What Devices do Data Centers Need?

Cedric F. Lam, Hong Liu, Ryohei Urata
Google, 1600 Amphitheatre Pkwy, Mountain View, CA 94043, USA
{clam, hongliu, ryohei}@google.com

Abstract: We discuss the trend in fiber optic technology developments to fulfill the scaling requirements of datacenter networks.

OCIS codes: (060.0060) Fiber Optics and Optical Communications; (250.0250) Optoelectronics

1. Introduction

Cloud computing has been driving the need for larger and larger scale datacenters [1] with higher and higher bandwidth network fabrics. As datacenter networks scale, optics is becoming ever more important from ultra-long-distance transmission between datacenters to short-reach interconnects inside datacenters. In this paper, we discuss the roles of fiber optic devices and the developing trend in these devices to fulfill the scaling requirements of datacenter networks.

2. Inter-datacenter Network Architecture

Figure 1 shows the generic architecture of a typical wide-area datacenter network. The bottom layer is a private backend backbone, which provides sparse, high-capacity and ultra-long haul point-to-point transport links between mega-scale datacenters [2]. This network transports machine generated traffic or data copies between datacenters and is not directly facing the public Internet. Above the private backend network is a transport backbone, which interconnects datacenter operators to the public Internet through peering so that users can gain access to datacenter services and facilities. It should be emphasized that the private backend backbone network is usually architecturally simple with point-to-point links, but is much larger in capacity compared with the publicly facing backbone network [3].

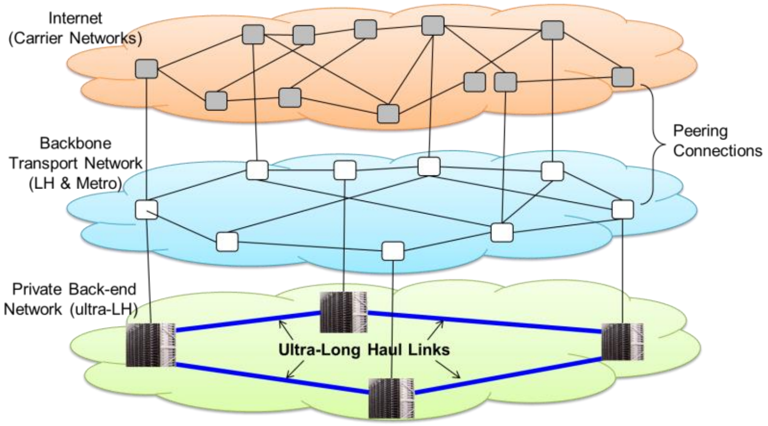

Figure 1 Layered Inter-datacenter Network Architecture

This inter-datacenter private backend network employs scarce and expensive long-distance fiber for transmission. The public facing transport network, on the other hand, contains many high-capacity metro transport links to interconnect with other carrier networks. Metro transport networks also serve to connect carrier networks to edge cache systems used by datacenter operators and content providers to improve content distribution experiences with faster access without burdening the expensive backbone transport network. Fast growing OTT (over-the-top) services such as YouTube and Netflix in the recent years are accelerating the deployment of edge cache and metro optical transport systems. In addition, hardware and software technologies that maximize the utilization of backbone fiber infrastructure resources and simplify backbone transport network operation are highly desirable. In terms of new physical technologies, spectrally efficient coherent transponders with flexible bit rate [4] will not only maximize the fiber capacity, but also simplify operation of transport networks by reducing the variety of transceivers operators have to maintain. These transponders automatically adapt the transmission rate to channel conditions and maximize the link capacity accordingly. Studies have also shown that such transponders can produce significant cost savings in a reach-diverse environment [4]. Advances in high speed electronics and integrated coherent receivers in the recent years have made these transponders a reality. Moore's law helps to continuously drive down the cost and power of long haul transponders while offering new enhanced capabilities such as soft-decision error correction codes. Other technologies that help to maximize long-haul fiber link capacities include, (1) WDM multiplexing techniques leaving no guard bands in the optical spectrum in order to maximize the system capacity, (2) Raman amplifiers to improve the optical signal to noise ratio, and (3) large effective area fiber to reduce optical non-linearity [3]. In terms of network control and management, SDN (Software Defined Network) has been demonstrated [2] to significantly improve the overall system utilization and availability through centralized traffic engineering.

3. Network Fabrics inside Datacenters

3.1 Optical Interconnects

Inside a datacenter, there are vast numbers of servers working in unison to run each application. These servers are interconnected through networking fabrics with extremely large bi-sectional bandwidth. Modern datacenters scale out using switching fabrics formed by low-cost commodity switching silicon [5, 6] and state-of-the-art high speed interconnects. Figure 2 shows the fat-tree cluster fabric topology typically deployed inside datacenters. Realization of such scale-out network clusters with large bi-sectional bandwidth requires a vast number of efficient high-speed interconnect links. Optics plays a crucial role in forming these high-speed interconnects, not only in signal transmission performance, but also in economics and operation.

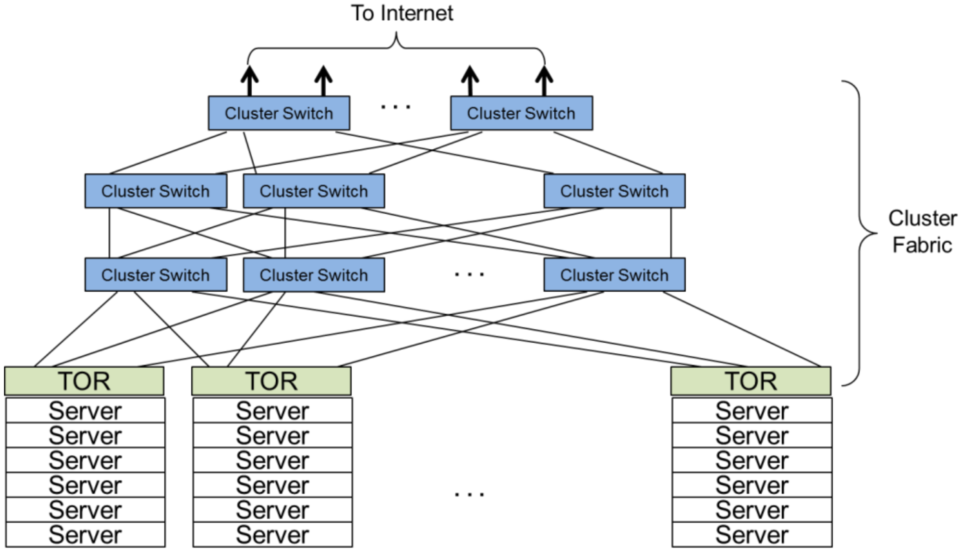

Figure 2 Intra-datacenter Network Architecture

Modern mega-data centers employee a large number of 10Gb/s interconnects from servers to switches and in between switches [5], for a distance from 2m (servers to switch connection) to 2km (for switch connections between buildings). At such speed and distance, it is difficult for copper connections to meet the performance requirements. As datacenter capacities increase, both the speed and number of interconnects need to increase. A scale-out datacenter fabric typically involves thousands of interconnect links [6]. For the same bisectional network bandwidth, the efficiency and cost of the fabric strongly depend on the port count on each switch chassis, which is in turn limited by the bandwidth of each switch silicon and front panel transceiver density. To achieve the best front panel I/O density, low-cost, low-power and low-profile optical transceivers are thus of utmost importance in scaling out datacenter infrastructures.

In order to maintain low cost and low power, it is important to not overdesign the performance of optical interconnects used inside datacenters. For short reach interconnects up to a few hundred meters at 10Gb/s, VCSEL based multi-mode optical transceivers are very low-cost and consume very low power. To achieve the 2km transmission distance requirements, single mode transceivers are preferred. Moreover, because of the large number of interconnects and much longer reaches involved in a scale-out cluster fabric implementation, the cost of optical fiber itself is significant in modern mega-scale datacenters. The cost of single mode fiber is intrinsically much lower than that of multi-mode fibers. So single-mode transceivers are becoming a trend for datacenter interconnect to not only save the cost of fibers but also improve the cabling efficiency and provide much longer reaches and future-proofing inside datacenters.

3.2 Scaling Optical Interconnects

To scale the interconnect speed beyond 10Gbps, transmitting signals in a single lane becomes exponentially more difficult. Parallel-lane transmission in the form of ribbon cable or WDM interfaces [5] help to realize higher bandwidth transmissions. The ribbon fiber approach increases the cost of interconnect cable infrastructure and makes upgrades difficult because of the necessity to install new fiber ribbons when the interconnect speed and number of transmission lanes increase. This is especially true in clusters prewired with structured cables between racks. Single mode fiber not only solves the transmission distance problem, but lends itself easily to capacity upgrade through WDM parallel lanes without introducing new fiber cables. As a matter of fact, technology advancement in the past a few years has significantly reduced the cost of optical transceivers such that the cost of fiber cables has now become a dominating part of interconnect infrastructure [7] and low-cost integrated WDM array transceivers with built-in multiplexer/demultiplexer will play a significant role to reduce the overall system cost and improve the performance and efficiency of the networking fabric.

The design considerations for short-reach WDM optical interconnect is very different from long haul WDM transmission links. The choice of wavelength plan and spacing for short-reach WDM integrated transceivers will directly impact the cost, size and power consumption of the resulting transceiver module. As an example, an uncooled solution is preferred to eliminate the thermo-electric cooler (TEC) and reduce power.

For 2km transmission distances with baud-rate less than 10Gbaud, direct modulation with on-off keying is simple, low power and cost-effective. Dispersion is usually not a limiting factor for transmission distances less than 2km. However, as link speed increases from 10G to 100G (4x25Gb/s), and 400G (16x25Gb/s, or 4x100Gb/s), direct modulation and on-off keying may no longer be the most effective way to support the transmission rate and reach. Novel modulation schemes and digital signal processing (DSP) will be needed for datacenter interconnects, also leveraging off previous work done for long haul transmission. Tradeoffs between cost, power consumption and complexity will eventually determine the most optimal transmission scheme at higher transmission speeds.

To overcome the extra insertion loss introduced by multitude of patch panels used in structured wiring inside datacenters, transceivers used in datacenter need to support higher losses than usually needed for the distances that they need to cover. At low baud rate, datacenter transceiver performances are usually loss budget limited. As baud rate scales up, dispersion penalties will no longer be ignorable and need to be considered in interconnect designs.

4. Outlook of Optics in Next Generation Systems

Innovations in photonic integration circuit (PIC) and optical packaging techniques are necessary to maintain the scalability of next generation datacenter systems. Optics has evolved from long distance communication to short reach interconnects linking servers to switches, and switches to switches in modern datacenters. As the speed of datacenter switching systems increase, signal speed will keep increasing, leading to even more challenges in integration and packaging - higher speed means higher baud rate and/or more signal lanes, as well as higher power consumptions. Optical interconnects outperform their electrical counterparts at longer distances for higher speed signals [7], and will become more & more ubiquitous as data rates increase. However, compared to electronic design and manufacturing, optics is still lagging far behind in integration and automation. The pent-up demand in bandwidth will drive innovations on integrated photonics and packaging designs. In the long run, photonic integration circuits hold the key to enhance system functionality and reduce the size and power consumptions of optical transceivers, from short reach interconnects to long-haul coherent transmission [8].

5. References

[1] L.A. Barroso & U. Hoelzle, *The datacenter as a computer – an introduction to the design of warehouse-scale machines*, Morgan & Claypool Publishers, 2009.

[2] S. Jain, et al, “B4: experience with a globally-deployed software defined WAN,” SIGCOMM 2013.

[3] C. Lam, et al., “Fiber Optic Communication Technologies: what’s needed for datacenter network operators,” *IEEE Communications Magazine*, July 2010, pp32-39.

[4] X. Zhou, “Rate-Adaptable Optics for Next Generation Long-Haul Transport Networks,” IEEE Communications Magazine, March 2013, pp41-49.

[5] H. Liu, et al., “Optical Interconnects for Scale Out Data Centers,” Chapter 2 in *Optical Interconnects for Future Datacenter Networks*, Springer, 2013.

[6] A. Vahdat, M. Al-Fares, N. Farrington, R.N. Mysore, G. Porter, S. Radhakrishna, “Scale-Out Networking in the Data Center,” IEEE Micro, 23/7/10, pp29-pp41.

[7] H Liu, et al., “Scaling optical interconnect in datacenter networks – opportunities and challenges for WDM,” *IEEE Hot Interconnect* 2010.

[8] ECOC 2013 workshop, “Low-cost access to photonic ICs 5th European photonic integration forum,” http://www.ecoc2013.org/workshop-proposals.html#ws2