Optical Interconnection Networks in Data Centers: Recent Trends and Future Challenges

Christoforos Kachris, Konstantinos Kanonakis, and Ioannis Tomkos, Athens Information Technology

ABSTRACT

Warehouse-scale data center operators need much-higher-bandwidth intra-data center networks (DCNs) to sustain the increase of network traffic due to cloud computing and other emerging web applications. Current DCNs based on commodity switches require excessive amounts of power to face this traffic increase. Optical intra-DCN interconnection networks have recently emerged as a promising solution that can provide higher throughput while consuming less power. This article provides an update on recent developments in the field of ultra-high-capacity optical interconnects for intra-DCN communication. Several recently proposed architectures and technologies are examined and compared, while future trends and research challenges are outlined.

INTRODUCTION

The rise of cloud computing and other emerging web applications has created the need for more powerful warehouse-scale data centers. These data centers comprise hundreds of thousands of servers that need to communicate with each other via high performance and low latency interconnection networks. Such intra-data center networks (DCNs) are currently based on commodity Ethernet switches connected in a fat-tree topology as shown in Fig. 1. The servers of each rack are interconnected using a top-of-rack (ToR) switch. ToR switches are connected through aggregate switches, and aggregate switches are in turn interconnected by means of high-performance core switches, usually via link aggregation of multiple 10GE interfaces. However, these networks suffer from limited throughput, high latencies, and high power consumption, and will not be able to sustain the projected future DCN traffic demands without consuming excessive amounts of power. Optical interconnection networks have been proposed as a promising solution to address those deficiencies by performing (part of the) switching at the optical domain, thus eliminating the need for power-intensive electrical switches. However, the design of high-performance optical interconnects meeting data center requirements constitutes an extremely challenging inter-disciplinary research area, requiring expertise from several fields such as photonics, computer networks, and computer architectures.

In this article we follow the rise of optical interconnection networks for data centers in order to meet network traffic requirements and reduce the power consumption of data centers. We present the main network architectures that have been proposed so far, and discuss the main benefits and drawbacks of each approach. Finally we identify the current trends and future challenges in the domain of optical interconnection networks.

DATA CENTER TRAFFIC REQUIREMENTS

Figure 2 depicts the projected increase of DCN traffic until 2016 according to Cisco [1]. As shown in this figure, the majority of network traffic is within data centers. The main reason is that most applications hosted in data centers are based on parallel programming frameworks such as MapReduce. In such environments, high interaction between the distributed processing and storage nodes is required for handling large sets of data, which translates to significant communication requirements between servers [2].

Another significant reason for the projected increase in intra-data center network traffic is the advent of new high-performance server processors that host new interfaces supporting 10 Gb/s data rates. Furthermore, as more and more processing cores are integrated into a single processor, higher-bandwidth interfaces will be required for the communication of these cores with other cores residing on separate racks. In conjunction with the ramp of new servers hosting several cores, 10 GbE port shipments are expected to become the majority of server ports by 2014, and continue to increase as a portion of total ports through 2016. To meet this demand, data center operators are quickly adopting 100 Gb/s Ethernet, which is estimated to grow at a

 0163-6804/13/$25.00

rate of 170 percent over the next five years, according to Infonetics Research Inc.

Power Consumption Requirements

A major concern in the design of data centers is power consumption of the infrastructure mainly due to the associated operational expenditure (OPEX) costs. According to relevant studies, data center networks (including ToR, aggregate, and core switches) consume around 10–20 percent of the total IT power consumption of data center sites, and this is expected to increase in the near future [3].

Table 1 illustrates the projected increase of the total node bidirectional interconnection bandwidth according to IBM [4]. As shown in this table, while traffic is expected to increase from 1 to 400 Pbytes/s, the projected allowable power consumption of the systems will only increase from 5 to 20 MW. Given the aforementioned estimation regarding the percentage of DCN consumption, the affordable DCN power consumption in 2020 could be as low as around 2 MW for 400 Pbytes/s. In other words, this means that DCN total energy efficiency will have to drop from a few milliwatts per gigabit per second to less than a milliwatt per gigabit per second [5]. It is thus clear that in order to sustain much higher network traffic but with a limited amount of power consumption, much more energy-efficient interconnection networks are required than the current ones based on commodity Ethernet switches.

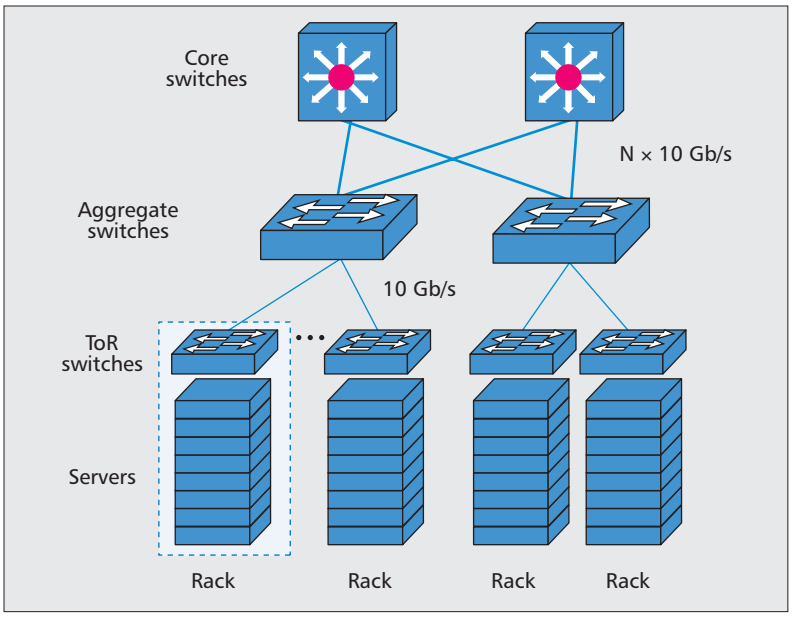

Figure 1. *Typical intra-data center network architecture.*

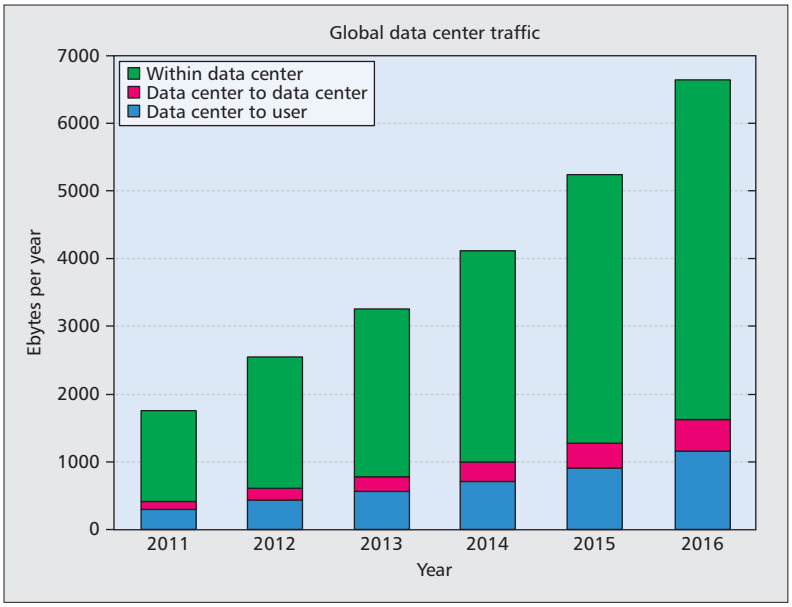

Figure 2. *Projected global data center traffic growth. Source: [1].*

Until now optical technology in DCNs has been utilized mainly in the form of point-to-point communication between the switches using optical fibers and optical transceivers. Low-cost multimode fibers operating at 850 nm are used, while optical transceivers are based on either SFP (1 Gb/s) or SPF+ (10 Gb/s) technology. For any communication, optical transmitters have to convert packets to optical signals, and optical receivers are used to transform the signal back to the electrical domain, since packet switching is performed using electronic switch fabrics. The main obvious drawback of this approach is that a lot of power is wasted in the electrical to optical (E/O) and optical to electrical (O/E) conversions at the transceiver modules as well as on the switch fabrics for the packet switching, while conversions also contribute to increased latency. Moreover, latency is further aggravated due to the electrical buffering required at the switch for addressing packet contention.

All-optical interconnects are therefore a promising solution in order to significantly reduce power consumption in the DCN as well as communication latency. At the telecommunications networks side, opaque networks based on point-to-point optical fibers have recently been replaced by all-optical (transparent) networks wherein switching is performed purely in the optical domain, bypassing power-hungry electrical switches. As depicted in Fig. 3, a similar paradigm shift should expected to take place in the case of future intra-DCNs.

Circuit vs. Packet-Switching

In the last years several research efforts have focused on the design of an energy-efficient optical interconnection network that can provide high throughput and reduced latency. A major distinction when it comes to optical interconnects is whether are based on circuit or packet switching. Circuit-based schemes mainly target DCNs in which long-term bulky data transfers are required between racks. For this reason, they are usually based on micro electro-mechanical switches (MEMS), which have increased reconfiguration time (in the orders of few milliseconds). It should be noted, though, that in order to guarantee all-to-all communication, significant overprovisioning might be required.

On the other hand, packet-based optical switching (Fig. 4b) is conceptually more similar to the networking schemes used in today's data centers. Packet-based switching assumes either an array of fixed lasers or fast tunable transmitters for addressing a specific destination port by

selecting the appropriate wavelength. Tunable wavelength converters (TWCs) and transceivers can be configured much faster than optical MEMS, and therefore, packet-based optical networks can achieve much faster switching times than circuit-based ones. Thus, packet-based optical switching fits better to data center networks with burstier traffic and all-to-all connectivity.

Hybrid vs. All-Optical Architectures

Most of the proposed optical interconnection networks require complete replacement of commodity electrical switches, meaning that they have to provide significantly improved characteristics in order to justify the increased capital expenditure (CAPEX) of the replacement. Therefore, some schemes follow a hybrid approach in which optical interconnects work in parallel with commodity switches. Hybrid schemes (Fig. 4a) offer the advantage of an incremental upgrade of an operating data center with commodity switches, thus reducing the associated CAPEX. ToR switches can be enhanced by adding optical modules, which will increase bandwidth and reduce latency, while at the same time the already deployed Ethernet network is used for all-to-all communication just as before. Thus, a portion of the traffic demands consists of traffic that lasts long enough to compensate for the reconfiguration overhead, so the overall network bandwidth can be enhanced significantly at reduced cost.

Feature	2012	2016	2020
(Bidi) bandwidth	1 Pbytes/s	20 Pbytes/s	400 Pbytes/s
Overall power consumption	5 MW	10 MW	20 MW
Network power consumption	0.5 MW	1 MW	2 MW optical interconnection networks

Table 1. *Data center bandwidth and power consumption projections [1].*

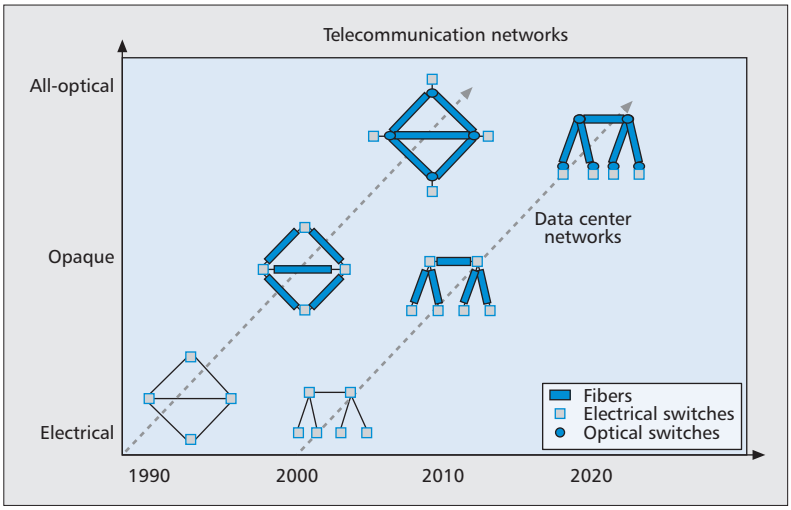

Figure 3. *From electrical to all-optical networks.*

Recent Optical Interconnection Architectures

A typical example of a hybrid circuit-based data center network based both on optical and commodity switches is the Helios architecture [6]. Helios was proposed by the University of California, San Diego and follows a typical two-layer DCN network architecture based on wavelength-division multiplexed (WDM) links. It consists of ToR switches (called pod switches) and core switches. Pod switches are typical packet switches, while core switches can be either electrical switches or optical circuit switches. Electrical packet switches are used for fast all-to-all communication between pod switches, while optical circuit switches are used for high-bandwidth, slowly varying, and usually long-lived communication between pod switches. The same group has also recently proposed a novel hybrid optical circuit/electrical packet network called Microsecond Optical Research Datacenter Interconnect Architecture (Mordia) [7]. This hybrid network uses an optical circuit switching (OCS) architecture based on a wavelength-selective switch (WSS). Each node is connected to both a standard 10G Ethernet network and the OCS. Each node is assigned its own wavelength, and the wavelengths are added or dropped using the OCS. The main advantage of the proposed scheme is that it can provide an average network reconfiguration time of as low as 11.5 μs and can be scaled efficiently to a high number of nodes.

A typical example of a packet-based optical interconnection network is the DOS architecture [8]. Switching in the DOS architecture is based on an arrayed waveguide grating router (AWGR) that allows contention resolution purely in the wavelength domain. The cyclic wavelength routing characteristic of the AWGR is exploited, which allows different inputs to reach the same output simultaneously using different wavelengths. Apart from the AWGR, the optical switch fabric also consists of an array of TWCs, one per node. Each node can access any other node through the AWGR by appropriately configuring the transmitting wavelength of its TWC. The scalability of the DOS scheme depends on the AWGR scalability as well as the tuning range of the TWC. Its main advantage is that latency is almost independent of the number of input ports and remains low even at high input loads. This is due to the fact that packets have to traverse only one optical switch and thus avoid the delay of the electrical switch's buffers. For contention resolution, a loopback buffer based on typical DRAM is used in which packets are temporarily stored. An enhanced scheme has also been presented, called **LIONS** [9], in which shared buffers are eliminated by using a novel negative acknowledgment scheme, thus significantly reducing the overall power consumption required for temporarily buffering the congested packets.

The **OSA** architecture, proposed by the University of Illinois and NEC, is based on WSS switch modules and an optical switching matrix based on MEMS [10]. Each ToR switch has several optical transceivers operating at different wavelengths. The optical wavelengths are com-

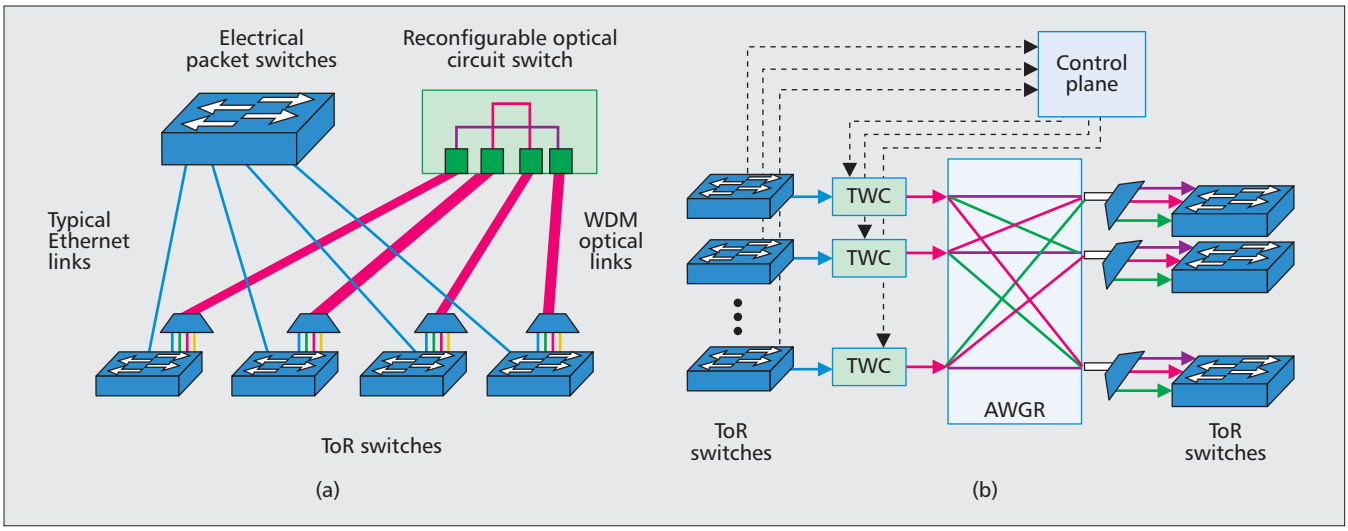

Figure 4. *Examples of proposed architectures: a) hybrid network with optical circuit switching; b) optical packet switching using tunable wavelength converters and an arrayed waveguide grating router.*

bined using a multiplexer and are routed to a WSS. The WSS multiplexes wavelength in up to k different groups, and each group is connected to a port in the MEMS optical switch. Thus, a point-to-point connection is established between the ToR switches. On the receive path, all of the wavelengths are demultiplexed and routed to the optical transceiver. The switching configuration of the MEMS determines which set of ToRs are connected directly. When a ToR switch has to communicate with a ToR switch with which it is not directly connected, hop-by-hop communication is employed. Thus, the OSA architecture must ensure that the entire ToR graph is connected when performing the MEMS reconfiguration. The main advantage of this architecture is that although it is based on circuit switching, it provides all-to-all communication through the use of multiple hops when two nodes are not directly connected. However, scalability is constrained by the number of ports of the MEMS switches.

Another scalable optical networking concept targeting data centers is the **Fission** architecture [11]. Fission comprises fiber rings that are used to interconnect the core data center switches by deploying ultra-dense WDM (UDWDM) technology. Each of the nodes (called electrical-optical switches, EOSs) connected to the fiber ring utilizes a WSS that is used to add and/or drop wavelengths into and from the fiber ring, respectively. The main advantage of the proposed scheme is that it can be scaled efficiently by supporting several rings depending on the number of nodes and the bandwidth requirements.

In [12], a novel space-time interconnection architecture (STIA) is proposed. STIA utilizes the space domain to switch packets and the time domain to switch packets to different nodes. A space switch is used as a central node for the switching of packets. Furthermore, the wavelength domain is used in order to increase the throughput by encoding packets on multiple wavelengths. The STIA architecture can be scaled efficiently using folded Clos or flattened butterfly topologies and can provide more than an order of a magnitude lower power consumption compared to Ethernet-based networks.

Although optical links can provide high throughput, in some cases the communication between some nodes (servers or switches) may require lower bandwidth than the one provided by a 10 Gb/s or 40 Gb/s link. To address this problem by providing fine grain bandwidth allocation, a flexible-bandwidth optical orthogonal frequency-division multiplexing (OFDM)-based DCN architecture has recently been presented [13]. OFDM can achieve high spectral efficiency via the parallel transmission of spectrally overlapping lower-rate subcarriers where the signals are mathematically orthogonal over one symbol period. The architecture of the scheme is depicted in Fig. 5. Like some of the aforementioned schemes, the central basic optical component for the proposed architecture is a cyclic AWGR. As mentioned above, an $N \times N$ AWGR can be thought of as a passive MIMO optical multiplexer/demultiplexer that routes different wavelengths from N different input ports to N different output ports in a cyclic manner. In this scheme, though, each wavelength can be shared by multiple ToR-ToR connections. Each node (i.e., ToR switch) hosts an OFDM modulator and an array of directly modulated lasers (DMLs). Depending on the exact ToR-to-ToR traffic requirements, each ToR transmitter allocates a specific number of OFDM subcarriers per destination ToR. Then the electrical OFDM signal (by means of a DML) modulates the wavelength that will be routed via the AWGR to the specific destination ToR. On the receiving side, a single photodetector PD at each ToR can receive *all* WDM channels simultaneously through parallel signal detection (PSD) technology and then electrically decode the subcarriers. Thus, as long as there is no subcarrier overlapping among the wavelengths arriving from different sources (which is guaranteed by an overlying control plane), all-to-all (multiple-input multiple-output, MIMO) operation is possible.

The main advantage of this architecture is that it can achieve fine-grained bandwidth allo-

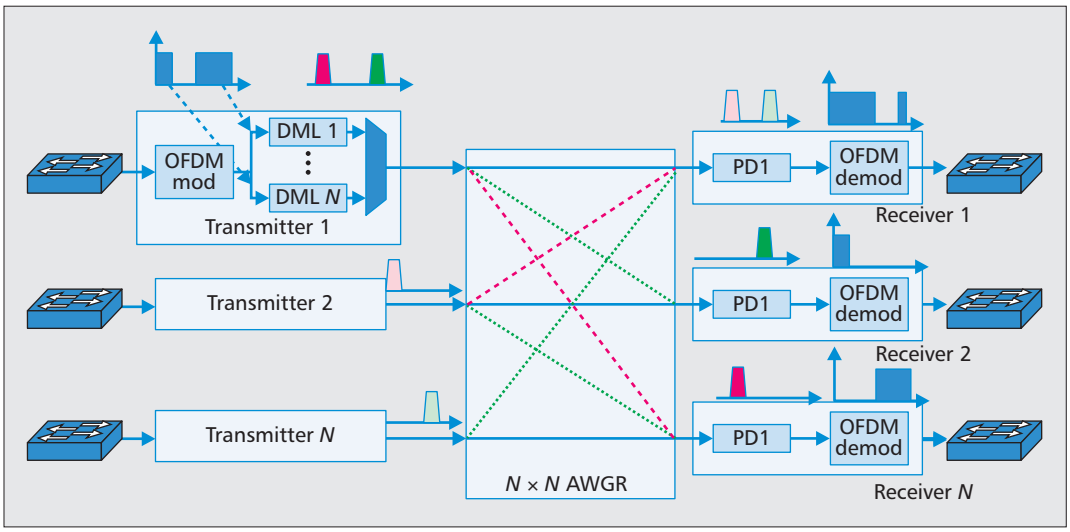

Figure 5. *The optical MIMO OFDM-based architecture [13].*

Most of the optical interconnection networks are proposed by academia and research centers. However, some companies have recently made commercial products available that target DCNs based on all-optical interconnects.

cation using the OFDM allocation scheme. This architecture also ensures that all switched signals take exactly one hop (i.e., passing through the AWGR only once), so latency is kept almost constant between the source ToR and the destination ToR. Furthermore, the proposed scheme can achieve low power consumption using the AWGR as the main passive switching component, while it also makes use of low cost modules.

Commercial Architectures

Most of the optical interconnection networks are proposed by academia and research centers. However, some companies have recently made commercial products available that target DCNs based on all-optical interconnects. For example, Calient Technologies is among the first companies that has commercialized optical interconnection networks explicitly for data centers. Calient is offering a hybrid packet-circuit solution in which the network consists of both packet switching and OCS [14]. Short non-persistent data flows use typical ToR switches, while large persistent data flows utilize OCS, providing very low latency (less than 60 ns) and high throughput. This hybrid approach requires the adoption of a software defined network (SDN), which can separate the control plane from the data plane. Calient is currently using the OpenFlow standard for the SDN infrastructure.

Plexxi is a startup company that has recently introduced an optical switch targeting DCNs. Plexxi's switch integrates Ethernet switching with a centralized SDN-based for the control plane. Plexxi switches are basically interconnected in a ring topology using LightRail optical multiplexing technology. The main advantage of this approach is that it replaces traditional switched hierarchies with a scalable high-bandwidth low-latency network. The flat ring architecture enables linear scaling, with each additional switch adding fabric capacity. In fact, although Calient and Plexxi provide different architectures, these two architectures can be combined in order to provide more flexible topologies and scalable solutions with even less latency [15]. In such hybrid architecture, data center switches would be connected through the Plexxi switch ring, which would in turn be connected to the Calient optical fabric via 10GbE or 40GbE ports. Plexxi switches are connected to each other via an optical interconnect, and when there is a large traffic flow, Plexxi switches bypass it from the ingress switch to Calient's optical fabric in order to reduce latency. The main advantage of this approach is that it protects the network from congestion and also guarantees a high-bandwidth low-latency path for high-volume flows.

Some other companies provide all-optical architectures that are based on advanced optical switching technologies. For example, Polatis also offers an optical interconnection network for intra-data centers [16]. The Polatis optical switch is based on piezo-electric OCS and beam steering technology. Hence, the provided scheme is based on a centralized optical switch that can be reconfigured based on the network traffic demand. The most important features of the provided switch are relatively low power consumption and the capability of being data rate agnostic, which means it can support 10, 40, and 100 Gb/s. The only drawback of this commercial scheme is that it is based on optical MEMS switches, and thus has increased reconfiguration time (according to the data sheets the maximum switching time is less than 20 ms).

Classification: Qualitative Comparison

Figure 6 depicts the types of optical interconnects in terms of connectivity (circuit-based vs. packet-based) and technology (all-optical vs. hybrid). As described above, circuit-based approaches mainly target data centers in which the hosted applications usually require long-lived traffic flows that transfer big chunks of data to other nodes. On the other hand, packet-based circuits can provide all-to-all communication, and due to the fast switching time, they can support both long-lived and short-lived traf-

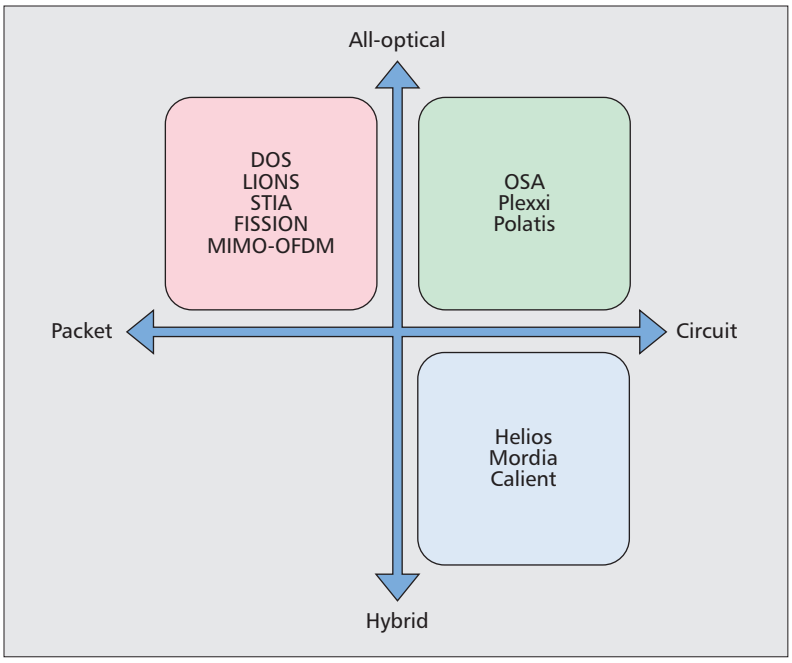

Figure 6. *DCN optical interconnections categorization.*

fic flows. The main advantage compared to electrical packet-based networks is that packets do not need to carry additional data for the destination port (i.e., headers) since switching can be performed based on the transmitted wavelengths.

Optical packet-based approaches can in essence support all types of traffic and are therefore proposed as a total replacement for current DCNs. This is the reason they have not been proposed in the context of hybrid schemes, in contrast to circuit-based architectures. In the case of circuit switching, a combination with current data center infrastructures can ensure all-to-all communication and the opportunity to avoid increased connection reconfiguration times for bursty traffic.

Future Challenges

As has been discussed, optical interconnects can provide a viable approach in order to address the need for high-performance networking inside data centers, by providing high throughput, low latency, and reduced energy consumption. There are, however, remaining challenges that need to be addressed for optical interconnection networks to be widely adopted by data center operators.

Software Defined Networking

Currently, DCNs are based on Ethernet, specifically on some Ethernet modifications such as converged enhanced Ethernet (CEE) and data center Ethernet (DCE) that try to overcome the limitations of Ethernet. However, several data center operators have already moved to specialized control plane schemes based on the SDN concept.

A prominent example is the OpenFlow protocol, which provides an open framework for centrally programming the flow tables across the network switches instead of using legacy distributed protocols. The main advantage of OpenFlow is that it decouples the control plane from the data plane. A similar approach could be followed in the optical interconnection networks discussed above, whereby scheduling and bandwidth allocation would be performed at a centralized controller, while the data plane would still be in the optical domain. However, this also implies that all optical equipment (tunable transmitters, AWGRs, etc.) should be SDN-compliant and that proper abstractions (possibly requiring extensions to the OpenFlow standard) should be defined regarding their capabilities and configuration.

Scalability

The main advantage of current DCNs is that they can be scaled cost efficiently to fat-tree topology hosting thousands of servers, leveraging the economies of scale of commodity Ethernet switches. On the contrary, most of the optical interconnection networks that have been proposed are based on centralized architectures (e.g., using a central AWGN or central MEMS switch), which obviously limits their scalability. This relates on one hand to the cost of those devices, but also to the technical challenges for enhancing their capabilities. Therefore, a major burden for the adoption of optical interconnects by data center operators in the years to come is the efficient deployment of scalable solutions that can host the hundreds of thousands of servers in a warehouse-scale data center.

Resilience

This is a major concern in data centers, currently addressed by means of redundancy. For example, as shown in Fig. 1, each aggregate switch can be connected to multiple core switches, not only for providing increased connectivity but also for enhancing resilience in case of a link or interface failure. The same could happen between ToR and aggregate switches. Optical interconnects based on centralized architectures (whereby a single ultra-high-capacity optical switch covers a very large number of servers) are thus vulnerable to single-point-of-failure issues, and therefore optimized variants of the proposed architectures should be developed to address relevant issues.

Conclusions

The emerging field of optical interconnection networks has opened up new horizons for ultra-high-capacity data center networks that can provide low latency and reduced power consumption. The architectures proposed until now promise to provide significant advantages over the current DCNs based on commodity switches. However, in order to be widely adopted by data center operators, optical interconnects have to overcome several major challenges, such as the need for enhanced scalability and resilience as well as reduced cost.

References

[1] Cisco Global Cloud Index: Forecast and Methodology, 2011–2016, Cisco White Paper, 2011.

[2] T. Benson, A. Akella, and D. A. Maltz, "Network Traffic Characteristics of Data Centers in the Wild," *Proc. 10th Annual Conf. Internet Measurement*, New York, NY, pp. 267–80.
[3] A. Greenberg *et al.*, "The Cost of a Cloud: Research Problems in Data Center Networks," *Proc. ACM SIGCOMM CCR*, Jan. 2009.
[4] M. Taubenblatt, "Optical Interconnects for High–Performance Computing," *J. Lightwave Tech.*, 30, 2012, pp. 448–57.
[5] P. Pepeljugoski *et al.*, "Low Power and High Density Optical Interconnects for Future Supercomputers," *OFC 2010*.
[6] N. Farrington *et al.*, "Helios: A Hybrid Electrical/Optical Switch Architecture for Modular Data Centers," *Proc. ACM SIGCOMM '10*, 2010, pp. 339–50.
[7] N. Farrington *et al.*, "A 10 ms Hybrid Optical–Circuit/Electrical–Packet Network for Datacenters," *OFC 2013*.
[8] X. Ye *et al.*, "DOS: A Scalable Optical Switch for Datacenters," *Proc. 6th ACM/IEEE Symp. Architectures for Networking and Commun. Sys.*, 2010, pp. 24:1–12.
[9] Y. Yin *et al.*, "LIONS: An AWGR–Based Low Latency Optical Switch for High–Performance Computing and Data Centers," *IEEE J. Sel. Topics Quantum Electronics*, vol. 19, no. 2, 2013.
[10] K. Chen *et al.*, "OSA: An Optical Switching Architecture for Data Center Networks with Unprecedented Flexibility," *10th USENIX Symp. Networked Sys. Design and Implementation*, 2012.
[11] A. A. Gumaste *et al.*, "Architectural Considerations for the FISSION (Flexible Interconnection of Scalable in Systems using Integrated Optical Networks) Data–Center," *Int'l. Conf. Optical Network Design and Modeling '13*, Brest, France, Apr. 2013.
[12] I. Cerutti *et al.*, "Designing Energy–Efficient Data Center Networks Using Space–Time Optical Interconnection Architectures," *IEEE J. Sel. Topics in Quantum Electronics*, vol. 19, no. 2, 2013.
[13] P. N. Ji *et al.*, "Design and Evaluation of a Flexible–Bandwidth OFDM–Based Intra Data Center Interconnect," *IEEE J. Sel. Topics Quantum Electronics*, vol. 19, no. 2, doi:10.1109/JSTQE.2012.2209409.
[14] "The Hybrid Packet Optical Circuit Switched Datacenter Network," White paper, Calient Inc., 2012.
[15] "Affinities in Action: Plexxi and Calient," Data sheet, Plexxi, Inc., Mar. 2013.
[16] "The New Optical Data Center," Polatis Data Sheet, Polatis Inc., 2009.

BIOGRAPHIES

CHRISTOFOROS KACHRIS (kachris@ait.edu.gr) is a senior researcher at AIT since 2010. He received his Ph.D. in computer engineering in 2007 from the Technical University of Delft. From 2009 to 2010 he was a visiting lecturer at University of Crete and a visiting researcher at FORTH where he coordinated the Interconnects cluster of High Performance Embedded Architectures and Compilers (HiPEAC). His main research interest is in the area of reconfigurable computing (FPGAs), high-speed network processing, multicore embedded systems, computer architecture, and interconnects.

KONSTANTINOS KANONAKIS (kkan@ait.edu.gr) was awarded his Ph.D. and Dipl.-Ing degrees in 2008 and 2004, respectively, both from the National Technical University of Athens (NTUA), Greece. His main research interests are in the area of architectures and control protocols for broadband access and optical core networks. He has co-authored more than 50 papers that appeared in international peer-reviewed journals and conferences, and has participated in several EU-funded projects.

IOANNIS TOMKOS (itom@ait.edu.gr) has been with AIT since September 2002. At AIT he founded and serves as the head of the High Speed Networks and Optical Communication (NOC) Research Group. He represented AIT as Principal Investigator and has had a consortium-wide leading role in over 20 EU projects. He has published in excess of 450 publications in journals and conference proceedings, and his work has received about 3000 citations (h-factor = 28). For his scientific achievements, he was named a Distinguished Lecturer of IEEE (2007) and a Fellow of the IET (2010) and OSA (2012) for "outstanding scientific contributions in the field of transparent optical networking."

In order to be widely adopted by data center operators, optical interconnects have to overcome several major challenges, such as the need for enhanced scalability and resilience as well as reduced cost.