State-of-the-Art 800G/1.6T Datacom Interconnects and Outlook for 3.2T

Xiang Zhou, Cedric F. Lam, Ryohei Urata, and Hong Liu

Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043, USA, zhoux@google.com

Abstract: We review state-of-the-art datacenter technologies for 800G, 1.6T and beyond interconnect speeds, focusing on 200G per-lane IM-DD (intensity modulated-direct detect) and 800G-LR1 coherent-lite transmissions.

OCIS codes: (060.2330) Fiber optics communications (060.4080) Modulation

1. Introduction

As of today, most of the popular Internet applications, from search, online interactive maps, social networks, to public clouds, are running in hyperscale datacenter (DC) infrastructures, which typically consists of tens of thousands of compute/storage/accelerator nodes interconnected through electrical packet switch (EPS)-only networking architectures or a hybrid EPS- and optical circuit switch (OCS)-based network [1]. A hyperscale DC campus network may span multiple datacenter buildings, each comprising one or more of these compute clusters.

The evolution of the DC optical interconnect technology for these networks is mainly driven by the need to match the switch electrical I/O speed while improving cost, power efficiency, bandwidth and density. As described in [2], five generations of optical interconnect technologies have been developed to meet Google's ever-growing DC network bandwidth demands, from the first-gen 10G SFP+ to the latest 800G OSFP, with bandwidth increased by a factor of 80, energy efficiency improved by a factor of 6, and linear density improved by a factor of 24. Through this evolution, the intra-DC (<1km) and campus interconnect (<10km) have leveraged the same technology, to benefit from shared volumes and economies of scale.

This paper intends to provide an overview on the state-of-the-art for DC interconnect technologies beyond 100G per wavelength, aiming for cost efficient 800G, next-gen 1.6T, and future 3.2T connection bandwidth scaling. This includes discussion on potential bifurcation of the intra-DC and campus interconnect roadmap, due to the growing challenge of physical impairments at high data rates and increasing campus reach.

2. State-of-the-Art 800G/1.6T: 200G Lane IM-DD Technology

Fundamentally, there are three orthogonal axes to scale bandwidth: 1) higher baud rate; 2) more spectrally-efficient modulation formats; and 3) more parallel lanes (both spatial and wavelength). These three axes have all been used to scale the bandwidth, from 10G using 10Gbaud, PAM2 and a single lane, to 800G using 50Gbaud, PAM4 and 8 lanes. Bandwidth cost ($/bit) reduction was achieved by serial data rate scaling through 1) and/or 2) without increasing optical/electrical component counts. To further improve DC bandwidth cost and power efficiency, scaling the serial rate to 200G by doubling the baud rate using the same PAM4 modulation format is thus a very attractive solution.

Doubling baud rate requires higher bandwidth components. From a 200G-lane components readiness survey conducted in 2021 [3], all the critical optical and electrical components for 200G PAM4 should be ready in 2023. For example, InP EML (externally modulated laser), InP and Silicon Photonic (SiP) waveguide photodetector (PD) can achieve a 3dB bandwidth >55GHz, while the CMOS DAC/ADC (digital to analog/analog to digital converter) achieve ~55GHz bandwidth with effective number of bits (ENOB) >5.5 using 5 nm CMOS technology. With such component bandwidths, low-power linear equalization techniques would be sufficient.

Doubling baud rate, however, worsens the receiver power sensitivity due to increase of the receiver noise bandwidth. Additionally, optical channel impairments, such as fiber chromatic dispersion (CD), polarization mode dispersion (PMD) and four wave mixing (FWM) nonlinear effects become limiting factors. For Intra-DC use cases where the reach is typically less than 1km, the link budget loss due to reduced receiver sensitivity and optical channel impairments could be compensated by introducing higher-gain and low-latency concatenated FEC [4] and moderately increasing the laser power. Nevertheless, for some campus use cases where the reach can be up to 10km, managing impairments from fiber CD, PMD [6,7] and FWM [8] becomes much more challenging.

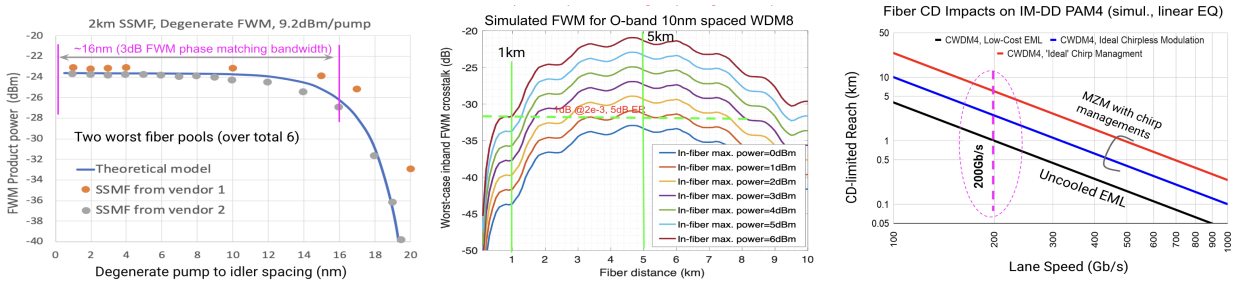
Four-wave Mixing (FWM): System impacts of FWM depends on WDM wavelength grid, maximum per channel launch power and fiber reach. Fig.1a shows the modeled 'worst-case' O-band degenerate FWM phase matching bandwidth and its experimental verification. Fig. 1b shows the simulated FWM crosstalk versus fiber distance under

different per channel launch power for a 1.6T 10 nm-spaced WDM8 system [8]. At a 1dB penalty for BER@2e-3, the allowable maximum Tx power is reduced from ~6dBm/ch at 1km to ~1dBm/ch at 5km reach (green dashed line in figure).

Chromatic Dispersion (CD): System penalty from fiber CD scales as the square of the baud rate. Fig. 1c shows a simulated study on CD-limited reach vs baud rate, defined at 1dB penalty using KP4 FEC and linear equalization. At 200G lane speed, the supported reach would be less than 1km when using low-cost uncooled EMLs for 800G CWDM4. With a mach zehnder modulator (MZM) and ideal chirp management, the supported reach can be extended to 6km (dashed pink line in figure). The impact of fiber CD can be reduced with tighter channel spacing, for example by using 800GHz-spaced LAN-WDM4 for 800G total data rate, or 400GHz-spaced DWDM8 for 1.6T speed. But tighter channel spacing increases FWM impairment. For 800G LAN-WDM4, polarization management with XYYX arrangement could help mitigate FWM effects [9], but this method can not scale to 1.6T DWDM8.

Polarization Mode Dispersion (PMD): System penalty from PMD scales linearly with the baud rate. From a recent study [6], a worst-case 10km IEEE PMD spec (5ps differential group delay) will introduce a ~3.4dB penalty for BER@2e-3 by using low-power linear equalization technology. PMD penalty can be reduced to 0.7dB with MLSE based nonlinear equalization[7].

Considering all the above impairments, with the requirement of CWDM4 compatibility, the 200G per lane IM-DD solution would be restricted to a 6km reach.



(a) (b) (c)

Fig. 1. Fiber FWM and CD impacts on 200G per lane and beyond IM-DD systems, where the in-fiber max power in Fig. c denotes per ch power

3. State-of-the-Art 800G/1.6T: Coherent Options

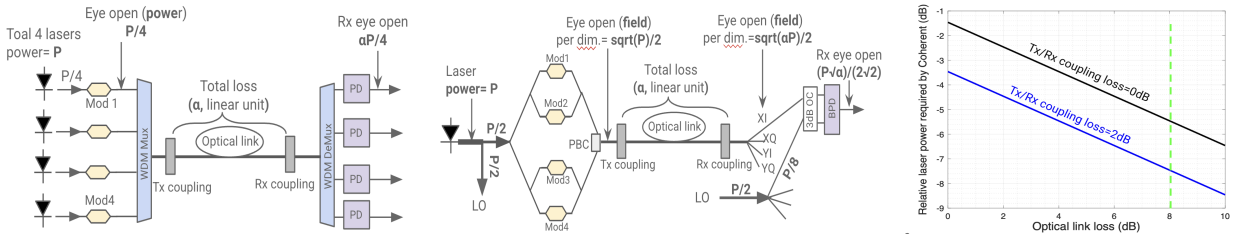
800G-LR1 based coherent-lite technology [5] could be a more efficient solution (than LAN-WDM4) to support campus reach >6km. Unlike the traditional digital coherent technology optimized for long haul and metro, coherent-lite greatly simplifies both optical and DSP designs. It uses low-cost DFB lasers with fixed wavelength and wider linewidth. Compared to the 4×200G LAN-WDM4, 800G-LR1 requires fewer lasers (4 vs 1). 800G-LR1 also enables higher optical link budget (without using optical amplifier) due to two fundamental reasons: 1) Coherent allows bipolar modulation, which effectively doubles the constellation Euclidean distance (in the field, under the same laser power as compared to IM-DD); and 2) the LO, which is much greater than the received signal, acts (equivalently) as a pump to amplify the received small signal. Despite the benefits of coherent optics, it still faces the challenges in modulator efficiency and DSP optimization.

Modulator efficiency: A coherent system benefits more from the improvement of modulator efficiency than an IM-DD system. For example, if we keep the MZM modulator drive swing a constant $2.5V_{dd}$, improving the MZM Vpi from 5V to 2.5V only results in 1.5dB link budget gain for a IM-DD system, but will result in a 5.3dB link budget gain for a coherent system. If we can further improve the MZM Vpi to 1.25V, then the coherent system can gain an additional 3dB link budget, while there is no additional link budget gain for the IM-DD.

Coherent DSP optimization: 800G-LR1 DSP power can be greatly reduced by 1) moving the wavelength band from the traditional C-band to the O-band wavelength [10], and 2) adopting baud-rate sampling and equalization technology [2]. By moving the wavelength band from the C-band to O-band, the worst-case fiber CD can be reduced from ~200ps/nm to ~13ps/nm, allowing removal of the dedicated and power-hungry CD compensation DSP. For 10km campus reach, fiber PMD is also much smaller than the Metro/LH, thus the 2×2 MIMO equalizer could be greatly simplified. With much reduced fiber CD and PMD, oversampling, as is needed for Metro/LH, may no longer be needed for <10km campus reach. This will further reduce 800G-LR1 DSP power. The continued advancement of CMOS technology will also help lower coherent DSP power.

Fundamental performance gain: Fig. 2 shows a fundamental performance comparison between a 4-lane IM-DD system and a coherent system by assuming ideal components. We assume a moderate LO power so TIA thermal noise dominates in both systems, and the system performance is characterized by the received vertical eye opening

under identical total laser power *P* (see Fig. 2a and b). From the relative eye openings analysis, to support the same total optical loss α (in linear unit), a coherent system can reduce the required laser power by a factor $\sqrt{2/\alpha}$. As can be seen from Fig. 2c, to support 8dB campus link loss, and assume practical 2dB Tx/Rx coupling loss, the coherent technology can reduce the required laser power by about 7.5dB.



a. 4-lane IM-DD system b. 4-dimensional coherent system c. Relative laser power by coherent

Figure 2. Fundamental performance comparison between a 4-lane IM-DD system (a) and a four-dimensional coherent system (b), in terms of relative laser power requirement to support the same optical link loss (c), where BPD denotes balanced PD

4. Avenues Toward 3.2T

Scaling to 3.2T needs to balance power efficiency, cost and performance. Continual advancement in photonic integration, high baud rate and low-power coherent optics technology will be critical for 3.2Tb/s bandwidth scaling.

Photonic Integration: with low-cost monolithic or hybrid silicon photonic integration technology, 16 lanes of 200G could be one of the options to achieve 3.2T interface, but major technology innovation is needed to improve modulator efficiency and wavelength Mux/DeMux performance.

High baud rate: serial data rate scaling is the most cost-effective technical solution to scale the bandwidth. To achieve 3.2T interface rate, doubling per lane speed from 200G to 400G by using the simple IM-DD technology could still be preferable. From Fig. 1c, by using MZM with appropriate chirp-management, the penalty from fiber CD could be manageable up to 1km reach. To enable 400G per lane, innovations are required to improve the bandwidth of critical optical components (modulator, PD) and electrical components (driver, the TIA, and CMOS ADC/DAC). Tight integration of electrics and optics is also critical for high speed performance. The required bandwidth is about ~110GHz for PAM4, but can be reduced to ~90GHz with PAM6. More powerful FEC and interference management technology [12] could be used for PAM6 to compensate for SNR penalty and mitigate higher optical interference such as MPI [2].

Coherent technology: beyond 1km, coherent optics may have to be considered. Conventional 800G per wavelength coherent-lite could be a potential candidate. The main advantages include the relatively low components and packaging bandwidth requirements (similar to 200G lane IM-DD), and high tolerance towards major optical channel impairments such as the fiber CD, PMD, FWM and inband optical interferences. If ~110GHz optical and electrical components bandwidth is achievable, 1.6T per wavelength coherent-lite could be a lower cost and lower power 3.2T solution to support beyond 1km reach.

5. Conclusions

For Intra-DC use cases, 200G lane IM-DD is the preferable 800G/1.6T solution. If components and E/O interface bandwidths can scale, 400G lane IM-DD could still be the preferable 3.2T solution for <1km reach. For 10km campus use cases, however, a coherent solution may have to be considered starting from 800G. Innovations in photonic integration, high baud rate and low-power coherent optics technology will be required to enable 3.2T bandwidth scaling.

References

[1] L. Poutievski, et al, "Jupiter Evolving: Transforming Google's Datacenter Network via Optical Circuit Switches and Software-Defined Networking, SIGCOMM 22, August 22-26, 2022, Amsterdam, Netherlands

[2] X. Zhou, R. Urata, and H. Liu, "Beyond 1Tb/s Intra-Data Center Interconnect Technology: IM-DD OR Coherent?" J Lightwave Technol., Vol. 38, No. 2, pp. 475 – 484

[3] C.Lam, X. Zhou, and H. Liu, "200G per Lane for beyond 400GbE, IEEE 802.3 B400G SG Meeting, March 2021

[4] S. Yin, X. Zhou, and C. Lam, 'FEC Requirements for 800GbE/1.6TbE Optics,' IEEE802.3df, July 2022

[5] X. Zhou and C. Lam, "800G LR1 Use Cases and Requirements Revisit ", OIF2021.369.02, 2021

[6] H. Zhang and P. Liao, "On PMD tolerance of 200 Gb/s PAM4," IEEE802.3df, July 2022

[7] M. Kuschnerov, N. Stojanovic and Y. Lin, "800G LR4 DGD penalty and fiber specifications," IEEE802.3df, July 2022

[8] X. Zhou and C. Lam, "Four-Wave Mixing Penalty for WDM-based Ethernet PMDs in O-band," IEEE802.3df, May 2022

[9] X. Liu, Q. Fan, T. Gui, and K. Huang, "Effective suppression of inter-channel FWM for 800G-LR4 and 1.6T-LR8 based on 200Gb/s PAM4 channels", EEE802.3df, July 2022

[10] X. Zhou and C. Lam, "800G LR1 Wavelength Consideration: O-band or C-band ?" OIF2022.364.0, August 2022

[11] T. Gui, L. Li, and X. Liu, "Consideration on the optical Spec for 800LR," OIF2022.471.00, Nov 2022

[12] X. Zhou, R. Urata, E. Mao, H. Liu, C. L. Johnson, "In-band optical interference mitigation methods for direct detection optical communication systems", US patent US10084547B2, https://patents.google.com/patent/US10084547B2/en, Jan 2016