Chapter 9
Optically Interconnected High Performance Data Centers

Keren Bergman and Howard Wang

9.1 Introduction

Over the years, advances in optical technologies have enabled unprecedented data transmission capacities through the engineering and exploitation of a number of extremely advantageous physical properties inherent to photonic media. Wavelength division multiplexing (WDM), which is enabled by the enormous bandwidth of guided optics (nearly 32 THz in optical fiber [1]), represents a highly disruptive capability enabling information transmission across a single physical photonic channel at data rates many orders of magnitude greater than its copper-based counterpart, with demonstrated capacities exceeding 20 Tb/s [2] in single mode fiber. The characteristically low loss of optical media further enables extremely high bandwidth-distance and bandwidth-energy products, lifting previously unyielding architectural constraints dictated by the physical limitations of electronic interconnects [3]. Moreover, optical fiber media can sustain much smaller bending radii with significantly lower volume and weight, resulting in much more tenable and robust physical cabling.

As a result, photonic media has recently seen appreciable penetration into large-scale cluster computing systems, where unprecedented growth in application scale and hardware parallelism has combined to impose increasingly intractable communications requirements on its underlying interconnection network. Given the immense number of computation and storage elements, performance of these systems is reliant upon the effective exchange of vast amounts of data among end nodes (e.g., processors, memory, and storage). Therefore interconnects capable of

K. Bergman (✉) • H. Wang (✉)
Department of Electrical Engineering, Columbia University, New York, NY 10027, USA
e-mail: bergman@ee.columbia.edu; howard@ee.columbia.edu

C. Kachris et al. (eds.), *Optical Interconnects for Future Data Center Networks*,
Optical Networks, DOI 10.1007/978-1-4614-4630-9_9,

supporting high-bandwidth low-latency communication across the scale of these highly distributed machines have become a nearly ubiquitous requirement for large-scale systems [4].

Accordingly, system designers have begun to embrace optical interconnects in production large-scale systems in the form of point-to-point links [3, 5]. While point-to-point interconnects have received significant attention and acceptance commercially, they can only partially alleviate the burgeoning bandwidth and power constraints plaguing modern day systems. Conventional electronic switches are still required at the terminus of each optical link. As these switches scale in port count and capacity, they are reaching fundamental performance limits. Worse still, the power consumed by electronic switches is already prohibitively high and continues to grow super-linearly with port count and bandwidth.

Therefore, in order to effectively address the power, bandwidth, and latency requirements imposed by these systems, optically switched networks have been proposed as a possible solution. By providing end-to-end optical paths from the source to the destination, all-optical networks can forgo costly translations between the electronic and optical domains. Photonic switches operate on the principle of routing lightpaths. This is a fundamentally different operation than that of electronic switching, which must store and transmit each bit of information individually. By doing so, a conventional electronic switch dissipates energy with each bit transition, resulting in power consumption proportional to the bitrate of the information it carries. However, a lightpath through an optical switch ideally remains transparent to the information it carries, a critical characteristic known as bit rate transparency [6]. Consequently, unlike an electronic switch, the power consumed by an optical switch is independent of the bitrate of the information it is routing. Therefore, scalability to significantly higher transmission bandwidths (using techniques such as WDM) can be achieved, enabling extremely low power-per-unit bandwidths through all-optically switched interconnection networks.

While bit-rate transparency is an eminently advantageous property for the design of high-bandwidth, energy-efficient switches, there are other fundamental properties of optical technology that represent significant challenges toward the realization of all-optical switches. Depending on the design and technology employed in optical switches, signal impairment and distortion due to effects such as noise and optical nonlinearities must be carefully considered. More critically, inherent limitations of the optical medium give rise to two architectural challenges that must be addressed: namely, the lack of effective photonic memories and the extremely limited processing capabilities realizable in the optical domain. Electronic switches are heavily reliant upon random access memories (RAM) to buffer data while routing decisions are made and contention resolution is performed. Effective header parsing and processing in electronics is simply assumed. As no effective photonic equivalent of RAM or processing exists, critical functionalities such as contention resolution and header parsing will need to be addressed in a manner unique to optical switching, dictating photonic network designs that are similarly unique. In the following sections, we describe two network architectures explicitly designed to leverage the capacity and latency advantages of all-optical switching while utilizing unique system-level solutions to the photonic buffering and processing problems.

9.2 Data Vortex

The data vortex architecture [7], specifically designed to be implemented as an all-optical packet switched topology, is comprised of simple 2 × 2 all-optical switching nodes (Fig. 9.1). Each node utilizes two semiconductor optical amplifier (SOA) devices, which perform the switching operation. Given the wide gain-bandwidth of the SOAs, the network utilizes a multi-wavelength striped packet format (Fig. 9.2), with high bit-rate payload data segmented across multiple channels and low bit-rate addressing information encoded on dedicated wavelengths, one bit per wavelength per packet. Passive optical splitters and filters within the node extract the relevant routing information (a frame bit to denote the presence of a packet and a header bit to determine the switch's configuration), which are subsequently detected by low-speed receivers. The SOA pair is controlled via high-speed electronic decision circuitry, and routes the packet based on the recovered header information.

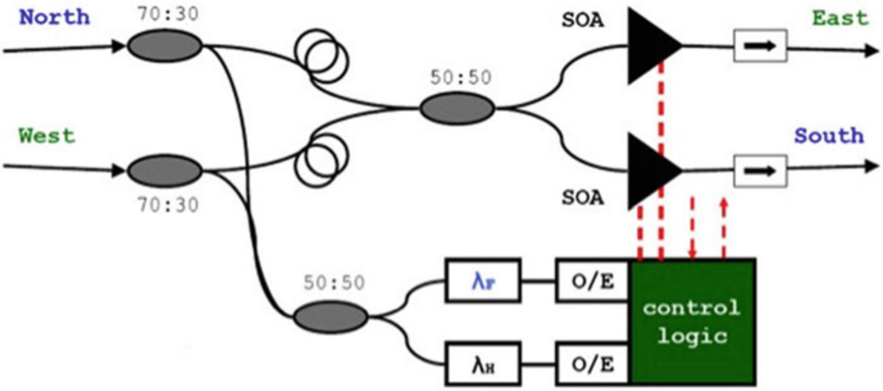

Fig. 9.1 (a) 2 × 2 data vortex switching node design

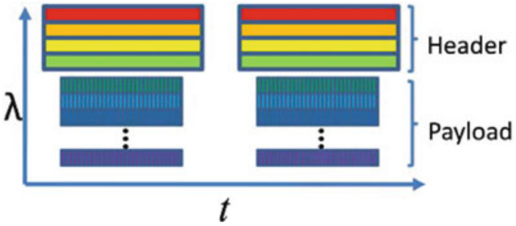

Fig. 9.2 Multi-wavelength striped packet format

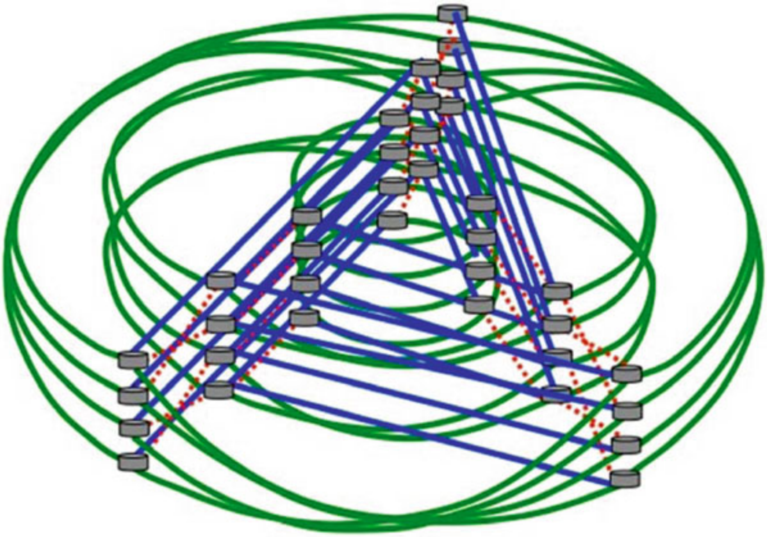

Fig. 9.3 Topology of a 12×12 data vortex all-optical packet switch consisting of 36 2×2 switching nodes. Green lines represent deflection fibers while blue lines represent ingression fibers

In a data vortex topology, the 2×2 switching nodes are organized as concentric cylinders and addressed according to their location within the topology, represented by their cylinder, height, and angle (C, H, A) (Fig. 9.3). The switching elements are arranged in a fully connected, directed graph with terminal symmetry but not complete vertex symmetry. The single-packet routing nodes are wholly distributed and require no centralized arbitration. The topology is divided into C hierarchies or cylinders, which are analogous to the stages in a conventional banyan network (e.g., butterfly). The architecture also incorporates deflection routing, which is implemented at every node; deflection signal paths are placed only between different cylinders. Each cylinder (or stage) contains A nodes around its circumference and $H = 2^{C-1}$ nodes down its length. The topology contains a total of $A \times C \times H$ switching elements, or nodes, with $A \times H$ possible input terminal nodes and an equivalent number of possible output terminal nodes. The position of each node is conventionally given by the triplet (c, h, a), where $0 \leq c \leq C$–1, $0 \leq h \leq H$–1, $0 \leq a \leq A$–1.

The switching nodes are interconnected using a set of ingression fibers, which connect nodes of the same height in adjacent cylinders, and deflection fibers, which connect nodes of different heights within the same cylinder. The ingression fibers are of the same length throughout the entire system, as are the deflection fibers. The deflection fibers' height crossing patterns direct packets through different height levels at each hop to enable banyan routing (e.g., butterfly, omega) to a desired height, and assist in balancing the load throughout the system, mitigating local congestion [8–11].

Incoming packets are injected into the nodes of the outermost cylinder and propagate within the system in a synchronous, time-slotted fashion. The conventional nomenclature illustrates packets routing to progressively higher numbered cylinders as moving inward toward the network outputs. During each timeslot, each node either processes a single packet or remains inactive. As a packet enters node (c, h, a), the cth bit of the packet header is compared to cth most significant bit in the node's height coordinate (h). If the bits match, the packet ingresses to node $(c+1, h, a+1)$ through the node's south output. Otherwise, it is routed eastward within the same cylinder to node $(c, G_c(h), a+1)$, where $G_c(h)$ defines a transformation which expresses the abovementioned height crossing patterns (for cylinder c) [10, 11]. Thus, packets progress to a higher cylinder only when the cth address bit matches, preserving the c–1 most significant bits. In this distributed scheme, a packet is routed to its destination height by decoding its address in a bitwise banyan manner. Moreover, all paths between nodes progress one angle dimension forward and either continue around the same cylinder while moving to a different height, or ingress to the next hierarchal cylinder at the same height. Deflection signals only connect nodes on adjacent cylinders with the same angular dimension; i.e. from $(c+1,\ h, a)$ to a node at position $(c, G_{c+1}(h), a)$.

The paths within a cylinder differ depending upon the level c of the cylinder. The crossing or sorting pattern (i.e., the connections between height values defined by $G_c(h)$) of the outermost cylinder $(c = 0)$ must guarantee that all paths cross from the upper half of the cylinder to the lower half of the cylinder; thus, the graph of the topology remains fully connected and the bitwise addressing scheme functions properly. Inner cylinders must also be divided into $2c$ fully connected and distinct subgraphs, depending upon the cylinder. Only the final level or cylinder ($c = C$–1) may contain connections between nodes of the same height. The cylindrical crossing must ensure that destinations can be addressed in a binary tree-like configuration, similar to other binary banyan networks.

Addressing within the data vortex architecture is entirely distributed and bitwise, similar to other banyan architectures: as a packet progresses inward, each successive bit of the binary address is matched to the destination. Each cylinder tests only one bit (except for the innermost one); half of the height values permit ingression for 1 values, and half for 0 values, arranged in a banyan binary tree configuration. Within a given cylinder c, nodes at all angles at a particular height (i.e., $(c, h,)$) match the same $c+1$st significant bit value, while the paths guarantee preservation of the c most significant address bits. Thus, with each ingression to a successive cylinder, progressively more precision is guaranteed in the destination address. Finally, on the last cylinder $c = C$–1, each node in the angular dimension is assigned a least significant value in the destination address so that the packets circulate within that cylinder until a match is found for the last $\sim \log_2(\mathrm{A})$ bits (so-called angle-resolution addressing) [8].

The data vortex all-optical packet switch is the result of a unique effort towards developing a high-performance network architecture designed specifically for photonic media. The overall goals were to produce a practical architecture that leveraged wavelength division multiplexing (WDM) to attain ultra-high bandwidths and reduce routing complexity, while maintaining minimal time-of-flight latencies by keeping packets in the optical domain and avoiding conventional buffering [7]. In keeping with these objectives, a functional prototype of a 12-port data vortex was implemented and demonstrated [8]. Physical layer scalability was analyzed and demonstrated in [12, 13], and further experimental studies of the optical dynamic range and packet format flexibility were performed [14, 15]. Sources of signal degradation in the data vortex were investigated in [16, 17] and data resynchronization and recovery was achieved using a source synchronous embedded clock in [18]. Extensible and transparent packet injection modules and optical packet buffers for the data vortex were presented in [19]. Alternative architectural implementations and performance optimizations were explored in [20–23].

9.3 SPINet

Based on an indirect multistage interconnection network (MIN) topology, SPINet (Scalable Photonic Integrated Network) [24], designed to be implemented using photonic integration technology, exploits WDM to simplify the network design and provide very high bandwidths. SPINet does not employ buffering, instead resolves contention by dropping contending messages. A novel physical-layer acknowledgment protocol provides immediate feedback, notifying the terminals whether their messages are accepted, and facilitates retransmissions when necessary in a manner resembling that in traditional multiple-access media.

A SPINet network is composed of 2×2 SOA-based bufferless photonic switching nodes [25, 26]. The specific topology can vary between implementations, and switching nodes of higher radices can be used if they are technologically available. The network is slotted and synchronous, and messages have a fixed duration. The minimal slot duration is determined by the round-trip propagation time of the optical signal from the compute nodes to the ports of the network. A slot time of 100 ns can, therefore, accommodate a propagation path of nearly 20 m.

A possible topology for SPINet is the Omega, an example of a binary banyan topology [4]. An $N_T \times N_T$ Omega network consists of $N_S = \log(N_T)$ identical stages. Each stage consists of a perfect-shuffle interconnection followed by $N_T/2$ switching elements, as Fig. 9.4a shows. In the Omega network, each switching node has four allowed states (straight, interchange, upper broadcast, and lower broadcast). In this implementation, we have modified the switching nodes by removing the broadcast states and introducing four new states (upper straight, upper interchange, lower straight, lower interchange). In these four states, the node passes data from only one input port to an output port and drops the data from the other port (see Fig. 9.4b).

At the beginning of each slot, any terminal may start transmitting a message, without a prior request or grant. The messages propagate in the fibers to the input ports of the network and are transparently forwarded to the switching nodes of the first stage. At every routing stage when the leading edges of the messages are

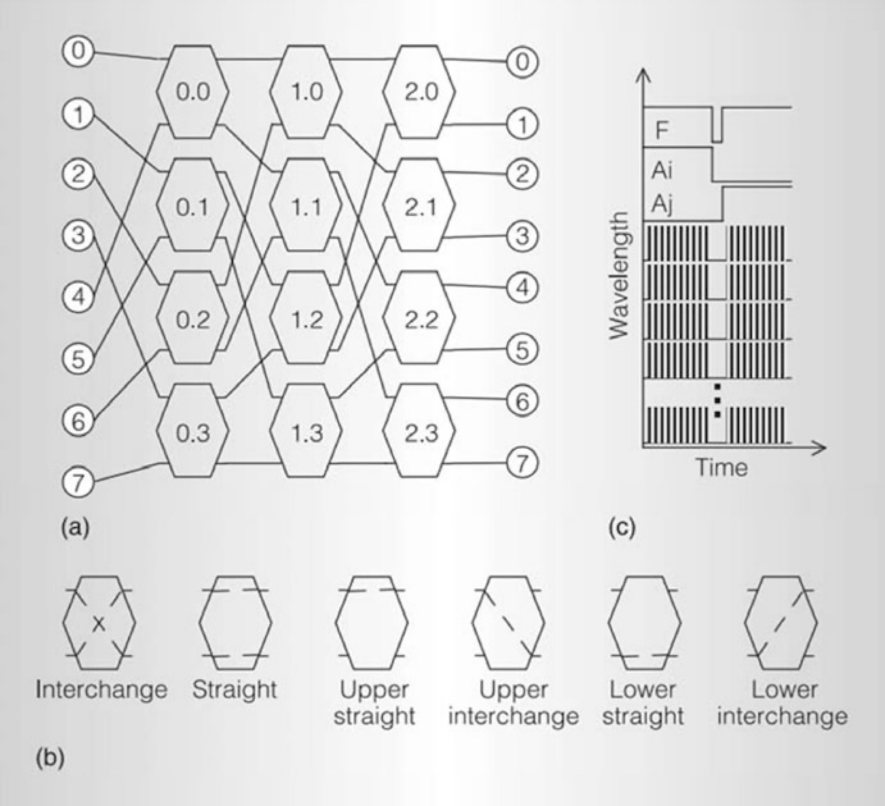

Fig. 9.4 An 8 × 8 Omega network (a); switching nodes' six states (b); and wavelength-parallel messages (c). Header bits and payload are encoded on dedicated wavelengths [25]

received from one or both input ports, a routing decision is made, and the messages continue to propagate to their requested output ports. In the case of output-port contention in a switching node, the network drops one of the contending messages. The choice of which message to drop can be random, alternating, or priority-based. Because the propagation delay through every stage is identical, all the leading edges of the transmitted messages reach all the nodes of each stage at the same time.

The nodes' switching states, as determined by leading edges, remain constant throughout the duration of the message, so the entire message follows the path acquired by the leading edge. Because the propagation delay through the network, which is ideally implemented via integrated photonics, is very short compared to the duration of the messages, the messages stretch across the PIC, effectively creating transparent lightpaths between the inputs and outputs. When the messages reach the output modules, they are transparently forwarded on the output fibers to the appropriate terminals; at the same time, the destination terminal generates an acknowledgment optical pulse and sends it on the previously acquired lightpath in the opposite direction. Because the node's switching elements preserve their states and support bidirectional transmission, the source terminal receives the acknowledgment pulse, which serves as confirmation of the message's successful reception.

When the slot time is over, all terminals cease transmission simultaneously, the switching nodes reset their switching states, and the system is ready for a new slot. The slot duration is set to ensure that the acknowledgment pulses are received at the source terminals before the slot ends. Hence, before the beginning of the next slot, every terminal knows whether its message was accepted; when necessary, it can choose to immediately retransmit the message.

Leveraging ultralow-latency signal propagation through the network, SPINet eliminates the need for central scheduling, instead employing the distributed computing power of the switching nodes to produce an input–output match at every slot. This process of implicit arbitration enables scalability to large port counts without burdening a central arbiter with computations of complex maximal matches. Because SPINet uses blocking topologies to reduce hardware complexity, the network's utilization is lower than that of a traditional maximum-matched nonblocking network (as in switching fabrics for high-performance Internet routers). Techniques that exploit the properties of integrated photonics can increase utilization by adding a small number of stages.

SPINet uses the wavelength domain to facilitate a routing mechanism in the switching nodes that can instantly determine and execute the routing decision upon receiving the leading edges, without any additional information exchange between the switching nodes. The mechanism also maintains a constant switching state for the duration of the messages. The messages are constructed in a wavelength-parallel manner, similar to that used in the data vortex architecture, trading off a part of the enormous bandwidth of optical fibers to simplify the switching-node design. As Fig. 9.4c shows, the routing header and the message payload are encoded on separate wavelengths and are received concurrently by the nodes. The header consists of a frame bit that denotes the message's existence and a few address bits. Each header bit is encoded on a dedicated wavelength and remains constant throughout the message duration. When a binary network is used, a single address bit is required at every stage, and therefore the number of wavelengths required for address encoding is the number of routing stages in the network, or $\log_2$ of the number of ports. The switching nodes' routing decisions are based solely on the information extracted from the optical header, as encoded by the source. The switching nodes neither exchange additional information nor add any to the packet. The payload is encoded on multiple wavelengths at the input terminal, which segments it and modulates each segment on a different wavelength, using the rest of the switching band. A guard time is allocated before payload transmission, accommodating the SOA switching time, clock recovery in the payload receivers, and synchronization inaccuracies.

9.4 Networking Challenges in the Data Center

The aforementioned networks, in addition to other implementations of all-optical networks, are capable of the ultra-high bandwidths and low power densities necessary for enabling the continued scaling of large cluster computing systems. However, these systems represent a wide range of computing classes, ranging from highly specialized designs to commodity and cost-driven computing environments. For example, the increasing popularity of cloud-based services continues to drive the creation of larger and more powerful data centers. As these services scale in both number and size, applications oftentimes extend well beyond the boundaries of a single rack of servers. Moreover, the continued advancements in computational density enabled by increasing parallelism in contemporary microprocessors and chip multiprocessors (CMP) have resulted in substantial off-chip communication requirements. As a result, in a similar fashion to their supercomputing counterparts, the performance of modern data centers are becoming increasingly communication-bound, requiring upwards of hundreds of thousands of ports supporting petabits per second of aggregate bandwidth [27].

However, due to the superlinear costs associated with scaling the bandwidth and port density of conventional electronic switches, network oversubscription is common practice. Consequently, data-intensive computations become severely bottlenecked when information exchange between servers residing in separate racks is required. Unlike high-performance computing systems, the very nature of the data center as a pool of centralized computational resources gives rise to significant application heterogeneity. The resultant workload unpredictability produces significant traffic volatility, precluding the efficacy of static capacity engineering in these oversubscribed networks.

Energy efficiency has also emerged as a key figure-of-merit in data center design [28]. The power density of current electronic interconnects is already prohibitively high—on the order of hundreds of kilowatts—and continues to grow exponentially. As it stands, the power consumption of a single switch located in the higher network tiers can reach upwards of tens of kilowatts when also considering the dedicated cooling systems required. Moreover, measurements on current data center deployments have recorded average server utilization as low as 30% [29], indicating significant wasted energy due to idling hardware starved for data.

As a result, alleviating inter-rack communication bottlenecks has become a critical target in architecting next-generation data centers. The realization of a full bisection-bandwidth, "all-servers-equidistant" interconnection network will not only accelerate the execution of large-scale distributed applications, but also significantly reduce underutilization by providing sufficient network performance to ensure minimal idling of power-hungry compute elements. In addition, the increased connectivity between computing and storage resources located throughout the data center will yield more flexibility in virtualization, leading to further enhancements in energy efficiency.

Despite continued efforts from merchant silicon providers towards the development of application-specific integrated circuits (ASICs) for high-performance switches and routers, the sheer scale of the data center and the relentless demand from data-intensive applications for increased connectivity and bandwidth continues to necessitate oversubscription in hierarchical purely packet-switched electronic interconnection networks. While there have been significant efforts focused on architectural and algorithmic approaches towards improving the overall performance of data center networks [30, 31], these proposals are ultimately constrained by the fundamental limitations imposed by the underlying electronic technologies.

Recently, in the context of production data center environments, there have been a number of efforts exploring the viability of circuit-switched optics as a cost-effective means of providing significant inter-rack bandwidth. Helios [32] and C-Through [33] represent two data center network architectures proposing the use of micro-electro-mechanical system (MEMS)-based optical switches. By augmenting existing oversubscribed hierarchical electronic packet-switched (EPS) networks, each implementation realizes a hybrid electronic/optical architecture that leverages the respective advantages of each technology. These initial proposals have successfully demonstrated the potential for utilizing photonic technologies within the context of data center traffic to provide significantly increased network capacities while achieving reduced complexity, component cost, and power in comparison with conventional electronic network implementations. Another network architecture, called Proteus, combines both wavelength-selective switching and space switching to provide further granularity in the capacity of each optical link, varying between a few gigabits per second to a hundreds of gigabits per second on-demand [34].

While network traffic is characteristically unpredictable due to application heterogeneity, communication patterns where only a few top-of-the-rack (ToR) switches are tightly coupled with long, extended data flows have been observed in production data centers [35]. Therefore, the utility of the aforementioned architectures are reliant on the inherent stability of traffic patterns within such systems. Nevertheless, further bandwidth flexibility remains a key target for future data center networks as applications require even higher capacities with increased interconnectivity demands. When studied under the communication patterns imposed by a richer, more representative set of realistic applications, the efficacy of architectures utilizing purely commercial MEMS-based switches, which are limited to switching times on the order of milliseconds, becomes ambiguous [36].

9.5 Conclusions

Traditional supercomputers usually employ specialized top-of-the-line components and protocols developed specifically to support highly orchestrated distributed workloads that support massively parallel, long-running algorithms developed to solve complex scientific problems. As such, these applications impose stringent latency requirements on processor-to-processor and processor-to-memory transactions, which represent the major bottleneck in these highly specialized systems.

On the other hand, data centers, which are primarily deployed by enterprises and academic institutions, predominantly run general-purpose user-facing cloud-based applications and are largely composed of commodity components. The majority of the messages being passed across a data center consist of very short random transactions. However, there typically exist a small number of long extended flows, which account for a majority of the data being transmitted through the network. Furthermore, traffic at the edges of the network is often bursty and unpredictable, leading to localized traffic hotspots across the system that leads to network congestion and underutilization. While it is apparent that bandwidth restrictions result in significant performance degradation in data centers, the latency requirements of these systems are relatively relaxed in comparison with that of supercomputers.

Consequently, the application demands and traffic patterns characteristic of these systems vary widely, resulting in highly variegated network requirements. Therefore, in addition to the improved capacity, power consumption and bandwidth-distance product delivered by a photonic interconnect medium, capacity flexibility is a key requirement for enabling future optically interconnected high-performance data centers and supercomputers.

References

1. Agrawal GP(2002) Fiber-optic communication systems. Wiley, New York
2. Gnauck AH, Charlet G, Tran P, Winzer PJ, Doerr CR, Centanni JC, Burrows EC, Kawanishi T, Sakamoto T, Higuma K (2008) 25.6-Tb/s WDM transmission of polarization-multiplexed RZ-DQPSK signals. IEEE J Lightwave Technol 26:79–84
3. Benner AF, Ignatowski M, Kash JA, Kuchta DM, Ritter MB (2005) Exploitation of optical interconnects in future server architectures. IBM J Res Dev 49(4/5):755–775
4. Dally WJ, Towles B (2004) Principles and practices of interconnection networks. Morgan Kaufmann, San Francisco
5. Kash JA, Benner A, Doany FE, Kuchta D, Lee BG, Pepeljugoski P, Schares L, Schow C, Taubenblatt M (2011) Optical interconnects in future servers. In: Optical fiber communication conference, Paper OWQ1. http://www.opticsinfobase.org/abstract.cfm?URI=OFC-2011-OWQ1
6. Ramaswami R, Sivarajan KN (2002) Optical networks: a practical perspective, 2nd edn. Morgan Kaufmann, San Francisco
7. Liboiron-Ladouceur O, Shacham A, Small BA, Lee BG, Wang H, Lai CP, Biberman A, Bergman K (2008) The data vortex optical packet switched interconnection network. J Lightwave Technol 26 (13):1777–1789
8. Shacham A, Small BA, Liboiron-Ladouceur O, Bergman K (2005) A fully implemented 12 × 12 data vortex optical packet switching interconnection network. J Lightwave Technol 23(10):3066–3075
9. Yang Q, Bergman K, Hughes GD, Johnson FG (2001) WDM packet routing for high-capacity data networks. J Lightwave Technol 19(10):1420–1426
10. Yang Q, Bergman K (2002) Traffic control and WDM routing in the data vortex packet switch. IEEE Photon Technol Lett 14(2):236–238
11. Yang Q, Bergman K (2002) Performance of the data vortex switch architecture under nonuniform and bursty traffic. J Lightwave Technol 20(8):1242–1247

12. Liboiron-Ladouceur O, Small BA, Bergman K (2006) Physical layer scalability of a WDM optical packet interconnection network. J Lightwave Technol 24(1):262–270
13. Liboiron-Ladouceur O, Bergman K, Boroditsky M, Brodsky M (2006) Polarization-dependent gain in SOA-Based optical multistage interconnection networks. IEEE J Lightwave Technol 24(11):3959–3967
14. Small BA, Lee BG, Bergman K (2006) Flexibility of optical packet format in a complete 12 × 12 data vortex network. IEEE Photon Technol Lett 18(16):1693–1695
15. Small BA, Kato T, Bergman K (2005) Dynamic power consideration in a complete 12 × 12 optical packet switching fabric. IEEE Photon Technol Lett 17(11):2472–2474
16. Small BA, Bergman K (2005) Slot timing consideration in optical packet switching networks. IEEE Photon Technol Lett 17(11):2478–2480
17. Lee BG, Small BA, Bergman K (2006) Signal degradation through a 12 × 12 optical packet switching network. In: European conference on optical comm., We3.P.131, pp 1–2, 24-28. doi: 10.1109/ECOC.2006.4801324 http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4801324&isnumber=4800856
18. Liboiron-Ladouceur O, Gray C, Keezer DC, Bergman K (2006) Bit-parallel message exchange and data recovery in optical packet switched interconnection networks. IEEE Photon Technol Lett 18(6):770–781
19. Shacham A, Small BA, Bergman K (2005) A wideband photonic packet injection control module for optical packet switching routers. IEEE Photon Technol Lett 17(12):2778–2780
20. Shacham A, Bergman K (2007) Optimizing the performance of a data vortex interconnection network. J Opt Networking 6(4):369–374
21. Liboiron-Ladouceur O, Bergman K (2006) Hybrid integration of a semiconductor optical amplifier for high throughput optical packet switched interconnection networks. Proc SPIE 6343–121, doi: 10.1117/12.708009
22. Liboiron-Ladouceur O, Bergman K (2006) Bistable switching node for optical packet switched networks. In: Proceedings 19th Annual Meeting of the IEEE Lasers and Electro-Optics Society (LEOS), 2006. Paper WW5, pp 631–632. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4054342&isnumber=4054019
23. Yang Q (2005) Improved performance using shortcut path routing within data vortex switch network. Electron Lett 41(22):1253–1254
24. Shacham A, Bergman K (2007) Building ultralow latency interconnection networks using photonic integration. IEEE Micro 27(4):6–20
25. Shacham A, Lee BG, Bergman K (2005) A scalable, self-routed, terabit capacity, photonic interconnection network. In: Proceedings of 13th Ann. IEEE Symp. High-Performance Interconnects (HOTI 05). IEEE CS Press, pp 147–150. doi: 10.1109/CONECT.2005.6 http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1544590&isnumber=32970
26. Shacham A, Lee BG, Bergman K (2005) A wideband, non-blocking, 2x2 switching node for a SPINet network. IEEE Photonic Technol Lett 17(12):2742–2744
27. Vahdat A, Al-Fares M, Farrington N, Mysore RN, Porter G, Radhakrishnan S (2010) Scale-out networking in the data center. IEEE Micro 30(4):29–41
28. Abts D, Marty MR, Wells PM, Klausler P, Liu H (2010) Energy proportional datacenter networks. In: Proceedings of 37th annual international symposium on computer architecture (ISCA'10), pp 338–347 ACM, New York, NY, USA http://doi.acm.org/10.1145/1815961.1816004
29. Meisner D, Gold BT, Wenisch TF (2009) PowerNap: eliminating server idle power. In: Proceedings of the 14th international conference on architectural support for programming languages and operating systems (ASPLOS'09), New York, NY, USA pp 205–216. http://doi.acm.org/10.1145/1508244.1508269
30. Al-Fares M et al (2008) A scalable, commodity data center network architecture. SIGCOMM Comp Comm Rev 38(4):63–74
31. Greenberg A et al (2009) Vl2: a scalable and flexible data center network. SIGCOMM Comp Comm Rev 39(4):51–62

32. Farrington N, Porter G, Radhakrishnan S, Bazzaz HH, Subramanya V, Fainman Y, Papen G, Vahdat A (2010) Helios: a hybrid electrical/optical switch architecture for modular data centers. In: SIGCOMM ‘10 proceedings of the ACM SIGCOMM 2010 conference on SIGCOMM. ACM, New York, pp 339–350
33. Wang G, Andersen DG, Kaminsky M, Papagiannaki K, Ng TE, Kozuch M, Ryan M (2010) c-Through: part-time optics in data centers. In: SIGCOMM ‘10 proceedings of the ACM SIGCOMM 2010 conference on SIGCOMM. ACM, New York, pp 327–338
34. Singla A, Singh A, Ramachandran K, Xu L, Zhang Y (2010) Proteus: a topology malleable data center networks. In: Hotnets ‘10 proceedings of the ninth ACM SIGCOMM workshop on hot topics in networks. ACM, New York, article 8
35. Benson T, Anand A, Akella A, Zhang M (2009) Understanding data center traffic characteristics. In: Proceedings of the 1st ACM workshop on research on enterprise networking, Barcelona, Spain, 21 August 2009. WREN ‘09. ACM, New York, pp 65–72
36. Bazzaz HH, Tewari M, Wang G, Porter G, Ng TSE, Andersen TG, Kaminsky M, Kozuch MA, Vahdat A (2011) Switching the optical divide: fundamental challenges for hybrid electrical/optical datacenter networks. In: Proceedings of SOCC’11: ACM symposium on cloud computing, Cascais, Portugal, Oct 2011