Low Power DSP-Based Transceivers for Data Center Optical Fiber Communications (Invited Tutorial)

Radhakrishnan Nagarajan, *Fellow, IEEE, Fellow, OSA*, Ilya Lyubomirsky, and Oscar Agazzi, *Life Fellow, IEEE*

(Tutorial)

***Abstract*—In this tutorial, we discuss the evolution of the technology deployed for optical interconnects and the trade-offs in the design of low complexity, low power DSP and implementation for direct detect and coherent, pluggable optical modules for data center applications. The design trade-offs include the choice of modulation format, baud rate, optical link design, forward error correction, signal shaping and dispersion compensation.**

***Index Terms*—Digital signal processing, forward error correction, optical fiber communication, optical fiber networks, optical interconnections.**

I. Introduction

THE use of digital signal processing (DSP) in optical links has about a 20-year history [1]–[4]. To use the powerful DSP techniques, the analog optical signal had to be digitized first. This required a low power, high-speed ADC (analog to digital converter) at the receiver. This, together with the development of a high-speed DAC (digital to analog converter) at the transmitter for signal conditioning, enabled the growth of the modern-day high-speed optical interconnects.

The first use of DSP was in MLSE (maximum likelihood sequence estimation) implementation for chromatic dispersion compensation in intensity modulated direct detect (IMDD) 10 Gbit/s systems [2], [3]. This was not very power efficient and provided only limited mitigation of chromatic dispersion for long-haul applications; hence MLSE was not deployed widely in 10 Gbit/s IMDD systems. The development of high-speed coherent receivers enabled the linear detection of both signal amplitude and phase. The use of DSP in coherent applications to compensate for a variety of linear optical impairments, as well as enabling QAM (Quadrature Amplitude Modulation), was a very powerful innovation which led to the widespread deployment of DSP hardware in optical links [1], [3].

In 2015, Inphi was the first to develop and commercialize a PAM4 (4 level Pulse Amplitude Modulation) DSP for direct detect (DD), 100G, intra data center (DC) applications [5], [6]. Based on this DSP, and a highly integrated silicon photonics platform, we successfully developed and deployed an inter data center (DCI), 100G switch pluggable module for 80 km reach [7]. Recently, we have developed a low power coherent DSP, in the 7 nm CMOS node, and successfully demonstrated a silicon photonics based 400ZR coherent pluggable module for DCI applications [8], [9].

Manuscript received March 14, 2021; revised May 6, 2021; accepted June 10, 2021. Date of publication June 16, 2021; date of current version August 30, 2021. *(Corresponding author: Radhakrishnan Nagarajan.)*

The authors are with the Marvell Semiconductor, Inc., San Jose, CA 95134, USA (e-mail: radha@marvell.com; ilyubomirsky@marvell.com; oagazzi@marvell.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JLT.2021.3089901.

Digital Object Identifier 10.1109/JLT.2021.3089901

II. Evolution of Transport Technology

Fig. 1 shows a bifurcated technology and product space for the 100G data center interconnect deployment. Intra (inside) data center (DC) optical interconnects were of the NRZ (Non-Return to Zero) format, based on analog CDR (Clock and Data Recovery). The intra DC deployment of 100G, in volume, started in 2015 [10]. Inter (between) DC interconnects, for up to 100 km, were a combination of IMDD PAM4 pluggable modules, which we introduced in 2017 [7], and a variety of coherent formats and rates in compact modular boxes, from different of vendors [11].

Specialized hardware [11] and lower complexity coherent formats, like 8QAM and QPSK, are commercially deployed for longer reaches.

Fig. 1 shows the nomenclature used to differentiate optical reach for the various transmission formats. Intra DC is nominally less than 2 km. Up to 10 km is labeled LR (long-reach) in IEEE standards. These are CWDM (O band) links, where the chromatic dispersion is a minimum in the standard single mode fiber (SSMF). We have used the prefix DCI (data center interconnect) in the tutorial. There is no single accepted definition for the "Campus/Edge" reach, other than it is outside the physical DC building. The definitions vary between the different network configurations.

Metro (metropolitan) and long-haul designations are used for backbone connections between large geographic distances.

As shown in Fig 2, DSP-based optical interconnects are being deployed at 200G/400G intra DC applications. This transition was in tandem with the evolution of the port level data rates in the DC switches from 25G NRZ to 50/100G PAM4 (25/50 Gbaud) [12].

At 400G, for inter DC applications, there was an opportunity for a common coherent pluggable module hardware to span a variety of distances as shown in Fig. 2. We designed a multi-format, multi-rate, wide-reach DSP ASIC, with a common

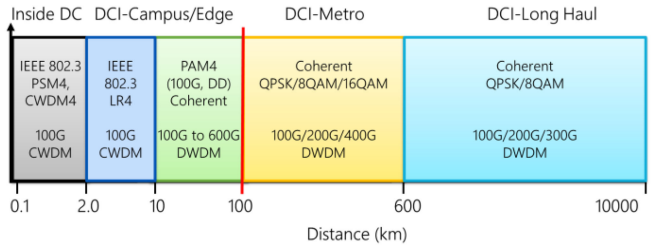

Fig. 1. Reach and transmission formats for a variety of DC links at the time of 100G deployment. These are not strict demarcations. These have also evolved over time and vary between service providers and fiber plants.

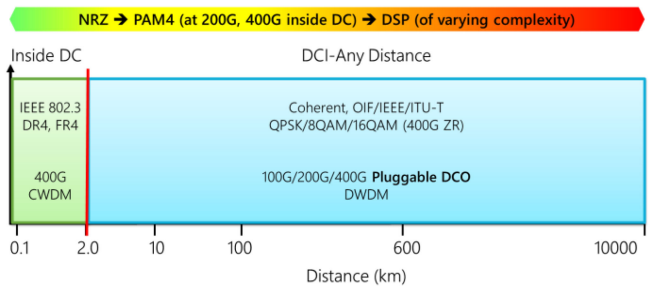

Fig. 2. Evolution of the DCI reach and technology configuration for the 200G/400G deployments. The color grading shows the DSP complexity for the various applications and reach (with green being the least complex and red being the most).

silicon photonics optical subsystem to accomplish this, without much compromise in performance or power consumption.

The transmission format and FEC (forward error correction) have been standardized in the coherent 400ZR [13], to enable an eco-system of inter-operable module and system vendors. Although a similar module form factor may be used, for longer reaches, higher functionality FEC, and a larger chromatic and polarization mode dispersion compensation will be needed.

A good review of the geographic reach, diversity of DCI, with a variety of span lengths and system deployments may be found in Ref. [14].

The commonly used pluggable module form factors, for 100G and above, are shown in Fig. 3. The QSFP-28 (Quad Small Form-Factor Pluggable, with 4 input electrical lanes) 100G PAM4 and QSFP-DD 2 (QSFP-Double Density, which has 8 input electrical lanes, and 2 in the name refers to the longer version of the QSFP-DD form factor) 400ZR coherent DCI modules, have the same width, hence the faceplate density in a switch, at 4x the data rate. The OSFP (Octal Small Factor Pluggable) module is also used for 400G inter and intra DC applications.

A 1RU (1 rack unit is 1.75" in height) switch chassis, in a 19" (width) rack, will accommodate 32 (which is a common switch radix in data centers) [12], [15] modules of the QSFP, QSFP-DD and OSFP form factors. CFP2 [16] (C Form-Factor Pluggable) is a wider module form factor and can dissipate more power than the QSFP-DD or OSFP modules. They are generally found in systems for enterprise and service provider market segments.

Fig. 4, shows a power consumption trend for MSA, CFP and CFP2 coherent modules over time, normalized to 100G [17]. The data are for larger modules (lower card density than the QSFP-28 or DD modules). They have higher power consuming

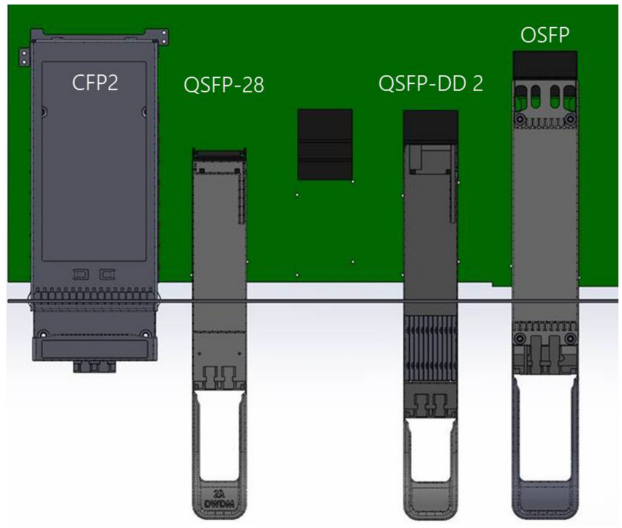

Fig. 3. Commonly used pluggable module form factors for data center optical interconnects.

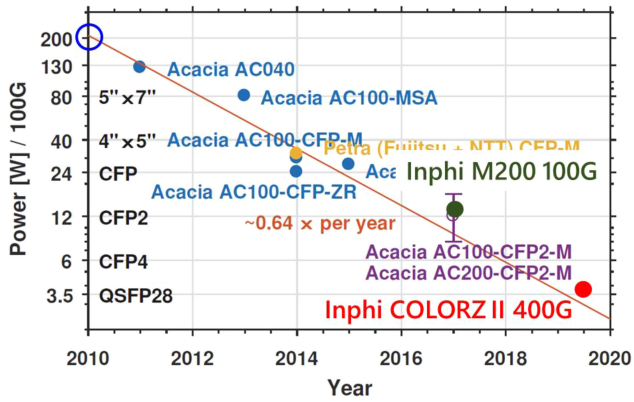

Fig. 4. Adapted from Ref [17]. The power consumption for our 100G, and 400G modules are overlaid on the data from other manufacturers as in [17].

DSP's from earlier CMOS generations. In the latest generation, we have been able to design a 400G module in a much smaller QSFP-DD form factor, with a 7 nm DSP, dissipating 4W/100G typically, at 16W for the 400ZR module.

The factor 50 reduction in power consumption, in a coherent module, over the last 10 years is simply amazing! This partly due to the lower complexity DSP for DC applications and the ever-shrinking CMOS nodes. The data in Fig. 4, is also in line with other published data [18].

Fig. 5 shows a nominal breakdown of power consumption for a 400ZR module [19]. Although this may vary between different vendors, we find this is to be generally true for a variety of DSP based modules (for both IMMDD as well as coherent). The power breakdown is also in line with other published data [20]. The DSP ASIC power is nominally about 50% of the module. Since the optical modules typically operate from a 3.3 V supply, and most high-end ASIC's operate at supply voltages less than 1V, there is a power conversion loss (segment labeled "power overhead") that is between 10% and 15%.

For the DSP power to be nominally 50% of the total module power, the complexity of the DSP and optics need to scale

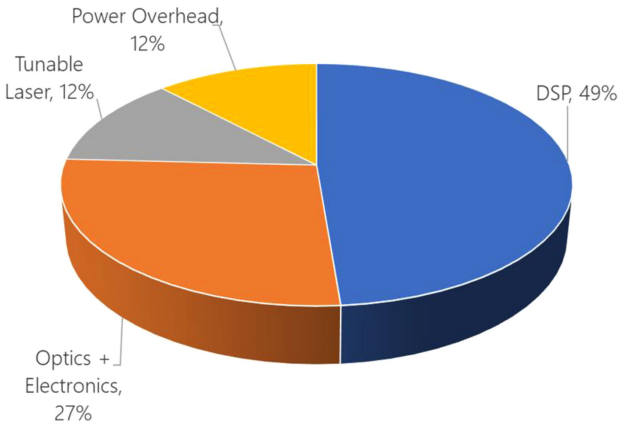

Fig. 5. Normalized power breakdown for the 400ZR module for the various internal blocks.

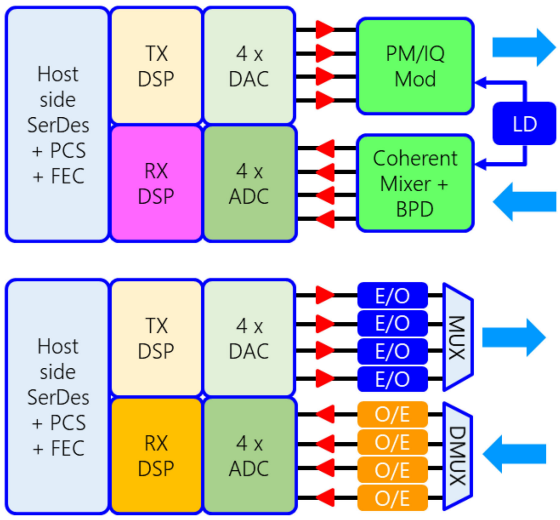

Fig. 6. Similarities between a 16QAM coherent interconnect for inter DC and DR4/FR4 PAM4 interconnect for intra DC applications.

with applications. For coherent applications, the laser is full C-band tunable, and its power consumption is in the slice labeled "tunable laser". In the 100G PAM4 DWDM applications, we used fixed wavelength lasers in the C-band which consume lower power [7].

The slice labeled "optics + electronics" has the modulator driver, TIA (transimpedance amplifier), and the integrated silicon photonics chip which has the high-speed Mach Zehnder modulators and photodetectors.

III. Low Complexity DSP Design Methodology

As the starting point for reducing the DSP complexity, we have compared the block diagrams for the implementation of the PAM4 and 16QAM (which is essentially PAM4 in the I and Q dimensions) modulation formats in Fig. 6 [21]. At 100 Gbit/s per lane (/per dimension), the baud rates for the two systems are similar, although 16QAM typically requires slightly higher overhead for pilots and 400ZR framing.

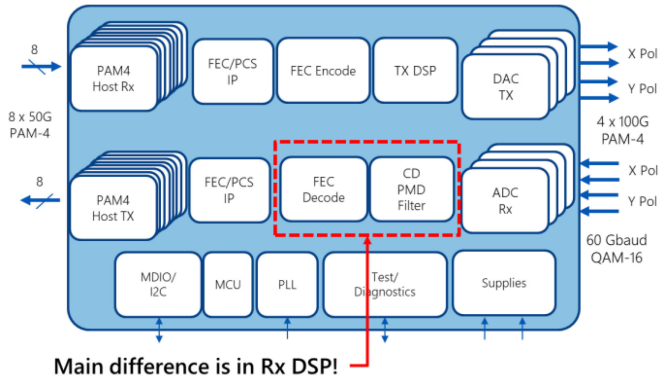

Fig. 7. Simplification of the receiver DSP block is the key to power reduction.

The host side SerDes (serializer/de-serializer), FEC and PCS (physical coding sublayer) interface to the switch or some other host system is identical for both implementations. The line or optical output of the transmit (TX) path are 4 streams of PAM4 data that are either individually encoded using Electrical to Optical converters and then optical multiplexed or optically converted to generate a single carrier, 16QAM signal using a polarization multiplexed IQ (PM/IQ) modulator structure. The TX DSP path, for all practical purposes, is identical.

The complexity and main differences between PAM and coherent DSP are in the receive path, for chromatic dispersion (CD) and polarization mode dispersion (PMD) compensation, frequency and polarization tracking, and I/Q decoding. Power goes up with transmission distance, as CD and PMD go up. There are also significant differences in the timing recovery schemes due to differences in data path latency. The key to a lower complexity DSP is the simplification of these blocks.

The first level details of the TX and RX DSP blocks are in Fig. 7 [21]. The FEC block needs to be optimized as well for a low power DSP. The CFEC (concatenated FEC) we developed for ZR is optimized for the highest NCG (net coding gain) at the lowest power consumption for fixed FEC overhead [22]. CFEC is based on concatenating a Staircase hard decision outer code with a simple Hamming soft decision inner code. In intra DC applications, the FEC latency also needs to be low, driving different FEC architectures. One way to achieve lower latency in the concatenated architecture is to use a Reed-Solomon FEC, such as IEEE standard RS (544,514), for the outer code, while keeping the Hamming soft decision inner code [21].

Fig. 8 shows the simplified PAM4 RX DSP architecture for a single lane. A baud rate ADC with baud rate timing recovery are used to keep the power consumption as low as possible. This requires the digital timing recovery loop to have a low enough latency to achieve a CDR bandwidth of 4 MHz or better. The key analog component in the PLL is the voltage-controlled oscillator (VCO); the VCO must have low jitter to meet the IEEE 802.3 specifications on receiver jitter tolerance (JTol). The PLL circuitry typically consumes a small fraction ~10–20% of the total Rx analog power.

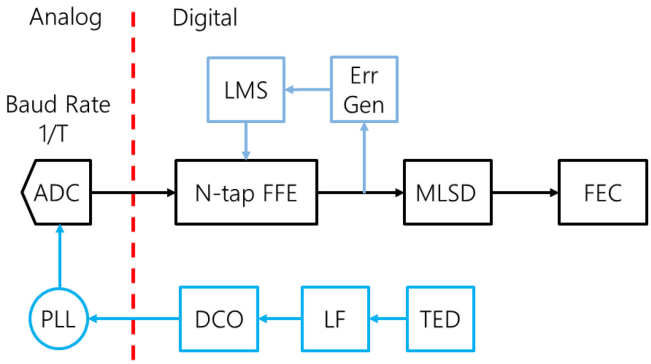

Fig. 8. Simplified receiver architecture for a PAM4 DSP.

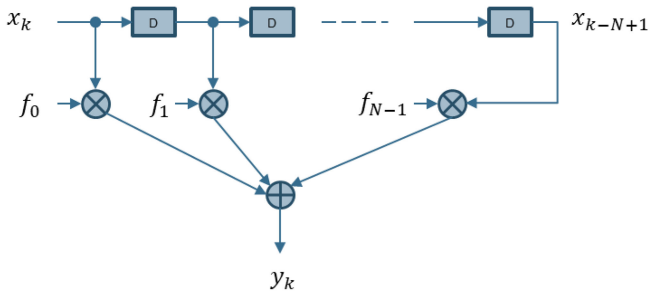

Fig. 9. Simplified block diagram for an N-tap digital feedforward equalizer.

A. PAM4 DSP Architecture

Most of the heavy lifting in equalizing the channel bandwidth limitations is performed by a linear feed forward equalizer (FFE) with 10-30 taps (Fig. 9). The FFE is typically a major contributor to the DSP power consumption in PAM4 receivers, and requires a careful digital design, e.g., exploiting properties of the channel response to minimize complexity/power. For the toughest channels, an additional maximum likelihood sequence detector (MLSD) [2], [23] can significantly improve performance. MLSD is especially useful in direct detect PAM4 systems to mitigate CD penalty. Low power MLSD designs are critical and leverage classic DSP design techniques such as reducing the memory of target response appropriately and set partitioning to simplify the branch metric computations.

B. Simplified Coherent DSP Architecture

Details of a conventional coherent DSP architecture for traditional coherent applications are discussed in (see Figs. 1 and 2) [24]. Fig. 10 shows a simplified Rx DSP architecture which may be used for system application in $<$10 km intra data center interconnects [21].

In contrast to direct detect systems where only the signal intensity needs to be detected, coherent systems must detect the optical signal amplitude, phase, and polarization. This drives additional complexity for coherent DSP, including a MIMO (multiple input multiple output) equalizer for demultiplexing the 2 optical polarization states while equalizing PMD, and carrier frequency compensation (CFC) and carrier phase compensation (CPC) blocks.

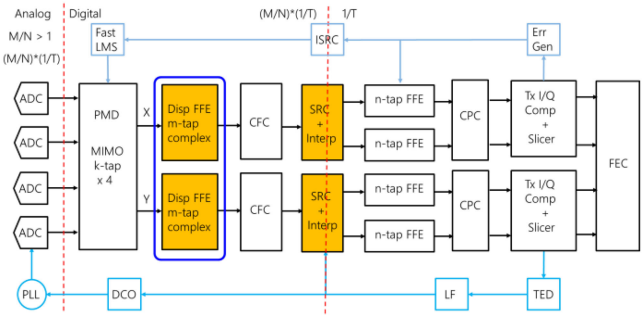

Fig. 10. Simplified DSP architecture for coherent receivers.

The architecture of Fig. 10 differs in some important respects from that of coherent transceivers for more traditional applications such as metro and long haul [24], [25]. For example, typically the CD equalizer is located before the MIMO equalizer because, among other reasons, the CD equalizer would introduce a large latency in the adaptation loop of the former. However, in intra data center interconnects the CD is low and the loop latency is not a significant consideration. Placing the MIMO equalizer at the input enables it to compensate not only PMD and polarization rotation, but also impairments of the RX optical and analog front ends, such as I/Q skew, quadrature and amplitude errors. Notice that the TX I/Q skew, quadrature, and amplitude errors must be compensated after all the channel impairments have been compensated. That is the reason the block labeled "TX I/Q Comp + Slicer" is located at the tail end of the DSP. It is also interesting to mention that the FFE in Fig. 10 does not have the traditional "butterfly" architecture, since at this point in the receiver data path the PMD and the CD have been compensated, and the I and Q components are essentially independent. The FFE serves primarily the purpose of compensating the bandwidth limitations of the analog and optical front ends.

Finally, the CFC and CPC blocks are typically based on pilot symbols, although algorithms such as PLLs supplemented by Blind Phase Search (BPS) and Maximum Likelihood (ML) could also be used.

Pilot symbols are used in 400ZR standard, to simplify the CFC/CPC implementations. For CPC, a secondary clean up stage (based on Maximum Likelihood estimation) is useful to enable low cost lasers with larger linewidth. The relatively low dispersion in intra-data center links can be efficiently compensated by short time domain complex FFE.

To simplify timing recovery, an oversampled ADC is used with all-digital timing recovery implemented using digital interpolating techniques in the sampling rate converter (SRC) block. An oversampling of $\sim$20% is typically needed to enable efficient implementation of the SRC digital interpolation filters. The slicer block may also include a final "clean up" stage of the QAM constellation by compensating some of the transmitter IQ modulator imperfections, efficiently implemented using decision directed adaptation. The MIMO and complex FFE blocks may be adapted "blindly" using CMA (constant modulus algorithm) in the bring-up mode, switching to decision directed LMS (Least Mean Square Error) in the operation mode.

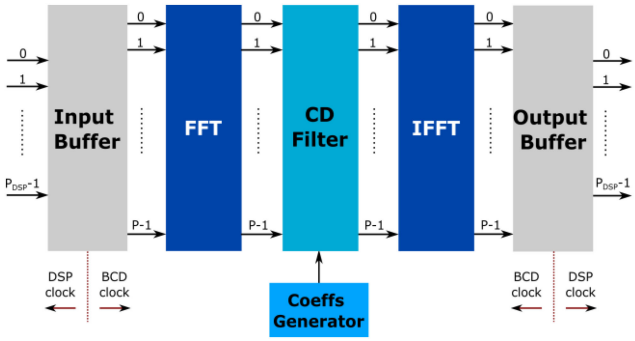

Fig. 11. Block diagram for bulk chromatic dispersion compensation.

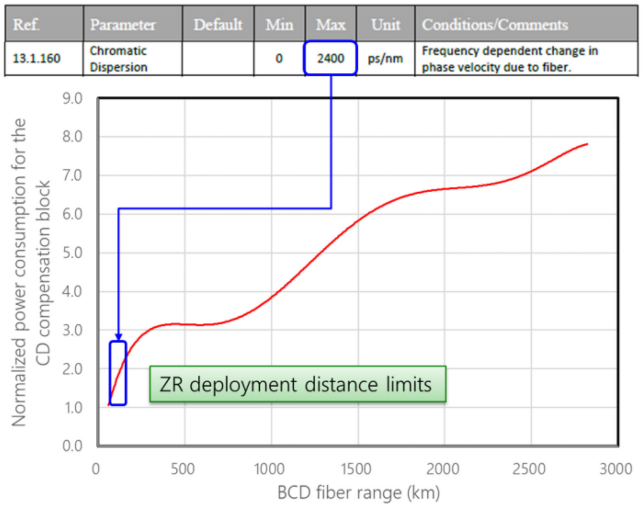

Fig. 12. Normalized power consumption for CD compensation as a function of fiber length.

When the dispersion values are large, e.g., metro DWDM networks, the compensation is more efficiently done in the frequency domain. As shown in Fig. 11, this DSP circuit is for an FFT/IFFT operation of various array sizes depending on the amount of dispersion that needs to be compensated. This bulk chromatic dispersion (BCD) compensation block is usually inserted in addition to the FFE block in the DSP (highlighted by the blue box in Fig. 10). A large fraction of the DSP power consumption comes from the BCD block.

In the ZR deployment distance range, 120 km at 20 ps/nm/km worst case chromatic dispersion (CD) in SSMF in the C band for a total of 2400 ps/nm of CD. The amount of CD compensation that can be handled by the FFE block alone is a function of the span length (Fig. 12).

BCD power consumption goes up quite dramatically with distance. Fig. 12 shows the measured power consumption (normalized) with BCD fiber range. The steps in the power consumption curve occur when there is a change in the FFT/IFFT block size with the CD compensation levels. Elimination of the BCD block simplifies the DSP for DC applications and reduces the overall power consumption in inter DC applications [26]. Ref. [24] has a comprehensive discussion of the implementation complexity of the BCD equalizer.

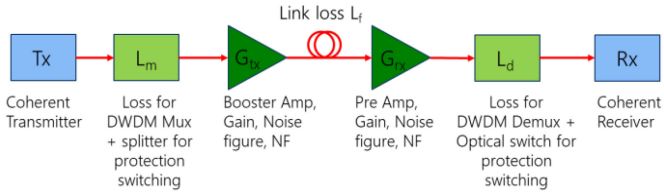

Fig. 13. 400ZR, single span, coherent reference link model.

IV. 400ZR Standardization and Switch Pluggable Silicon Photonics Coherent Modules

OIF (optical internetworking forum) 400ZR [13] was the first attempt to develop an inter-operable coherent transmission standard in the industry. It involved several industry partners working over 3 years to bring to fruition.

It's stated goals (excerpts) are, "… to enable inter-operable, cost-effective, 400Gb/s implementations based on single-carrier coherent DP-16QAM modulation … low power DSP … a Concatenated FEC (C-FEC) with a post-FEC error floor <1.0E-15 …".

A. 400ZR Modulation Format

There are number of choices for implementing the dual polarization single carrier 400G signal. Shannon's theorem that looms large in information theory is recast, in terms of spectral density (information rate, C, divided by the signal bandwidth, B) in Eqn. 1. Scaling the symbol rate as the channel bandwidth has no impact on the SNR (signal to noise ratio) or a linear impact on OSNR (optical SNR which is defined with respect to a fixed noise bandwidth) of the signal, while increasing the complexity of the modulation format leads to an exponential increase in the SNR [24], [27].

$$S/N = 2^{\left(C/B\right)} - 1 \tag{1}$$

60 Gbaud, 16QAM was chosen as a compromise between the analog bandwidth needed in the components and the OSNR needed in the line system [13], [27].

B. ZR Reference Link Model

Fig. 13 shows the reference link model for the inter DC 400ZR applications. It is a single span, nominally a 64-wavelength system at 75 GHz channel spacing, for a total fiber capacity of 25.6 Tbit/s per fiber pair for maximum span reach of 120 km. There are other variants of this system being implemented in practice.

Table I shows the nominal system parameters assumed to model this link. The output power of the ZR modules, per wavelength, is −10 dBm.

Some line system implementations may have a 3 dB power splitter at the transmitter and an optical switch at the receiver for line protection switching. The power specification at the receiver, where the link had to close, was −12 dBm. The booster and receive EDFA's (erbium doped fiber amplifiers) were assumed to have a NF's (noise figures) of 6 dB.

TABLE I
SYSTEM PARAMETERS USED TO MODEL THE LINK IN FIG. 13

Parameter	Value
Transmitter module output power	> -10 dBm (spec)
Pout in the line (after the booster EDFA)	< 3 dBm/λ
Booster EDFA Psat (64λ's)	~ 21dBm
Booster EDFA Gain, Gtx	~ 22 dB
Additional transmit losses, Lm	9 dB
Mux Loss < 5.5 dB	
Protection switching splitter < 3.5 dB	
Booster, Receive EDFA, NF	6 dB
Lf, 0.2 dB/km*Distance	24 dB (spec)
Gain Ripple	±0.5 dB
Other line OSNR penalties	< 1.0 dB

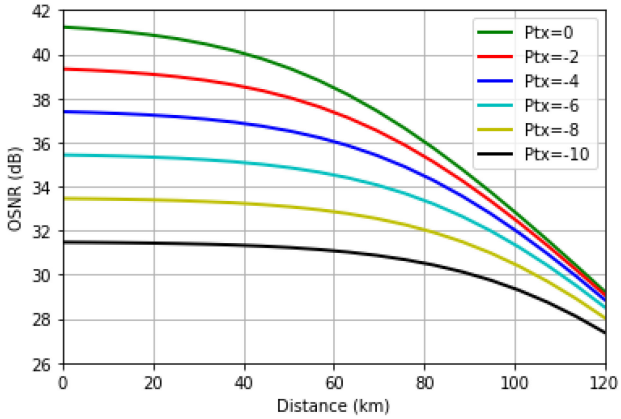

Fig. 14. OSNR as function of ZR reach for various transmitter output power levels.

Fig. 14 shows the OSNR as a function of reach for the ZR link. The OSNR curves are computed using the analytical link model described in Ref. [28]. At 120 km and −10 dBm module output power, the link should close at 26 dB OSNR. This is within the reach of the 16QAM modulation format at 400G. The model also shows that for 120 km reach higher module output power does not have a large impact on the link OSNR.

Further work, where the booster EDFA has a variable optical attenuator (VOA) at the output [29], represents how these systems will most likely be deployed. The conclusions are essentially the same that there is OSNR margin available in these systems at 120 km reach, and any benefits of higher module output power are greatly diminished at these distances.

There is a scenario where one can consider eliminating one of the EDFA's (booster or pre-amp). For optimal launch OSNR, it is better to keep the booster at the transmitter and eliminate the pre-amp at the receiver. In this scenario, the fiber reach will be reduced. For the 400ZR module receive power specification of −12 dBm and the same fiber launch power after the booster amp, a span reach of 10 dB or more can be achieved.

One can also consider integrating a miniature EDFA inside the module. This has been done in CFP2 modules, which are larger than the QSFP-DD modules (Fig. 3). Adding an EDFA, increases the size, cost, complexity, power consumption and thermal dissipation of the pluggable module and of the host switch chassis, and is not typically done.

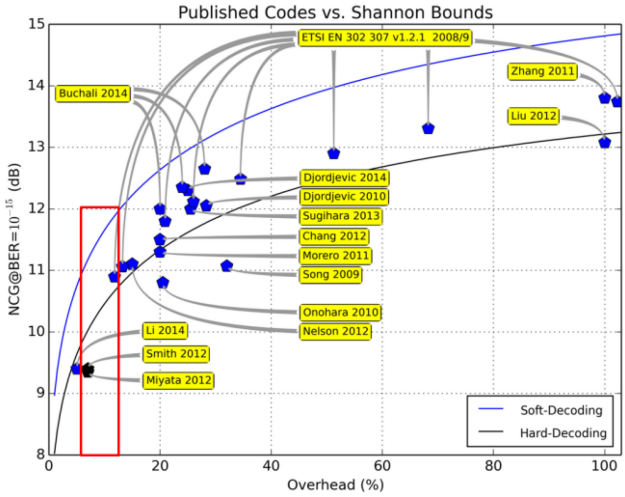

Fig. 15. Performance of the widely published FEC code adapted from Ref. [24]. The references in the figure, labeled by the first author's names and year of publication, are Ref. [31] to [45].

C. FEC Overview

FEC has a rich history in optical communications, being first deployed in submarine communications [30]. The goal here is not to review the details of the various FEC codes but point out the trade-off between the correction efficiency and complexity of the various options.

We have reproduced Fig. 15 from Ref. [24] which provides a good overview of FEC development. This chart shows the NCG (net coding gain) of the various FEC implementations as a function of the overhead (the increase in the data rate due to the addition of the FEC). NCG (effective gain) accounts for the performance loss as the result of increased receiver noise at higher bandwidths needed for larger FEC overheads. Generally, the coding gain increases with the complexity of the code (higher overhead) with the gain being offset by the increase in the symbol rate.

For algorithms commonly used in practice, the NCG starts to saturate beyond about 20% FEC overhead. The complexity of the FEC code is insufficient to offset the system impairments due to the larger overhead. Although, we have not explicitly considered the increased latency and power consumption with FEC complexity, the sweet spot is between 7% and 15% FEC overhead.

D. 400ZR CFEC

A concatenated FEC (CFEC) is used in 400ZR [13], [22]. A concatenated FEC uses two or more FEC codes in tandem. The inner code (code closest to the channel) is well-tuned for the channel. The outer code "cleans up" the errors left by the inner code. Overall efficiency of the code is the product of the two [46].

TABLE II
COMPARISON OF COMMON FEC CODES USED IN INTER AND INTRA DC

FEC type	KP4 RS (544,514,15)	Staircase FEC	Concatenated Hamming + Staircase FEC
Overhead	5.8%	6.7%	14.8%
Net Coding Gain (PAM4/16QAM)	6.9 dB	9.76 dB	10.8 dB
FEC Threshold (BER = 1.e-15)	2.3E-4	4.5E-3	1.22E-2
Power	P	~ 2.5*P	~ 3.0*P
Latency	t	~ 100*t	~ 100*t

For the CFEC, the inner code is a soft decision Hamming (128,119) code. The outer code is a hard decision Staircase code (255,239). The NCG for the CFEC is 10.8 dB in the 16QAM mode. The FEC overhead is 14.8%. CFEC design is a trade-off between NCG and power consumption. It is an ultra-low power FEC at 420 mW (for 7 nm CMOS node) at 400G. CFEC burst tolerance is 1024 bits, and the latency is 4 μs at 400G.

Generalized Hamming code is a single-error-correcting binary linear code. Beyond the textbook Hamming code, we can add additional parity bits to improve the minimum distance of the code. In the extended Hamming codes, the increase in overhead is outweighed by the improvement in decoding performance. The (128,119) Hamming code used for CFEC is a doubly extended (i.e., two "extra" parity bits) version of the native (127,120) Hamming code. In CFEC, we use a soft decoder for the Hamming code. In the context of decoding, this significantly increases the coding gain, but necessitates a slightly more complicated decoder implementation. Hamming is a good choice for concatenating with the Staircase FEC, because the Hamming code has a low-complexity soft-decision decoder, yet still provides excellent overall performance.

Using the CFEC and the 400ZR framing structure, the optical signal is 59.8 Gbaud 16QAM.

E. Standard IEEE FEC Codes

For intra DC applications, IEEE has adopted the KR4/KP4 FEC, based on Reed-Solomon (RS) coding, in the 802.3 standards. RS FEC encoding introduces redundancy into the codeword. A block of k data symbols becomes a codeword of n symbols, (n, k, T). The FEC decoding finds the decoded codeword that is closest to the received codeword. The FEC decoding is guaranteed to correct $\boldsymbol{T}$ errored symbols in a received codeword [47].

Table II shows a comparison between the FEC codes commonly used in DC applications at equivalent single lane data rates. It is generally preferred that the latency in the FEC does not dominate the overall latency in the system. Intra DC, the time of flight in a 100 m SSMF is ~500 ns. The latency of KP4 FEC at 100G (per lane) ~100 ns [48]. On the other hand, the time of flight in a 100 km intra DC link is 500 μs. Hence, higher latency, higher NCG FEC's can be easily justified for links where the fiber transmit time dominates the overall link latency.

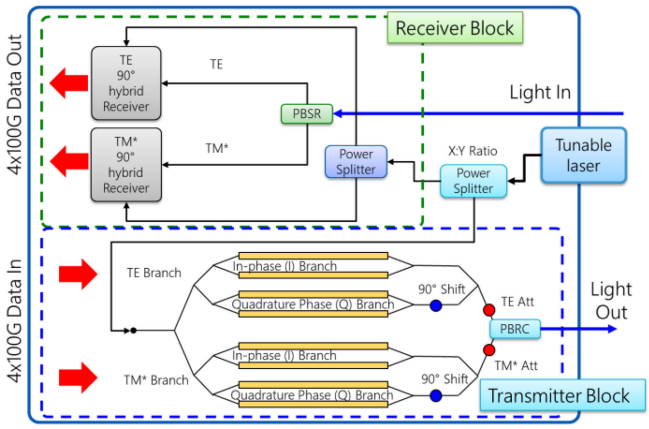

Fig. 16. Block diagram for the transmitter and receiver blocks of the integrated, coherent silicon photonics chip.

F. Silicon Photonics for 400G Applications

Fig. 16 shows coherent silicon photonics chip block diagram. The transmit and receive blocks are integrated onto the same chip. The modulator driver, at the input, and the TIA, at the output, are adjacent (to the left) of the silicon photonics chip.

A single, external tunable laser is split between the transmit and receive blocks. The power split ratio is a trade-off between the module output power and receiver LO (local oscillator) power. The tunable laser is TE polarized. The silicon photonics chip is designed for TE polarization. Polarization rotator followed by a combiner (PBRC) is used in the transmitter to generate the dual polarization output. The dual polarization, together with the IQ modulation, gives the 16QAM the 4x increase in density compared to PAM4. Polarization beam splitter followed by a rotator (PBSR) is used in the receiver to convert the incoming light to single polarization for processing.

The transmit block consists of a "nested" IQ modulator structure. The inner modulators are biased at their null point. One set of modulators is for the TE and the other for the final TM output. The outer modulators are biased at the quadrature point (90°) to combine the I and Q blocks. There are attenuators in the TE and TM paths to balance the loss and minimize the PDL (polarization dependent loss) in the system.

On the receive side, a 90° hybrid is used to separate the I and Q components. This is followed by an integrated Ge, balanced photodetector circuit and transimpedance amplifiers (TIA) for conversion to an output voltage signal. Integrated photonic integrated circuits have been covered in detail in literature [49].

Silicon photonics traveling wave Mach-Zehnder modulators (MZM) are used in the transmit path. These are depletion mode diode structures, which may be reversed biased for bandwidth enhancement.

Electrical-electrical (EE) small signal response shown in Fig. 17, is the measurement of the microwave loss at the end of the MZM transmission line. To calculate the equivalent electrical-optical (EO) response, we need to "break" the MZM into small segments and integrate the phase change along the entire MZM. The phase change is larger near the MZM input but is smaller at the end of MZM due to microwave loss. After

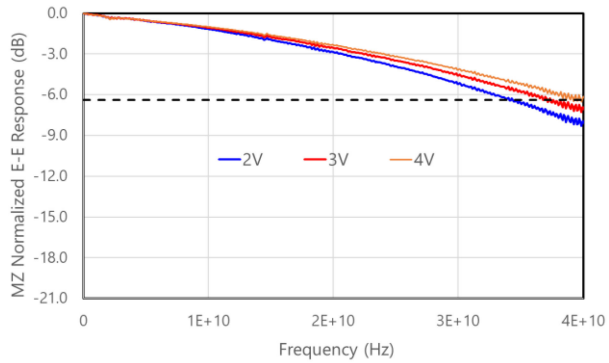

Fig. 17. Normalized, small signal electrical-electrical response of the MZM as a function of frequency. The curves are for the various reverse bias voltages applied to the MZM.

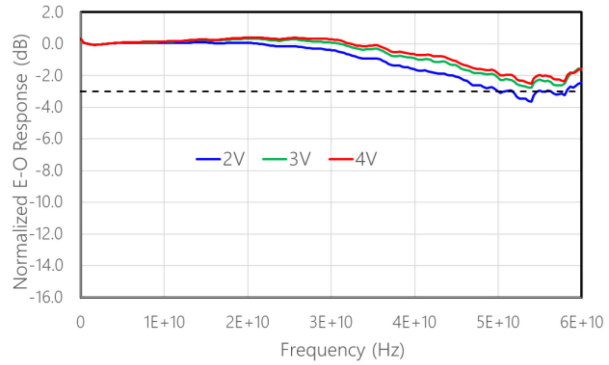

Fig. 18. Normalized, small signal electrical-optical response of the high-speed photodetector as a function of frequency for various reverse bias voltages.

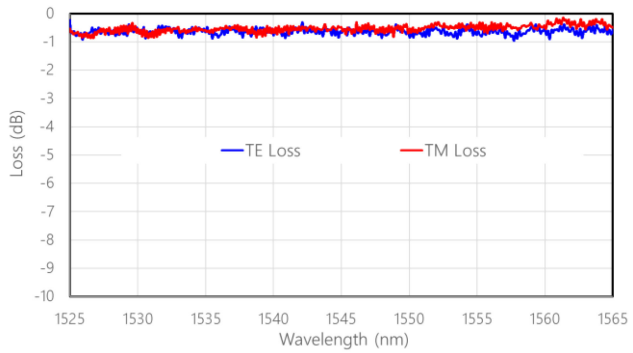

Fig. 19. Insertion loss performance of the polarization beam splitter/combiner across the C band of the SSMF.

accounting for the total (integrated) phase change, we get the relationship between the EE (6.4 dB) and EO (3 dB) bandwidths. The equivalent EO bandwidth (6.4 dB line), as shown in Fig. 19, of the MZM is in excess of 40 GHz. Vpi (the voltage applied for a π phase shift in the MZM) is in the range of 4 V to 6 V.

Fig. 18 shows the small signal frequency response of the integrated Ge photodetector (PD). PD with 3 dB bandwidth in excess of 50 GHz can be built. Such a large bandwidth is not needed for a 60 Gbaud signal and is optimized for minimizing noise in the receiver optical chain.

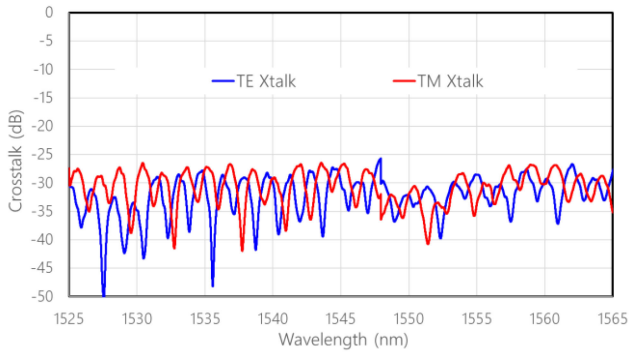

Fig. 20. Crosstalk performance of the polarization beam splitter/combiner across the C band of the SSMF.

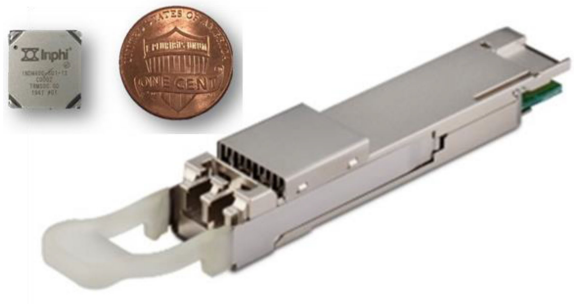

Fig. 21. 400ZR module in the QSFP-DD form factor. The DSP chip is shown as the inset.

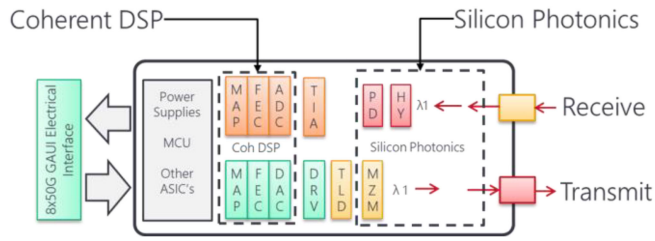

Fig. 22. Block diagram for the 400ZR module implementation.

The driver has a bandwidth in excess of 45 GHz, and the TIA has a bandwidth in excess of 43 GHz.

PBSR and PBRC are based on large optical bandwidth (entire C band), low loss, polarization beam splitter/combiner design. In such designs, the insertion loss in the signal port and the crosstalk in the adjacent port need to be minimized.

Figs. 19 and 20 show the performance of the basic polarization processing circuit with insertion loss of <1 dB and crosstalk <-25 dB across the C band [50].

G. Pluggable 400ZR Modules

A pluggable 400ZR module is shown in Fig. 21. The block diagram for the module construction is shown in Fig. 22. The coherent DSP chip has the host side interface to the switch and the line side interface to the silicon photonics chip. Our 100G PAM4 modules have a similar block level implementation [50].

The DSP chip is designed with the advanced 7 nm CMOS node [51]. It has 8 electrical inputs at 26.5625 Gbaud (53.125 Gbit/s) PAM4 signal (IEEE 802.3bs). The signal is then mapped into the

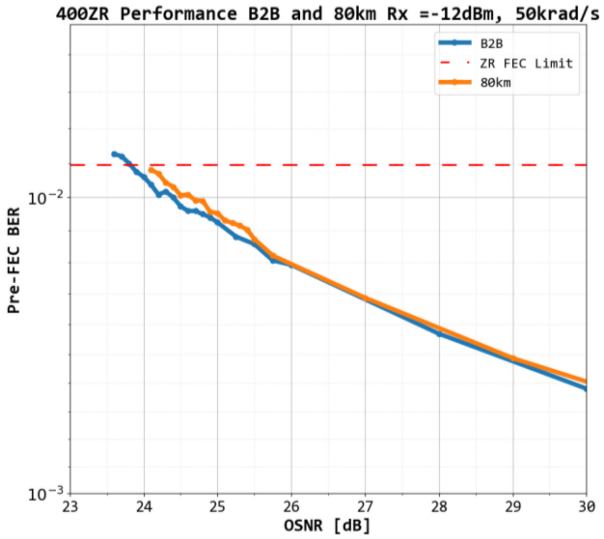

Fig. 23. Performance of the 400ZR module with polarization scrambling in the input signal.

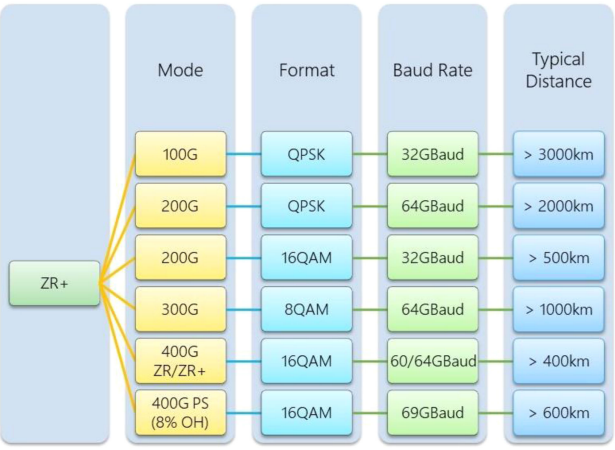

Fig. 24. Variety of non ZR modes supported by the same DSP. Data rate, modulation format, baud rate and nominal reach for the different modes.

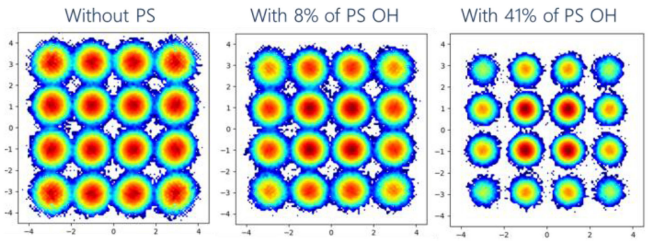

Fig. 25. 400G 16QAM constellations with varying levels of probabilistic shaping overhead.

400ZR framing structure and encoded with CFEC. The output DAC feeds the driver IC which then drives the MZM on the silicon photonics chip.

On the receive side, the output of the TIA is sampled using the front-end ADC of the DSP. After signal processing and FEC decoding, the data is remapped into the ethernet frame and sent to the switch. A variety of power supplies, a micro-controller and control circuits fill rest of the module.

The pre-FEC performance of the 400ZR module, as a function OSNR, is shown in Fig. 23. The noise loaded input signal is polarization scrambled at 50 krad/s (as per the 400ZR spec [13]). The CFEC correction limit is 1.25E-2. The OSNR penalty at -12 dBm receiver power is $<$0.2 dB at 80 km.

Unlike coherent systems, inter DC PAM4 links need external DCM (dispersion compensation module) [52]. For PAM4 inter DC systems, already deployed with DCM's, the coherent signals can be overlaid in the same line system with little impact [53]. This may reduce the DSP power for CD compensation even further.

V. Extending the Reach of Pluggable Coherent Modules

In the Introduction, we had hinted at the possibility of using a single pluggable module to span a wide reach of DC links. The additional reaches and modes that may be supported by the same small form factor pluggable module, in addition to the ZR mode, have been collectively called the ZR+ or Metro modes in the industry. Although it varies with vendors, a sample of these modes are shown in Fig. 24.

ZR+ offers an attractive feature of being operable directly from a switch or a router. The chassis and the module must be designed to dissipate the higher power that will be dissipated by some of the higher complexity and longer reach modes.

One of the advanced features of building a DSP chip with complex capabilities is ability to use probabilistic shaping (PS) [51], in addition to FEC, for longer reach links. In PS, constellation symbols are transmitted with a non-uniform probability density function, usually a Maxwell-Boltzmann distribution. Fig. 24 shows a continuous progression of transmission formats and coding to enable higher baud rates and longer reaches.

Fig. 25 shows, from left to right, a 16QAM constellation without shaping, with 8% and 41% PS overhead (OH). The PS OH is the increase in symbol rate required to keep a constant data rate when PS is applied. According to information theory, constellation shaping provides "shaping gain," whose maximum ideal value is 1.53 dB. In practice, achieved shaping gain is less than the theoretical value. However, constellation shaping also provides a form of coding gain, which results from the increase in Euclidean distance among constellation points when shaping is applied, and the transmitted power is kept constant. Ref. [54] has a comprehensive discussion of the benefits of PS in optical communications.

The OSNR gain using PS is shown in Fig. 26. Just like FEC encoding, PS introduces a data rate OH. Even with that overhead, we have a $>$1.0 dB OSNR gain using 8% PS for 400G, 16QAM [51]. Although in principle, PS can be incorporated at *increasing levels* to obtain better performance, just like FEC, there is a practical limit imposed by the modulation complexity and higher baud rate on the DSP design, and power consumption.

VI. Conclusion

There are a whole array of constraints and trade-offs when designing low power, DSP based pluggable transceivers for data center applications. Total power utilization is critical for

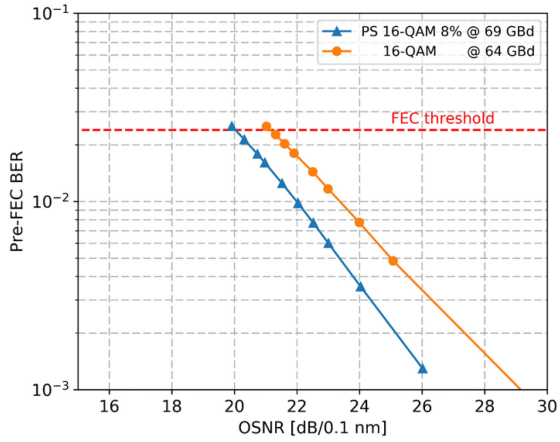

Fig. 26. Pre-FEC BER as a function of OSNR for 400G 16QAM, with and without probabilistic shaping.

successful data center operations. Data center infrastructures have power supplied to them by the utilities. As the data rates grow, the growth in the overall services provided by the data center is limited by the maximum power available, which does not scale over time for an existing building. This imposes an upper limit on the power that can be allocated to the optical interconnects.

The goal of the tutorial was to present the trade-offs that have been implemented and are being considered for data center optical interconnects that increasingly use digital signal processing for both direct detect and coherent transmission formats. With increasing data rates, the coherent transmission formats become more relevant to short distance applications. In the not too distant future, coherent links may displace direct detect formats even for intra data center applications.

ACKNOWLEDGMENT

The authors would like to thank the exceptional engineering team at Inphi (and now Marvell) which contributed enormously to this work.

REFERENCES

[1] D. Crivelli, H. Carrer, and M. Hueda, "Adaptive digital equalization in the presence of chromatic dispersion PMD and phase noise in coherent fiber optic systems," in *Proc. Global Telecom. Conf. (GLOBECOM)*, Nov. 2004, vol. 4, pp. 2545–2551.

[2] O. Agazzi *et al.*, "A 90 nm CMOS DSP MLSD transceiver with integrated AFE for electronic dispersion compensation of multi-mode optical fibers at 10Gb/s," *J. Solid-State Circuits*, vol. 43, no. 12, pp. 2939–2957, Dec. 2008.

[3] K. Roberts and I. Roberts, "DSP: A disruptive technology for optical transceivers," in *Proc. 35th Eur. Conf. Opt. Commun.,* Vienna, Austria, 2009, pp. 1–4.

[4] K. Kikuchi, "Fundamentals of coherent optical fiber communications," *J. Lightw. Technol.*, vol. 34, no. 1, pp. 157–179, Jan. 2016.

[5] K. Gopalakrishnan *et al.*, "A 40/50/100Gb/s PAM-4 ethernet transceiver in 28nm CMOS," in *Proc. IEEE Int. Solid-State Circuits Conf.*, San Francisco, USA, 2016, pp. 62–63.

[6] S. Bhoja, "PAM4 signaling for intra-data center and data center to data center connectivity (DCI)," in *Proc. Opt. Fiber Commun.*, Los Angeles, USA, 2017, pp. 1–54.

[7] R. Nagarajan, M. Filer, Y. Fu, M. Kato, T. Rope, and J. Stewart, "Silicon photonics-based 100 Gbit/s, PAM4, DWDM data center interconnects," *J. Opt. Comm. Netw.*, vol. 10, no. 7, pp. 25–36, Jul. 2018.

[8] R. Nagarajan, "Pluggable coherent transport modules for datacenter interconnects," in *Proc. IEEE Summer Top. Mtg . (V irtual Conf.)*, 2020, Art. no. MA2.3.

[9] R. Nagarajan, "Small form factor modules for coherent optical interconnects," in *Proc. IEEE Photon. Conf. (Virtual Conf.)*, 2020, Art. no. MH4.2.

[10] *High-Speed Ethernet Optics*, 10th ed., Eugene, OR, USA: LightCounting, Sep. 2019, pp. 8.

[11] S. Wilkinson and A. Schmitt, "Transport applications report (2Q20)," Cignal AI, Boston, MA, USA, Sep. 2020.

[12] R. Chopra, "Looking beyond 400G – A system vendor perspective," IEEE 802.3 Beyond 400 Gb/s Ethernet Study Group, Feb. 2021, [Online]. Available: https://www.ieee802.org/3/B400G/public/21_02/chopra_b400g_01_210208.pdf

[13] Implementation Agreement 400ZR, Optical Internetworking Forum, Fremont, CA USA, OIF-400ZR-01.0, Mar. 2020.

[14] M. Filer, J. Gaudette, Y. Yin, D. Billor, Z. Bakhtiari, and J. Cox, "Low-margin optical networking at cloud scale," *J. Opt. Comms. Netw.*, vol. 11, no. 10, pp. C94–C108, Oct. 2019.

[15] R. Nagarajan and I. Lyubomirsky, "Next-Gen data center interconnects: The race to 800G," Consortium for On-Board Optics webcast, Jan. 2021, [Online]. Available: https://www.onboardoptics.org/the-race-to-800g-inphi

[16] CFP Multi-Source Agreement (MSA), [Online]. Available: http://www.cfp-msa.org/

[17] F. Frey, R. Elschner, and J. Fischer, "Estimation of trends for coherent DSP ASIC power dissipation for different bitrates and transmission reaches," in *Proc. Photonic Netw.; 18. ITG-Symp.*, Leipzig, Germany, 2017, pp. 1–8.

[18] X. Zhou, R. Urata, and H. Liu, "Beyond 1 Tb/s intra-data center interconnect technology: IM-DD or coherent?," *J. Lightw. Technol.*, vol. 38, no. 2, pp. 475–484, Jan. 2020.

[19] R. Nagarajan, "Datacenter interconnect systems with coherent detection (Tutorial)," in *Proc. Fibre-Opt. Cable*, San Diego, USA, 2019, Art. no. M3H.1.

[20] C. Fludger, "Performance orientated DSP design for flexible coherent transmission (Tutorial)," in *Proc. Opt. Fiber Commun.*, San Diego, CA, USA, 2020, Art. no. Th3E.1.

[21] I. Lyubomirsky, "Coherent vs. direct detection for next generation intra-datacenter optical interconnects," in *Proc. IEEE Summer Top. Mtg. (Virtual Conf.)*, 2020, Art. no. WA2.3.

[22] B. Smith, J. Riani, A. Farhood, and S. Bhoja, "16QAM DSP & FEC proposal for 400G-ZR," *Opt. Internetworking Forum*, Apr. 2017, OIF2017.200.02.

[23] O. Agazzi, M. Hueda, D. Crivelli, and H. Carrer, "Maximum-likelihood sequence estimation in dispersive optical channels," *J. Lightw. Technol.*, vol. 23, no. 2, pp. 749–763, Feb. 2005.

[24] D. Morero, M. Castrillón, A. Aguirre, M. Hueda, and O. Agazzi, "Design tradeoffs and challenges in practical coherent optical transceiver implementations," *J. Lightw. Technol.*, vol. 34, no. 1, pp. 121–136, Jan. 2016.

[25] S. Faruk and S. J. Savory, "Digital signal processing for coherent transceivers using multilevel formats," *J. Lightw. Technol.*, vol. 35, no. 5, pp. 1125–1141, Mar. 2017.

[26] T. Kupfer, A. Bisplinghof, T. Duthel, C. Fludger, and S. Langenbach, "Optimizing power consumption of a coherent DSP for metro and data center interconnects," in *Proc. Opt. Fiber Commun.*, Los Angeles, USA, 2017, Art. no. Th3G.2.

[27] P. Winzer, S. Chandrasekhar, and X. Liu, "Modulation formats and receiver concepts for optical transmission systems," in *Proc. Opt. Fiber Commun.*, Los Angeles, USA, 2012, Art. no. SC105.

[28] I. Lyubomirsky, "OSNR link budget methodology," in *Proc. IEEE 802.3cn Ad-Hoc Meet.*, Oct. 2018. [Online]. Available: https://www.ieee802.org/3/cn/public/adhoc/18_1025/lyubomirsky_3cn_01_181025.pdf

[29] B. Zhang and K. Kota, "800ZR reference links," *Opt. Internetworking Forum*, Dec. 2020, OIF2020.472.01.

[30] K. Onohara and T. Mizuochi, "Forward error correction: A powerful and indispensable technology for ultra high-speed transmission," in *Proc. SubOptic Conf.*, Yokohama, Japan, 2010, Tutorial 5.

[31] K. Onohara *et al.*, "Soft-decision-based forward error correction for 100 Gb/s transport systems," *J. Sel. Topics Quantum Electron.*, vol. 16, no. 5, pp. 1258–1267, Sep./Oct. 2010.

[32] B. Smith, A. Farhood, A. Hunt, F. Kschischang, and J. Lodge, "Staircase codes: FEC for 100 Gb/s OTN," *J. Lightw. Technol.*, vol. 30, no. 1, pp. 110–117, Jan. 2012.

[33] D. Chang *et al.*, "LDPC convolutional codes using layered decoding algorithm for high speed coherent optical transmission," in *Proc. Opt. Fiber Commun.*, Los Angeles, CA, USA, 2012, Paper OW1H.4, pp. 1–3.

[34] L. Nelson *et al.*, "WDM performance and multiple-path interference tolerance of a real-time 120 Gbps Pol-Mux QPSK transceiver with soft decision FEC," in *Proc. Opt. Fiber Commun.*, Los Angeles, CA, USA, 2012, pp. 1–3.
[35] D. Morero, M. Castrillon, F. Ramos, T. Goette, O. Agazzi, and M. Hueda, "Non-Concatenated FEC codes for ultra-high speed optical transport networks," in *Proc. GLOBECOM*, Houston, TX, USA, 2011, pp. 1–5.
[36] Y. Miyata, K. Kubo, K. Onohara, W. Matsumoto, H. Yoshida, and T. Mizuochi, "UEP-BCH product code based hard-decision FEC for 100 Gb/s optical transport networks," in *Proc. Opt. Fiber Commun.*, Los Angeles, CA, USA, 2012, pp. 1–3.
[37] I. Djordjevic, L. Xu, and T. Wang, "Reverse concatenated coded modulation for high-speed optical communication," *Photon. J.*, vol. 2, no. 6, pp. 1034–1039, Dec. 2010.
[38] K. Liu, Q. Huang, S. Lin, and K. Abdel-Ghaffar, "Quasi-cyclic LDPC codes: Construction and rank analysis of their parity-check matrices," in *Proc. Inf. Theory Appl. Workshop*, San Diego, CA, USA, 2012, pp. 227–233.
[39] S. Song, B. Zhou, S. Lin, and K. Abdel-Ghaffar, "A unified approach to the construction of binary and nonbinary quasi-cyclic LDPC codes based on finite fields," *Trans. Commun.*, vol. 57, no. 1, pp. 84–93, Jan. 2009.
[40] L. Zhang, S. Lin, K. Abdel-Ghaffar, Z. Ding, and B. Zhou, "Quasi-Cyclic LDPC codes on cyclic subgroups of finite fields," *Trans. Commun.*, vol. 59, no. 9, pp. 2330–2336, Sep. 2011.
[41] K. Sugihara *et al.*, "A spatially-coupled type LDPC code with an NCG of 12 dB for optical transmission beyond 100 Gb/s," in *Proc. Opt. Fiber Commun.*, Anaheim, CA, USA, 2013, Paper OM2B.4, pp. 1–3.
[42] J. Li, K. Liu, S. Lin, and K. Abdel-Ghaffar, "Algebraic quasi-cyclic LDPC codes: Construction, low error-floor, large girth and a reduced-complexity decoding scheme," *Trans. Commun.*, vol. 62, no. 8, pp. 2626–2637, Aug. 2014.
[43] F. Buchali, A. Klekamp, L. Schmalen, and T. Drenski, "Implementation of 64QAM at 42.66 GBaud using 1.5 samples per symbol DAC and demonstration of up to 300 km fiber transmission," in *Proc. Opt. Fiber Commun.*, San Francisco, CA, USA, 2014, Paper M2A.1, pp. 1–3.
[44] I. Djordjevic, "Advanced coded modulation for ultrahigh speed optical transmission," in *Proc. Opt. Fiber Commun.*, San Francisco, CA, USA, 2014, Paper W3J.4, pp. 1–41.
[45] Digital Video Broadcasting, ETSI EN Standard 302 307, Rev. 1.2.1, 2009.
[46] F. Kschischang, "Introduction to forward error correction," in *Proc. Opt. Fiber Commun.*, San Francisco, USA, 2017, p. SC390.
[47] H. Zhang, B. Jiao, Y. Liao, and G. Zhang, "PAM4 signaling for 56G serial link applications – A tutorial," Design Con, Santa Clara, CA, USA, 2016, [Online]. Available: https://www.xilinx.com/publications/events/designcon/2016/slides-pam4signalingfor56gserial-zhang-designcon.pdf
[48] S. Bhoja, V. Parthasarathy, and Z. Wang, "FEC codes for 400 Gbps 802.3bs," IEEE P802.3bs 200 GbE & 400 GbE Task Force, Nov. 2014. [Online]. Available: https://www.ieee802.org/3/bs/public/14_11/parthasarathy_3bs_01a_1114.pdf
[49] R. Nagarajan, C. Doerr, and F. Kish, "Semiconductor photonic integrated circuit transmitters and receivers," in *Opt. Fiber Telecommun.*, vol. VI A I. Kaminow, T. Li and A. Willner, ed.s., New York, NY, USA: Elsevier, 2013, pp. 25–98.
[50] R. Nagarajan and M. Filer, "10. Silicon photonics based PAM4, DWDM datacenter interconnects," in *Datacenter Connectivity Technologies: Principles and Practice,* vol. 1, F. Chang, ed., Denmark: River Publishers, 2018, pp. 405–430.
[51] A. Castrillón *et al.*, "First real-time demonstration of probabilistic shaping 400G transmission enabling high-performance pluggable module applications," in *Proc. IEEE Photon. Conf. (Virtual Conf.)*, 2020, Paper MG1.3, pp. 1–2.
[52] S. Searcy, G. Brochu, S. Boudreau, F. Trépanier, M. Filer, and S. Tibuleac, "Statistical evaluation of PAM4 data center interconnect system with slope-compensating fiber bragg grating tunable dispersion compensation module," *J. Lightw. Technol.*, vol. 38, no. 12, pp. 3173–3179, Jun. 2020.
[53] J. Downie, J. Hurley, R. Nagarajan, T. Maj, H. Dong, and S. Makovejs, "100 Gb/s wavelength division multiplexing four-level pulse amplitude modulated transmission over 160 km using advanced optical fibres," *Electron. Lett.*, vol. 54, no. 11, pp. 699–701, May 2018.
[54] J. Cho and P. J. Winzer, "Probabilistic constellation shaping for optical fiber communication," *J. Lightw. Technol.*, vol. 37, no. 6, pp. 1590–1607, Mar. 2019.

Radhakrishnan Nagarajan (Fellow, IEEE) received the B.Eng. degree (1st Class Hon.) in electrical engineering from the National University of Singapore, Singapore, the M.Eng. degree in electronic engineering from the University of Tokyo, Tokyo, Japan, and the Ph.D. degree in electrical engineering from the University of California, Santa Barbara, Santa Barbara, CA, USA.

He is currently the SVP & CTO with Marvell Semiconductors. Previously, he was a Fellow with Infinera, where he led the development of multiple generations of InP based large scale photonics integrated circuits. Prior to that he was with SDL, which was acquired by JDS Uniphase, where he developed high power pump laser modules for EDFA applications. He has authored or coauthored more than 190 publications in journals and conferences, and five book chapters in the area of high-speed optical components and photonics integration.

He is a Fellow of the Optical Society and a Fellow of the Institution of Engineering and Technology. He was the recipient of the 2006 IEEE LEOS Aron Kressel Award for his contributions to the commercialization of Large Scale Photonic Integrated Circuits. He has been awarded 200 U.S. patents.

Ilya Lyubomirsky received the dual B.S. degrees in electrical engineering and mathematics from the University of Maryland, College Park, MD, USA, in 1991, and the M.S. and Ph.D. degrees in electrical engineering from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 1994 and 1999, respectively. In 2017, he joined Marvell Semiconductors as a Distinguished Engineer to work on coherent DSP for 400ZR. He is currently a Senior Technical Director, leading the Systems and DSP Architecture team on the research and development of 400G/800G PAM4 DSP ASICs.

He has more than 20 years of experience in digital communications, signal processing, DWDM systems, and data center interconnects. Prior to Marvell, he led the engineering team on Facebook's Voyager project, successfully demonstrating the world's first 800 Gb/s DWDM white box coherent transponder. Prior to that, he held various engineering and academic roles with Finisar, Infinera, University of California, Ciena and Telcordia. During his career in industry and academia, He made contribution extensively to IEEE 802.3 Ethernet standards, has authored or coauthored more than 50 peer-reviewed journal and conference papers, and is an inventor or co-inventor on 18 patents.

Oscar Agazzi (Life Fellow, IEEE) received the Electrical and Electronic Engineer degree and the Licentiate in physics from the University of Cordoba, Cordoba, Argentina, and the Ph.D. degree in electronic engineering from the University of California at Berkeley, Berkeley, CA, USA. He is the VP, Engineering, of the Coherent DSP BU with Marvell Semiconductors. He is a recognized expert in the fields of communications and signal processing. Previously, he was the Chief Architect of ClariPhy Communications family of coherent and direct detection optical transceivers. Prior to joining ClariPhy, he was with Broadcom Corp., San Jose, CA, USA, where he was the Chief Architect of a family of Gigabit Ethernet (1000BASE-T) transceivers. He also played an instrumental role in the development 10-Gigabit Ethernet optical and copper transceivers. Prior to joining Broadcom, he was with Lucent Technologies Bell Laboratories, where he did research and development in the areas of ISDN transceivers, magnetic recording for read-channel devices, and document recognition systems. With Lucent, he was appointed Bell Labs Fellow, the most prestigious distinction awarded by Bell Labs to its research staff. He has more than 150 patents and has authored or coauthored more than 60 technical papers in journals and conferences.