Beyond 1 Tb/s Intra-Data Center Interconnect Technology: IM-DD OR Coherent?

Xiang Zhou, *Fellow, OSA*, Ryohei Urata, and Hong Liu, *Fellow, OSA*

(Invited Paper)

***Abstract*—We discuss technology options and challenges for scaling intra-datacenter interconnects beyond 1 Tb/s bandwidths, with focus on two possible approaches: pulse amplitude modulation (PAM)-based intensity modulation-direct detection (IM-DD) and baud-rate sampled coherent technology. In our studies, we compare the performance of various orders of PAM modulation (PAM4 to 8). In addition to these fixed PAM signaling options, a flexible PAM (FlexPAM) technique leveraging granularity in spectral efficiency (SE) is proposed to maximize link margin. For baud-rate sampled coherent technology, we propose a simplified digital signal processing (DSP) architecture to bring down power consumption of the coherent approach closer to that of IM-DD PAM. We also propose two new phase noise tolerant 2D coherent modulation formats to relax the laser linewidth requirement. In closing, a comparative study of fixed IM-DD PAM versus coherent polarization multiplexed-quadrature amplitude modulation (PM-QAM) is presented for a 1.6 Tb/s solution (200 Gb/s per dimension), with consideration of link loss/reach budget, power consumption, implementation complexity, as well as fan-out granularity.**

***Index Terms*—Coherent detection, coherent modulation, datacenter, direct detection, DSP, fiber, FlexPAM, IM-DD, interconnect, modulation format, optical, PAM, QAM.**

I. Introduction

IN THE past decade, datacenters (DCs) have become the key technology enabler for internet-based applications. Most of the popular Internet applications today are running in DC infrastructure - from search, online interactive maps, social networking, video streaming, to the Internet of Things (IoT). The pivotal role of the DC will be further enhanced by wider adoption of cloud computing, wherein a significant portion of compute and storage is migrated to shared DCs. The growth in application diversity and functionality can be directly tied to corresponding increases in DC capabilities (improved search, language translation with machine learning (ML) as recent examples). For intra-DC networking, the bisection bandwidth of Google's DC networks has increased by a factor of one thousand over the past decade [1]. Furthermore, the fast adoption of ML-based applications not only fuels traditional DC network bandwidth growth, but also drives the need for new network topologies with very high bandwidths [2].

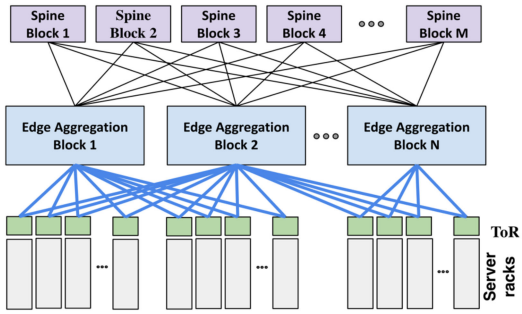

Fig. 1. Schematic illustration of intra-DC network fabric and interconnects (source figure refers to [10, Fig. 1b]).

Fiber-optics-based optical interconnects have become the technology of choice to scale out DC networks, covering link distances from a few meters to thousands of kilometers. However, technology requirements for intra-DC optical interconnects are quite different from traditional long-distance telecommunication transport systems, where achieving higher per fiber capacity on scarce fiber resources is more critical.

Intra-DC interconnects have much shorter reach (typically $<$2 km) with a large number of connections (Fig. 1), and thus their cost is largely dominated by the transceivers. Besides the fanout/rich connectivity needed to realize a scaled-out Clos fabric [1], intra-DC optics have more stringent requirements on power consumption, density, and cost due to their sheer volume [1]–[10]. Good serviceability, cabling efficiency, and low latency are also important metrics which must be considered.

Given the above constraints, scaling the interconnect (interface) bandwidth efficiently from 400 Gb/s per fiber to beyond 1 Tb/s per fiber will be a challenge. This paper reviews and discusses these challenges and potential technical solutions, with a special focus on PAM (Pulse Amplitude Modulation)-based direct-detection (DD) technology and a baud-rate-sampling-based coherent detection technology.

The remainder of this paper is organized as follows. In Section II, we review the evolution of DC optics technology, as well as the challenges for beyond 1 Tb/s bandwidth scaling. In Section III, we introduce the concept of FlexPAM, and its potential advantages for achieving 200 Gb/s per lane (1.6 Tb/s with 8 lanes) bandwidth scaling. Section IV discusses a low-power

Manuscript received June 9, 2019; revised September 11, 2019 and November 7, 2019; accepted November 21, 2019. Date of publication November 29, 2019; date of current version January 23, 2020. *(Corresponding author: Xiang Zhou.)*

The authors are with Google Inc., Mountain View, CA 94043 USA (e-mail: zhoux@google.com; ryohei@google.com; hongliu@google.com).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JLT.2019.2956779

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

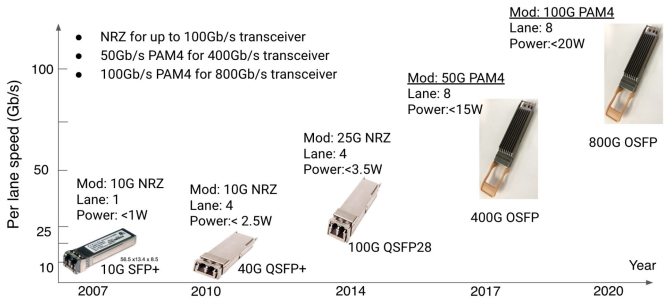

Fig. 2. Google DC optics technology evolution. NRZ: non-return-to-zero.

TABLE I
SUMMARY OF FIVE GENERATIONS OF GOOGLE DC OPTICS

	Gen 1	Gen 2/3	Gen 4	Gen 5
Interface (Gb/s)	10	40/100	400	800
Form factor	SFP+	QSFP	OSFP	OSFP
Modulation	PAM2	PAM2	PAM4	PAM4
Baud rate (GHz)	10	10/25	25	50
Lane number	1	4	8	8
Mux type for SR	N/A	SDM	SDM	SDM
Mux type for LR	N/A	CWDM	CWDM	CWDM
Laser/fiber for SR	VCSEL MMF	VCSEL MMF	VCSEL MMF	EML/ SMF
Laser/fiber for LR	N/A	DML/SMF	EML/SMF	EML/SMF

baud-rate sampling coherent digital signal processing (DSP) technology. In Section V, we introduce two new phase noise tolerant coherent modulation formats, specifically optimized for shorter, intra-DC links. Section VI presents a comparative study between PAM-based IM-DD technology and baud-rate-sampling-based coherent technology for 200 Gb/s per dimension bandwidth scaling. We conclude in Section VII.

II. TECHNOLOGY EVOLUTION AND SCALING CHALLENGES

Fig. 2 and Table I summarize the evolution of Google's intra-DC optical interconnect technology, which has been driven by the need to match the switch ASIC (application-specific integrated circuit) electrical I/O (input/output) speed [1] while improving cost, power, and density. The first generation operated at 10 Gb/s SFP+ (Small Form-Factor Pluggable) using PAM2 modulation, direct detection, and a single wavelength. The second generation 40 Gb/s QSFP (Quad Small Form-Factor Pluggable), was achieved by scaling the optical lanes to four (10 Gb/s per lane). Space division multiplexing (SDM) with vertical cavity surface emission lasers (VCSELs) and multiple-mode fiber (MMF) technology was used for $<$100 m, short reach (SR) applications. Coarse wavelength-division multiplexing (CWDM) with uncooled directly modulated lasers (DMLs) and single mode fiber (SMF) technology was used for $<$2 km, longer reach (LR) applications. The third generation 100 Gb/s QSFP28 scaled the lane speed to 25 Gb/s while the number of lanes remained at 4.

The fourth generation 400 Gb/s OSFP (Octal Small Form-Factor Pluggable), employs more bandwidth-efficient PAM4 and doubles the optical lanes from 4 to 8 and data rate from 25 Gbits/s to 50 Gbits/s, while the baud rate remains at 25 Gbaud/s (excluding FEC overhead). To improve the modulation extinction ratio (ER) for better optical multipath interference (MPI) tolerance, uncooled externally-modulated lasers (EMLs) were also introduced at this interface rate. By increasing the baud rate from 25 Gbaud to 50 Gbaud/s, single lane 100 Gb/s could be achieved for SMF transceivers (more challenging with VCSEL/MMF technology) using optical and electrical components with higher bandwidth and better linearity. This will enable 800 Gb/s bandwidth in an OSFP form factor.

It will be a challenge to further double the data rate from 800 Gb/s to 1.6 Tb/s due to bandwidth constraints on the optical and electrical components and channel impairments (fiber dispersion, loss, MPI, etc). Fundamentally, there are only three axes of design freedom to scale interconnect bandwidth: 1) the symbol rate per lane; 2) the number of parallel lanes, where the parallelization can be in space, polarization, or frequency (wavelength) domains; and 3) more bits encoded per symbol. Each of these three axes has its advantages and constraints. Historically, symbol rate is the most cost-effective bandwidth scaling method, but this would double the net symbol rate from 50 Gbaud to 100 Gbaud and thus require $>$50 GHz optical and electrical component bandwidths. Scaling in the parallelization axis requires doubling the number of optical and electrical components (assuming no changes in encoding or detection techniques). This will result in an approximately linear increase in cost and power. With 16 optical lanes, the achievable yield is also a concern even for silicon photonics based optical integration technique, as the overall yield may drop off exponentially with increasing number of lanes. Extremely high single-lane yield is required for reasonable 16-lane aggregate yield. Alternatively, some redundant lanes can be incorporated to manage yield loss, but this technique also has complications in design and manufacturing. Finally, encoding more bits per symbol by using even higher order PAM signaling such as PAM8 could alleviate the bandwidth requirement of electrical and optical components. However, this is achieved at the expense of tolerance to noise and other channel impairments [3]. Thus, a judicious combination of the three techniques is the most likely path forward.

III. IM-DD WITH PAM

Until recently, direct detection has been the clear technology choice for intra-datacenter interconnects due to ease of implementation and low power consumption. There are two technical paths to scale direct-detection technology from 100 Gb/s to 200 Gb/s per lane: A) Double the baud rate with the same modulation format PAM4; or B) switch to a higher-order modulation format such as PAM8 to lower the baud-rate requirement. There are advantages and disadvantages for each solution. Solution A) can achieve higher link budget but requires higher component bandwidth, which can be a significant challenge. Operating at

higher baud rate also increases the power consumption and reduces fiber chromatic dispersion (CD) tolerance. Solution B) can lower component bandwidth requirements but this is achieved at the cost of reduced link loss budget (assume Tx laser power is kept as a constant) since a higher-order PAM requires a higher SNR (signal to noise ratio) to achieve the same bit error ratio (BER). Increasing modulation order also greatly reduces the tolerance toward optical MPI, which is a major optical channel impairment in a direct detection PAM system [4])

To achieve better trade-offs between the two technical solutions, a FlexPAM concept [5], [6] was recently proposed. For this technique, multiple PAM modulation modes with fine SE granularity (and corresponding different baud rates) would be implemented in a single DSP chip. The choice of PAM signaling for each individual link could then be determined by end-to-end performance based on actual module component bandwidth and link characteristics which may vary by module and link due to manufacturing and/or deployment variation. Such a technique could be used to increase link margins and/or to lower the overall interconnect network power consumption. For example, we may implement both PAM4 and PAM6 into a single DSP chip. For a specific link, if PAM6 performs better than PAM4 due to reasons such as lower transceiver component bandwidths and/or higher fiber CD, we could select PAM6 for this specific link to achieve a higher link margin. On the other hand, if PAM4 performs better than PAM6 due to reasons such as higher transceiver components bandwidth and/or more severe link MPI, we could select PAM4 as the operating modulation format. If both modulation formats can close the link with enough margin, we could select the modulation format with the lowest power to reduce the interconnect power consumption.

In Fig. 3 we show the impact of optical transceiver component bandwidth and link MPI on the achievable performance of various PAM modes. Fig. 3a shows the link diagram used for this modeling. Fig. 3b shows simulated receiver power sensitivity to achieve 200 Gb/s throughput by using Shannon mutual information theory with the assumption that TIA thermal noise and component bandwidth are the two dominant performance limiting factors. For this simulation, identical 3-dB bandwidth is assumed for the digital to analog converter (DAC), the driver, the Mach-Zehnder modulator (MZM), the transimpedance amplifier (TIA) and the analog to digital converter (ADC), where a fifth-order Bessel filter is used for the DAC, the driver and the ADC model, while first-order and fourth-order Butterworth filters are used for the MZM and the photodetector (PD)/TIA model, respectively. ENOB (effective number of bits) for both the DAC and ADC is assumed to be 5.5. Transmitter (Tx) side DSP includes a 3-tap baud-rate-spaced feedforward equalizer (FFE) while the receiver (Rx) side DSP includes a 17-tap baud-rate-spaced FFE. The 3-tap Tx FFE is used for pre-equalizing the Tx-side band-liming effects from the DAC, the driver and the MZM. TIA input-referred noise current is assumed to be 16 pA/$\sqrt{\text{HZ}}$. The peak to peak modulation depth, which is defined as the ratio of the peak to peak drive swing to the Vπ of the MZM, is assumed to be 0.6.

For each PAM mode, we first calculate the achievable mutual information (in terms of bits per symbol) for a range of PAM

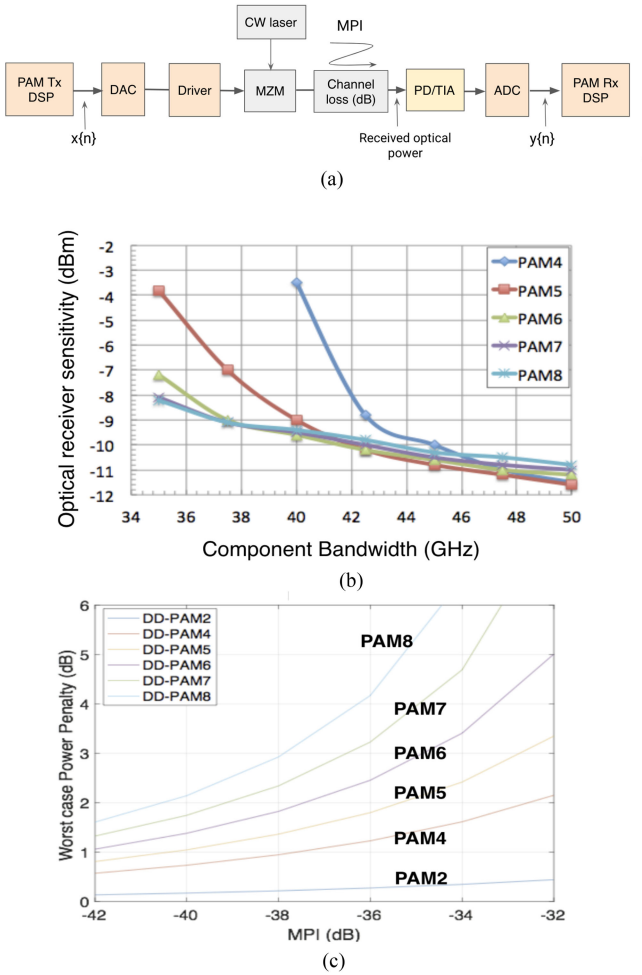

Fig. 3. (a) Link diagram used for simulation studies of component bandwidth impacts. (b) Simulated results of achievable Rx power sensitivity as a function of component bandwidth for a throughput of 200 Gb/s. (c) Plot showing the impact (power penalty) of optical MPI for different PAM modulation orders with modulation ER of 5 dB. CW: Continuous wave. Source figures of Fig. 3b and c refer to [6, Fig. 2].

baud rates and average receiving optical powers. Then we select the lowest receiving optical power that can achieve 200 Gb/s throughput as the achievable receiver power sensitivity. The mutual information between the transmitted signal x and the received signal y is given by the following classic formula [7]

$$I_{xy} = \iint P(y|x)\, P(x) \log \left\{ \frac{P(y|x)}{P(y)} \right\} dxdy \qquad (1)$$

where P(y|x) denotes the conditional probability of y given x, while P(x) and P(y) denote the probability of the transmitted signal and the received signal, respectively.

MPI-induced power penalties shown in Fig. 3c are calculated based on a worst-case analytical model [8], given by

$$P_{MPI}\,(dB) = 10\log_{10}\left[1/\left(1-\gamma\right)\right] \qquad (2)$$

$$\gamma \cong 4\,(m-1)\,\sqrt{MPI}\left(\frac{ER}{ER-1}\right) \qquad (3)$$

where m denotes the PAM modulation level, ER denotes the modulation extinction ratio, and MPI is defined as the power ratio of the interfering signal to the original signal. From Eq. 2 and Eq. 3, one can clearly see that MPI-induced penalty increases with modulation levels.

From Fig. 3b one can see that, when link performance is limited by transceiver component bandwidths, switching to a higher-order modulation format can greatly improve the receiver sensitivity. For this simulated system, if the 3-dB component bandwidths are limited to be 40 GHz, the receiver sensitivity can be improved by 6 dB using PAM6 (5.5 dB using PAM5) compared to PAM4. Further increasing the modulation order provides negligible sensitivity gain. On the other hand, if the performance is limited by channel impairments such as the detrimental MPI, switching to a lower-order PAM could help to improve the overall link performance as can be seen from Fig. 3c. For a link MPI of −35 dB, switching from PAM6 to PAM4 can reduce receiver power penalty (due to MPI) by more than 1.5 dB. Since actual transceiver component bandwidth and link condition may vary by module and link, optimal modulation format for each individual link could be different. The proposed FlexPAM technique allows link by link performance optimization, thus could enable better performance (statistically) than current (fixed) single PAM signaling designs.

Note that the sensitivity results shown in Fig. 3b are obtained by Shannon's mutual information theory, which inherently assumes optimal forward error correction (FEC) being used for each PAM mode. As a result, PAMs with different modulation levels can achieve similar performance when component bandwidths are not constrained (a higher order PAM can afford more powerful FEC with more overhead). If different PAM modes share the same FEC, a higher-order PAM such as the PAM8 will perform significantly worse than a lower order PAM such as the PAM4 if there is no bandwidth-limiting effect.

While FlexPAM helps to maximize link margin under the constraints of limited component bandwidths, there are some challenges for FlexPAM implementation. These are: (a) Multiple clock sources may be needed to achieve an equivalent throughput, (b) Coordination between the transmitter and the receiver is needed, which complicates the module firmware and overall system/link bring-up, c) Incorporating multiple modulation formats into a common chip will increase the chip size and power, and d) the non-power-of-2 PAMs such as the PAM5 (if used) require complicated/not-trivial bit mapping.

Given the above implementation complexities facing FlexPAM, a fixed PAM signaling design remains a viable option for 200 Gb/s per lane scaling if the link loss/reach budget can be satisfied with the available components.

IV. Low-Power Baud-Rate Coherent DSP

Unlike the IM-DD system where the signal is modulated over only one dimension of light (intensity), a coherent system modulates the signal over four dimensions of light: the in-phase and quadrature phase space for both X and Y polarizations. Coherent technology offers significant advantages over IM-DD in terms of receiver power sensitivity, spectral efficiency, as well as the tolerance to several optical impairments, including MPI [9], fiber chromatic dispersion (CD), and polarization mode dispersion (PMD). Coherent detection has been widely used only in long-haul (LH) and metro optical networks, and not in short reach links because of the higher power footprint (attributable to receiver DSP), the stringent requirement on laser phase noise as well as the much higher cost and larger size of coherent transceivers.

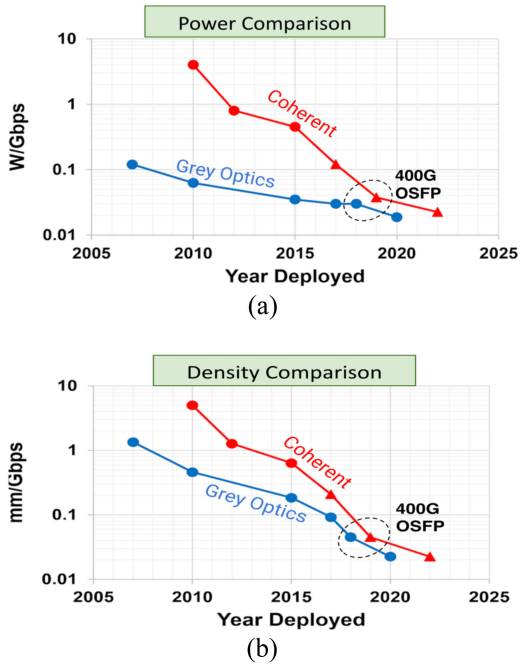

Fig. 4. Comparison of Intra-DC IM-DD and coherent transceiver (a) power/Gbps and (b) linear density/Gbps versus time of technology adoption. Triangles indicate shorter reaches (<100 km). This figure is an updated version of [10, Fig. 2].

However, with the continual advancement of optical and electrical technologies, there has been an enormous reduction in cost, power, and spatial footprint of coherent technology over the last decade [10], [11]. The evolution of coherent technology is illustrated in Fig. 4, which shows the comparison of power/Gbps and linear density in terms of mm/Gbps for coherent transceivers and intra-DC grey optics based on non-coherent (IM-DD) technologies. These curves are derived from available (or projected) commercial coherent and gray optics transceiver specs. Although coherent transceivers still consume ∼1.5x more power, the same linear density is achieved at 400 Gb/s (<100 km 400G ZR [12]). Here the linear density is defined as the required transceiver width per Gb/s, which is a main parameter to determine how many transceivers can fit on the faceplate of a standard 1RU (rack unit) switch.

The performance of coherent transceiver for DC reach (typically <2 km) could be further optimized by using lower-power baud-rate sampling DSP [13]–[18] and shedding unnecessary DSP functions [19], [20], such as the high fiber CD and polarization-mode dispersion (PMD) compensation. The overall power consumption of DC transceiver based on coherent technology could approach the level of an ADC-enabled direct detection PAM system (still 10 to 20% higher). Baud-rate-sampling

coherent technology digitizes the received optical signal at the baud rate, which not only reduces ADC power, but also reduces equalization power since lower-power baud-rate-spaced equalization can be implemented without sampling rate conversion, while the conventional coherent technology uses oversampling and fractionally-spaced equalization technique.

Baud-rate sampling and baud-rate spaced equalization techniques have been used in short reach IM-DD PAM4 systems, but baud-rate sampling and equalization (BRSE) techniques have never been used in real coherent systems for following reasons. Firstly, unlike the IM-DD PAM system where there are distinguishable levels for analog based clock recovery, the received optical signal in a coherent receiver is much more distorted (no eye-diagram at all) since the incoming signal is temporally varying due to reasons such as frequency offset (and phase noise processes) between the signal's carrier and the local oscillator as well as polarization rotations, and thus the traditional analog based clock recovery schemes do not work for a coherent system. In order to perform digital clock recovery prior to phase recovery, oversampling is thus needed. Secondly, the BRSE technique is sensitive to ADC sampling time [15], [21], and optimal performance can only be achieved by sampling at the center of each signal pulse. Such a requirement can be challenging to meet for coherent systems that require joint processing of four received signals, including in-phase and quadrature components in two orthogonal polarizations. There are unknown and variable time delays or skews between the four signals. For oversampled systems, timing skews are compensated using interpolation based digital methods after ADC sampling. This method does not work for baud-rate sampled coherent systems. Thirdly, the BRSE technique has limited tolerance for fiber CD and PMD [17], [18], which are unacceptable for traditional coherent use cases for metro/LH. However, for intra-DC use cases with typical reach <2 km, fiber CD and PMD are much smaller, especially when the common O-band wavelengths are used.

The concept of baud-rate sampling coherent receiver has been demonstrated in several offline-DSP based experiments by introducing an additional low-pass anti-aliasing filter either at the Transmitter side [14] or at the Receiver side [18] to help 1) reduce the impact of sampling phase sensitivity and 2) to increase the CD/PMD tolerance. A 2×2 MIMO equalization architecture optimized for baud-rate clock recovery is also reported in [16].

Fig. 5 shows a new DSP architecture for a baud-rate sampled coherent receiver. The low-power is enabled by two new concepts. The first concept is an integrated skew compensation functional block in the baud-rate clock recovery loop, in which the phase of each individual ADC sampling clock is adjusted to compensate the timing skew. The required timing delay for each of the four ADC clocks could be individually or jointly optimized by monitoring phase errors detected by the baud-rate clock phase error detector. One sample per symbol clock phase error detection can be achieved by the classic Mueller-Mueller algorithm [22] using phase-recovered QAM signals. The second new concept is an equalization functional block using 1-tap 4×4 real-valued multiple-input and multiple-output (MIMO) equalizer for both polarization recovery and Inphase/quadrature (I/Q) phase error compensation. While the implementation complexity of a 4×4 real-valued MIMO equalizer is the same as the 2×2 complex-valued MIMO equalizer, an additional I/Q phase error compensation functional block is needed for conventional 2×2 complex-valued MIMO equalizer. In addition, two mixed-value single-input and single output (MV-SISO) equalizers could be used for bandwidth equalization and fiber CD compensation (if needed). The equalizer input signal to MV-SISO equalizer is a complex-valued signal, and the equalizer coefficient for each tap needs to be complex-valued number for fiber CD compensation, and could be real-valued number for bandwidth equalization.

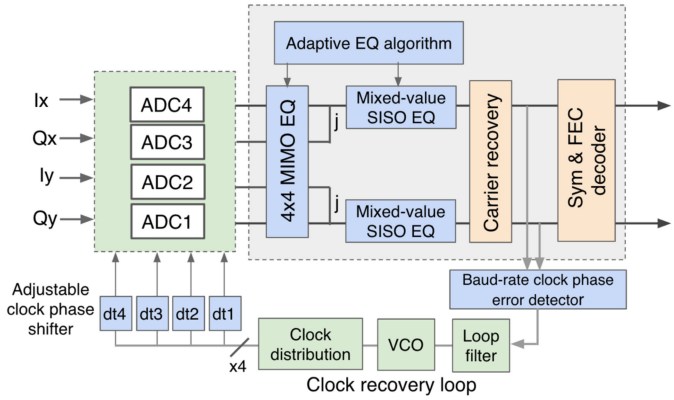

Fig. 5. A baud rate sampled low-power coherent DSP architecture showing the key functional blocks.

When the fiber CD is negligible for intra-data center applications with <2 km reach and the use of O-band wavelength, real-valued coefficients could be applied to all equalizer taps in MV-SISO equalizer. The implementation complexity of the two MV-SISO equalizers is simplified close to that of four real-valued SISO equalizers used for IM-DD PAM receiver. As compared to an IM-DD receiver, a coherent receiver still requires an additional 1-tap 4×4 real-valued MIMO equalizer and a carrier recovery functional block. The implementation complexity of a 1-tap 4×4 real-valued MIMO equalizer is equivalent to four 4-tap real-valued SISO equalizers. The implementation complexity of the carrier recovery functional block is equivalent to four 2-tap real-valued SISO equalizers if pilot symbol-based phase recovery is used [23]. The additional DSP required by a coherent receiver is thus equivalent to four 6-tap real-valued SISO equalizers, which, to the first-order estimation, only consume about 10 to 20% of the total power of a typical four-channel ADC based PAM4 ASIC [24].

V. Phase Noise Tolerant Coherent Modulation

For a single laser to transmit 800 Gb/s, there are design trade-offs in performance, power and cost for various coherent modulation formats. 5 bits per symbol (per polarization) 2-dimensional (2D) coherent modulation format, in which the transmitted data is encoded on both the amplitude and the phase dimensions of a light, is attractive for DC reach applications. A modulation format of 16 QAM with 4 bits/ symbol enables better link performance but it needs higher component bandwidths

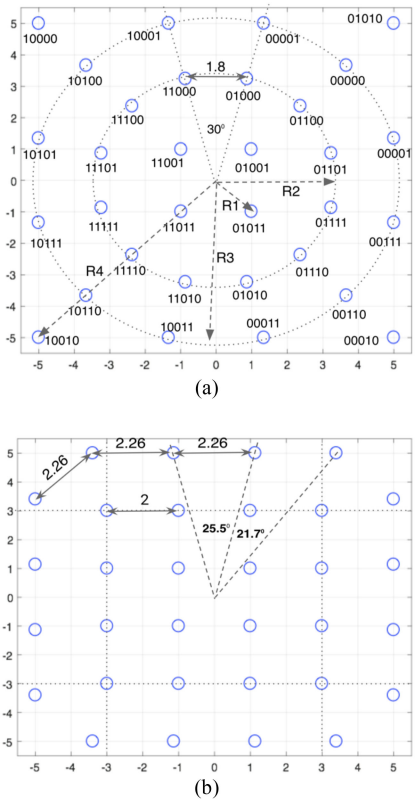

Fig. 6. Constellations for two new five bits per symbol coherent modulation formats. a) Square-32APSK with an exemplary bit to symbol mapping scheme. b) Modified Cross-32QAM.

(due to the need of a higher baud rate), while a higher order modulation format of 64QAM with 6 bits/ symbol needs lower component bandwidth but link performance won't be able to close 2 km transmission distances for DC applications (under realistic laser powers and modulator driving swings).

Traditional coherent transponders have tight laser phase / frequency requirements which make the transponder more complex and expensive (due to the need of frequency-stable and narrow linewidth lasers such as external cavity lasers). In order to reduce the laser linewidth requirement and to enable the use of the common distributed feedback (DFB) lasers, we propose two novel 2D coherent modulation formats with 5 bits per symbol: a square-32APSK (amplitude phase shift keying) and a modified cross-32QAM. The concept is shown in Fig. 6a and 6b, respectively. The traditional cross-32QAM is optimized for metro/LH application with better tolerance toward additive Gaussian noise under 2D average power (including both in-phase and quadrature signal components) constrained communication systems. Such systems include the typical LH and metro optical transmission systems where optical amplifiers are used to boost the average optical signal power prior to transmission, thus the average launch optical signal power remains a constant for different modulation formats. For intra-DC reach optical communication systems without post-modulation optical amplification, however, the signal SNR is largely constrained by the modulator electrical drive-swing (assume with the same laser, PD and TIA). Such a communication system can be modeled as a per dimension peak power constrained communication system. The two modulation formats shown in Fig. 6 are specifically optimized for per dimension peak power constrained short-reach optical communication systems. The square-32APSK in Fig. 6a is aiming to increase laser phase noise tolerance with 'minimal' additive Gaussian noise tolerance degradation while the modified cross-32 QAM in Fig. 6b is intended to increase both laser phase noise tolerance and additive Gaussian noise tolerance.

The Square-32APSK can be decomposed into four ring-based constellations: the inner and the outer ring (with radius R1 and R4, respectively) both have a QPSK (quadrature phase-shift keying) constellation, and the two middle rings (with radius R2 and R3, respectively) have the same 12-PSK constellations. Both QPSK constellations have equal in-phase and quadrature components. The ratio between R1, R2, R3 and R_4 is given by: R2/R1≈2.4, R3/R1≈3.75; and R4/R1≈5.12. The two 12-PSK and the two QPSK constellations are arranged in a way that the outer 16 constellation points make a square. Such a square constellation arrangement allows greater minimum Euclidean distance and phase spacing for per dimension peak power constrained communication systems. Compared to the conventional cross-32QAM, the square-32APSK increases the minimum phase spacing from 19° to 30° (a ∼53% increase), while the minimum Euclidean distance is only reduced from 2 to 1.8, (∼11% reduction).

For the modified cross-32QAM modulation format, the inner 16 constellation points are the same as the conventional cross-32QAM, but the outer 16 constellation points are rearranged to have equal Euclidean distance between any neighboring constellation points of them (the outer 16 constellation points of a conventional cross-32QAM are more concentrated toward the center to minimize the average power). As a result, for a peak power constrained shorter reach optical communication system, the modified cross-32QAM will increase the Euclidean distance for the outer 16 constellation points by 13%, and increases the minimum phase spacing by about 11%. This will translate into improved tolerance toward both the phase noise and additive Gaussian noise. Note that the same symbol to bit mapping schemes developed for the conventional cross-32QAM can be used for the modified cross-32QAM as well.

Fig. 7 shows the simulated laser phase noise tolerance results under a fixed peak electrical SNR per dimension of 21.9 dB (to obtain a baseline BER close to 1e-3). The peak electrical SNR is defined as the ratio of received peak electrical signal power (excluding noise, after PD) to average noise power. For comparison, results for the common cross-32QAM are also displayed. One can see that, the modified cross-32QAM performs best when the laser linewidth is smaller than 1.5 MHz, and the square-32APSK performs best when the laser linewidth is more than 1.5 MHz. For this simulation, training-assisted two-stage phase recovery [25] is used for phase estimation, where training symbols (outermost QPSK symbols) are periodically inserted after every 31 signal symbols. Coarse phase is directly estimated from the inserted training symbols and then a maximum likelihood phase recovery stage is followed to refine the phase estimation.

The concept of flexible/adaptive modulation described in Section III for FlexPLM could be extended to coherent QAM as

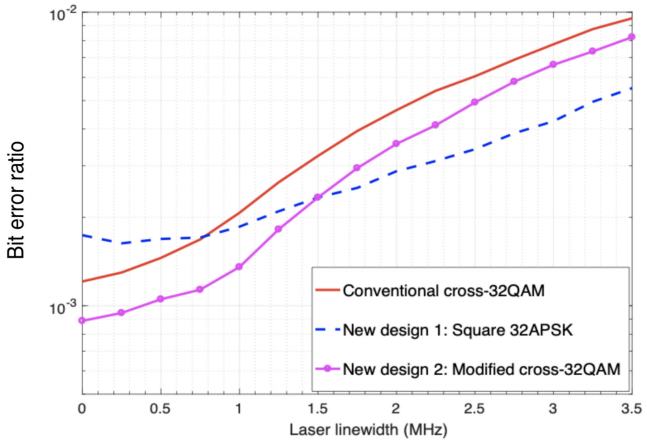

Fig. 7. Simulated laser linewidth tolerance for the conventional cross-32QAM, the proposed square-32APSK, and the modified cross-32QAM.

well: polarization multiplexed (PM) 16QAM, 25QAM, Square-32APSK and modified cross-32QAM may be implemented in a single ASIC chip, and the optimal modulation format could be chosen based on actual module component bandwidth, linewidth, and link characteristics.

Several flexible modulation methods with fine granularity in spectral efficiency (SE) have been proposed for long-haul (LH) coherent transmission systems, including Time-Domain Hybrid QAM [25] and Probabilistically-Shaped QAM [26]. Both these methods are highly effective for LH optical transmission to achieve the optimal performance (gap to Shannon limit) and easily traverse the trade-off between capacity and distance. In these LH systems, the performance is constrained by the average signal power unlike in intra-DC links which are unamplified and hence constrained by the peak signal power. Since the peak signal power is limited by modulator drive swing and laser power, neither time-domain hybrid modulation nor probabilistically-shaped modulation applied to PAM/QAM would be effective in increasing performance [5].

VI. IM-DD vs. Coherent

200 Gb/s per lane (for IM-DD PAM) or per dimension (for coherent PM-QAM) is likely the candidate for 1.6 Tb/s and beyond interface bandwidth scaling. In this section, we compare IM-DD versus coherent technology for 200 Gb/s per lane/dimension throughput scaling in terms of achievable link budget, power consumption, implementation complexity, as well as fan out granularity. Modulation formats with identical per dimension bandwidth/spectral efficiency of 2 bits/symbol, 2.5 bits/symbol and 3 bits/symbol are used for comparison, corresponding to 16QAM, 32QAM and 64QAM for coherent and PAM4, PAM6 and PAM8 for IM-DD. For a fair comparison, external optical MZM (Mach-Zehnder modulator) is used for both coherent QAM and IM-DD PAM modeling. In addition, Nyquist bandwidth for all the components (DAC, driver, MZM, PD/TIA, and ADC) and the input referred TIA thermal noise is assumed to be 16 pA/$\sqrt{\text{Hz}}$. Except for PD shot noise, other optical impairments such as MPI, laser RIN and fiber CD are neglected for the

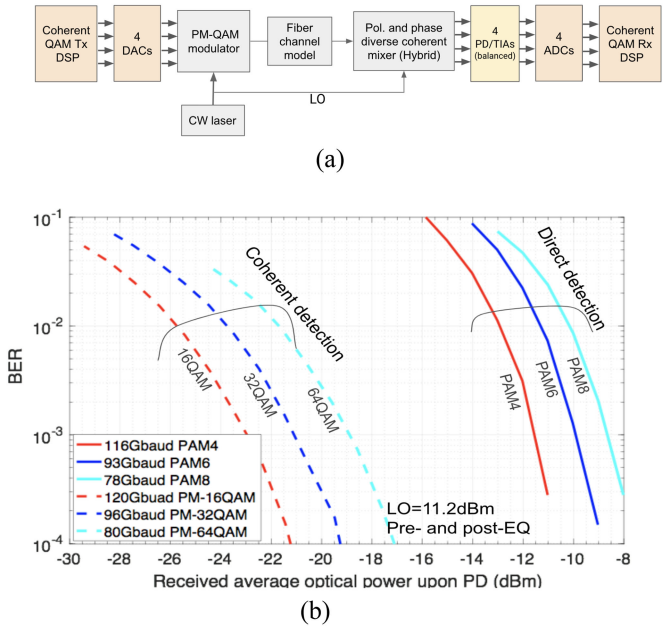

Fig. 8. (a) Link diagram used for Coherent PM-QAM baseline performance modeling. (b) Receiver sensitivity comparison between coherent PM-QAM with 200 Gb/s per dimension and IM-DD PAM with 200 Gb/s per lane.

baseline link loss budget modeling. Ideal MZM is assumed for all simulations through this section.

Fig. 8a shows the link diagram (functional blocks) used for coherent PM-QAM baseline performance modeling (the Link diagram used for IM-DD PAM baseline performance modeling is shown in Fig. 3a). Fig. 8b shows the simulated results on the achievable receiver power sensitivity for both the PM-QAM and the IM-DD PAM systems at identical 200 Gb/s per dimension throughput, which corresponds to 800 Gb/s throughput per laser for the coherent PM-QAM systems and 200 Gb/s per laser for the IM-DD PAM systems. For coherent PM-QAM systems, a single laser is used for both the signal source and the local oscillator (LO) source (16 dBm total laser power and 1/3 of laser power used for LO). MZM drive swing of full Vπ with ideal MZM nonlinearity compensation is used for both the PM-QAM and the IM-DD PAM systems. From Fig. 8b, one can see that coherent PM-16, 32 and 64QAM can achieve 13 dB, 13 dB and 12 dB better receiver power sensitivity (at BER 1e-2) than the IM-DD PAM4, PAM6 and PAM8, respectively. Note that here the received power for the coherent system is defined as the total power incident on the four balanced PDs.

Although coherent detection can improve the receiver power sensitivity, 2D coherent modulation also introduces larger modulation loss at the Tx side. Fig. 9 shows modulation loss versus MZM drive swing for both coherent QAM and IM-DD PAM. Optical modulation amplitude (OMA) loss is used for IM-DD PAM where the MZM is biased at the quadrature point and the average power is independent of the modulator drive swings. Average optical power loss is used for coherent QAM, where the MZM is biased at the null point. While the coherent QAM modulation loss is strongly dependent on the drive swing applied to the MZM, coherent QAM modulation loss is 8 dB higher

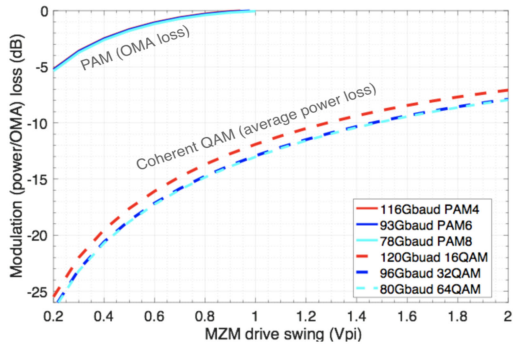

Fig. 9. Modulation loss comparison between coherent QAM (including 3 dB intrinsic IQ modulator loss) and IM-DD PAM.

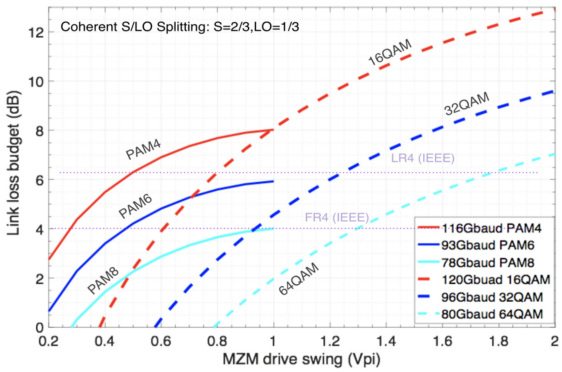

Fig. 10. Supported link loss budget comparison between coherent PM-QAM and IM-DD PAM.

TABLE II
TX AND RX LINK LOSS PARAMETERS USED FOR LINK BUDGET CALCULATION

	2x800Gb/s PM-QAM	8x200Gb/s PAM
Laser number	2	8
Per Laser power (dBm)	16	16
MZM IL (dB)	4	4
Tx path loss(dB)	7.8[a]	4
Rx path loss (dB)	4	2
Mux+DeMux (dB)	1	4
Implement. penalty (dB)	5/5.5/6 (for 16/32/64QAM)	4/4.5/5 (for PAM4/6/8)

[a]: including 1.8 dB signal/LO splitting loss (signal = 2/3, LO = 1/3)

TABLE III
OVERALL COMPARISON BETWEEN THE 200 GB/S PER DIMENSION COHERENT QAM TECHNOLOGY AND THE 200 GB/S PER LANE IM-DD PAM TECHNOLOGY (ASSUME 4 DB LINK LOSS BUDGET, 0.8 Vπ DRIVE, AND 1.6 TB/S THROUGHPUT)

	Coherent PM-QAM	IM-DD PAM
Lasers per 1.6Tb/s	2	8
Per laser power	16.9dBm (32QAM)	14.4dBm (PAM6)
Total laser power	19.9dBm (32QAM)	23.4dBm (PAM6)
Laser requirement	Cooled Linewidth<1MHz	Uncooled
MZMs and Drivers	8	8
PD/TIAs	8 (balanced)	8 (single-ended)
Relative Tx DSP[b] power	**1**	**1**
Relative Rx DSP[b] power	1.1 to 1.2	**1**
MPI/RIN/CD Tolerance	Good	Less tolerant
2km with CWDM8	Yes	Challenging (CD)
2km with LAN WDM8 OR PSM8 (O-band)	Yes	Yes
Fan-out granularity	800Gb/s	200Gb/s
Scale to 3.2Tb/s and beyond	Yes (more optimization freedoms)	Challenging (Laser yield and fiber CD)

[a]: link budget and required laser power are calculated based on loss parameters shown in Table II
[b]: include both the analog front ends and the digital portions.

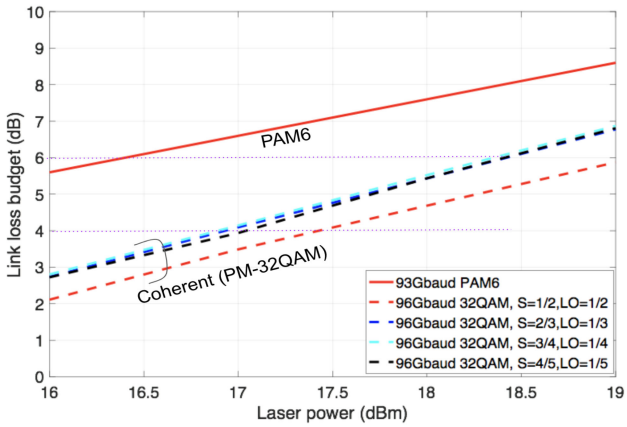

Fig. 11. Effectiveness of increasing laser power on link loss budget (MZM drive swing = 0.8 Vpi).

than IM-DD PAM even at full 2 Vπ drive swing. At 0.5 Vπ drive swing, the coherent QAM system introduces 15 dB more modulation loss than the IM-DD PAM system.

Fig. 10 shows the achievable link loss budget versus MZM driver swing based on realistic Tx and Rx optical path loss assumptions listed in Table II, assuming silicon photonics-based technology (due to forward-looking optical and electrical integration capability). In the link budget analysis, assuming the same FEC capability, but 1 dB higher implementation penalty for the coherent PM-QAM system due to additional phase noise related penalty. Fig. 10 shows that the supported link loss strongly depends on MZM drive swings, especially for the coherent technology. Under identical per laser power of 16 dBm, the IM-DD system can support larger link loss if the drive swing is below 1 Vπ. For example, at a drive swing of 0.8 Vπ, PAM6 can support 5.6 dB link loss while coherent PM-32QAM can only support 2.7 dB link loss. The coherent system performs better if we allow larger drive swings: PM-32QAM can support 9.6 dB link loss with full 2 Vπ drive swing while PAM6 can only support 6 dB link loss with full Vπ drive wing. The reason that the coherent system performs poorer at lower drive swing is mainly due to the large modulation loss as is shown in Fig. 9.

Fig. 10 assumes identical per laser power, so the total laser power for the 1.6 Tb/s IM-DD PAM system is 4x higher. Since a coherent PM-QAM system uses 4x fewer lasers, increasing per laser power could be a more power-efficient way than increasing drive swing to get the required link loss budget. In Fig. 11

we show the supported link loss budget versus per laser power for both coherent PM-32AM and IM-DD PAM6 systems. For coherent PM-32QAM, we also show the results with different laser power splitting ratios (between the signal source and the LO). At the optimal laser power splitting ratio ($\sim$3/4 for the signal and $\sim$1/4 for the LO), 1 dB increase in laser power will gain about 1.37 dB link loss budget. For IM-DD PAM6, however, 1 dB laser power gain only translates into 1 dB link loss budget gain. The reason is due to the fact that a coherent system uses a single laser as its signal source as well as its LO through power splitting, increasing laser power by 1 dB will translate into 1 dB signal power gain plus 1 dB LO power gain. The 1 dB additional LO power will improve the receiver power sensitivity by about 0.37 dB.

Under identical 0.8 Vπ drive swing and 16 dBm per laser power, the coherent PM-32QAM system supports 2.9 dB less link loss budget than the PMA6 system, but the coherent PM-32QAM system can support the same link loss budget by only increasing the laser power by 2.1 dB. Note that a 1.6 Tb/s ($2 \times$ 800 Gb/s PM-QAM) coherent system still requires 3.9 dB lower total laser power than an $8 \times$ 200G IM-DD PAM system. Even when considering the need for thermo-electric coolers (TEC) for the coherent system, the Tx power consumption of a coherent PM-QAM system could be comparable to an IM-DD PAM system, depending on TEC efficiency and operating temperature range.

With the 200 Gb/s per lane IM-DD PAM technology, 1.6 Tb/s interface bandwidth scaling needs 8 CWDM wavelengths. Fiber CD could be a problem for certain LR applications with O+E bandwidth wavelengths at 20 nm grid spacing. The worst fiber dispersion for a 2 km SMF reach is 19.3 ps/nm at 1417.5 nm [27]. From simulation, such a dispersion value will introduce 1.5 dB penalty at BER = 1e-3 for IM-DD PAM6 even using ideal, chirpless optical modulation. Higher CD penalty can be expected for realistic MZM with finite DC extinction ratio (ER), which may introduce positive transient chirp.

The performance comparison between the IM-DD PAM and the coherent PM-QAM technology is summarized in Table III, where PSM denotes parallel single mode fiber, and LAN-WDM denotes local area network wavelength division multiplexing. Both the IM-DD PAM and the coherent PM-QAM could enable 1.6 Tb/s bandwidth scaling. For reach less than 1 km, IM-DD PAM with CWDM8 or PSM8 has the advantage of slightly lower power and 200 Gb/s fan-out granularity. For reach up to 2 km, however, LAN-WDM or additional fiber CD management technology is needed for IM-DD PAM. The coherent PM-QAM technology requires one-fourth the number of lasers, although the laser requires very narrow linewidth. Coherent PM-QAM technology has higher performance potential and is also more tolerant toward several major optical impairments including MPI, laser RIN (relative intensity noise), and fiber CD. Moving forward, coherent technology can scale to and beyond 3.2 Tb/s, but IM-DD PAM faces significant challenges going beyond 1.6 Tb/s due to the number of lasers needed (yield/cost concerns) as well as increasing fiber CD. In addition, the potentially larger link loss and reach supported by the coherent technology may also help the introduction of optical switching technology into the datacenter networks in the future.

Depending on datacenter network topologies and deployment strategies, however, backward compatibility could be a problem when we switch from IM DD PAM to coherent technology for brownfield deployments. Innovations on coherent/IM-DD dual operation optical transceiver technology are needed.

VII. Conclusions

In the past decade, intra-DC optical interconnect technology has been successfully scaled from 10 Gb/s to 800 Gb/s per port, nearly two orders of magnitude. All three degrees of design freedom (symbol rate scaling, parallel optics scaling, and bits per symbol scaling) have been utilized to sustain this growth.

To scale Intra-DC interface bandwidth to 1.6 Tb/s or beyond, 200 Gb/s per lane or per dimension is likely needed although it could become more challenging to obtain sufficient optical and electrical component bandwidths. A flexible PAM (FlexPAM) with fine SE granularity (in terms of bit/symbol) helps to maximize link margin under the constraints of limited component bandwidths. Fixed PAM signaling benefits from a simpler implementation, and remains an attractive option if performance can be met.

Highly sensitive coherent detection technology presents another option. To lower the coherent receiver DSP power close to the level of IM-DD PAM technology, a new DSP architecture is proposed with baud-rate ADC sampling and baud rate spaced equalization techniques. We also proposed two new phase noise tolerant 5 bits per symbol 2D coherent modulation formats to relax the laser linewidth requirement.

Finally, we presented a detailed performance comparison on the 200 Gb/s per lane IM-DD PAM technology and the 200 Gb/s per dimension coherent PM-QAM technology in terms of link loss, power consumption, implementation complexity, and fan-out granularity. For 1.6 Tb/s interface bandwidth, the IM-DD PAM technology has some advantage for SR applications where finer bandwidth granularity is needed, while baud-rate sampled coherent technology has potential advantages for $>$2 km LR reach applications. For 3.2 Tb/s and beyond, coherent technology extends these advantages for both SR and LR applications.

References

[1] A. Singh *et al.*, "Jupiter rising: A decade of clos topologies and centralized control in Google's datacenter network," *Commun. ACM*, vol. 59, no. 9, pp. 88–97, 2015.

[2] L. Barroso, U. Hölzle, and P. Ranganathan, *The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines*, 3rd ed. San Rafael, CA, USA: Morgan & Claypool, 2018.

[3] X. Zhou, H. Liu, R. Urata, and S. Zebian, "Scaling large data center interconnects: Challenges and solutions," *Opt. Fiber Technol.*, vol. 44, pp. 61–68, 2018.

[4] W. Way and T. Chan, "MPI penalties for 400GBASE-FR8/LR8 links," in *Proc. IEEE802.3bs 400 GbE Meeting*, Sep. 2015, pp. 1–8.

[5] X. Zhou and H. Liu, "Constellation shaping: Can it be useful for datacenter reach communication," in *Proc. Eur. Conf. Opt. Commun.*, 2017, pp. 1–9.

[6] X. Zhou, R. Urata, and H. Liu, "Beyond 1 Tb/s datacenter interconnect technology: Challenges and solutions," in *Proc. Opt. Fiber Commun. Conf. Exhib.*, San Diego, CA, USA, 2019, Paper Tu2F.5.

[7] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*. Urbana, IL, USA: Univ. Illinois Press, Sep. 1998.
[8] 2016. [Online]. Available: http://www.ieee802.org/3/bs/public/16_01/bhatt_3bs_01a_0116.pdf
[9] F. Zhu, Y. Wen, and Y. Bai, "Component BW requirement of 56 Gbaud modulations for 400 GbE 2 & 10 km PMD" in *Proc. IEEE 802.3bs 400 GbE Task Force Plenary Meeting*, Jul. 2014, pp. 1–13.
[10] R. Urata, H. Liu, X. Zhou, and A. Vahdat, "Datacenter interconnect and networking: From evolution to holistic revolution," in *Proc. Opt. Fiber Commun. Conf. Exhib.*, Los Angeles, CA, USA, 2017, Paper W3G.1.
[11] R. Urata, X. Zhou, and H. Liu, "Beyond 400G: Business as usual or coherent convergence?" in *Proc. OFC Workshop Talk: Beyond 400G Hyperscale DCs Workshop*, 2019, pp. 1–7.
[12] 2019. [Online]. Available: https://www.oiforum.com/technical-work/hot-topics/400zr-2/
[13] A. Gorshtein, O. Levy, G. Katz, and D. Sadot, "Coherent compensation for 100G DP-QPSK with one sample per symbol based on antialiasing filtering and blind equalization MLSE," *IEEE Photon. Technol. Lett.*, vol. 22, no. 16, pp. 1208–1210, Aug. 2010.
[14] D. Millar, D. Lavery, R. Maher, B. C. Thomsen, P. Bayvel, and S. J. Savory, "A baud-rate sampled coherent transceiver with digital pulse shaping and interpolation," in *Proc. Opt. Fiber Commun. Conf. Expo. Nat. Fiber Opt. Engineers Conf.*, Anaheim, CA, USA, 2013, Paper Tu2I.2.
[15] G. Lu, T. Sakamoto, and T. Kawanishi, "Experimental investigation of sampling phase sensitivity in baud-rate sampled coherent receiver for Nyquist pulseshaped high-order QAM signals," in *Proc. Conf. Lasers Electro-Opt., Laser Sci. Photon. Appl.*, San Jose, CA, USA, 2014, Paper SW1J.2.
[16] A. Gorshteina, D. Sadota, and G. Dormanb, "MIMO equalization optimized for baud rate clock recovery in coherent 112 Gbit/sec DP-QPSK metro systems," *Opt. Fiber Technol.*, vol. 22, pp. 23–27, Mar. 2015.
[17] X. Zhou and H. Liu, "Pluggable DWDM: Considerations for campus and metro DCI applications," in *Proc. Eur. Conf. Opt. Commun.*, 2016, Paper WS3.
[18] C. R. S. Fludger *et al.*, "Coherent equalization and POLMUX-RZ-DQPSK for robust 100-GE transmission," *J. Lightw. Technol.*, vol. 26, no. 1, pp. 64–72, Jan. 2008.
[19] K. Matsuda, R. Matsumoto, and N. Suzuki, "Hardware-efficient adaptive equalization and carrier phase recovery for 100 Gb/s/λ-based coherent WDM-PON systems," in *Proc. Eur. Conf. Opt. Commun.*, 2017, Paper Th.1.B.2.
[20] J. Cheng, C. Xie, M. Tang, and S. Fu, "A low-complexity adaptive equalizer for digital coherent short-reach optical transmission systems," in *Proc. Opt. Fiber Commun. Conf. Exhib.*, San Diego, CA, USA, 2019, Paper M3H.2.
[21] J. G. Proakis, *Digital Communication*, 3rd ed. New York, NY, USA: McGraw-Hill, 1995, ch. 6.
[22] K. H. Mueller and M. S. Müller, "Timing recovery in digital synchronous data receivers," *IEEE Trans. Commun.*, vol. COM-24, no. 5, pp. 516–531, May 1976.
[23] B. Smith, J. Riani, I. Lyubomirsky, and S. Bhoja, Interleaving and pilot insertion for CFEC oif2017.535.00.
[24] S. Bhoja, "PAM4 signaling for intra-data center and data center to data center connectivity (DCI)," in *Proc. Opt. Fiber Commun. Conf. Exhib.*, Los Angeles, CA, USA, 2017, Paper W4D.5.
[25] X. Zhou *et al.*, "High spectral efficiency 400 Gb/s transmission using PDM time-domain hybrid 32–64 QAM and training-assisted carrier recovery," *J. Lightw. Technol.*, vol. 31, no. 7, pp. 999–1005, Apr. 2013.
[26] F. Buchali, F. Steiner, G. Böcherer, L. Schmalen, P. Schulte, and W. Idler, "Rate adaptation and reach increase by probabilistically shaped 64-QAM: An experimental demonstration," *J. Lightw. Technol.*, vol. 34, no. 7, pp. 1599–1609, Apr. 2016.
[27] "400G CWDM8 MSA 2 km optical interface technical specifications," Rev. 1, Feb. 13, 2018.

Xiang Zhou received the Ph.D. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 1999. He is currently with Google Datacenter Optics Group, leading next-gen optical interconnect technologies and roadmap development. Prior to joining Google, He was with AT&T Labs-Research, conducting research on various aspects of long-haul optical transmission and networking technologies.

He has authored or co-authored papers extensively in top Journals and conferences in his field, including several record-setting 'hero' results. He is the author of several book chapters and the holder of more than 50 U.S. patents. He is currently an Associate Editor for *Journal Lightwave Technology* and is an OSA Fellow. He was also on the editorial board of Optics Express and is on the Program Committees of a variety of technical conferences.

Ryohei Urata received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA. He is currently a Technical Lead/Manager with the Google Platforms Optics Group, responsible for Google's datacenter optical technologies and corresponding roadmap. Prior to joining Google, he was a Research Specialist with NTT Laboratories from 2004 to 2010. He has authored or co-authored more than 135 patents/publications in the areas of optical interconnect, switching, and networking.

Hong Liu received the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA. She is currently a Distinguished Engineer with Google Technical Infrastructure, where she is involved in the system architecture and interconnect for a large-scale computing platform. Her research interests include interconnection networks, high-speed signaling, optical access, and metro design. Prior to joining Google, she was a Member of Technical Staff with Juniper Networks, where she worked on the architecture and design of network core routers and multichassis switches. She is an OSA Fellow.