Bridging the Last Mile for Optical Switching in Data Centers

Hitesh Ballani Paolo Costa* Istvan Haller Krzysztof Jozwik
Kai Shi Benn Thomsen Hugh Williams
Microsoft Research
Corresponding author: paolo.costa@microsoft.com

Abstract: Optical switches promise to revolutionize data centers by providing high bandwidth and low latency at low cost. This paper discusses some of the remaining challenges that need to be solved to make this technology successfully deployed in production.

OCIS codes: 060.1810, 130.4815

1. Introduction

The Moore's law for networking—roughly every two years network switches double their bandwidth at the same cost—has allowed data center providers to keep up with increasing server demands while maintaining the network cost low and relatively steady over the years [1]. This favorable trend, however, is being challenged by two potential disruptions expected in the next few years. First, the move towards hardware-accelerated applications [2] and disaggregated workloads [3] will greatly boost the demands in terms of network throughput and latency. Today's network costs are kept relatively low (approximately 10% of the overall expenditure [4]) as network fabrics are heavily oversubscribed [1]. Meeting the requirements of these new emerging applications will need much higher bisection bandwidth, which would greatly inflate costs.

Second, and perhaps more worryingly, current electrical switch technology is expected to hit a wall in two generations from now (>25.6 Tbps) due to the inability to increase the pin density on the Ball Grid Array (BGA) package [5]. While new architectures based on Silicon Photonics (SiPh) and 2.5D/3D packaging or monolithic integration are being investigated, a number of challenges have to be solved before this technology can become viable [6, 7], e.g., the complexity of packaging an external laser source and fiber coupling, not to mention the high manufacturing costs (including packaging and testing). Even if these challenges were to be solved, this technology will ultimately suffer from the difficulty of keeping increasing transistor density due to the limitations of CMOS scaling.

In principle, optical switching has the potential to overcome both these issues. First, it does not require any opto-electrical (OEO) conversion, which significantly reduces the number of expensive transceivers and, hence, cost. Second, by being transparent to modulation format, it can support much higher throughput than today's switches and at much lower latency due to the lack of buffering or packet inspection overhead. Finally, it does not use electronics for switching, thus sidestepping the electrical switch scaling wall and offering a potentially future-proof solution.

To date, many solutions employing optical switches have been proposed [8], using various building blocks, including 2D and 3D microelectrical mechanical switches (MEMS), Liquid Crystal on Silicon (LCoS) display matrices, tunable lasers and arrayed waveguide grating routers (AWGRs), semiconductor optical amplifiers (SOAs), Mach-Zehnder interferometers (MZIs), and microring resonators (MRRs). While these proposals greatly improved over the state of the art and brought us closer to our final target, we argue that bridging the last mile and making this technology ready for production requires a joint effort between the networking and optics communities. With this paper we hope to raise the awareness in both communities on the remaining challenges and to foster the collaboration between the two.

2. Design Challenges

The networking and optics communities have been working on optical switching for many years but each community has addressed the problem from a different perspective. The networking community has focused on designing solutions that can scale to the entire data center and addressed system-level issues such as scalable flow scheduling, demand estimation, and coexistence with legacy transport protocols such as TCP. On the other hand, they usually opted for commercially-available optical switches, typically based on MEMS [9, 10], which exhibit a relatively high reconfiguration time (tens of microseconds to milliseconds). Unfortunately, since the vast majority of flows in data center workloads are smaller than 10 KB [11], much shorter reconfiguration times (tens of nanoseconds or less) are required in order to maintain high network utilization. Consequently, many of the solutions proposed by the networking community included a hybrid design in which a slow (milliseconds or higher) optical network is used for long flows

(i.e. larger than few MBs) while an electrical one is used for short flows. Albeit potentially attractive for some specific domains (e.g. HPC), these approaches, however, are hard to deploy in production large-scale data centers due to the complexity of running two separate fabrics and distributing the traffic across the two. In contrast, the optics community has developed a number of technologies that achieve nanosecond-level switching time, e.g. [12–15]. However, they mostly focused on designing the individual devices (or the collection thereof) while devoting little attention to addressing some of the other challenges, e.g., scalable scheduling or fault-tolerant time synchronization.

Thanks to the combined work of these two communities, we believe that the goal of deploying optical switching in production is now within reach. However, this requires integrating and, in some cases, rethinking the solutions that have been proposed independently by each community. In the rest of this section, we outline some of these remaining *technical* challenges, highlight the pros and cons of existing proposals, and discuss some promising approaches. Finally, we conclude the section with a list of *deployment* challenges that should also be accounted for in the design.

Scalable control plane. One of the main architectural differences between electrical packet switches and optical switches is that the latter are typically *bufferless*. While the lack of buffers avoids any queuing delay and, hence, reduces end-to-end latency, it significantly complicates the design as flows must be carefully orchestrated to prevent any collision, which would result in signal degradation and packet loss. Usually this is achieved by means of a centralized scheduler that given a demand traffic matrix computes the flow assignment and the optical-switch configuration.

This seemingly simply design, however, poses a number of challenges. First, to not inflate the flow completion time of short flows, the scheduler should be able to compute the schedule in a sub-microsecond. Given the scale of today's data centers, which typically comprise a few thousand racks and hundreds of thousands servers, this requires a non-trivial amount of ingenuity and custom hardware support. This task is further complicated by the fact that typically low-radix devices are interconnected in a hierarchical fashion, e.g., using rearrangeably non-blocking topologies, to be able to connect all racks or servers in a data center. Therefore, beside solving the bi-partite matching problem, the scheduler also has to devise a valid routing assignment. Second, the computation throughput of the scheduler must be matched by an equivalent network capacity to collect the demands from racks (or end hosts) and distribute assignments. This is usually implemented using a separate, electrically-switched, control-plane network. This solution, however, increases the deployment and management complexity. Finally, estimating demands beyond the next few packets is rather challenging because applications lack a standardized way to specify the total length of the flow and using the status of the current queues at the NICs or racks is not very accurate given their small sizes.

One promising way to sidestep some of these issues is to adopt a load-balanced switch architecture [16, 17], which relies on a fixed schedule. This, however, extends the average path length with negative implications both in terms of higher latency and reduced throughput (up to 50% in the worst case), which need to be addressed in the final design.

Time synchronization. For the computed schedule to be effective, the end points connected to the optical switch must be time-synchronized at a very fine granularity (ideally few nanoseconds). Any inaccuracy in the time synchronization must be compensated with an accordingly sized interpacket gap, which would reduce the overall throughput.

Recent work demonstrated that it is possible to achieve sub-nanosecond time synchronization at data center scale [18]. This approach, however, assumed frequency synchronization through an electrical-switched network and the availability of bidirectional fiber, which might not be easy to replicate in a production setting. Further, its feasibility and cost implications at 25 Gbps and beyond remain to be investigated.

Burst Clock Data Recovery (CDR). In an electrically-switched network, in the absence of failures or manual intervention, the end points of a Layer 0 (L0) connections never change. In contrast, with optical switches, new L0 connections are created every time the switch configuration changes. This implies that the receiver has to continuously recover the clock of the current transmitter to properly sample the incoming signal and recover the data. The longer this process takes, the lower the network throughput will be as no valid data can be received before the CDR has completed. Off-the-shelf transceivers usually require hundreds of nanoseconds to recover the clock, which is two to three orders of magnitude higher than our target switching time.

Burst CDR has been extensively studied in the context of Passive Optical Networks (PONs) and architectures based on over-sampling or gated oscillators have been shown to achieve sub-nanosecond locking time [19]. These techniques, however, increase the complexity and cost of the transceiver design, and need to be re-evaluated for higher data rates, although initial results are encouraging [20].

We now discuss some of the deployment challenges that arise when transferring a research technology into production. It is crucial that these are not addressed as afterthought but, instead, are considered from the start.

Cost. One of the big advantages of electrical switches is that they can leverage the well-established CMOS ecosystem,

thus benefiting from economies of scale. This is not the case for optical switches, whose manufacturing process is less mature and requires expensive packaging and testing procedures. Further, beside the cost and power of the single switch device, it is important to consider all additional costs derived from implementing an operational optical network, e.g., the need for a separate control-plane network or custom transceiver logic to implement time synchronization or fast CDR. Finally, depending on the technology used, network architects may need to overprovision the number of optical switches or transceivers to compensate for the loss of throughput due to switching time or the interpacket gap. To make optical switches cost-effective against their electrical counterpart, all these contributing factors should be properly accounted for during the design and their cost minimized.

Reliability. Although cost is a critical metric, the top priority for cloud providers is to ensure high availability and uninterrupted service. Even minor outages negatively impact first- and third-party businesses and ultimately result in severe money loss and reduced market share. Due to the tighter coupling exhibited by optical-switched networks, they are intrinsically more prone to failures and particular care must be taken to defend against such scenarios. For example, centralized schedulers or a control-plane network with no (or limited) redundancy represent single points of failure and should be avoided. Similarly, the time synchronization protocol should be designed so as to be robust against individual node failures or network partitions.

Incremental deployment. Despite the attractiveness of clean-slate designs, in practice optical switches will need to be first deployed in a subset of the data center (e.g., a row of racks) before being able to replace the entire network gear. This means that their design must be able to seamlessly integrate with existing network hardware and software stacks (e.g., Ethernet or TCP). Further, it should be possible to regularly upgrade/service individual parts of the network fabric with only minimal (if any) impact on overall performance.

3. Opportunities Ahead

Despite all the challenges distilled in the previous section, we believe that optical switching is becoming mature and it has the potential to revolutionize the cloud infrastructure by providing predictable and uniform high performance (bandwidth and latency) across the entire data center, thus breaking today's silos (e.g., a single server or a rack) with significant benefits in terms of fault tolerance, resource management, and application performance.

Further, some of the technology developed to support optical switching could also create new exciting opportunities to rethink other parts of the stack. For instance, a tightly scheduled network would reduce the reliance on distributed congestion control protocols like TCP or DCQCN, thus making it easier to implement sophisticate QoS policies and provide guaranteed performance to applications and services running in the cloud.

References

1. A. Singh *et al.*, "Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network," in "ACM SIGCOMM," (2015).
2. A. Putnam *et al.*, "A Reconfigurable Fabric for Accelerating Large-Scale Datacenter Services," in "ISCA," (2014).
3. P. X. Gao *et al.*, "Network Requirements for Resource Disaggregation," in "USENIX OSDI," (2016).
4. J. Hamilton, "Overall Data Center Costs," http://bit.ly/1JzfqTq.
5. H. J. S. Dorren, E. H. M. Wittebol, R. de Kluijver, G. G. de Villota, P. Duan, and O. Raz, "Challenges for Optically Enabled High-Radix Switches for Data Center Networks," Journal of Lightwave Technology **33** (2015).
6. S. Rumley, M. Bahadori, R. Polster, S. D. Hammond, D. M. Calhoun, K. Wen, A. Rodrigues, and K. Bergman, "Optical interconnects for extreme scale computing systems," Parallel Computing **64** (2017).
7. D. Thomson *et al.*, "Roadmap on silicon photonics," Journal of Optics **18** (2016).
8. F. Testa and L. Pavesi, eds., *Optical Switching in Next Generation Data Centers* (Springer, 2017).
9. N. Farrington, G. Porter, S. Radhakrishnan, H. H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, and A. Vahdat, "Helios: A Hybrid Electrical/Optical Switch Architecture for Modular Data Centers," in "ACM SIGCOMM," (2010).
10. M. Ghobadi, R. Mahajan, A. Phanishayee, N. Devanur, J. Kulkarni, G. Ranade, P.-A. Blanche, H. Rastegarfar, M. Glick, and D. Kilper, "ProjecToR: Agile Reconfigurable Data Center Interconnect," in "ACM SIGCOMM," (2016).
11. S. Kandula, S. Sengupta, A. Greenberg, P. Patel, and R. Chaiken, "The nature of data center traffic: measurements & analysis," in "ACM IMC," (2009).
12. C. P. Chen, X. Zhu, Y. Liu, K. Wen, M. S. Chik, T. Baehr-Jones, M. Hochberg, and K. Bergman, "Programmable Dynamically-Controlled Silicon Photonic Switch Fabric," Journal of Lightwave Technology **34** (2016).
13. Q. Cheng, A. Wonfor, J. L. Wei, R. V. Penty, and I. H. White, "Low-energy, high-performance lossless 8x8 SOA switch," in "OSA/IEEE OFC," (2015).
14. Y. Yin, R. Proietti, X. Ye, C. J. Nitta, V. Akella, and S. J. B. Yoo, "LIONS: An AWGR-Based Low-Latency Optical Switch for High-Performance Computing and Data Centers," IEEE Journal of Selected Topics in Quantum Electronics **19** (2013).
15. A. Funnel, K. Shi, P. Costa, P. Watts, H. Ballani, and B. Thomsen, "Hybrid Wavelength Switched-TDMA High Port Count All-Optical Data Centre Switch," Journal of Lightwave Technology **35** (2017).
16. C.-S. Chang, D.-S. Lee, and Y.-S. Jou, "Load Balanced Birkhoff-von Neumann Switches, Part I: One-stage Buffering," Computer Communications **25** (2002).
17. W. M. Mellette, R. McGuinness, A. Roy, A. Forencich, G. Papen, A. C. Snoeren, and G. Porter, "RotorNet: A Scalable, Low-complexity, Optical Datacenter Network," in "ACM SIGCOMM," (2017).
18. M. Lapinski *et al.*, "White Rabbit: a PTP Application for Robust Sub-nanosecond Synchronization," in "IEEE ISPCS," (2011).
19. M. Hsieh and G. Sobelman, "Architectures for Multi-Gigabit Wire-Linked Clock and Data Recovery," IEEE Circuits and Systems Magazine **8** (2008).
20. A. Rylyakov *et al.*, "A 25 Gb/s Burst-Mode Receiver for Low Latency Photonic Switch Networks," IEEE Journal of Solid-state Circuits **50** (2015).