The Emerging Optical Data Center

Amin Vahdat[1,2], Hong Liu[1], Xiaoxue Zhao[1] and Chris Johnson[1]
[1] Google, Inc., Mountain View, CA 94043, USA
[2] UC San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
Author e-mail address: vahdat@google.com

Abstract: We review the architecture of modern datacenter networks, as well as their scaling challenges; then present high-level requirements for deploying optical technologies in datacenters, particularly focusing on optical circuit switching and WDM transceivers.

OCIS codes: (200.0200) Optics in computing; (060.4253) Networks, circuit-switched

1. Introduction

An increasing fraction of computing and storage is migrating to a planetary cloud of warehouse-scale data centers [1]. While substantial traffic will continue to flow between users and these data centers across the Internet, an increasing fraction of overall data communication is taking place within the data center [2]. For example, a data center with 100,000+ servers, each capable of 40 Gb/s of bandwidth, would require an internal network with 4 Petabits/sec of aggregate bandwidth to support full-bandwidth communication among all servers. While seemingly outlandish, the technology, both on the software [3] and hardware [4,5,6] side, is available today.

However, leveraging existing datacenter switching and interconnect technology makes it difficult and costly to realize such scale and performance. While beyond the scope of this paper to describe in detail, there are many limitations with existing technology and architectures, just to name a few: i) the number of electrical packet switches (EPS) would substantially complicate management and OpEx; ii) the cost of EPS ports and optical transceivers would dominate the overall cost of network equipment; and iii) millions of meters multimode fiber would be required, presenting a likely insurmountable deployment and operational overhead.

Optics plays a critical role in delivering on the potential of the data center network and addressing the above challenges. However, fully realizing its potential in the data center network will require a rethinking of the optical technology components traditionally used for telecom and will require optimizations targeting the specific data center deployment environments. In this paper, we present an overview of current data center network deployments, the role played by optics in this environment, and opportunities for developing variants of existing technologies specifically targeting large-scale deployment in the data center. In particular, we consider wavelength division multiplexing (WDM) technology optimized for data center deployments along with the benefits of incorporating optical circuit switching (OCS) alongside EPS in the data center [7].

2. Background: Data Center Network Architecture

We begin by exploring some of the communication and network requirements in emerging large-scale data centers. The first question is the target scale. While economies of scale suggest that data centers should be as large as possible, typically sized by the amount of power available for the site; data centers should also be distributed across the planet for fault tolerance and latency locality. The second question is the total computing and communication capacity required by a target application. Consider social networking as an example. Their sites must essentially store and replicate all user-generated content across a cluster worth of machines. The network requirements supporting such services are also significant. For each external request, many hundreds or even thousands of servers must be contacted in parallel to satisfy the request. The last question is the degree that individual servers are multiplexed across applications and properties. For instance, a portal such as Yahoo! may host hundreds of individual user-facing services along with a similar number of internal applications to support bulk data processing, index generation, ads placement, and general business support.

While no hard data is available in answering these questions, on balance we posit a trend to increasing compute densities in data centers certainly at the level of tens of thousands of servers. It is of course possible to partition individual applications to run on dedicated machines with a dedicated interconnect, resulting in smaller-scale networks. However, the incremental cost of scaling the network will ideally be modest [8] and the flexibility benefits of both shifting computation dynamically and supporting ever-larger applications are large. Hence, we consider interconnects that must roughly scale with the number of servers in the data center.

Figure 1(a) shows the architecture of typical data center networks. Individual racks house tens of servers, which connect to a top-of-rack (TOR) switch via copper links. TOR switches then connect to a core switching layer via optical transceivers, typically 10G SFP+ SR. To achieve the largest scale networks, each TOR switch would connect to all available core switches. If each TOR employs *u* uplinks, then the network as a whole can support *u* core switches. The port count *c* of each core switch then determines the total number of TORs that may be supported. If each TOR employs *d* downlinks to hosts, then the network scales to *cxd* total ports (with an *oversubscription ratio* of *d:c*). If the scale of this two-stage architecture is insufficient, then additional layers may be added to the hierarchy [5], at the cost of increased latency and larger overhead for internal network connectivity.

Figure 1(b) shows an emerging data center architecture [7,9] that employs OCS as a first-class entity. We replace some fraction of the core electrical switches with optical circuit switches. Multiple 10G SFP+ transceivers are replaced with integrated CWDM transceivers (e.g., 4x10G QSFP-LR4) to aggregate electrical channels with a common destination. While OCS cannot perform per-packet switching, it can switch more long-lived flows between aggregation points. The per-port cost of an OCS is competitive with, if not inherently cheaper than, the comparable EPS. However, it has more capacity through wavelength bundling and lower power consumption. WDM also reduces cabling complexity, a significant challenge in the data center. Finally, OCS eliminates some fraction of the optical transceivers and EPS ports by eliminating a subset of the required OEO conversions.

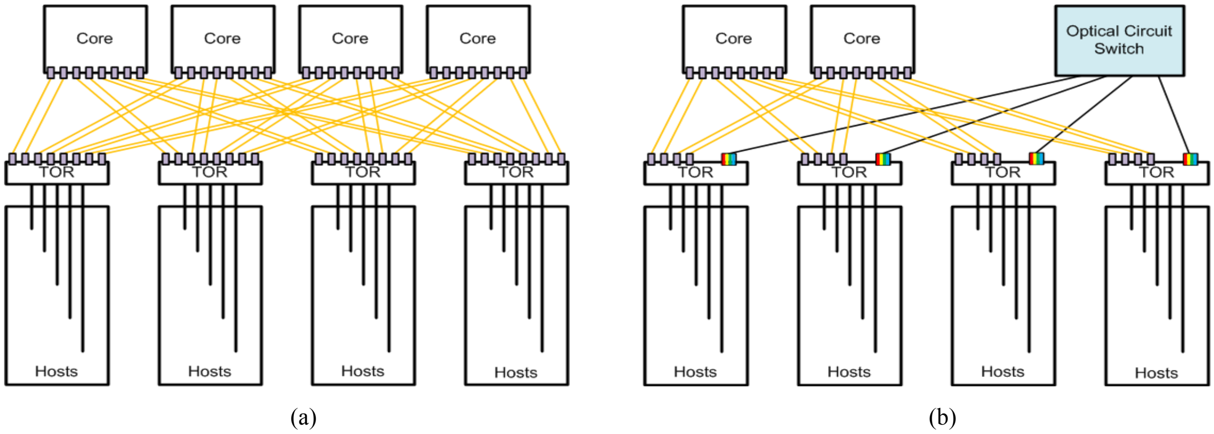

Figure 1: (a) Traditional hierarchical data center organization and (b) emerging architecture incorporating OCS and WDM.

3. Optical Circuit Switching

Native optical packet switching (OPS) has long been a goal of the optics community. However, a number of fundamental challenges leave this vision a breakthrough away from widespread commercial adoption. While awaiting such a breakthrough, OCS promises to dramatically alter the face of the data center. OCS holds a number of benefits relative to EPS. OCS is (largely) data rate agnostic and extremely energy efficient. MEMS-based OCS simply reflects light from one port to another port; so as the data rate improves as well as the number of per-port wavelengths increases, an OCS can scale without replacement. Similarly, since there is no per-packet processing, there is no added latency, and per-bit energy consumption can be orders of magnitude lower than EPS counterparts.

Data center economic, scale and performance challenges impose a number of requirements for OCS hardware:

- *Lower Cost*: The cost of integrated MEMS-based OCS is currently a barrier to entry in the data center. At the same time, the underlying chip technology is inherently inexpensive. We argue that just as commodity Ethernet switch silicon have created a large market in the data center, commodity MEMS-based OCS modules and chips will spur demand and form the basis for a variety of inexpensive network solutions.
- *Larger scale*: The largest OCS we are aware of currently supports few hundreds of duplex ports. For integration into data centers of even moderate scale, OCS must scale to thousands or even tens of thousands of ports.
- *Faster switching time*: Commercial OCS switching time is typically between 10-20 ms. Such switching times are largely driven by the requirements of the telecom industry, which only requires failover in less than 50 ms. On the other end of the spectrum, per-packet switching would require switching times measured in nanoseconds. We argue that in the data center, there are significant opportunities for large-scale OCS supporting less than 100 μs switching time.
- *Lower insertion loss*: Currently, insertion loss varies depending on the exact port pair and coupling technique used in a large-scale OCS, but goes as high as 5dB. Supporting larger-scale optical circuit

switches and integrating cost-effective optical transceivers with moderate link power budget into the data center requires driving down the insertion loss through the OCS, ideally to below 2 dB.

4. WDM Optical Transceivers

At 10 Gb/s speeds and beyond, passive and active copper cables are infeasible beyond a few meters of reach, because of their bulk, error rates, and power consumption. The emergence of cheap short-reach optics (e.g., LightPeak) changes the equation in the data center. In the next few years, we will see commodity network interface cards (NICs) with cost-effective *n*x10G optical interfaces. EPS will also have native PHY and accept 10G serial connections to further reduce cost and power.

Low-power, inexpensive VCSELs and multimode fiber (MMF) already play a critical role for communication within the data center. However, overcoming the reliability and yield hurdles to scale VCSELs significantly beyond 10 Gb/s link speed has thus far proven difficult. Further, VCSELs have limited reach, today insufficient to cross a single data center building. This maximum reach shrinks rapidly with higher data rates. Maintaining VCSEL bandwidth at 10 Gb/s means that higher speed links require VCSEL arrays, each with a dedicated MMF cable. The associated ribbon fiber and MPO connectors can incur a significant portion of the entire datacenter network cost [4]. Commodity VCSELs are intrinsically incompatible with WDM technology. Without WDM, employing VCSEL transceivers with OCS will face scaling challenges, as each data lane will consume one OCS port. Finally, MMF is typically not compatible with modern MEMS-based optical circuit switches due to its large beam size.

To reduce the cabling overhead, to scale with increasing link bandwidth, and to leverage optical circuit switching, spectrally efficient optics needs to be employed in next-generation data center transceivers [4]. However, meeting data center economies and scale requires WDM performance without an associated explosion in power and cost, as outlined below:

- *Power consumption:* Transceivers with large power consumption present thermal challenges and limit EPS chassis density. In the data center, non-retimed, un-cooled solutions are preferred. Photonic integrated circuits (PIC), low-threshold lasers with better temperature stability (e.g., quantum dot laser) and silicon photonic modulators with low switching energy hold promise for further reducing power.
- *Optical link budget:* Data center transceivers must account for multi-building span reaching 1km and optical loss from OCS and patch panels.
- *Bandwidth and speed:* Photonics highway must align seamlessly with the electrical switch fabric in bandwidth and speed. Today 10G, 4x10G LR4 and 10x10G LR10 provide cost-effective and power-efficient WDM transceiver solutions. Moving forward, we require further integration in the transceiver to align with the bandwidth and speed from EPS, with the availability of *n*x10G, *n*x20G or *n*x25G native electrical link speeds.
- *Spectral efficiency:* There will continue to be a tension between spectral efficiency, power consumption, OCS port count, path diversity and cabling complexity. For the intra-building network, a rich-mesh topology is desirable; hence, lower spectral efficiency can be traded for lower power, cheaper transceiver cost and richer network fabric. While at higher aggregation layers or the inter-building network, bandwidth is more concentrated over point-to-point links and dark fiber is expensive to procure; hence, DWDM with higher spectral efficiency is preferred.

5. Conclusions

Optics has already had a significant impact on the data center. However, we are at the cusp of a transformation of data center network architecture fueled by emerging optical technology and components. We present components of and requirements for data center networking, with a focus on the role of optical circuit switching and WDM transceivers in the data center.

6. References

[1] L. Barroso, et al, “The Datacenter as a Computer - an Introduction to the Design of Warehouse-Scale Machines,” May 2009.
[2] C. F. Lam, et al, “Fiber Optic Communication Technologies: What’s Needed for Datacenter Network Operations,” IEEE Comm. (July 2010).
[3] R. N. Mysore, et al, “PortLand: A Scalable Fault-Tolerant Layer 2 Data Center Network Fabric,” In ACM SIGCOMM’09, pp. 39-50.
[4] H. Liu, et al, “Scaling Optical Interconnects in Datacenter Networks,” in 18th IEEE Hot Interconnects (August 2010), pp. 113-116.
[5] M. Al-Fares, et al, “A Scalable, Commodity, Data Center Network Architecture,” In ACM SIGCOMM’08, pp. 63-74.
[6] P. B. Chu, et al, “MEMS: The Path to Large Optical Crossconnects,” in IEEE Comm. Magazine, (March 2002), pp. 80-87.
[7] N. Farrington, et al, “Helios: A Hybrid Electrical/Optical Switch Architecture for Modular Data Centers,” in SIGCOMM ‘10, pp. 339–350.
[8] A.Vahdat, et al, “Scale-Out Networking in the Data Center,” in IEEE Micro, (July/August 2010), pp. 29-41.
[9] G. Wang, et al, “c-Through: Part-time Optics in Data Centers,” in ACM SIGCOMM ’10, pp. 327-338.