Datacenter optics: requirements, technologies, and trends

(Invited Paper)

Xiang Zhou*, Hong Liu, and Ryohei Urata

Platforms, Google Inc., Mountain View, CA 94043, USA

**Corresponding author: zhoux@google.com*

Received December 7, 2016; accepted February 24, 2017; posted online March 17, 2017

We review over a decade of technology evolution and advancement of intra-datacenter optical interconnect, mainly driven by the explosive bandwidth growth of web and cloud-based services. Emerging trends and technology options to scale interface bandwidth beyond 400 Gb/s will also be discussed.

OCIS code: 060.0060.

doi: 10.3788/COL201715.120008.

Over the past decade, datacenters and their networks have become technology enablers for a number of internet-based applications. As of today, most of the popular internet applications, from the traditional search, online interactive maps, social networks, video streaming, and the Internet of Things, are running in datacenters. The pivotal role played by the datacenter will be further heightened by a wider adoption of cloud computing, where a significant portion of computing and storage is migrating into shared datacenters. This is already occurring at a rapid pace today with a number of large cloud providers leading the way. This has resulted in a dramatic increase of datacenter capability. For example, the bisection bandwidth (BW) of Google's datacenter cluster networks have increased by a factor of one thousand over the past decade[1,2].

A datacenter is essentially a massively parallel supercomputing infrastructure, consisting of clusters with several thousands of servers interconnected together in each cluster (Fig. 1). The early small-scale datacenter networks used copper-based interconnects, but as BW requirements scaled, fiber-optics-based optical interconnects were introduced and quickly established as the most cost-effective, deployable solution to scale out the datacenter. As of today, high BW and low power optical interconnects have been ubiquitously used in datacenters for any link distance beyond a few meters. This Review intends to review over a decade of technology evolution of datacenter optics. Emerging trends and technology options to scale BWs beyond 400 Gb/s will also be discussed.

For a typical intra-datacenter network adopting a Clos topology (Fig. 1), a massive number of interconnection links are required to implement the large fan out and corresponding high bisection BW[1]. Thus, the number one consideration for interconnect is the BW cost. To minimize the total interconnection cost, different technologies are adopted at different segments of the network. For example, a backplane printed circuit board (PCB) and copper interconnects are typically used for intra-rack interconnection for a reach of less than a few meters, while fiber-based optical interconnects are used for interconnection between the top of rack (TOR) switch and the edge switch, as well as the between the edge aggregation switch and the spine switch, with a link distance ranging from a few meters up to 2 km. For a link distance of less than 100 m, vertical cavity surface emitting laser (VCSEL) and multimode fiber (MMF)-based technologies have proven to give the best overall link cost (transceiver cost plus fiber cost). Beyond 100 m, however, more expensive single mode fiber (SMF) transmission technologies usually have to be chosen due to the following reasons: (1) the BW of commercially available VCSELs have been limited (about 20 GHz as of today), (2) the BW of an MMF reduces as the distance increases [for example, the BW is limited to be about 20 GHz for 100 m of optical multimode (OM3) fiber], and (3) the higher cost of an MMF, although the cost of a VCSEL transceiver is significantly lower than an edge emitting laser-based SMF transceiver[3].

The second important criteriom is power consumption. From an aggregate energy consumption point of view, the power consumption of networking is only a modest piece of the power consumed by a datacenter (less than 10%[4]). But, the power efficiency of optical transceivers is essential for the front panel density (the allowable transceiver size is largely determined by its power envelope). Hence, without power efficient transceivers, there is no optimal way to take advantage of the full capacity of the switch application-specific integrated circuit (ASIC).

The third important criterion is serviceability. Since the reliability of the typically used active optical components are not very high, it is better to design the optical transceiver in a way that it can be easily replaced or serviced. In this regard, pluggable optics is preferred over on-board optics[5], although on-board optics enables a higher front panel density. Due to the need of serviceability, we expect pluggable optics will continue to be the mainstream switch-to-switch optical transceiver choice for the

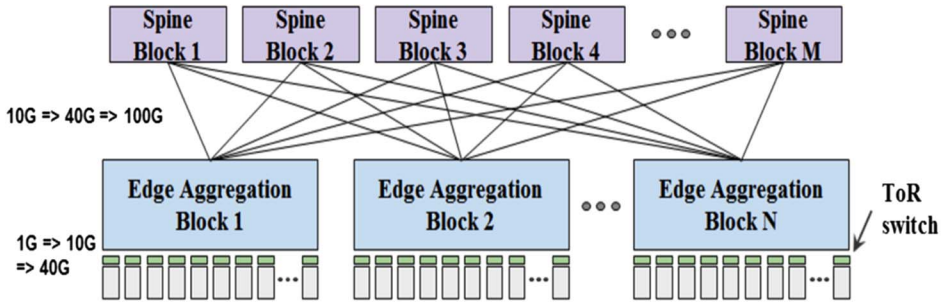

Fig. 1. Google intra-datacenter network.

foreseeable future. An additional advantage for using pluggable optics is that it allows us to optimize the cost for different reaches (e.g., copper for a few meters, MMF for <100 m, and SMF for >100 m). Finally, for datacenter optics, cabling efficiency as well as transmission latency also needs to be considered.

Back in 2004, our first intra-datacenter network design adopted a four-post cluster architecture implemented with commercial switches and a copper-based interconnect. As BW h requirements scaled, a large Clos network architecture was introduced using custom switch hardware and software. To improve the network BW and the overall interconnection cost, in 2007, optical interconnects were deployed at a large scale in Google datacenters, which was an industry first. The viability in cost of 10 Gb/s VCSEL/MMF-based interconnects [with small-form-factor-pluggable (SFP) transceivers] tipped the scales away from copper interconnects, which were bulky and too difficult to deploy at the number and length scales required. In 2012, we achieved the first large-scale deployment of SMF-based interconnects in the datacenter with a 40 GbE quad-SFP (QSFP) for Jupiter Networks, capable of up to 1.3 Pb/s of bisection BW[1]. Specifically, we used 4×10 Gb/s of uncooled coarse wavelength division multiplexing (CWDM) technology with direct modulated lasers.

Figure 2 illustrates datacenter optics technology evolution over time. The first generation of datacenter optics operates at 10 Gb/s, using two-level non-return-to-zero (NRZ), direct detection, and a single wavelength (850 nm for an MMF and 1300 nm for an SMF). The second generation, which operates at 40 Gb/s using a QSFP form factor, was enabled by scaling the optical lanes: space division multiplexing (SDM) for VCSEL/MMF transceivers and uncooled CWDM technology for SMF transceivers. The current generation (i.e., the third generation) operates at 100 Gb/s using a QSFP28 form factor. The higher BW is enabled by scaling the lane speed from 10 to 25 Gb/s. Such a lane speed increase is made possible thanks to the advancement of high-speed direct optical modulation technologies as well as the progress made in high-speed electrical processing technologies. Note that simple pulse amplitude modulation (PAM2) and direct detection techniques were used in all three generations. For the next generation 400 G, a more BW-efficient higher-order modulation format PAM4 has been introduced to scale lane speed from 25 to 50 Gb/s. By increasing the baud rate from 25 to 50 Gbaud, single lane 100 Gb/s could also be achieved for SMF transceivers by using higher BW optical components (external modulators, photodetector). But, it will be challenging to scale to 100 Gb/s per lane for VCSEL/MMF (a major technology advancement is needed to increase VCSEL BW to 30 GHz or beyond without compromising the required reliability and manufacturability).

In Fig. 3(a), we show the front panel BW using 10, 40, and 100 G pluggable optics technology. For comparison, the switch input/output (I/O) BW trend line is also

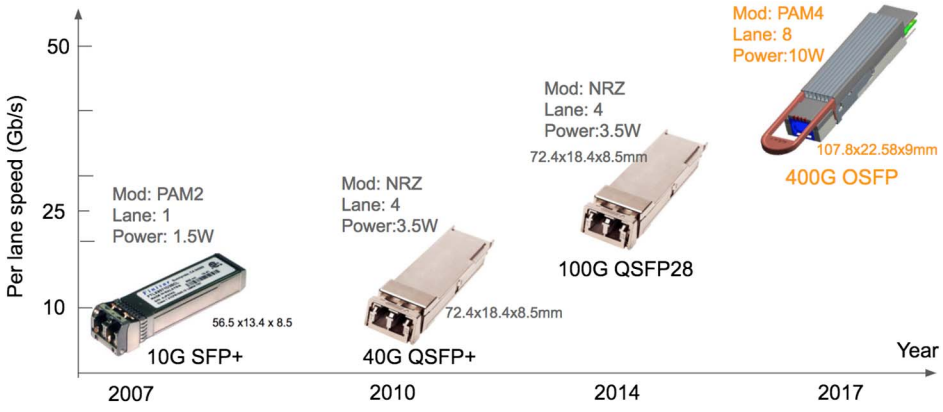

Fig. 2. Datacenter optical transceiver evolution.

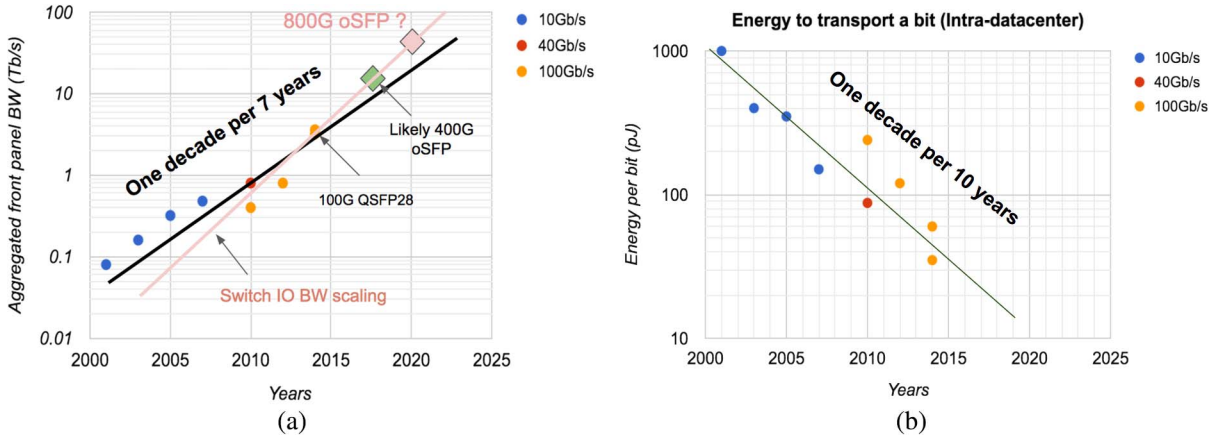

Fig. 3. Historical view on the (a) pluggable optics front panel BW and (b) energy efficiency scaling.

drawn in the figure. Historically, a pluggable optics front panel BW density improves about one decade per seven years. Such an improvement rate is slightly slower than the switch capacity growth rate, which is about one decade per five years. Until 100 G, there has been no scaling problem as the switch capacity has been smaller than what can be supported by the pluggable optics density. Beyond 100 G, however, pluggable optics will need to scale faster to keep up with switch capacity growth: 400 G at octal SFP (OSFP)[6] or a similar form factor may be needed by the 2017/2018 time frame, while 800 Gb/s at the same form factor may be needed by 2020/2021. In Fig. 3(b), we show the corresponding energy efficiency improvement over time and technologies. The energy efficiency improves about one decade per 10 years, which is slower than the actual BW density improvement. Better thermal designs for the pluggable optics was used to sustain the faster BW density improvement in the past, and we expect that more aggressive thermal management is needed to meet the more stringent BW density requirements for 400 Gb/s and beyond.

Fundamentally there exists three degrees of design freedom to allow us to scale interconnect BW as is illustrated in Fig. 4: (1) increasing the symbol rate per lane, (2) increasing the number of parallel optical lanes, and (3) encoding more bits into each symbol (i.e., higher-order modulation formats). Each of the three orthogonal

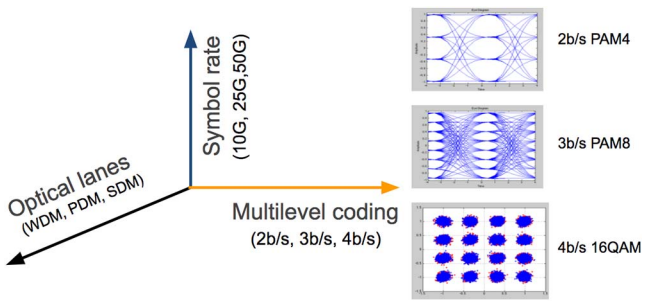

Fig. 4. Technology options to scale the BW. WDM, wavelength division multiplexing; PDM, polarization division multiplexing.

technology choices has its advantages and constraints. Scaling in the symbol rate axis is the most cost-effective method, but its potential is limited by electrical I/O capability as well as the achievable optical component BW. For example, a 16 nm CMOS is needed to enable power efficient 25 Gbaud performance, and a 7 nm CMOS is needed to enable an efficient 50 Gbaud electrical performance. For optical components, the commercially available lower cost directly modulated laser (DML) BW is limited to about 20–25 GHz, and the more expensive external modulated laser (EML) or Mach–Zehnder modulator (MZM) is limited to about 30 to 40 GHz. A 7 nm CMOS could enable power efficient 50 Gbaud systems (100 Gb/s per lane with PAM4) with EML/MZM, but a major innovation in optical modulator technology is needed to scale beyond 50 Gbaud.

Scaling BW using the parallel channel axis is very effective in terms of increasing the aggregate data rate per interface, but the downside is that the required number of optical and electrical components increases linearly with the number of optical lanes. Furthermore, the use of an increased number of optical components, especially the active optical components, will impact the total yield and cost. The use of mature photonic integration technology may help alleviate this problem. Tighter optoelectrical integration and/or packaging are also critical for reducing the total power (otherwise the worst case power increases linearly with the number of lanes). Scaling in space is undesirable due to a higher fiber cost and volume.

Finally, encoding more bits into each symbol allows us to scale the serial bit rate without imposing higher component BW requirements, but such a lower BW requirement is achieved at the expense of the signal-to-noise ratio (SNR) as well as inter-symbol interference (ISI) and other channel impairments (such as various optical and electrical interferences). For example, as compared to PAM2, PAM4 requires about a 9.4 dB higher SNR than PAM2 (assuming it is operating at the same baud rate), and it is about 9 dB less tolerant toward optical multipath interference (assuming it is operating at an identical bit error

rate)[7]. To reduce the ISI penalty, more advanced digital equalization may be needed. Higher coding gain forward error correction (FEC) can be used to compensate or partly compensate for the increased SNR requirement. However, management of optical interference can be more challenging with direct intensity modulation formats such as PAM4 and PAM8[8].

Coherent modulation and detection techniques allow us to encode more bits into each symbol and have enjoyed great success in long-distance optical networks. There has been enormous reduction in cost, power, and density on the coherent technologies over the last decade[2]. If we further optimize the coherent system for datacenter reach by (1) using baud-rate sampling digital signal processing (DSP)[9], (2) shedding all unnecessary DSP functions, such as high coding gain soft decision FEC, excessive chromatic dispersion, and polarization mode dispersion compensation, and (3) trading the higher receiver sensitivity of a coherent receiver for even higher-order modulation format (to lower the sampling rate), the required coherent ASIC power may eventually drop to a level that can be considered for intra-datacenter applications at certain CMOS nodes (likely 7 nm or below).

Compared to direct detection, coherent detection offers the following advantages: (1) higher spectral efficiency, which could enable single wavelength 400 Gb/s (two wavelength 800 G or beyond) using common components of about 30 GHz BW; (2) better receiver sensitivity with the potential to support a larger link loss; and (3) more tolerance toward optical impairments[7]. But, in addition to higher DSP power, the coherent technology also requires frequency-stable and narrow linewidth lasers as well as a more complex (and lossy) I/Q modulator and receiver front end[5].

To better understand the potential benefit of the coherent technique, in Fig. 5 we give a simulation study on the supported link loss under various modulator driving conditions for two exemplary 400 G designs; one using the single wavelength coherent polarization-multiplexed (PM) 16 quadrature amplitude modulation (QAM), and the other one using PAM4 with four CWDM wavelengths. The first-order estimation of the required relative driving power under different drive swings is also given in this figure. For this study, we have assumed a moderate hard-decision FEC code with a raw coding gain of 9.2 dB as well a set of realistic optical/electrical component parameters. From Fig. 5, one can see that coherent technology can support more link loss when the modulator drive swing is more than 0.7 Vpi. With a full 2 Vpi drive swing, the coherent technology can support about 5 dB more link loss than using the direct intensity modulation technique. However, increasing the driving swing from a moderate 0.5 to 2 Vpi will increase the power consumption by a factor of about 16. So the development of a highly power efficient (with ultra-low Vpi) and low-cost modulator technology is one of the key innovations needed for moving coherent technology into the datacenter.

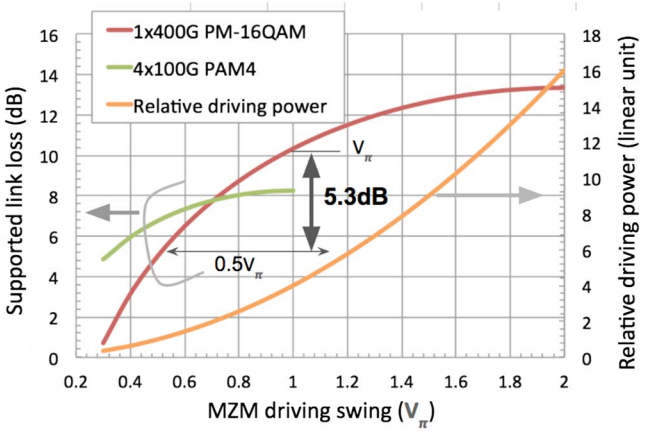

Fig. 5. Impact of modulator drive swing on supported link loss and relative power consumption for two 400 G designs: 4 × 100 Gb/s PAM4 and 1 × 400 Gb/s PM-16QAM.

Cost, energy efficiency, and serviceability are the three most important performance metrics for datacenter optics. Until now, pluggable optics were still the preferred choice to meet these requirements. From 10 to 100 G, BW density scaling was mainly achieved by Moore's Law and the advancement in optical component technologies as well as the use of parallel optics. More advanced multilevel direct modulation enables us to scale BWs to 400 Gb/s but, its high SNR requirement limits its scaling potential. More advanced coherent modulation could enable single wavelength 400 Gb/s and two wavelength 800 Gb/s or beyond, but this technique requires highly efficient optical modulators and application optimized low-power DSP to scale the power. In addition to energy-efficient high BW (and low cost) optical modulation, much tighter optoelectronic integration and/or packaging are also critical for scaling datacenter optics beyond 400 Gb/s.

References

1. A. Singh, J. Ong, A. Agarwal, G. Anderson, A. Armistead, R. Bannon, S. Boving, G. Desai, B. Felderman, P. Germano, A. Kanagala, H. Liu, J. Provost, J. Simmons, E. Tanda, J. Wanderer, U. Hölzle, S. Stuart, and A. Vahdat, Commun. ACM **59**, 88 (2015).
2. R. Urata, H. Liu, X. Zhou, and A. Vahdat, in *OFC* (2017).
3. H. Liu, C. F. Lam, and C. Johnson, in *2010 18th IEEE Symposium on High Performance Interconnects* (2010).
4. L. Barroso and U. Hölzle, *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines* (Morgan & Claypool, 2009).
5. X. Zhou and H. Liu, in *Frontier in Optics* (2016), paper FTh5E.1.
6. http://osfpmsa.org/.
7. F. Zhu, Y. Wen, and Y. Ba, in *IEEE 802.3bs 400 GbE Task Force Plenary Meeting*, San Diego, CA, July 14–18, 2014 (2014).
8. A. Farhood, V. Bhatt, B. Smith, and S. Anderson, "Improved MPI upper bound analysis," http://www.ieee802.org/3/bm/public/nov12/.
9. X. Zhou and H. Liu, in *ECOC* (2016), paper WS3.