Photonic Switching Technologies, Architectures, and Integrated-Systems for Future Disaggregated and Optically Reconfigurable Data Centers

S. J. Ben Yoo  
Department of Electrical and Computer Engineering  
University of California  
Davis, California 95616, USA  
Email: sbyoo@ucdavis.edu

Abstract—This Tutorial covers technologies, architectures, and system-integration for future data centers with optical reconfigurability. Optical interconnects allow disaggregation of computing resources in the data centers thanks to distance-independent energy-efficient and high-throughput communications of photonics. Photonic switching can provide additional benefits of reconfigurability of the interconnection topologies without requiring electronic switches that accompany store-and-forward mechanisms. Hence, the primary motivation for considering photonic switching in data centers rises from the need for energy-efficient and scalable intra-data center networks to meet rapid increases in data traffic driven by emerging applications, including machine learning. To accommodate such traffic, today’s large-scale data centers employ cascaded stages of many power-hungry electronic packet switches interconnected across the data center network in fixed hierarchical communication topologies. Numerous research papers have predicted significant benefits in scalability, throughput, and power efficiency from deploying photonic switches in data centers. However, photonic switching is not yet widely deployed in commercial warehouse-scale data centers at the time of writing this Tutorial due to significant challenges. They are related to (1) cross-layer issues involving control and management planes together with data integrity during switching, (2) scalability to > 5000 racks (> a quarter-million servers), (3) performance monitoring required for reliable operation, (4) currently existing standards allowing limited power margin (3 dB), and (5) other practical (technology-dependent) issues relating to polarization sensitivity, temperature sensitivity, cost, etc. We will discuss possible solutions for future data centers involving cross-layer methods, new topologies, and innovative photonic switching technologies. Furthermore, the Tutorial broadly surveys state-of-the-art photonic switching technologies, architectures, and experimental results, and further covers the details of arrayed-waveguide-grating-router-based switch fabrics offering hybrid switching methods with distributed control planes towards scalable data center networking.

Keywords—data center networking, switching, optical switching, photonics, optical packet switching, optical burst switching, silicon photonics, electronic switching

I. INTRODUCTION

Our daily lives critically depend on data communications. Global data center IP traffic grew 11-fold over the past eight years [6] at a Compound Annual Growth Rate (CAGR) of 25%, exceeding 20 Zettabytes per year by 2021 [1]. More recently, driven by the rapid increases in AI and machine learning related traffic, some estimates indicate that the annual data traffic will increase by over 400× over the next 10 years corresponding to CAGR of 82%. At the same time, the global energy consumption in data centers reached 200 TWh in 2020 [2] with a CAGR of 4.4% [3].

Today’s data center network architectures heavily rely on cascaded stages of many power-hungry electronic packet switches interconnected across the data center network in fixed hierarchical communication topologies such as Fat-Tree within the data center (see Fig. 1(a))[4]. Due to the limited radix and bandwidth of the electronic switches, warehouse-scale data centers involve a large number of cascaded electronic switches where high energy consumption and latency compound due to repeated ‘store-and-forward’ electronic processes. These architectures are also designed with a fixed topology at fixed data rates. On the other hand, as Fig. 1(b) illustrates, employing a passive optical fabric or a reconfigurable optical switch fabric with distributed electronic switches (e.g. ToR) could greatly improve (a) scalability of the capacity and the number of compute nodes (or racks with ToRs), (b) energy-efficiency of the network, (c) modular upgradeability, and (d) cost savings by eliminating many large and power-hungry core electronic switches at the core while keeping the smaller and disaggregated electronic switches (e.g. ToR) at the edge nodes. This transformation not only flatten the interconnect topology of the data center networks with a reduced number of hierarchies, but it also brings the possibility of optical reconfigurability enhanced by wavelength division multiplexing (WDM) and silicon photonics. Fundamentally, an all-to-all interconnect topology (shown in Fig. 1(c)(Left)) can offer uniform and contentionless interconnections between the compute nodes. As actual data centers must handle data movements of nonuniform and dynamically changing traffic patterns, therefore, their interconnection topologies and bandwidth assignments should closely reflect those driven by the workflow. ‘Application-aware’ networking [5] of data centers, would then benefit from a reconfigurable interconnection platform which can, for example, represent a low-latency all-to-all topology (e.g. Fig. 1(c)(Left)) at certain
Fig. 1. (a) A fat tree topology using electronic switches at the core and at the aggregation edges of the network, (b) a flattened optically interconnected network example utilizing a passive optical fabric (such as arrayed waveguide routers) or a reconfigurable optical switch with electronic switches at the edges (e.g. ToR), and (c) [Left] all-to-all interconnection and [Right] bandwidth-steered interconnected topology after reconfiguration [4].

Fig. 2. Examples of applications showing very different data movement patterns. Map-reduce, deep neural network, and sorting applications all show all-to-all communication pattern at some point of the workflow but not throughout the entire workflow.

2022 International Conference on Optical Network Design and Modelling (ONDM)
can benefit from a rich set of processing capability by the electronic switch without having to include the guard time for switching.

Lack of viable optical buffers add challenges to both the control plane and the data plane of optically reconfigurable data center networks. The electrically reconfigurable data center networks can mitigate such challenges thanks to electronic buffers, despite inferior scalability and energy-efficiency. The hybrid switching networks with optically interconnected distributed electronic switches can possibly achieve the benefit of scalability, energy-efficiency, and agile reconfigurability. The question then is whether to add reconfigurability to this optical interconnection of distributed electronic switches.

III. DATA PLANE, CONTROL PLANE, AND MANAGEMENT PLANE

The introduction of TCP/IP and the availability of Layer 2 and Layer 3 protocols such as Ethernet, ATM, SONET, and OTN meant that hardware switches with proper protocols embedded in the linecards can readily achieve network switching since the control plane and management plane protocols can run on those protocol-specific linecards (and the switch fabric). Such switches employed distributed control and management planes, using the protocol-specific information embedded in the datagram.

Recent electronic switches have evolved towards better programmability, reconfigurability, and protocol independency. Adoptions of an open source programming language such as P4 [13], which allows fast reconfiguration and software-level programmability, greatly facilitates deployment of large-scale data centers and computing clusters, making it easy for operators to control and manage these complex systems. This also meant that optical switches developed for 2nd generation and 3rd generation optical networking could play an active role in data center networks with the SDN paradigm. Optical MEMS switches already developed for telecom more than a decade ago could be readily deployed in data centers. However, scalability of the centralized control and management planes becomes challenging if rapid reconfigurations are required at high load in a network with a large number of nodes. For this reason, optical switches face challenges if optical reconfigurations are required rapidly and frequently in a large data center network, while electronic switches can more readily support such reconfigurations due to the integrated electronic memory and switch fabric despite their high-power consumption and capacity limitations. Thus, for dynamic optical circuit switching data center networks with reconfigurable optical switches, cross-layer issues involving the application, transport, network, link, and physical layers inevitably become extremely important. However, this cross-layer issue remains as unsolved and too challenging for data center networks at scale.

IV. SCALING AND DISAGGREGATION OF DATA SWITCHING

Today’s data centers often employ many thousands of racks, and the scalability of the data centers is a compelling requirement while networking such a large-scale data center becomes an immense challenge seen both from the data plane and the control plane perspectives. Further, the multi-tenant data centers are becoming more popular running heterogeneous applications simultaneously. Hence, a scalable and disaggregated switching network is desired in the data plane, while controllability, manageability, and virtualization [14] of the data center network are necessary.

In an electronic data center network shown in Fig. 1 (a), scaling-up becomes challenging due to the limitations in the bandwidth, radix, and switching capacity of the electronic switches if electronics-only solutions are sought. In scale-up data center networks, each individual network devices must increase its capacity and bandwidth, which is difficult to achieve with electronics-only solutions. Scaling-out data center networks utilizing commodity electronics is far more attractive from both flexibility and energy-efficiency perspectives [15], as demonstrated by Facebook’s F16 networks [16].

On the other hand, as Fig. 1(b) illustrates, employing a passive optical fabric or a reconfigurable optical switch fabric with distributed electrical switches (e.g. ToR) at the edges could greatly facilitate scalability and disaggregation of the data center networks while offering significant energy, modular upgradeability, and cost savings by eliminating large and power-hungry electronic switches at the core. Fig. 3 shows one such example employing a N×N cyclic arrayed waveguide grating router (AWGR) with all-to-all interconnection capability by optical wavelength routing. As we will discuss later, since such an AWGR supports N² simultaneous optical circuits without contention, the switch capacity can scale to, for example 26.2 Pb/s interconnection capacity using 100 Gb/s transceiver per port for N=512 [17] (512²×100 Gb/s=26.2 Pb/s).

Scalability of all-optical switch fabrics are typically limited by the number of required switching elements (that may scale as \(O(\sqrt{N})\), \(O(N\log N)\), or \(O(2N)\)) or by the number of cascaded stages of optical switches. In some cases, such as c-Through [18] or Helios [19] networks, the authors proposed to use optical switches to supplement or partially replace electronic switches to improve communications in data centers.

Alternatively, hybrid switching consisting of wavelength routing and electronic switches achieves this arbitration-free all-to-all interconnection. Then, each node is interconnected in an all-to-all topology as shown in Fig. 3(a) where \(P\) nodes are directly optically interconnected to each other. Physically, such a network would require \(\frac{P(P-1)}{2}\) pairs of optical fibers. As Fig. 3 (b) illustrates, this interconnection can be greatly simplified by introduction of WDM and a wavelength routing device such as an AWGR with the well-known cyclic frequency routing characteristic, where an \(N\times N\) AWGR interconnects [10], [11],[22] for \(p\) number of nodes emitting \(N\) wavelengths, where \(N = p + \mu\). Hierarchical switching can be achieved as illustrated in Fig. 3 (c) [22]. Fig. 3 (d) and (e) illustrate wavelength routing properties of cyclic frequency routing AWGRs (shown is an \(N=5\) example) [23][24], and Fig. 3 illustrates how a data center network with a passive optical wavelength routing device, such as a cyclic frequency \(N\times N\) AWGR [23][24], to interconnect \(N\) racks with a Top of the Rack (ToR) switch with \(N\) wavelength WDM ports.

The scalability of supporting many compute nodes can be achieved in three ways. The first method is to introduce one very large \(N\times N\) AWGR. Although silicon photonic 512×512
AWGRs [17] and other large-scale cyclic-frequency AWGRs on PLCs have been demonstrated, this method is considered impractical because it would require a large number of wavelengths (N) and TRXs and it would induce a substantial amount of crosstalk. To address the wavelength and crosstalk issues, a Thin-CLOS architecture [25], [26] has been designed to achieve the same all-to-all interconnection by using many small $W \times W$ AWGRs so that the number of wavelengths and the amount of crosstalk would reduce significantly.

![Diagram](image)

Fig. 3. (a) Fully connected all-to-all interconnection network, (b) fully-connected all-to-all interconnection network utilizing wavelength routing by an Arrayed Waveguide Grating Router (AWGR), (c) Hi-LIONS with fully connected subnetworks that are interconnected with a reconfigurable optical switch, (d) all-to-all wavelength routing interconnection pattern of a $N \times N$ cyclic AWGR using $N$ wavelengths ($N = 5$ example), (e) wavelength routing property of the $N \times N$ cyclic AWGR ($N = 5$ example) offering all-to-all interconnects using $N$ wavelengths.

V. INTERCONNECTION NETWORK TOPOLOGIES

Ideally, the interconnection topology of the data center network should closely match the data flow pattern according to the workload of the data center at any given time. In practice, as Fig. 2 illustrates, the data flow pattern changes from workload to workload, and from one phase of the workload to another even within the same workload. Fig. 2 also shows that all-to-all communication is necessary at some point in time in all three application examples but not necessary all the time. On the other hand, supporting all-to-all communications in a network topology realized by interconnection of low-radix switches often cause elevated congestion and latency. For these reasons, a reconfigurable optical switch capable of supporting any arbitrary connectivity including all-to-all interconnection is desirable. Various interconnection topologies: Flattened Butterfly, FatTree, Dragonfly, 3D Torus, 3-stage CLOS, Hypercube, and SlimFly are considered typically for electronic switches, and hybrid data center interconnection topology involving both electronic and reconfigurable optical switches including are possible for c-Through [18], Helios [19], and Optical Switching Architecture (OSA) [27].

VI. TIME SCALES FOR RECONFIGURATION AND LIMITATIONS IMPOSED BY THE CONTROL PLANE

The previous section compels us to consider optical switches capable of configuring the data center interconnection topology that would be optimally matched to the data flow pattern for the given workload, or even reconfiguring during the run time of the application as the data flow pattern changes within the run time. As we will see later in this section, it is extremely challenging to realize scalable and low-latency control planes for such a reconfigurable optical circuit switching driven by the dynamicity of the changing traffic patterns.

In standard or dynamic circuit networks’ physical layers, 100 μs or longer timescale reconfiguration may be sufficient, but flow-switching or burst-switching should achieve reconfigurations at below 100 μs, while packet-switched networks must achieve switching at much faster time-scale than the length of the packets (< 1 ns). Seen from the applications or workload, job-level reconfigurations can be at timescales longer than 1 ms, while flow-level reconfigurations and packet-level reconconfigurations should achieve < 100 μs and < 10 ns respectively. In terms of the control plane, depending on the scale of the data center network and the scheduling algorithm, the centralized SDN control plane may be able to achieve and complete reconfiguration of the data center networks at 1 ms or longer, while faster reconfiguration should resort to distributed hardware control using FPGAs or ASICs.

The benefit of reconfigurations of communication networks in data and computing systems have been discussed from the perspective of efficiently and effectively utilizing the available resources (processing, memory, and communications) [27]–[30]. In particular, mitigating hot-spot creations in data centers [28], [31], [32] can be where optical reconfigurations can prove to be very useful.

These findings indicate that a static optical circuit network may not be effective for future data centers and a dynamically reconfigurable optical network should be considered. But the time scale of these bursts at < 25 μs casts serious challenges for scheduling and for control plane designs across the data center network. Even if optical switches can reconfigure in less than 1 ns, the control plane cannot achieve coordination between all the compute nodes in the data center in such a short amount of time. In a high-performance computing system running a single threaded application with predictable changes in traffic patterns (e.g. map-reduce application transition from map-phase to reduce-phase), such a reconfiguration is conceivable if a guard time is incorporated, but it is difficult to predict such traffic patterns in a data center running many heterogeneous workloads simultaneously. Some studies are underway to apply machine-learning methods to statistically predict data flow patterns within the data centers [33][34].

The challenges in implementing centralized control planes for optical reconfiguration also raise an interesting question regarding the viability of optical-packet switching (OPS) and optical-burst switching (OBS), in addition to dynamic optical circuit switching (OCS) in future data centers. For OCS based intra-datacenter networks, we assumed the central control plane triggering reconfiguration of optical switches for dynamical reconfiguration of optical circuits. The scalability of this method depends greatly on scheduling algorithms and on the dynamicity and the load of the traffic pattern. The NEPHELE project [35] introduced WDM ring network with optical reconfiguration based on TDMA time-slot allocated by
a SDN (OpenFlow) control plane to schedule these applications, and showed that the makespan can reach 48% when short-term load dynamicty is high [36]. Such centralized schedulers inevitably add scheduling delays which can become unacceptably high in large data center networks. Just to poll the traffic demands, the delay $T_D$ scales as $O(N^2)$ [37], which corresponds to 4 ms for $N=5000$ racks at $C_{BW}=100$ Gb/s. Ref [38] suggested, with some optimism, an observe-analyze-act framework including a number of intelligent algorithms while recognizing unsolved problems, while Ref. [37] declared that the central scheduling a dead-end unless (a) fixed scheduling without considering application awareness and with additional latency, or (b) distributed scheduling with less accurate or no coordination is adopted.

VII. SWITCHING TECHNOLOGIES

In considering optical switching technologies for data centers, there are countless attributes that must be considered. As discussed in [8], these attributes can be summarized in three categories: signal quality, configuration, and performance. Table 1 summarizes various photonic switching technologies. It is also possible to integrate the switching functions to all-to-all interconnects discussed as LIONS so that reconfiguration from all-to-all to arbitrary interconnection is possible. This new reconfigurable wavelength routing switch are called FlexLIONS [39]–[40].

VIII. SUMMARY AND FUTURE PROSPECTS

The accelerating trend of exponential growth in data traffic and the fact that the large portion of the data traffic reside in the data centers imply that photonic switching could play increasingly important roles in future scalable data and computing systems. However, there are significant challenges relating to scheduling of many concurrent applications with dynamically changing traffic patterns when attempting to introduce photonic switching technologies in large-scale data centers. Cross-layer design of scheduling and control will be important. Centralized control plane is effective only if it can handle dynamic high-capacity applications in a scalable manner. A combination of distributed and centralized control planes is expected to be necessary. Further, recently developed silicon photonic switches and the availability of foundry-based manufacturing and packaging exploiting CMOS electronic industry ecosystem can accelerate electronic-photonic integration and development of photonic switching embedded in compute nodes, backplanes, and racks.

Table 1. Summary of various optical switching technologies and their comparisons. (SNB: Strictly Nonblocking, WNB: Wide-sense Nonblocking, RNB: Rearrangeable Nonblocking).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>~4 ms</td>
<td>~55</td>
<td>&lt; 2</td>
<td>10,000</td>
<td>Point-to-Point Mesh</td>
<td>SNB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>~10 ms</td>
<td>~50</td>
<td>20</td>
<td>10,000</td>
<td>Point-to-Point Mesh</td>
<td>SNB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>240x240</td>
<td>~70</td>
<td>&gt; 3</td>
<td>10,000</td>
<td>Crossbar</td>
<td>SNB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>350 ns</td>
<td>~25</td>
<td>&gt; 3</td>
<td>10,000</td>
<td>Point-to-Point Mesh</td>
<td>SNB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>~10 µs</td>
<td>~50</td>
<td>20</td>
<td>10,000</td>
<td>Cylindrical Spanke-Benes</td>
<td>SNB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16x16</td>
<td>Single: -23</td>
<td>Single Pol.</td>
<td>~1 [on-chip]</td>
<td>Benes, Dilated-Benes, Dilated-Banyan, Crossbar</td>
<td>SNB; RNB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2x2 WOXC; 1x9 WSS</td>
<td>-35</td>
<td>0.2</td>
<td>10,000</td>
<td>For 2x2: SNB</td>
<td>SNB</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8x8 (crosstalk limited)</td>
<td>-28</td>
<td>Single Pol.</td>
<td>~1 [on-chip]</td>
<td>Crossbar, Benes, Dilated-Benes, etc.</td>
<td>Crossbar: SNB; Others: RNB</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
ACKNOWLEDGMENT

This work was supported in part by DoD #H98230-16-C-0820 and NSF grant # 1611560. The author would like to thank the many researchers around the world who contributed to this paper, especially R. Proietti, X. Xiao, G. Liu, and Yu Zhang.

IX. REFERENCES


