Silicon Photonics-Based 100 Gbit/s, PAM4, DWDM Data Center Interconnects

Radhakrishnan Nagarajan, Mark Filer, Yang Fu, Masaki Kato, Todd Rope, and James Stewart

***Abstract*—In this paper we discuss the nature of and requirements for data center interconnects. We then demonstrate a switch-pluggable, 4.5 W, 100 Gbit/s, silicon-photonics-based, PAM4, QSFP-28 module to transport Ethernet data directly over DWDM for layer 2/3 connection between switches at data centers up to 120 km apart, thereby eliminating the need for a separate optical transport layer. The module, based on the direct detect modulation format, is of much reduced complexity, power, and cost compared to the coherent systems that are currently being deployed for this application.**

***Index Terms*—Optical fiber communication; Optical interconnections; Silicon photonics; Wavelength division multiplexing.**

I. Introduction

Modern-day data center interconnects are not limited to distances within buildings. Although the data centers themselves are physical locations, limited in land area, the optical interconnects between them could span anywhere from tens of kilometers (km) of terrestrial distances to thousands of km of subsea routes.

Figure 1 shows the worldwide footprint of the Microsoft data center network as of mid-2017 [1]. This supports 42 Azure regions (which themselves consist of hundreds of data centers), 4500 points of presence (or peering points), and over 130 edge locations [2]. All these facilities are interconnected over vast distances.

Optical transmission distances of more than 2 km are inter-data-center interconnects, external to the physical data center, and are abbreviated as DCI in this paper. In general, DCI have the following characteristics.

1. Point to point interconnects with little optical add/drop or switching.
2. Ethernet is the framing protocol of choice.
3. Dense wavelength-division multiplexing (DWDM) is the optical transmission format.
4. They cover a wide range of geographic distances, including submarine links.

Direct detection and coherent detection are the choices of optical technology for the DCI. Both are implemented using the DWDM transmission format in the C band, 192–196 THz window, of the optical fiber. Direct detection modulation formats are amplitude modulated, have simpler detection schemes, consume lower power, cost less, and usually need external dispersion compensation. Coherent modulation formats are phase modulated, need a local oscillator for detection, need sophisticated digital signal processing, consume more power, have a longer reach, and are more expensive. It is not the goal of this paper to do a detailed comparison of these two formats. They both have applications in the DCI. We will only limit ourselves to the discussion of direct-detection-based optical transmission.

Practical direct detection schemes could either employ a binary level, non-return–to-zero (NRZ) modulation or a four-level pulse amplitude modulation (PAM4) [3,4]. Discrete multi-tone (DMT) format has not proven to be practically realizable with realistic optical signal-to-noise ratio (OSNR) requirements at 100 Gb/s bit rates in this application. The PAM4 modulation format has twice the capacity of NRZ. In this paper, we will discuss the application of a PAM4-based direct detection modulation format for a 100 Gbit/s DCI optical link.

Figure 2 shows the currently preferred optical technology as a function of link distance. In a fiber-rich environment inside the data center, where minimizing impairments due to chromatic dispersion and meeting tight link budgets without amplification are key design criteria, 1300 nm links are commonly deployed. The pluggable modules for the 100 Gbit/s links inside the data center, like the PSM4, LR4, and CWDM4, employ the direct-detect NRZ format. As the data rate progresses to 400 Gbit/s, for DR4 and FR4 modules, PAM4 direct-detect formats have become the standard [5].

Outside and in between data centers, as the transmission distances progressively increase and where maximizing spectral density is important, 1550 nm DWDM links are commonly deployed. The DWDM links are either direct-detect PAM4 or coherent quadrature amplitude

Manuscript received January 17, 2018; revised April 17, 2018; accepted April 19, 2018; published May 18, 2018 (Doc. ID 319719).

R. Nagarajan (e-mail: rnagarajan@inphi.com) is with the Inphi Corp, 2953 Bunker Hill Lane, Santa Clara, California 95054, USA.

M. Filer is with the Microsoft Corp, One Microsoft Way, Redmond, Washington 98052, USA.

Y. Fu, M. Kato, T. Rope, and J. Stewart are with the Inphi Corp, 2953 Bunker Hill Lane, Santa Clara, California 95054, USA.

https://doi.org/10.1364/JOCN.10.000B25

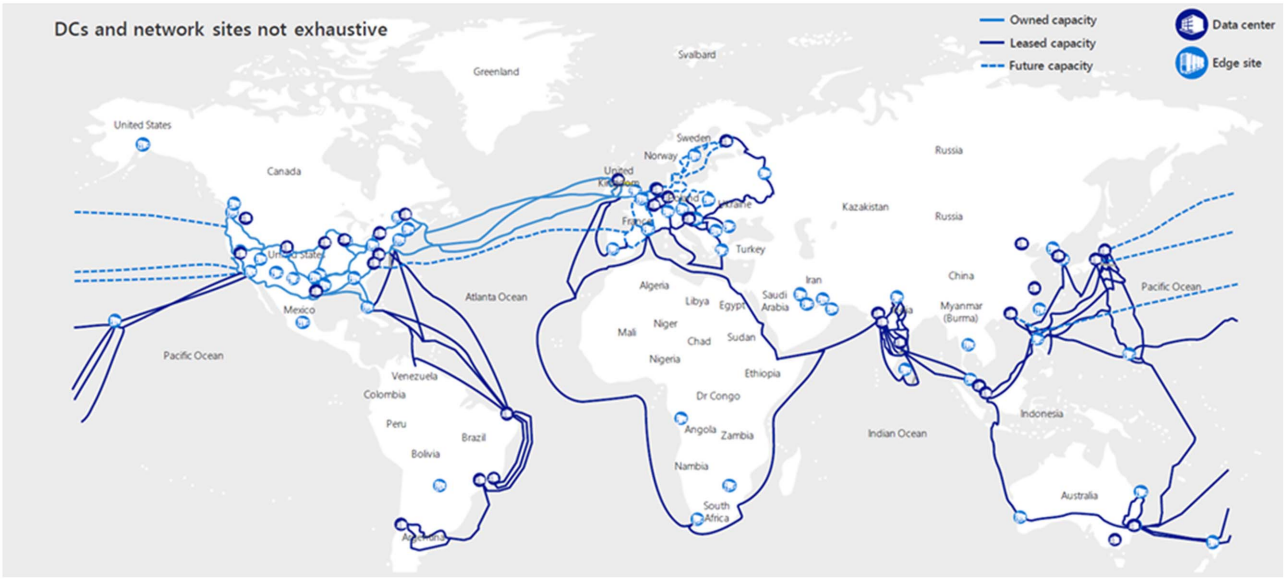


Fig. 1. Worldwide footprint of the Microsoft Data Center network.

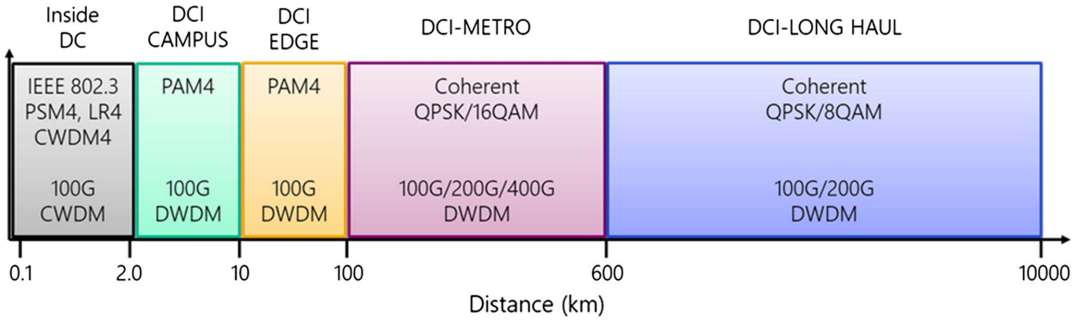


Fig. 2. Optical technology as a function of distance.

modulation (QAM, $m = 2^n$, where n is an integer > 2). For transmission distances less than 120 km, up to the limit of the DCI-Edge links PAM4 is a suitable modulation format.

The commonly used nomenclature for the various DCI reaches are as follows.

1. DCI-Campus: These connect data centers that are close together, as in a campus environment. The distances are typically limited to between 2 and 5 km, which may be easily covered without a need for external dispersion compensation at 28 Gbaud symbol rates. There is also an overlap of CWDM and DWDM links over these distances, depending on fiber availability in the environment.
2. DCI-Edge: The reaches for this category range from 2 to 120 km. These are generally latency limited and are used to connect regional, distributed data centers.
3. DCI-Metro/Long Haul: The DCI-Metro and DCI-Long Haul, as a group, lump fiber distances beyond the DCI-Edge up to 3000 km for terrestrial links and longer for subsea ones. Coherent modulation format is used for these, and the modulation type may be different for the different distances. A typical deployment scenario may be as follows:
   a. 16 QAM: <1000 km
   b. 8 QAM: >1000 km to 3000 km
   c. QPSK: >3000 km

II. Data Center Interconnect—Edge

The DCI space has become an area of increased focus for traditional DWDM system suppliers over the last few years. The growing bandwidth demands of cloud service providers (CSPs) offering SaaS (software as a service), PaaS (platform as a service), and IaaS (infrastructure as a service) capabilities have in turn driven demand for optical solutions to connect switches and routers at the different tiers of the CSP's data center network. Today, this requires solutions at 100 Gbit/s, which inside the data center can be met with direct-attach copper cabling (DAC), active optical cables (AOC), or 100 Gbit/s "grey" optics (non-DWDM, e.g., CWDM4, PSM4). For links connecting

data center facilities (Campus or Edge/Metro applications), the only choice that was available until recently was full-featured coherent transponder solutions, which are sub-optimal, as discussed below.

Along with the transition to the 100 Gbit/s ecosystem, there has been a shift in data center network architectures from more traditional data center models where all the data center facilities reside in a single, large "mega data center" campus. Most CSPs have converged on distributed regional architectures to achieve the required scale and provide cloud services with high availability as shown in Fig. 3.

Data center regions are typically located near large metropolitan areas with high population densities to provide the best possible service (latency and availability) to end customers nearest those regions. Regional architectures vary slightly among CSPs but fundamentally consist of redundant regional "gateways" or "hubs" that connect to the CSP's wide area network (WAN) backbone (and possibly to edge sites for peering, local content delivery, or subsea transport). Each regional gateway connects to each of the region's data centers, where the compute/storage servers and supporting switching fabrics reside. As the region needs to scale, it becomes a simple matter of procuring additional data center facilities and connecting them to the regional gateways. This enables rapid scaling and growth of a region, compared to the relatively high expense and long construction times of building new mega data centers, and has the side benefit of introducing the concept of diverse "availability zones" [1,6,7] within a given region.

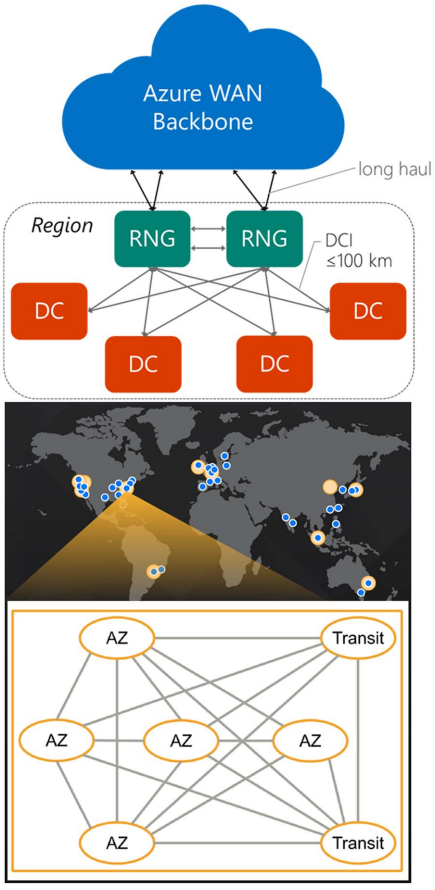


Fig. 3. Basic data center architecture of Microsoft Azure (top) and Amazon Web Services (bottom).

The transition to regional from mega data center architectures introduces some additional constraints that must be considered when choosing locations of gateways and data center facilities. One is that, to ensure the same customer experience (from a latency perspective), the maximum distances between any two data centers (via the common gateway) must be bounded. For Microsoft Azure, this is driven by a maximum round-trip server-to-server latency in the few-millisecond range. This is also true for the Amazon Web Services data center architecture, where the availability zones (AZs) are <2 ms apart [6]. This latency is primarily dominated by the physical fiber propagation time in the DCI portion of the network. Telecom-grade optical fibers have latencies of approximately 5 μs/km; to meet the latency service level agreements (SLAs), the practical maximum distance of a gateway to any data center is around 80 km. Requirements for other CSPs may differ slightly in this regard, but consensus is that DCI applications typically do not practically exceed 120 km. Supporting this statement, a distribution of fiber distances for DCI applications in Microsoft's network can be seen in Fig. 4.

Another consideration is that the spectral efficiency of grey optics is too low for interconnecting physically disparate data center buildings within the same geographic region. In the 100 Gbit/s era, this problem was typically solved using dense wavelength-division multiplexed (DWDM) coherent quadrature phase-shift keying (QPSK) transponders which offer soft-decision forward error correction (FEC)-enabled performance down to 11 dB OSNR, subsea-capable chromatic dispersion (CD) compensation of 250,000 ps/nm, power efficiencies on the order of 100 W per 100 Gbit/s, and capacities of 8–9.6 Tbit/s per fiber pair. For DCI links that are latency-constrained to 120 km or less (<2400 ps/nm), line systems can easily deliver OSNRs in the low 30 dB range. Rack space and power are typically limited and costly in these facilities, making power and space efficiency critical design goals. While fiber is not as abundant in these metro environments

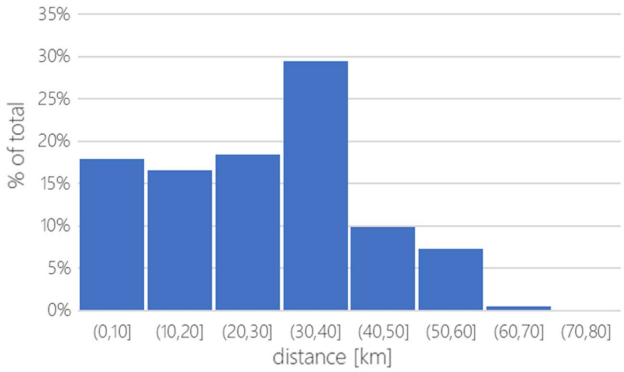


Fig. 4. Distribution of DCI fiber distances in Microsoft's network.

as within the data center itself, typically tens of fiber pairs are available at reasonable costs, relaxing ultimate spectral efficiency as a primary design criterion. With these considerations, today's coherent solutions are not an ideal fit for DCI applications.

In response, low-power, low-footprint, direct-detect solutions have been conceived, employing the PAM4 modulation format [8]. By utilizing silicon photonics technology, a dual-carrier transceiver featuring a PAM4 application specific integrated circuit (ASIC), with integrated digital signal processing (DSP) and FEC, was developed and packaged into a QSFP28 form-factor. The resulting switch-pluggable module enables DWDM transmission over typical DCI links at 4 Tbit/s per fiber pair and power consumption of 4.5 W per 100 Gbit/s. About half of the total module power consumption is for the PAM4 ASIC, and the rest is for the modulator driver, receiver TIA, DFB lasers, switching power supplies, micro-controller, and other electronics. The following sections cover this development in detail.

A. Switch-Pluggable 100 Gbit/s DWDM Module

The switch-pluggable QSFP28 module (SFF-8665 MSA compatible) is based on a highly integrated silicon photonics optical chip and a PAM4 ASIC as shown in Fig. 5 [9,10]. The four KR4 encoded 25.78125 Gbit/s inputs are first stripped of their host FEC. Then the two 25 Gbit/s streams are combined into a single PAM4 stream, and a more powerful line FEC called IFEC is added, making the line rate 28.125 GBaud per wavelength. The four 25 Gbit/s Ethernet streams are thus converted to two 28.125 GBaud PAM4 streams. In the 40 Gbit/s use case, there are four 10 Gbit/s inputs, from which a single 22.5 GBaud PAM4 stream is generated.

The line system configuration for the DCI link is shown Fig. 6. This is essentially a single-span DWDM link with an external (chromatic) dispersion compensation module (DCM). It has a booster erbium-doped fiber amplifier (EDFA) at the transmit side and pre-amplifier EDFA on the receive side. These are labeled as optical amplifier (OA) in Fig. 6.

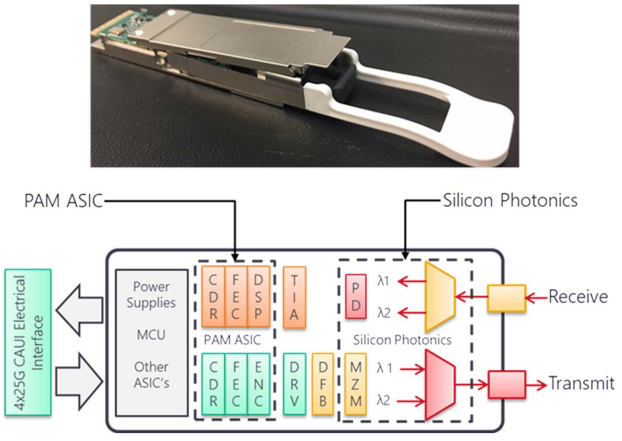


Fig. 5. 100 Gbit/s, QSFP-28, DWDM module architecture.

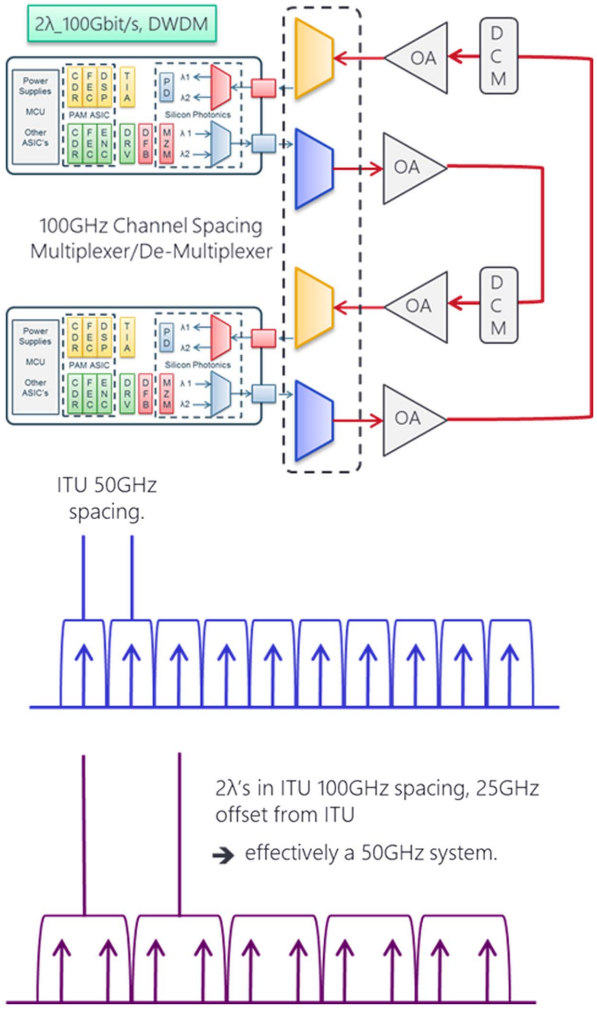


Fig. 6. Typical configuration of point-to-point DCI-Edge link.

The pluggable optical modules are meant to operate with commercially available external multiplexers and demultiplexers on the 100 GHz International Telecommunication Union (ITU) grid. The two wavelengths from the module are spaced 50 GHz apart, but they are not on the 50 GHz ITU grid. To transmit both wavelengths within the 100 GHz passband, the wavelengths are offset 25 GHz to either side of the standard ITU frequency on the 100 GHz grid. This enables the transmission of up to 40 100 Gbit/s channels or equivalently 80 50 Gbit/s carriers at 50 GHz spacing within the C band of the optical fiber transmission spectrum. The channel schema is depicted in the bottom half of Fig. 6.

The dispersion compensation could either be fixed or tunable [11], and is broadband to compensate all wavelengths simultaneously. The tunable DCMs are commonly based on either the tunable fiber Bragg grating or tunable etalon technologies, both of which are channelized to compensate 50 GHz spaced optical carriers. These modules could be easily incorporated into the mid-stage of the transmit or receive EDFA with minimal impact on the link OSNR.

B. PAM4 DSP ASIC

The PAM4 ASIC [12–14] is the electrical heart of the module. The ASIC architecture and its electrical performance are shown in Fig. 7. The ASIC supports a dual-wavelength 100 Gbit/s mode (with 4×25 Gbit/s inputs) and a single-wavelength 40 Gbit/s mode (with 4×10 Gbit/s inputs). The ASIC has MDIO and I2C management interfaces for diagnostics and device configuration. There is also an on-chip microcontroller to initialize the device, sequence the DSP, and monitor the link health.

The ASIC has a current-mode logic (CML) driver with CMOS backend as shown in the top of Fig. 7. The transmit block is clocked at half rate, and the receive block is clocked at 1/4 sampling rate.

The ASIC also provides the CAUI-4 host interface to the switch or another host. On the transmit side, 4×25.78125 Gbit/s input data are converted into 2×28.125 Gbaud PAM4 streams with the appropriate FEC overhead. The PAM4 streams drive the silicon photonics modulator via an external driver. The ASIC is capable of a combined PAM4 output or MSB/LSB constituent outputs of the PAM4 signal.

On the receive side, the output of the TIA is sampled in the PAM ASIC using 28 GS/s, 7 bit ADCs with a successive approximation register (SAR) core. The receive DSP has a multi-tap, feed-forward equalizer (FFE) and decision feedback equalizer (DFE) with calibration to recover the PAM4 signal. The equalizer taps are automatically adapted using a least-mean-squared (LMS) algorithm. The FEC decoder then corrects the errors and generates the original Ethernet data.

The PAM4 ASIC generates the clock from the input CAUI-4 data and does not need an internal reference. The bottom of Fig. 7 shows the electrical eye diagrams from the PAM4 ASIC's transmitter block. The NRZ and PAM4 outputs have the same swing and rate (25 Gbaud). The measured electrical signal-to-noise-distortion ratio (SNDR) is better than 33 dB [14].

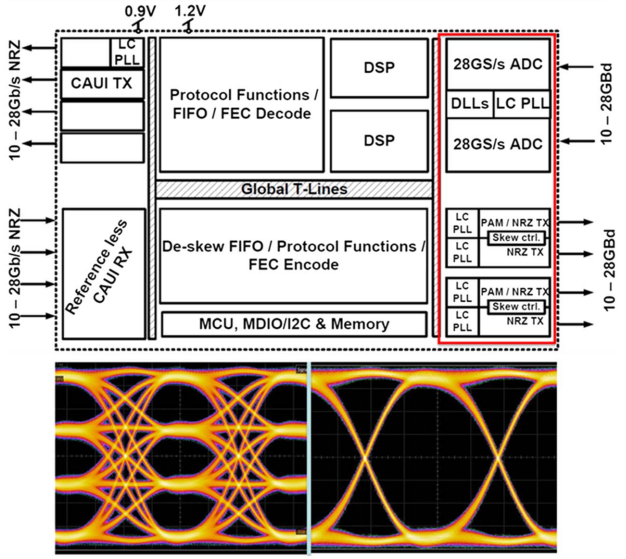


Fig. 7. Architecture and electrical performance of the PAM4 DSP ASIC.

C. Silicon Photonics

InP and Si are the commonly used platforms for large-scale integration of optical components [15], but the Si CMOS platform allows a foundry-level access to the optical component technology at larger (200 mm and 300 mm) wafer sizes [16,17]. Although the Si absorption is in the <1 μm wavelength region, photodetectors in the 1300 nm and 1550 nm wavelength range can be built by adding Ge epitaxy to the standard Si CMOS platform. Further, silica-based components may be integrated to fabricate low index contrast and temperature-insensitive optical components [15,18]. The Si photonics technology is sufficiently mature

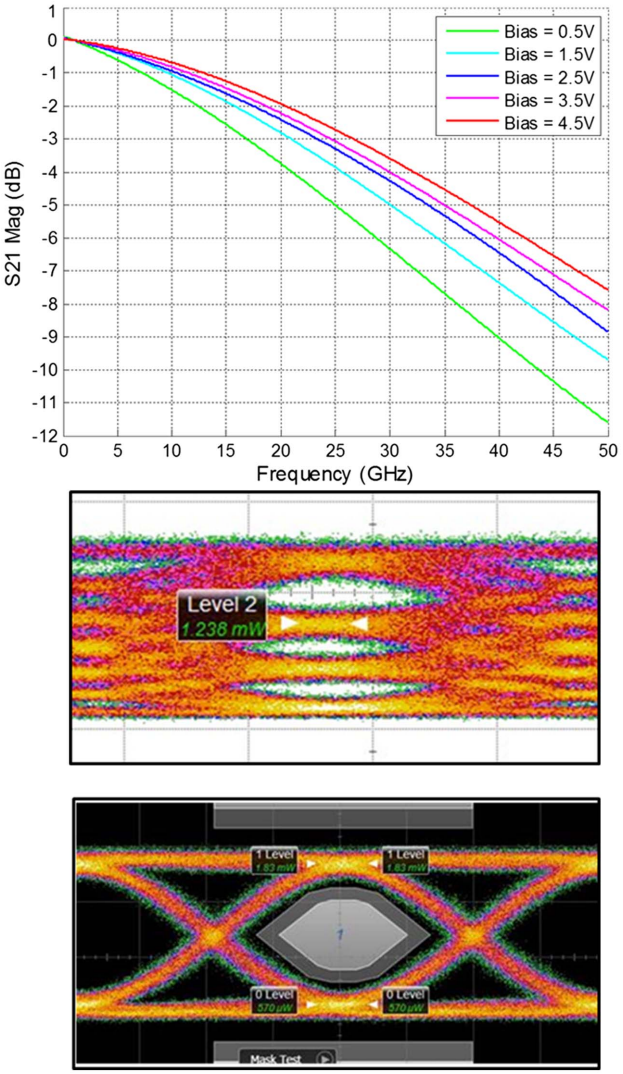


Fig. 8. Performance of the silicon photonics Mach–Zehnder modulator.

and the design methodology well established now that various tools are commercially available in this field [19,20].

The highly integrated Si photonics chip that is at the optical heart of the pluggable module contains a pair of traveling wave Mach–Zehnder modulators (MZM) in the output optical path, one for each wavelength. These are standard depletion-mode CMOS structures [21] with a small signal bandwidth of about 25 GHz as shown on the top of Fig. 8. The two wavelength outputs are then combined on-chip using an integrated 2:1 interleaver that functions as the DWDM multiplexer. The Si photonic circuit schematic is shown in Fig. 5.

The same Si MZM may be used for both NRZ and PAM4 modulation formats, with different drive signals, as shown for the 25.78 Gbit/s eye diagrams on the bottom in Fig. 8. The PAM4 eye diagram was measured after an EDFA. The signal-spontaneous beat noise causes the 11 level to broaden more than the 00 level.

There are a pair of integrated high-speed Ge photodetectors (PD) in the receive path. The PD has a small signal bandwidth of more than 25 GHz as shown in the top half of Fig. 9. The dual wavelength receive signal is separated using a de-interleaver structure that is similar to the one used in the transmitter. The reverse dark current at −2 V bias is typically less than 10 μA at 85°C. This is within the performance requirements for PAM4 link.

Both the MZM driver amplifier and the PD transimpedance amplifier (TIA) are wire bonded to the Si photonics chip. We use a direct-coupled MZM differential

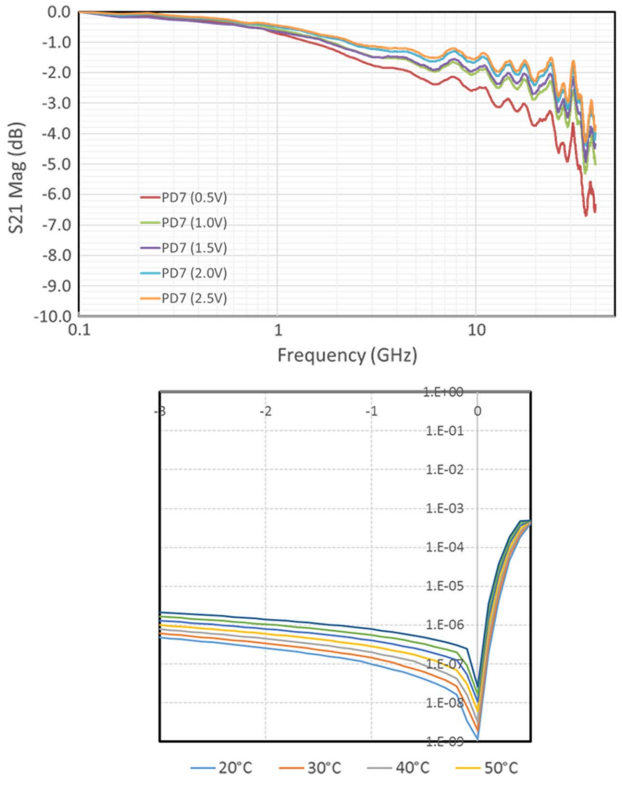


Fig. 9. Performance of the high-speed Ge photodetector.

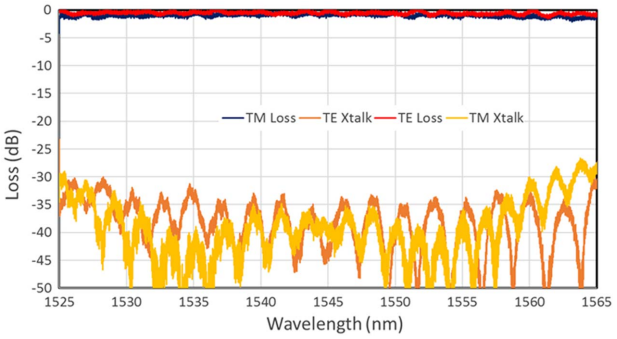


Fig. 10. Extinction ratio and loss of the polarization diverse path in the receiver.

driver configuration to minimize electrical power consumption [22].

The DFB lasers are external and edge coupled to the Si photonics chip. Unlike the vertical grating couplers, edge couplers have a wide optical bandwidth with low insertion loss [23]. The optical signal output and input are likewise fiber coupled to the Si photonics chip.

Since the polarization state of the incoming optical signal is not deterministic, the receive path is designed to be polarization diverse to eliminate polarization-induced signal fading. This is accomplished by using a low-loss polarization beam splitter (PBS) [24]. After the PBS, transverse-electric (TE) and transverse-magnetic (TM) paths are processed independently and then combined again at the PDs. The insertion loss of the PBS is <0.5 dB, and the crosstalk isolation is >25 dB across all of the C band, as shown in Fig. 10.

III. Module and Transmission Performance

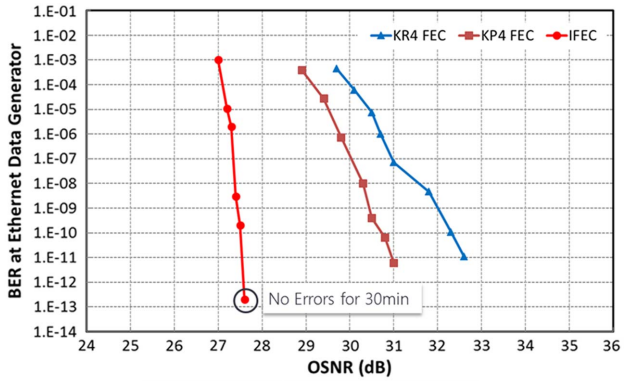
A. FEC Performance

The PAM4 ASIC supports two IEEE standard 802.3bj 100GBASE-KR4 and 100GBASE-KP4 FEC schemes. In addition, it also supports a proprietary IFEC mode. The details of the data rate overheads and the coding schemes for the three FEC modes are in Fig. 11. The ASIC is also capable of operating in the bypass mode, where the host FEC and framing are preserved.

IFEC is a low-power, multi-level, iterative code [25]. Its 10.5 dB coding gain is high compared to the 5–7 dB coding gains for the various optical and copper standard FECs. For the IFEC, the theoretical correction limit bit error rate (BER) is 1E-2, compared to 1E-5 for the KR4 FEC. The relative OSNR performance of the three FEC modes are given in Fig. 11. IFEC has more than 2 dB operating margin at a BER of 1E-3.

B. Dispersion Performance

The PAM4 format, like all optical modulation formats, is susceptible to chromatic-dispersion-induced impairments



FEC Code	Baud Rate (Gbaud)
IFEC	28.125
KP4	26.5625
KR4	25.78125

Fig. 11. Comparative performance of the various FEC modes in the PAM4 ASIC.

during fiber propagation. In coherent transmission systems, the recovery of the phase and amplitude information at the receiver enables the digital signal processor to compensate for the effects of chromatic dispersion, which is largely a linear impairment. The phase information is lost in the much simpler direct-detect PAM4 systems. The intrinsic dispersion tolerance of the 28.125 Gbaud PAM4 format is shown in Fig. 12. It has a ±100 ps/nm window. This is in line with the previously reported data [26]. The typical OSNR penalty in that window is about 0.8 dB.

In general, except for links that are <5 km (within the intrinsic dispersion tolerance of the 28.125 Gbaud PAM4 format) of standard single-mode fiber (SSMF), external chromatic dispersion compensation is needed for PAM4 systems. Figure 13 shows the results of a tunable DCM to compensate for chromatic dispersion impairments up to 80 km of SSMF.

Two implementation cases are discussed in Fig. 13. For 50 km and 80 km links, the case where all the dispersion is compensated for with a tunable dispersion compensation module (TDCM) [11] is compared to the case where a 40 km of fixed DCM is used together with a TDCM. In this test, the complete link configuration with the EDFAs,

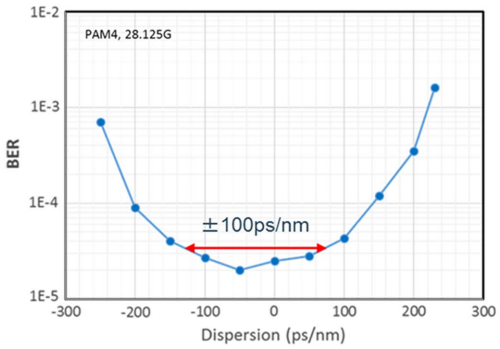


Fig. 12. Dispersion tolerance of the 28 GBaud PAM4 signal.

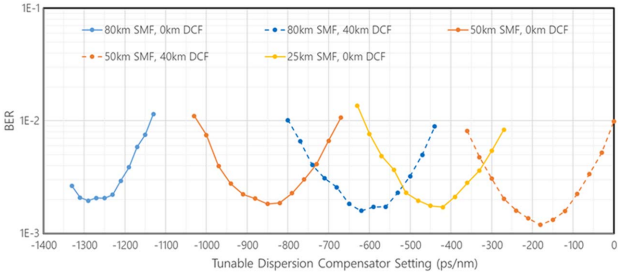


Fig. 13. Use of a single tunable dispersion compensation module up to 80 km links.

external multiplexers, demultiplexers, and TDCM are used. Fiber-Bragg-grating (FBG)-based TDCM is used in this experiment. The performance details of this tunable module may be found in Ref. [11].

At the OSNR used for these measurements, the baseline back-to-back BER is 1.1E – 3. The back-to-back BER with the TDCM set to zero is 1.8E – 3. There is some residual impairment due to the limited bandwidth and phase ripple of the fiber-Bragg-grating-based TDCM. Within the tolerance of the BER measurement, there is negligible penalty for using either dispersion compensation strategies. A single TDCM unit may be used to compensate for all the dispersion impairment due to 80 km of SSMF.

FBGs and etalons both have free-spectral ranges (FSR) equal to the spacing of the optical carriers but differ in slightly their characteristics. Both offer linear dispersion compensation across all channels. In addition to the linear compensation, FBGs offer dispersion slope compensation across the channels. However, this comes at the expense of a filter passband width, which becomes narrower with increasing dispersion compensation. This can have the undesired effect of clipping the edges of the modulated carriers, causing an inter-symbol interference (ISI) penalty at the receiver. Additionally, FBGs generally have higher uncontrolled group delay ripple compared to etalons, which can cause additional penalties. In contrast, etalons apply minimal passband filtering on the modulated signals—their FSR is primarily a phase effect and only has minimal impact on the passband, but they lack the dispersion slope compensating abilities of the FGBs. For the single-span, DCI-edge applications, this is typically sufficient. Additionally, etalons' dispersion compensation range is more limited than FBGs', and symmetric around 0 ps/nm, so for longer spans must be combined with a fixed DCM technology.

The PAM4 format was measured to have <0.8 dB OSNR penalty for polarization mode dispersion (PMD) up to 10 ps.

C. Nonlinear Tolerance

Nonlinear tolerance of the PAM4 format was first investigated in SSMF (Fig. 14). DWDM optical signals experience several nonlinear impairments as a function of launch power during fiber propagation [27]. Several different types of fibers have been developed over the years to mitigate the impact of nonlinear fiber propagation.

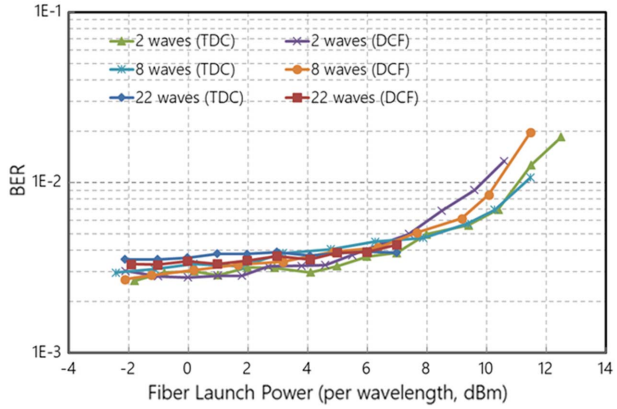


Fig. 14. Nonlinear tolerance of the PAM4 modulation format in SSMF.

In Fig. 14, the channels were running at 28.125 GBaud over 80 km of SSMF. There are two sets of dispersion compensation conditions considered in the experiment. Each set of propagating wavelengths, after 80 km, was compensated for either using a fixed dispersion compensation fiber (DCF) or a tunable DCM. The dispersion compensation was incorporated in the mid-stage of the receiver EDFA. The nonlinear impact on performance starts increasing after about 6–8 dBm launch power per wavelength, irrespective of the number of wavelengths. This shows that the nonlinear interactions between the wavelengths are negligible. The effect is solely governed by the individual launch power, which shows that the direct-detect PAM4 format seems to be primarily impacted by self-phase modulation. This is largely in line with the data that has been reported to date on the nonlinear tolerance of the PAM4 modulation format [26,28].

Then, a series of tests was performed over five widely deployed fiber types to quantify the impact of nonlinearity [29]. The experimental link, shown in Fig. 15, consisted of a single fiber span with an ADVA Corp. line system, comprised of a 100-GHz arrayed-waveguide grating mux/demux, pair of EDFAs, etalon-based TDCM, and optional dispersion compensation fiber (DCF). The system was fully loaded with 40, 100 Gbit/s, QSFP28 transceivers (for a total of 80 wavelengths).

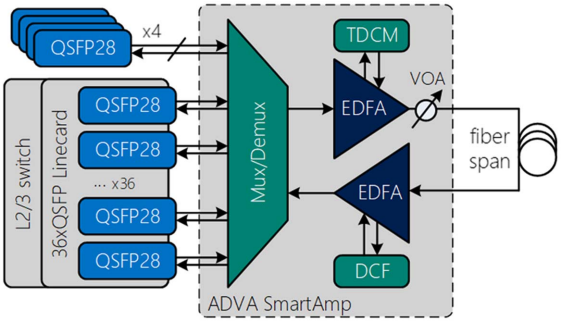


Fig. 15. Experimental setup with PAM4 modules and line system and different fiber types.

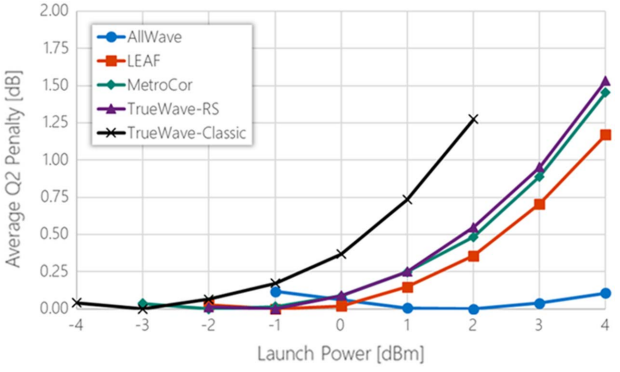


Fig. 16. Nonlinear tolerance of the PAM4 modulation format in multiple fiber types.

Tolerance to fiber nonlinearity was tested over five different fiber types: OFS AllWave, TrueWave-RS, TrueWave-Classic, Corning LEAF, and MetroCor. The spans ranged from 50 to 60 km with 11–13 dB loss (including connectors), such that the average received OSNR was in the same range (~34–35 dB, depending on launch power). For AllWave, a 40 km DCF was used, while for the other types no DCF was used and the TDCM performed all compensation.

Results are shown in Fig. 16 where the average Q^2 penalty across all monitored channels is plotted versus the average launch power per wavelength into the fiber span. There is a small increase of ~1 dB in received OSNR when moving from lowest to highest launch powers. For AllWave, the impact from fiber nonlinearity is negligible up to the maximum launch power of +4 dBm/λ. This nonlinear tolerance is higher than previously reported for similar systems, despite the significantly larger number of co-propagating channels [26]. The other NZ-DSF types all exhibit some nonlinear penalties at higher launch powers. LEAF, MetroCor, and TrueWave-RS all show similar performance, with nonlinear impact becoming significant starting at launch power +2 dBm/λ. TrueWave-Classic shows the strongest nonlinearity due to its small effective mode area and very low fiber dispersion. Significant impact starts at launch power 0 dBm/λ, yet even at this condition, all channels had >1 dB Q^2 margin.

D. 120 km Transmission

For a simple point-to-point transmission system with two EDFAs, the ultimate transmission distance is limited by the launch power at the transmitter. Figure 17 shows the results for a pair of wavelengths (100 Gbit/s combined) for a 120 km link of SSMF. The bottom of Fig. 17 shows the BER versus OSNR curve for the 120 km transmission for a total fiber loss of 26.2 dB. The launch power per wavelength was pushed close to 10 dBm at the transmit EDFA to achieve this nonlinearity limited transmission distance. The FEC correction limit is about 24 dB OSNR. Since the saturation output power of EDFAs typically used in commercial applications is limited to about 24 dBm, the

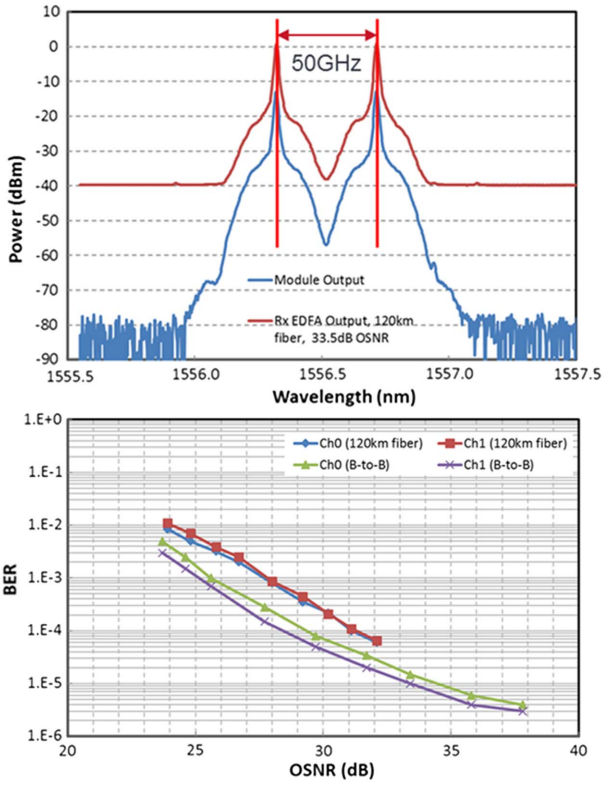


Fig. 17. OSNR performance of the longest link closed to date.

number of wavelengths that can be launched at such high powers will also be limited. There is a trade-off between reach and density in single span links limited by the transmit EDFAs used.

With dispersion compensation, there is ~1 dB OSNR penalty after 120 km transmission. The overlay on the optical spectrum plot (Fig. 17 top) shows the shot-noise-limited output spectrum of the module and the optical spectrum at the output of the receive EDFA. The OSNR at the output of the receive EDFA was 33.5 dB, and the signal was then noise loaded to generate the BER versus OSNR curve. Since the link had excess OSNR, one could have, in principle, increased the overall transmission distance.

E. Live Deployments

There have been several live DCI deployments of systems using various switch and line system configurations. Figure 18 shows a performance snapshot of one such system. In this deployment, there are four fiber pairs between data centers, each carrying 3.2 Tbit/s of data. Including both sides of the link, there is a total of 256 100 Gbit/s pluggable modules in this deployment. The total bisectional data rate between the data centers is 25.6 Tbit/s.

The top chart in Fig. 18 shows the normalized output power for both wavelengths for all 256 modules. The distribution is tight and is less than ±1 dB.

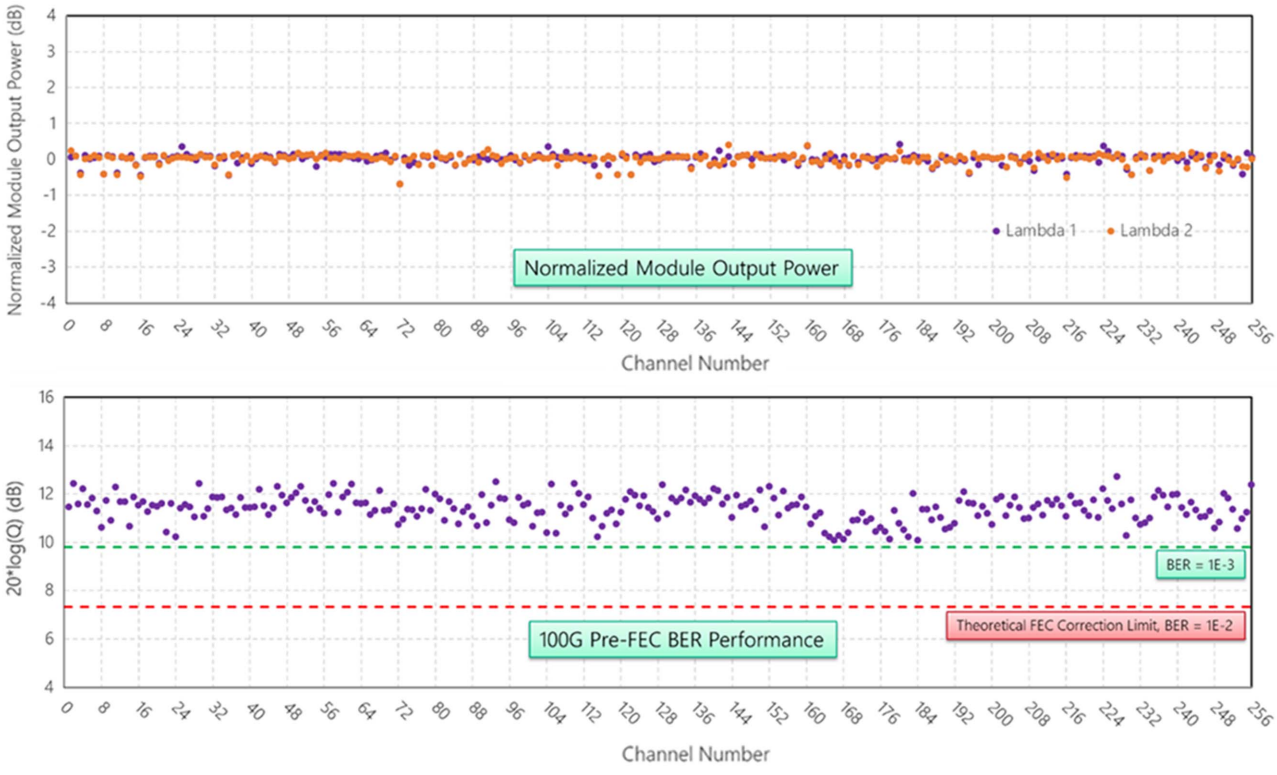


Fig. 18. Live deployment of a DCI-Edge system.

The bottom chart in Fig. 18 shows the pre-FEC BER (equivalent Q^2 values) at the receive end. Pre-FEC BER is one of the diagnostics that can be read directly from the module. Although the FEC correction limit is 1E – 2, the deployed BERs are all better than 1E – 3.

IV. Evolution to Switch-Pluggable 400 Gbit/s DWDM Module

As bandwidth demands in data center networks continue to grow, Moore's law dictates that advances in switching silicon will enable switch and router platforms to maintain switch chip radix parity while increasing capacities per port. At the time of this manuscript's writing, the next generation of switch chips are all targeting per-port capacities of 400 Gbit/s. Accordingly, work has begun to ensure that the optical ecosystem timelines coincide with the availability of next-gen switches and routers.

Toward this end, a project has been initiated in the Optical Internetworking Forum (OIF), currently termed 400ZR [30], to standardize next-gen optical DCI solutions and create and vendor-diverse optical ecosystem. The concept is, in principle, like that of the DWDM PAM4 technology described throughout, but scaled up to support 400 Gbit/s requirements. The trade-offs in spectral efficiency and complexity of this type of approach versus a single-laser, single-modulator coherent approach make using a coherent solution at 400 Gbit/s and beyond the obvious choice, especially when such a coherent solution can be developed to still fit in a faceplate pluggable form-factor at <4 W/100 G.

With 100 Gbit/s serial PAM4 already pushing the bandwidth limits of the supporting electro-optics, it was quickly determined in the OIF that next-gen solutions at 400 Gbit/s would need to leverage coherent technologies. However, the drawbacks of coherent technologies mentioned in Section II dictate that a different approach needs be taken with the sub-120 km application space in mind. A dedicated, DCI-targeted coherent chipset with very limited features (and therefore power consumption) would need to be developed independently from the "swiss-army knife" coherent solutions currently available—namely, a coherent DSP with limited dispersion and PMD compensation, baud rate options, flex-modulation support, and FEC modes. Additionally, to make it viable as a switch-pluggable solution like today's PAM4 modules, it was decided that the optics and electronics (including DSP chip) need to fit in a power envelope of 15 W and package sizes to achieve parity with today's solutions ($\sim \geq 32$ ports per 1RU). Finally, to ensure a robust supplier ecosystem, this technology must all be interoperable (including modulation, framing, pilot tones, and most importantly FEC), which is why the project is being carried out in the OIF.

Work on the OIF Implementation Agreement (IA) is underway at the time of writing, with several of the larger hurdles toward interoperability already passed and good support from the supplier community in validating the assumptions and models around product feasibility. At a high level, the solution will be based on coherent 16 QAM near 60 Gbaud with hybrid SD-FEC achieving moderate net coding gain (NCG), but sufficient to meet the requirements of the application space. Form factor possibilities include OSFP, QSFP-DD, and COBO options, but there is much work to be done there to ensure thermal and signal-integrity concerns around the desired form-factor are fully vetted out in the required time frames.

V. Conclusion

In this paper, we discussed the 100 Gbit/s, DWDM PAM4 data center interconnections that span regional distances of up to 120 km. The combination of Si photonics for the highly integrated optical components and high-speed Si CMOS for signal processing is critical for the implementation of low-cost, low-power, switch-pluggable optical modules that enable massive interconnections between vast regional data center deployments today.

There is work underway currently to extend this concept to low-power 400 Gbit/s switch-pluggable, 16 QAM coherent modules using the next generation Si photonics and Si CMOS.

Acknowledgment

We (RN) thank the extended engineering and operations team and the management at Inphi without whom the work on the module would not have been possible. There were major contributions to the PAM4 ASIC/DSP from Sudeep Bhoja, Pulkit Khandelwal, Jamal Riani, and Karthik Gopalakrishnan. We (MF) also thank the optical network and architecture teams at Microsoft, and in particular, Jeff Cox, for his uncanny vision and follow-through in transforming the DCI landscape. Additionally, thanks to Steven Searcy, Sorin Tibuleac, and the rest of the team at ADVA Optical Networking for the majority of the PAM4 transmission results presented within.

References

[1] M. Russinovich, "Inside Microsoft Azure datacenter hardware and software architecture—Microsoft Ignite," Sept. 2017 [Online]. Available: https://myignite.microsoft.com/videos/54962.

[2] J. Zander, "Cloud infrastructure: enabling new possibilities together—Microsoft Ignite," Sept. 2017 [Online]. Available: https://channel9.msdn.com/Events/Ignite/Microsoft-Ignite-Orlando-2017/GS05.

[3] H. Zhang, B. Jiao, Y. Liao, and G. Zhang, "PAM4 signaling for 56G serial link applications—A tutorial," in *DesignConn*, Jan. 2016.

[4] S. Bhoja, "PAM4 signaling for intra-datacenter and datacenter to datacenter connectivity (DCI)," in *Optical Fiber Communication Conf. and Exhibition/National Fiber Optic Engineers Conf. (OFC/NFOEC)*, Mar. 2017, paper W4D.5.

[5] *400GbE standard*, IEEE 802.3bs, Dec. 2017.

[6] W. Vogels, "The pace of innovation at AWS," in Amazon Web Services Summit, July 2015 [online]. Available: https://d0.awsstatic.com/events/aws-hosted-events/2015/israel/datacenter-innovation.pdf.

[7] "Overview of availability zones in azure" [online]. Available: https://docs.microsoft.com/en-us/azure/availability-zones/az-overview.

[8] ACG Research, "Connecting metro-distributed data centers: the economics and applicability of Inphi's ColorZ™ Technology," Mar. 2017 [online]. Available: http://www.acgresearch.net/wp-content/uploads/2017/03/INPHI_COLORZ_for_DCI_ACG.pdf.

[9] R. Nagarajan, S. Bhoja, and T. Issenhuth, "100 Gbit/s, 120 km, PAM 4 based switch to switch, layer 2 silicon photonics based optical interconnects for datacenters," in *Hot Chips 28 Symposium*, Aug. 2016, paper HC28.23.521.

[10] R. Nagarajan, "100 Gbit/s, switch pluggable, silicon photonics based PAM4 DWDM modules for 4 Tbit/s inter-datacenter links (Invited)," in *CLEO-PR/OECC/PGC/Photonics SG*, Aug. 2017, paper 3-4E-1.

[11] "Data center interconnect: taking the complexity out of dispersion management," TeraXion/Inphi Webinar, Oct. 2017 [online]. Available: http://www.teraxion.com/en/events, http://www.teraxion.com/images/stories/pdf/cs-tdcmx-sm-datasheet_cs-tdcmx-sm1.1.pdf.

[12] P. Khandelwal, J. Riani, A. Farhoodfar, A. Tiruvur, I. Hosagrahar, F. Chang, J. Wu, K. Gopalakrishnan, S. Herlekar, and S. Bhoja, "100 Gbps dual-channel PAM-4 transmission over datacenter interconnects," in *DesignConn*, Jan. 2016.

[13] F. Chang, S. Bhoja, J. Riani, I. Hosagrahar, J. Wu, S. Herlekar, A. Tiruvur, P. Khandelwal, and K. Gopalakrishnan, "Link performance investigation of industry first 100G PAM4 IC chipset with real-time DSP for data center connectivity," in *Optical Fiber Communication Conf. and Exhibition/National Fiber Optic Engineers Conf. (OFC/NFOEC)*, Mar. 2016, paper Th1G.2.

[14] K. Gopalakrishnan, A. Farhood, A. Ren, A. Tan, A. Tiruvur, B. Helal, C. Loi, C. Jiang, H. Cirit, I. Quek, J. Riani, J. Gorecki, J. Pernillo, J. Wu, L. Tse, M. Le, M. Ranjbar, P. Khandelwal, R. Narayanan, P. Wong, R. Mohanavelu, S. Bhoja, S. Herlekar, and V. Shvydun, "A 40/50/100 Gb/s PAM-4 ethernet transceiver in 28 nm CMOS," in *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, Feb. 2016, p. 3.4.

[15] R. Nagarajan, C. Doerr, and F. Kish, "Semiconductor photonic integrated circuit transmitters and receivers," in *Optical Fiber Telecommunications VIA: Components and Subsystems*, I. Kaminow, T. Li, and A. Willner, Eds., 2013, pp. 25–88.

[16] N. Izhaky, M. Morse, S. Koehl, O. Cohen, D. Rubin, A. Barkai, G. Sarid, R. Cohen, and M. Paniccia, "Development of CMOS-compatible integrated silicon photonics devices," *IEEE J. Sel. Top. Quantum Electron.*, vol. 12, no. 6, pp. 1688–1698, 2006.

[17] T. Liow, K. Ang, Q. Fang, J. Song, Y. Xiong, M. Yu, G. Lo, and D. Kwong, "Silicon modulators and germanium photodetectors on SOI: monolithic integration, compatibility, and performance optimization," *IEEE J. Sel. Top. Quantum Electron.*, vol. 16, no. 1, 307–315, 2010.

[18] T. Tsuchizawa, K. Yamada, T. Watanabe, S. Park, H. Nishi, R. Kou, H. Shinojima, and S. Itabashi, "Monolithic integration of silicon-, germanium-, and silica-based optical devices for telecommunications applications," *IEEE J. Sel. Top. Quantum Electron.*, vol. 17, no. 2, pp. 516–525, 2011.

[19] L. Vivien and L. Pavesi, Eds., *Handbook of Silicon Photonics*, CRC Press, 2013.

[20] L. Chrostowski and M. Hochberg, *Silicon Photonics Design*, Cambridge University Press, 2015.

[21] G. Reed, G. Mashanovich, F. Gardes, and D. Thomson, "Silicon optical modulators," *Nat. Photonics*, vol. 4, pp. 518–526, 2010.

[22] C. Pobanz, "Direct coupled driver for Mach-Zehnder optical modulators," U.S. patent 8,948,608 (Feb. 3, 2015).

[23] K. Shastri, "CMOS photonics," in *Asia Communication and Photonics Conf.*, Nov. 2009, paper TuR3.

[24] J. Zhang, H. Zhang, S. Chen, M. Yu, G. Lo, and D. Kwong, "A polarization diversity circuit for silicon photonics," in *Optical Fiber Communication Conf. and Exhibition/National Fiber Optic Engineers Conf. (OFC/NFOEC)*, Mar. 2011, paper JThA19.

[25] A. Farhood, "Optimal unipolar PAM solutions for 100G SMF link from channel capacity perspective," Sept. 2012 [online]. Available: http://www.ieee802.org/3/bm/public/sep12/farhood_01_0912_optx.pdf.

[26] N. Eiselt, J. Wei, H. Griesser, A. Dochhan, M. Eiselt, J. Elbers, J. Olmos, and I. Monroy, "First real-time 400G PAM-4 demonstration for inter-data center transmission over 100 km of SSMF at 1550 nm," in *Optical Fiber Communication Conf. and Exhibition/National Fiber Optic Engineers Conf. (OFC/NFOEC)*, Mar. 2016, paper W1K.5.

[27] G. Agrawal, *Nonlinear Fiber Optics*, 5th ed., 2012.

[28] S. Yin, T. Chan, and W. Way, "100-km DWDM transmission of 56-Gb/s PAM4 per λ via tunable laser 0061nd 10-Gb/s InP MZM," *IEEE Photon. Technol. Lett.*, vol. 27, no. 24, pp. 2531–2534, 2015.

[29] M. Filer, S. Searcy, Y. Fu, R. Nagarajan, and S. Tibuleac, "Demonstration and performance analysis of 4 Tb/s DWDM metro-DCI system with 100G PAM4 QSFP28 modules," in *Optical Fiber Communication Conf. and Exhibition/National Fiber Optic Engineers Conf. (OFC/NFOEC)*, Mar. 2017, paper W4D.4.

[30] "OIF 2016.463, 400ZR interoperability," [online]. Available: https://www.oiforum.com/get/48077.

Radhakrishnan Nagarajan (S'85–M'92–SM'97–F'08) received a B.E. (first class Hons.) in Electrical Engineering from the National University of Singapore, Singapore, in 1986, a M.E. in Electronic Engineering from the University of Tokyo, Tokyo, Japan, in 1989, and a Ph.D. in Electrical Engineering from the University of California, Santa Barbara, in 1992. He has been Chief Technology Officer, Optical Interconnect, of Inphi, Santa Clara, since June 2013. Prior to joining Inphi, he was with Infinera, as a Fellow, working on the design, development, and commercialization of large-scale photonic integrated circuits. From 1995 until 2001, he was with SDL (acquired by JDSU), where he managed the development of the new generation 980 nm EDFA pump module, which won the Photonics Circle of Excellence Award in 2000. He has authored/coauthored more than 185 publications in journals and conferences and four book chapters mainly on high-speed optical components. He has been awarded 135 U.S. patents. He is a Fellow of The Optical Society and a Fellow of the Institution of Engineering and Technology. He is the recipient of the 2006 IEEE LEOS Aron Kressel Award for his contributions to commercializing LS PICs. He has also in the past served as a Guest Editor of the Optical and Quantum Electronics and Applied Optics journals.

Mark M. Filer (M'08) received a B.S. and M.S. in Electrical Engineering (with highest honors) from the Georgia Institute of Technology, Atlanta, Georgia (USA) in 2000 and 2006, respectively. He has been in the role of Optical Network Architect, Azure Networking at Microsoft in Redmond, Washington (USA) since November 2014. Prior to joining Microsoft, he was at ADVA Optical Networking in Atlanta, Georgia (USA) working in DWDM system architecture, transmission, and engineering rules. He has authored or coauthored more than 50 publications in

journals and conferences in the areas of long-haul transmission, ROADM network architectures, and system impairments focused on nonlinear effects and crosstalk, and he has been awarded two US patents. He is a member of the IEEE and The Optical Society.

Yang Fu received a B.S. in Microelectronics from Peking University, Beijing, China in 2008, and a Ph.D. in Electrical Engineering from University of Virginia, Charlottesville, Virginia in 2012. He has been in the role of Optical Engineer at Inphi in Santa Clara, California since September 2014. Prior to joining Inphi, he was with JDSU in Milpitas, California working on integrated coherent receivers. His research work was focused on high-speed, high-power photodiodes for analog optic links. He has authored or co-authored more than 20 publications in journals and conferences.

Masaki Kato received a B.S., M.S., and Ph.D. in Electronic Engineering from the University of Tokyo, Tokyo, Japan in 1994, 1996, and 1999, respectively. In 1999, he joined the Department of Electrical Engineering, University of Tokyo, as a Research Associate, where he studied semiconductor optical devices and their applications for wavelength conversion/all-optical switching. In 2002, he joined Infinera Corporation, Sunnyvale, CA, as a Technical Staff Member, where he was involved in the development of LS PICs. In 2013, he joined Inphi Corporation, Santa Clara, CA, where he has been involved in the development of silicon photonics. He has authored/coauthored more than 60 journals and conferences. He has been awarded 60 U.S. patents. He is a member of the Japanese Society of Applied Physics.

Todd Rope received a B.S. in Physics from the California Institute of Technology in 1995. He is Senior Director of Software Architecture for Inphi. Prior to joining Inphi he was Chief Software Architect at Source Photonics, Vice President of Technology at MRV Communications and President of Zuma Networks. He has been designing and developing with pluggable optical modules since the late 1990s. Todd Rope has authored several articles on pluggable technology, and has been awarded 12 US patents.

James Stewart (M2007–SM2016) received his BS in mechanical engineering from the University of California Santa Barbara in 1990. He joined InPhi in 2014 and is the Associate Vice President of Engineering for optical interconnects. Prior to Inphi, he held management positions in optical module development at Finisar, Infinera and JDSU. James Stewart is a senior member of the OSA. He has authored numerous papers for journals and conferences and has been awarded over 25 US patents in the fields of high-power LEDs and optical communications.