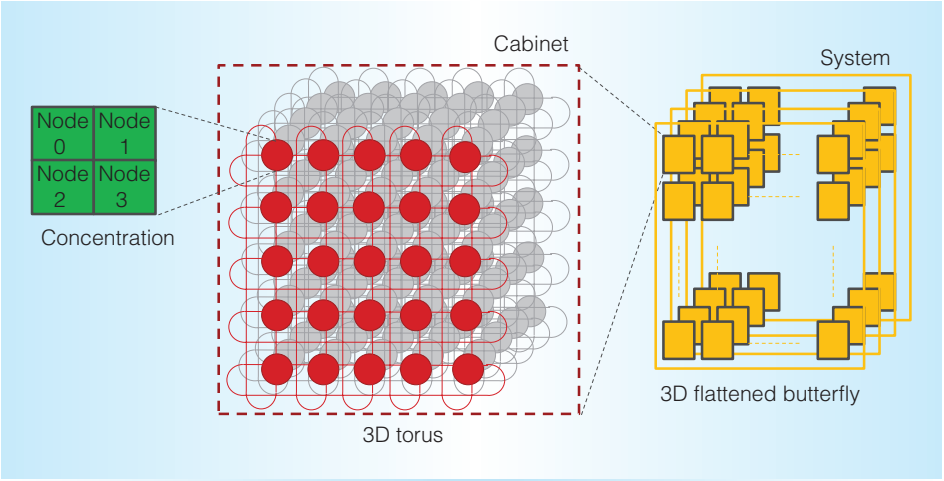

Figure 1. Multitier design in which supernodes are combined ($c = 4$) into a 3D torus ($k = 5$, $n = 3$), and each 3D torus is connected as a flattened butterfly network ($k = 6$, $n = 3$). With $c = 4$, $k = 5$, and $n = 3$, a total of 500 $(= 4 \times 5^3)$ nodes can be connected at the intracabinet level. Each of these cabinets is connected using a flattened butterfly network in three dimensions with $k = 6$ and $n = 3$, giving 216 $(= 63)$ total cabinets.

bisection, a minimal number of switches, and a nominal number of links? We implemented a similar optimization for each networks to design a 100,000-node system using a single-tier approach.

The motivation for using a multitier approach to design exascale networks is based on the fact that, generally, nodes are first combined to reduce the number of switches (via concentration) and then grouped to form a cabinet. Several cabinets are then connected to form the system. Networks must be modular at the cabinet level to allow for servicing in case of faults and for expansion with new cabinets when workloads demand more computing. This naturally splits the system design into two levels: intracabinet and intercabinet networks. At the intracabinet level, all the system supernodes (after concentration) can be connected via the backplane using optical waveguides. Each supernode can be connected directly to other cabinets at the intercabinet level using optical fibers. To reach 100,000 nodes, for example, we could connect approximately 200 cabinets that each contain 500 nodes.

Extending the example, consider a multitier design that combines a 3D torus and a flattened butterfly as shown in Figure 1. We chose a 3D torus instead of higher dimension to reduce the link penalty generally associated with higher dimensions. Four nodes are concentrated into one supernode (see the inset in the figure), and each supernode is connected using a 3D torus. With $c = 4$, $k = 5$, and $n = 3$, a total of 500 $(= 4 \times 5^3)$ nodes can be connected at the intracabinet level. Now each cabinet is connected using a flattened butterfly network in three dimensions with $k = 6$ and $n = 3$, giving 216 $(= 6^3)$ total cabinets. Figure 2 shows the actual connections between the cabinets. For clarity, only two dimensions out of three are shown. Each supernode within the 3D torus has exactly three outgoing connections to three other supernodes along each dimension.

As a second example, Figure 3 shows two levels of a dragonfly network. Unlike other topologies, there are several combinations for designing multitier dragonfly networks; for example, the first-level topology yields the total number of routers within a cabinet. To connect all routers (intercabinet), the global links can be connected with a 1:1 ratio or overprovisioned (1:2 or 1:4). This increases the global bisection width because there are more links at the intercabinet level when these are overprovisioned. To illustrate the overprovisioned example, consider the first level (intracabinet) with $a = 7$ and $c = h = 3$, as shown in Figure 3. The total number of nodes at the intracabinet level, $N' = ac(ah +$

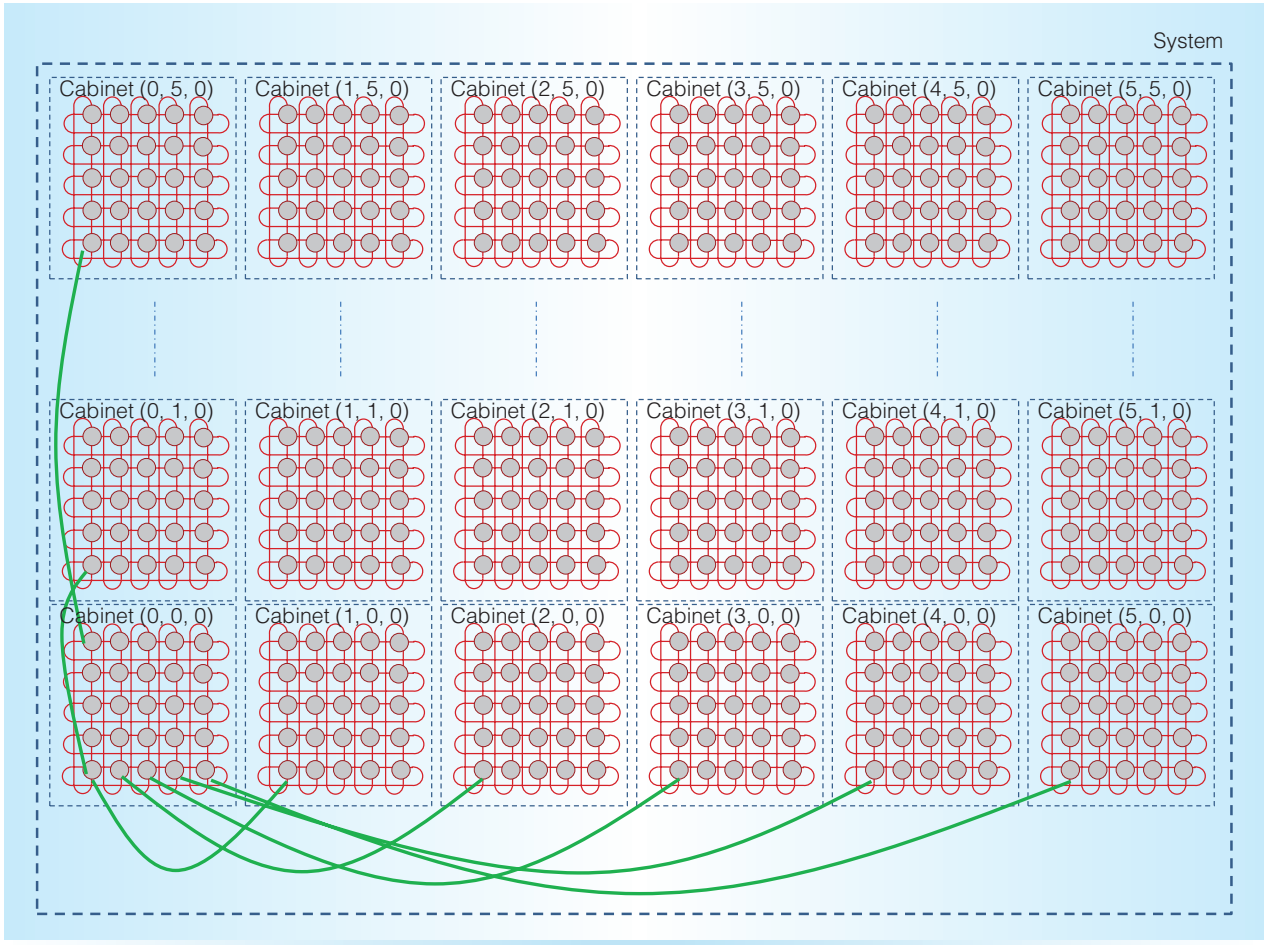

Figure 2. Connections between cabinets. Each supernode within the 3D torus is connected to three supernodes in each dimension (only two dimensions are shown). Each cabinet is numbered as (*x*, *y*, *z*), and all connections from cabinet (0, 0, 0) are shown for clarity.

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 connect to only 154 cabinets. However, if we overprovision 1:2, then we can connect all cabinets and are left with 92 extra global links. Although 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.

Table 2 shows the overall analysis of single-tier and multitier topologies. The evaluation shows the network parameters, degree or 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 are from our analysis on single-tier optimization. Higher-dimension k-ary n-cube networks ($k = 7$, $n = 6$) have a low degree, but a higher diameter and link cost. Flattened butterfly networks ($c = 4$, $k = 4$, $n = 8$) reduce the switch cost (because of concentration) and the average hop count, but have both higher link cost and larger degree requirements. Fat-tree topologies ($k = 10$, $n = 4$) show a high bisection width (50 percent of the total number of nodes) and low diameter but high switch counts. In a fat-tree topology, the number of levels dictates the total cost: more levels imply lower degree or smaller radix but a higher number of links, and fewer levels imply higher degree but a lower number of links. Dragonfly topologies ($a = 26$, $c = 12$, $h = 12$) show low diameter,

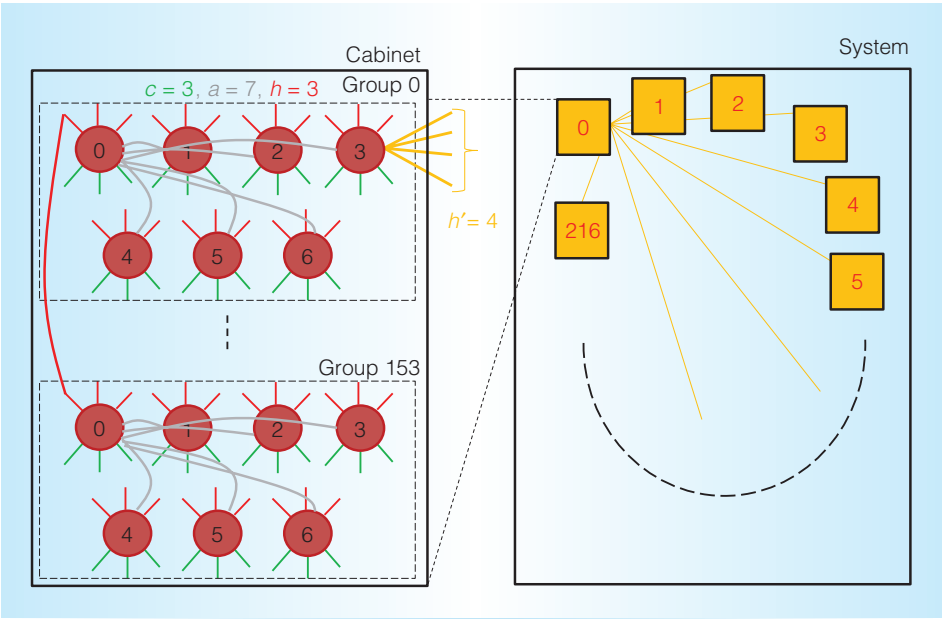

Figure 3. Multilevel dragonfly topology design ($a = 7$, $c = 3$, $h = 3$, and $h' = 4$). With $a = 7$, $c = 3$, and $h = 3$, the total number of nodes at the intracabinet $N' = 462$, and the total number of routers is 154. To reach 100,000 nodes, we need at least 216 cabinets; however, with 1:1 provision, we can only connect 154. With 1:4 overprovision ($h' = 4$), we can increase the global bandwidth.

average hop count, and low number of switches. However, dragonfly offers lower bisection width, but the switch's radix is high because it must support the local radii, concentration, and global links. Because the bisection is determined only by the global links, fewer global links imply less bisection (approximately 25 percent of the total number of nodes).

For the multitier designs shown, the total radix is reduced to below 20, the diameter is less than 10, and the average hop count is less than five. For (*kn* + *Fbfly*) and (*Fbfly* + *Fbfly*) topologies, the link costs dominate because of the flattened butterfly topology. For (*kn* + *Dfly*) and (*Fbfly* + *Dfly*) topologies, because the network topologies are not overprovisioned, the bisection is 25 percent of the injection bandwidth. Although there are several combinations of (*Dfly* + *Dfly*), three provide higher bisection bandwidth with reduced radix. More importantly, these topologies double the diameter, owing to the two dragonfly levels the packet must travel from source to destination. Multitier topologies (two-level *Dfly* or *Dfly* + *FT*) can provide the ideal combination of lower degree and average hop count with a lower number of links.

Photonic technology and router microarchitecture

VCSEL-based optical interconnects satisfy bandwidth and energy-efficiency requirements for exascale topologies. Current VCSELs support 16 Gbps of data rate per laser, and 25-Gbps prototypes are being tested.[6] Therefore, we expect 32-Gbps data rate VCSELs to be available in 2 to 3 years and available in volume in the 2020 timeframe. To achieve a link bandwidth of 256 Gbytes/second or 2 Tbps, we need 64 VCSELs running at 32 Gbps each. Each VCSEL array will contain four VCSELs and occupy 1 mm × 0.25 mm. This is similar to the area needed for the photodetector (PD) to receive the photonic signals.

Because these VCSEL and PD arrays would be attached to the bottom of a router's organic carrier, substrate pads need to be eliminated to fit the transceivers. The size of the package pad is 1 mm × 0.6 mm, and pitch (separation between pads) is 1 mm. Therefore, to pack four VCSEL arrays (16

Table 2. Evaluation of multitier hierarchical topologies.

Hierarchical design (topology 1 + topology 2)	Parameters	Degree or radix (*d*)	Diameter (*D*)	Bisection width (*Bc*) (in thousands)	Avg. hop count (H_{avg})	Total no. of links (*L*) (in thousands), long links percentage	No. of switches (*S*) (in thousands)
7-ary 6-cube (*kn*)	$k=7, n=6$	13	21	67	12	1,300 (46%)	100
Flattened butterfly (*Fbfly*)	$c=4, k=4, n=8$	32	8	65	4	800 (38%)	25
Fat tree (*FT*)	$k=10, n=4$	20	8	50	7.8	600 (40%)	50
Dragonfly (*Dfly*)	$a=26, c=12, h=12$	50	3–5	26	2	430 (25%)	8.7
3D torus + Fbfly (*kn* + *Fbfly*)	$c=4, k=5, n=3$; $k=6, n=3$	13	10.5	40	~5	670 (60%)	27
Fbfly + *Fbfly*	$c=4, k=5, n=3$; $c=6, n=3$	19	6	40.5	~3	830 (48%)	27
3D torus + *Dfly* (*kn* + *Dfly*)	$c=4, k=5, n=3$; $a=6, c=4, h=4$	14	8.5	25	~4	350 (28%)	27
Fbfly + *Dfly*	$c=4, k=5, n=3$; $a=6, c=4, h=4$	20	4	25	~2	500 (20%)	27
Dfly + *Dfly* ($a=7$)	$a=7, c=3, h=3; h'=4$	16	7	35	~3	530 (25%)	33
Dfly + *Dfly* ($a=8$)	$a=8, c=2, h=2; h'=3$	14	7	67	~3	660 (21%)	47
Dfly + *Dfly* ($a=5$)	$a=5, c=h=3; h'=5$	15	7	43	~3	500 (33%)	33
Dfly + *FT*	$a=7, c=h=3; k=16, n=2$	16	8	67	~4	550 (27%)	41

total VCSELs) and four PD arrays, we must eliminate 8 mm of package pads on the substrate edge.

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. The total area for two sets of 16 VCSELs and PDs is 60 mm^2.

We determined the total link power by adding power for the transmitter (42.5 mW per channel at 32 Gbps, which includes total laser driver, clock generation and distribution, and serializer) and the receiver (17 mW at 32 Gbps, which includes the three-stage transimpedance amplifier, clock and data recovery, and deserializer). The total achieved power is 60 mW at 32 Gbps or 1.86 pJ/bit.

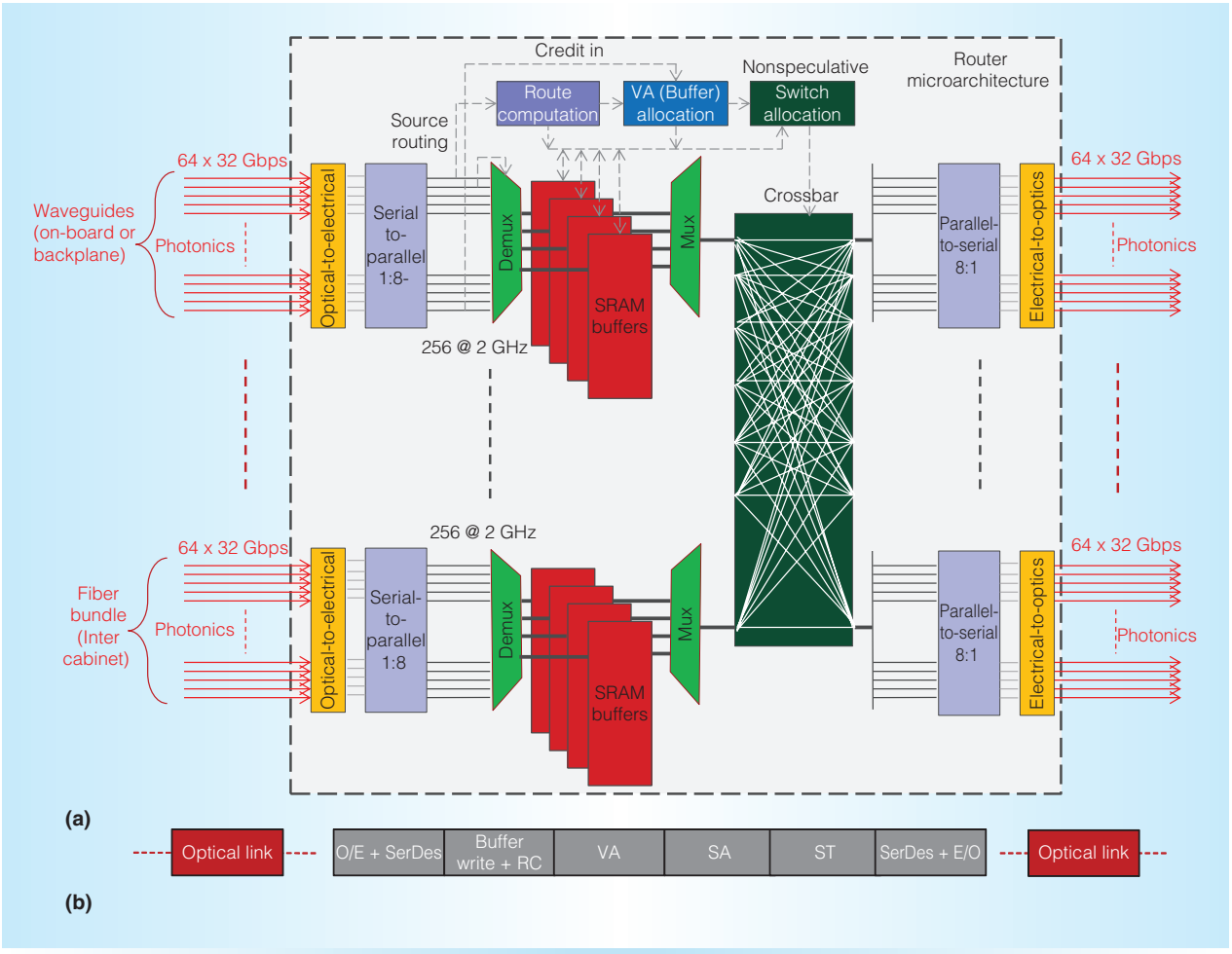

Figure 4. Router microarchitecture (a) and pipeline stages (b): optical-to-electrical conversion + serializer-to-deserializer (SerDes), buffer write + route computation, virtual channel allocation, switch allocation, switch traversal, and SerDes + electrical-to-optical conversion.

Figure 4a shows the router microarchitecture, and Figure 4b shows the router pipeline. Our proposed architecture has six router pipeline stages (optical-to-electrical, buffer write, virtual channel allocation, switch allocation, switch traversal, and electrical-to-optical). The proposed high-radix switch has either waveguides (from onboard and backplane) or fibers (from intercabinet) as I/O ports. Each port has 16 × 32 Gbps of photonic signals arriving and departing from the router. Each of the photonic links is converted to electrical signals via the optical-to-electrical blocks shown. This corresponds to the first router pipeline stage. This is followed by converting serial links into parallel using a 1:8 serializer-to-deserializer (SerDes) to generate 256 bits of data. The head flit is 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 static RAM (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 the buffer-write stage combined with route computation lookup. For the header flit, the information is then written to the VC state table maintained

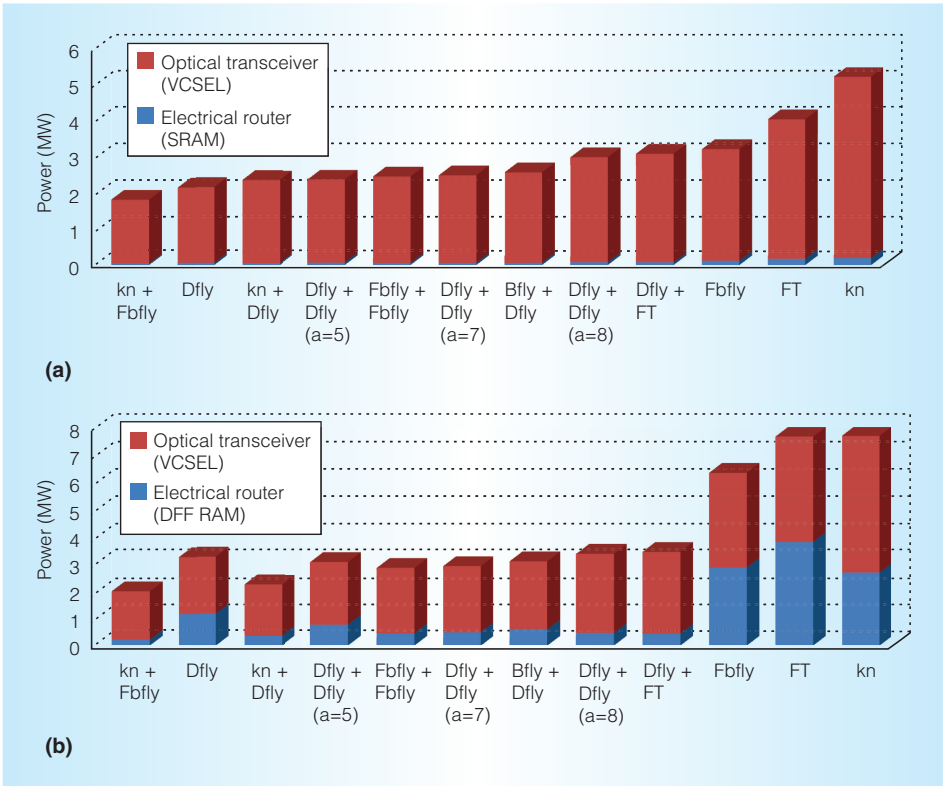

Figure 5. Network power estimation for 100,000-node exascale system with SRAM buffers (a) and DFF-RAM buffers (b).

at the router. This enables payload flits to follow the same route to the destination.

The next stage is the VC allocation 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 will also create credit information to be sent to the downstream router.

The next pipeline stage is switch allocation, when all the flits contend at the switch stage. We propose a nonspeculative switch allocation that follows the VC allocation. The next router pipeline stage is switch traversal, when the parallel data is converted back to serial data using a SerDes and then transmitted using the VCSELs. Finally, the electrical signals are converted to photonic links.

Performance evaluation: power, area, and throughput

In this evaluation, we estimate router and link power and area for different topologies. For CMOS router power, we used DSENT 0.9.[12] DSENT models D-flip-flop (DFF) for buffers and multiplexer-based crossbars. For a complete evaluation, we also considered two-stage switch allocation as well as clocking. For switch allocation, we used the first stage to arbitrate between VCs in the same input port, and the second stage to arbitrate between input ports. We modeled broadcast-based H-tree for clocking. We considered 32-nm bulk CMOS technology for modeling CMOS components. For on-chip communication, we use DFF for buffers. However, the round-trip latency for global channels in exascale topologies can be high because of long links; therefore, SRAM is a better technology choice. We modeled this only for the buffers using CACTI 6.0. Figure 5 shows the power breakdown for different topologies.

SRAM reduces the impact on overall power consumption compared to DFF because of 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 + 26 meters (from Titan). Except for dragonfly, all single-tier topologies

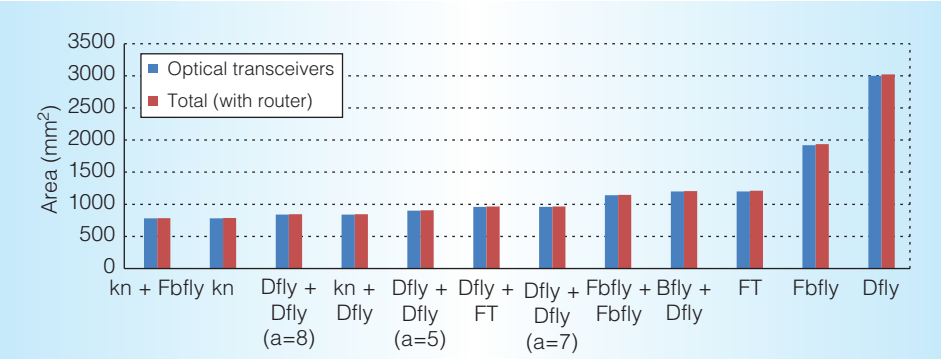

Figure 6. Area estimation for optical transceiver and total router area including buffers, crossbars, and allocators.

consume the most power for both SRAM and DFF-RAM. Single-tier dragonfly reduces the total power because there are fewer switches. The other scalable topology is *k*-ary *n*-cube, which also consumes less power because there are more local connections. Most of the combined topologies evaluated (*Dfly* + *Dfly*, *Bfly* + *Dfly*, *Dfly* + *FT*) consume less power than single-tier networks. We expect an exascale system to consume 20 MW of power, which includes the processor, memory, and network. Current petaflop machines consume 10 to 12 percent of the system power for the network, which implies that most of our multitier topologies (2 to 3 MW) are suitable candidates.

Figure 6 shows the overall area each router required for different topologies. Single-tier dragonfly consumes the most area simply because of the high radix, which is almost 3× more than other combined topologies. Most combined topologies have an area of 1000 mm^2, which when compared to the Aries router from Cray XC with an area of 16.6 mm × 18.9 mm (= 313.74 mm^2), is only 3× larger.

To gain insight into network performance, we modeled a small-scale network of 1,000 nodes. The proposed network evaluation merges single-level networks such as mesh, flattened butterfly, and dragonfly with multilevel 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). We restricted the analysis to direct topologies and therefore did not model the fat-tree network. For open-loop measurement, we varied the network load from 0.1 to 0.9 of the network capacity. The simulator was warmed up under load without taking measurements until a steady state was reached. A sample of injected packets was then labeled during a measurement interval. We allowed the simulation to run until all the labeled packets reached their destinations. We tested all designs with different synthetic traffic patterns such as uniform-random, bit-reversal, butterfly, matrix-transpose, complement, and perfect shuffle.

Figure 7 shows the latency plots for uniform-random and matrix-transpose traffic patterns. For uniform traffic, both dragonfly and flattened butterfly show the best performance (that is, best latency and bandwidth before saturation), which is almost 10 percent better compared to multilevel designs. We expected this because the hop count in multilevel topologies is greater than in dragonfly and flattened butterfly. 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 network's diameter. For matrix-transpose, dragonfly topology provides the best performance by saturating at 25 percent of network load. G_dragon-L_dragon performs the next best by saturating at 20 percent of network load. The multitier topologies have higher zero-load latency because of the larger network diameter.

Figure 8 shows the saturation throughput for various traffic patterns. The last column shows the geometric mean for various traffic patterns. The results follow the earlier

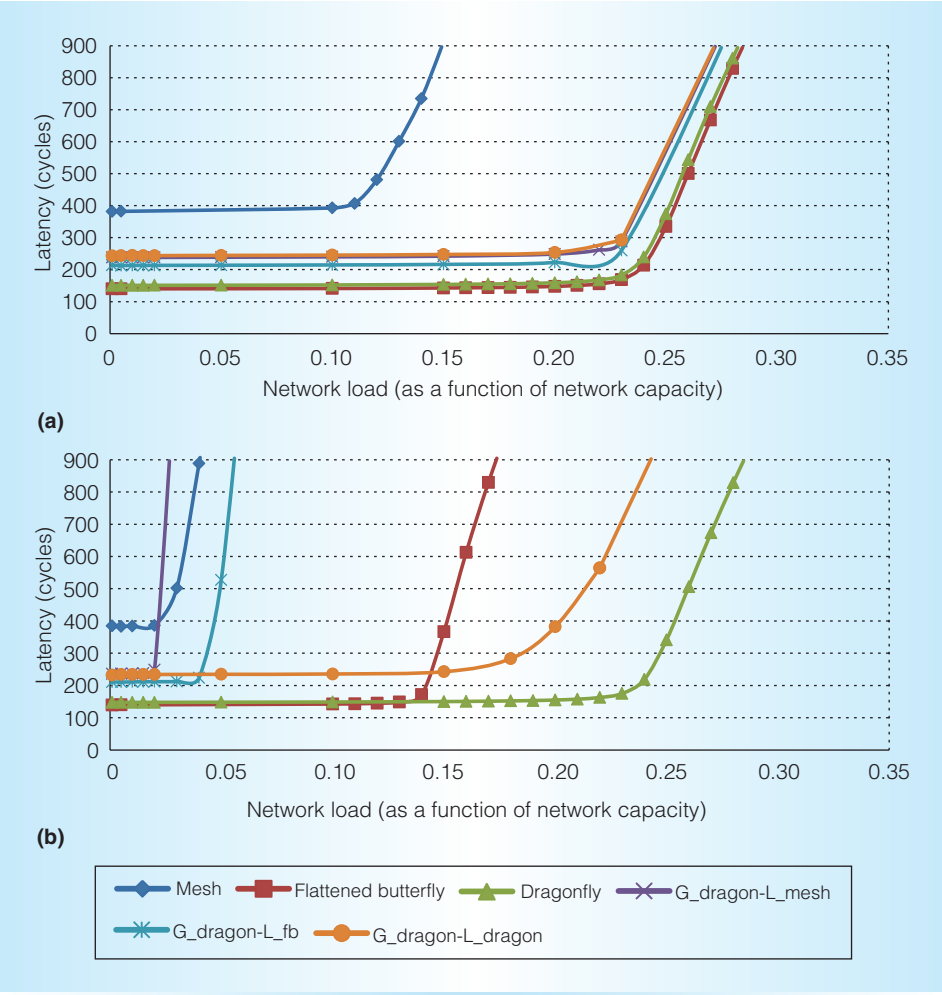

Figure 7. Average network latency evaluation for different topologies for 1,000 nodes under uniform-random (a) and matrix-transpose (b) traffic patterns operating at 2 GHz router clock (0.5 ns router clock cycle).

network saturation points; if a network topology saturates at a higher network load, then the throughput also is higher. For example, uniform-random, flattened butterfly, and dragonfly provide the highest throughput, which is marginally higher when compared to the multilevel topologies. The geometric mean of the averages provides some insights into how 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 we consider dragonfly and the remaining multitier topologies that implement dragonfly globally, multitier topologies perform 15 to 50 percent better than a single-tier dragonfly topology. One reason for the improvement is the increased path diversity in multitier topologies. Another reason is that the routing algorithm implemented in our models is mostly shortest-path, which could lead to increased congestion. Multitier topologies provide improved performance and do not have to resort to Valiant's algorithm for distributing the hot spot. Moreover, the same performance is delivered at a much-reduced radix, thereby decreasing the router complexity and cost.

Figure 9 shows the power dissipated for 1,000 nodes with SRAM buffers implemented

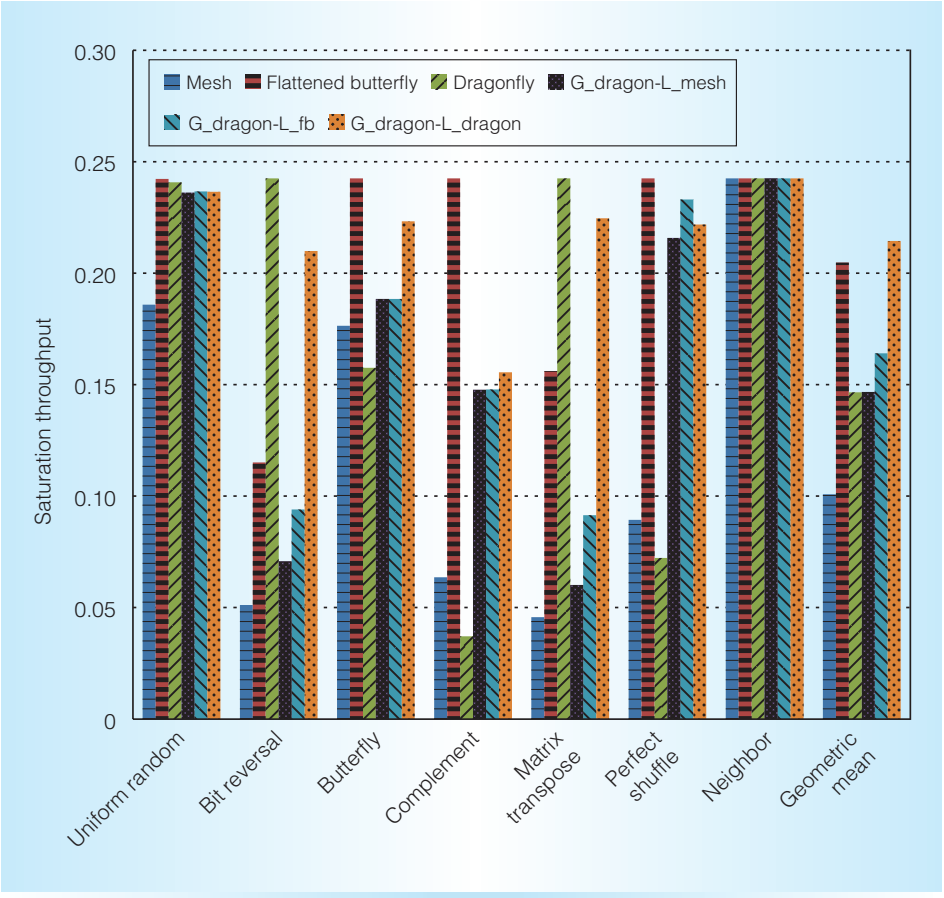

Figure 8. Saturation throughput for various traffic patterns for 1,000 nodes. The last column shows the geometric mean for various traffic patterns.

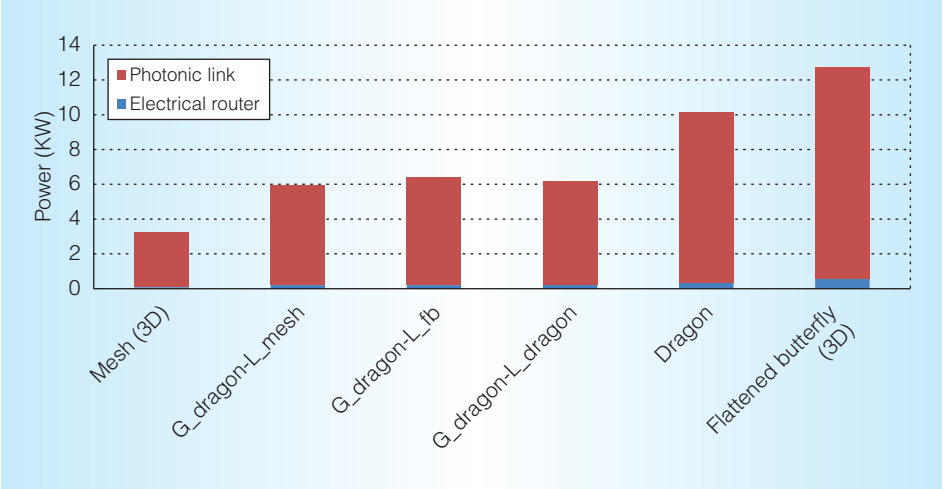

Figure 9. Power dissipated for 1,000 nodes for different topologies with SRAM buffers. The multitier topologies consume 40 percent less power than single-tier dragonfly and 60 percent less power than flattened butterfly.

for routers. The multitier topologies consume 40 percent less power than single-tier dragonfly and 60 percent less power than flattened butterfly. The higher radix of the router, as well as the optical links, contribute to increased router power use; however, as the network size increases, higher-radix topologies will have a power advantage because they can connect more routers directly than multitier topologies.

Overall, our work finds that multitier topologies for exascale and datacenter networks reduce router complexity and radix, both of which become critical when implemented with emerging optical interconnects. Although we have estimated the power of the VCSEL-based transceivers, more attention needs to be paid to the actual layout, which can have significant effects on power and area overhead. Electrical CMOS router and photonic transceivers do not scale in a similar fashion; this could result in underestimation of the transceiver power consumption. Given the likelihood that CMOS router and optical transceivers will be designed in future technology nodes, further evaluation is required to estimate or extrapolate the expected power consumption with technology scaling. Furthermore, newer technology such as silicon photonics should be carefully evaluated for designing optical transceivers. With silicon photonics, researchers should focus on three important issues: laser coupling efficiency; thermal stability of modulators and demodulators; and losses due to high-density waveguide crossings, bends, and crosstalk. While silicon photonics has the potential to reduce the area overhead and provide significant power efficiency, more research at the device and technology levels is necessary before the technology can be deployed for exascale networks. Finally, we need to evaluate latency and throughput when we scale the network to a large number of nodes (beyond 1,000 nodes). MICRO

Acknowledgments

We thank Petre Popescu for providing the analysis for VCSEL power estimation and Steve Reinhardt for valuable discussions involving the architecture design. This research was partially funded under US Department of Energy (DoE) contract no. DE-AC52-8MA27344 and subcontract B600716.

References

1. W. Dally and B. Towles, *Principles and Practices of Interconnection Networks*, Morgan Kauffman, 2004.
2. J. Kim, W. Dally, and D. Abts, "Flattened Butterfly: A Cost-Efficient Topology for High-Radix Networks," *Proc. 34th Ann. Int'l Symp. Computer Architecture* (ISCA), 2007, pp. 126-137.
3. J. Kim et al., "Technology-Driven, Highly-Scalable Dragonfly Topology," *Proc. 35th Ann. Int'l Symp. Computer Architecture* (ISCA), 2008, pp. 77-88.
4. F. Petrini et al., "The Quadrics High Performance Clustering Technology," *IEEE Micro*, vol. 22, no. 1, 2002, pp. 46-57.
5. B. Alverson et al., *Cray XC Series Network*, white paper, WP-Aries01-1112, Cray, 2012.
6. M.A. Taubenblatt, "Optical Interconnects for High-Performance Computing," *J. Lightwave Tech.*, vol. 30, no. 4, 2012, pp. 448-457.
7. A.V. Krishnamoorthy et al., "Computer Systems Based on Silicon Photonic Interconnects," *Proc. IEEE*, vol. 97, no. 7, 2009, pp. 1337-1361.
8. A. Dhodapkar et al., "SeaMicro SM 10000-64 Server: Building Datacenter Servers Using Cell Phone Chips," Hot Chips 23, 2011; www.hotchips.org/wp-content/uploads/hc_archives/hc23/HC23.19.7-Server/HC23.19.710-CellPhone-Lauterbach-SeaMicro.pdf.
9. N.R. Adiga et al., "Blue Gene/L Torus Interconnection Network," *IBM J. Research and Development*, vol. 49, nos. 2/3, 2005, pp. 265-276.
10. M. Al-Fares, A. Loukissas, and A. Vahdat, "A Scalable, Commodity Data Center Network Architecture," *Proc. ACM Sigcomm Conf. Data Communication*, 2008, pp. 63-74.
11. B. Arimilli et al., "The PERCS High-Performance Interconnect," *Proc. 18th Symp. High-Performance Interconnects*, 2010, pp. 75-82.
12. C. Sun et al., "DSENT—A Tool Connecting Emerging Photonics with Electronics for

Opto-Electronic Networks-on-Chip Modeling," *Proc. IEEE/ACM 6th Int'l Symp. Networks-on-Chip*, 2012, pp. 201-210.

Avinash Karanth Kodi is an associate professor of electrical engineering and computer science at Ohio University. His research interests include optical interconnects, networks on chips, datacenters, and chip multiprocessors. Kodi has a PhD in electrical and computer engineering from the University of Arizona. He is a senior member of IEEE.

Brian Neel is an electrical engineer at South Central Power Company. His research interests include optical interconnects, networks on chips, and datacenters. Neel has an MS in electrical engineering and computer science from Ohio University. He completed the work for this article while working as an intern at Advanced Micro Devices. He is a student member of IEEE.

William C. Brantley is a Fellow design engineer in the research division at Advanced Micro Devices, where he leads two research areas of AMD's Fast Forward project. His research interests include technologies for meeting the exascale challenge, such as high-bandwidth networks with low latency, power and thermal control, resilience, and techniques that efficiently exploit APUs from standard higher-level languages. Brantley has a PhD in electrical engineering from Carnegie Mellon University. He is a senior member of IEEE.

Direct questions and comments about this article to Avinash Karanth Kodi, 322D Stocker Center, Electrical Engineering and Computer Science, Ohio University, Athens, OH 45701; kodi@ohio.edu.