400G per Lane Ethernet PHY Whitepaper

Network Innovation and Development Alliance

March 27, 2026

Chairs: Jannik Hammel Nielsen (Huawei), Weiqiang Cheng (China Mobile)

Editors: Runlong Hu (China Mobile), Guangcan Mi (Huawei), Zushu Yan (Credo)

Contributors: Xuebo Wang (Huawei), Congshi Zou (Huawei), Zhiwei Yang (ZTE), Jie Zhou (Joywell), Kim Cao (Luxshare), Lu Liu (CAICT), Yu Xu (Huawei), Anbin Wang (Alibaba), Cathy Chen (TFC), Chaonan Yao (Ligent), Chongjin Xie (PhotonicX AI), Dennis Zhou (AFR), Haojie Wang (China Mobile), Hu Zhu (Accelink), Huijun Sha (Viavi), Junjie Wang (Centec), Keiwei Liu (CAICT), Loong Jin (Luxshare), Michael Liu (Joywell), Shenglei Hu (Tencent), Tao Pan (H3C), Wenbo Shen (Xinertel), Xiang He (Huawei), Xiaoyong Qiu (Longsight), Yunguo Li (Ruijie), Yunpeng Cao (Joywell), Zhidong Tian (Keysight)

Supported by Members from:

Abstract

The rapid scaling of artificial intelligence infrastructure, with training clusters exceeding 10,000 accelerators and communication overhead consuming up to 60% of training time, has made 400 gigabit per second per lane an inevitable requirement for next-generation Ethernet switches. With switch radix constrained by fundamental packaging and thermal limits to approximately 128 ports, achieving switch capacities exceeding 400 terabits per second demands lane speed of over 400 gigabits per second. This paper examines the technical landscape for 400G per lane implementation, addressing the emerging divergence between optical and electrical domains. While optical signaling has converged on PAM4 with multiple viable modulator platforms—Indium Phosphide, Silicon Photonics, and Thin Film Lithium Niobate—the electrical interface remains contested between high-baud PAM4 and reduced-baud PAM6 approaches. We analyze passive channel constraints, SerDes technology requirements, and Ethernet PHY architectures including pluggable optics, linear variants, and co-packaged solutions. Central to our analysis is the role of Forward Error Correction as the architectural anchor: with RS-FEC (KP4) fixed as the end-to-end outer code for backward compatibility, the choice between PAM4 with lightweight inner coding versus PAM6 with stronger concatenated FEC involves trade-offs among bandwidth requirements, signal-to-noise ratio, latency, power consumption, and silicon area. We conclude that passive link bandwidth capability—specifically whether the ecosystem can deliver connectors, cabling, and packaging sustaining bandwidths exceeding 112 GHz—will determine whether the industry favors the continuity of PAM4 or the complexity trade-offs of PAM6 for 400G per lane electrical interfaces.

Contents

1 Introduction

The advancement toward 400 gigabit per second per lane is driven by unprecedented demands from artificial intelligence infrastructure. Large language models have undergone rapid parameter scaling, with frontier models approaching or exceeding two trillion parameters [2, 3]. Training these models requires coordination across thousands of accelerator devices; Alibaba's HPN architecture supports training on more than 10,000 GPUs with high-efficiency parallel computing [4]. Industry deployments now regularly exceed 10,000 accelerators per cluster, with hyperscalers planning over 100,000 devices for next-generation models [5, 6].

The communication bottleneck in distributed training has emerged as a critical limiting factor for large-scale artificial intelligence workloads. Analysis of large-scale training systems demonstrates that all-reduce communication overhead can consume 30 to 60 percent of total training time for trillion-parameter models when network bandwidth is insufficient [1]. Similarly, analysis of large-scale training systems shows that communication-bound workloads achieve only 20 to 40 percent of peak computational efficiency without optimized networking [7]. This scale introduces a fundamental challenge: communication bandwidth between devices must keep pace with computational capacity to avoid becoming the bottleneck.

However, switch radix saturation limits constrain the scaling path. Switch ASICs package sizes and front-panel real estate constrain the maximum number of ports to 64 to 128 ports due to bump pitch, power delivery, and thermal dissipation constraints [8]. Flip-chip packaging technology is approaching fundamental limits at 128 ports for high-power ASICs, with 51.2 terabit per second switch chips consuming approximately 750 watts and reaching heat flux of 1.79 watts per square millimeter [9]. Cabling complexity compounds this constraint, as port counts increase quadratically, creating significant management overhead, airflow obstruction, and reliability concerns. Cable congestion in raised-floor plenums can reduce cooling fan energy efficiency, with unsealed cable cut-outs allowing 22 to 78 percent of conditioned air to escape without cooling IT equipment [10, 11]. Industry analysis indicates that switch radix scaling has reached a plateau, with further radix increases beyond 128 ports facing fundamental packaging and signal integrity barriers [12].

Given radix saturation, the industry has pivoted from more ports to more bandwidth per port as the primary scaling vector. Port bandwidth requirements have progressed from 25.6 terabit per second switches in 2022, to 51.2 terabit per second switches (Tomahawk 5) in 2023 to 2024, and now to 102.4 terabit per second switches (Tomahawk 6) entering volume production in 2024 to 2025 [8, 13]. Industry roadmaps project that artificial intelligence training workloads will require switch capacities exceeding 400 terabit per second in the 2026 to 2027 timeframe [14]. With radix saturation constraining practical port counts to 128, achieving such switch capacities demands per-port bandwidths of 3.2 terabits per second or higher. And given that the lane count is fixed at 8 lanes per port due to electrical complexity management, 400 gigabits per second per lane becomes the inevitable demand.

Historically, the evolution of Ethernet physical layer interfaces has maintained synchronization between electrical and optical domains. Across the past generations of signaling rates from 25 gigabit per second to 200 gigabit per second, both domains have shared common modulation schemes, progressing from NRZ to PAM4, and end-to-end FEC standards. This alignment has enabled simple interoperability and maximized the reuse of digital IPs and analog frontends across the industry.

However, as the industry advances toward 400 gigabit per second per lane, a divergence is emerging. On the optical side, the direction is comparatively clear: recent industry discussions converge on extended PAM4 viability for intensity-modulation direct-detection optics, supported by progress across multiple modulation and integration platforms. On the electrical side, feasible modulation formats (PAM4, PAM6, and PAM8) remain under active discussion, but practical viability is increasingly governed by passive channel bandwidth, discontinuities, and packaging transitions. The approximately 112 gigahertz bandwidth required for 400 gigabit per second PAM4 signaling (224 GBd Nyquist frequency) challenges PCB materials, connectors, and packaging transitions, with passive channel physics imposing hard limits on electrical domain scaling [15].

The remainder of this paper examines the technical landscape for 400 gigabit per second per lane implementation. Section II addresses optical transceiver component status, reviewing the three primary modulator platforms and their readiness for 400 gigabit per second per lane operation. Section III analyzes electrical channel technology status, addressing both passive link constraints and

SerDes technology considerations. Section IV presents Ethernet physical layer architecture options, from pluggable optics to co-packaged solutions. Section V explores FEC considerations for electrical links, analyzing concatenated FEC schemes and modulation-specific requirements. Section VI provides comparative analysis of PAM4 versus PAM6 electrical interfaces, including quantitative trade-offs in coding overhead and signal-to-noise ratio requirements.

2   Optical Interconnect Status

Optical interconnect has been the leading force of the 2-fold increasement of signaling rate throughout the Ethernet physical layer evolution from 25Gbps per lane to 200Gbps per lane. The technical foundation of this trend is the leading progress on optical component bandwidth exceeding the required Nyquist frequency of each signaling rate transition. The story will continue to repeat in the transition to 400Gbps signaling, where optical interconnect takes the first step and is facilitated by multiple credible implementation paths of the determining components, i.e., modulators and photo-diodes.

2.1   Indium Phosphide (InP) modulator

InP remains a strong candidate technology for high-speed transmitter. Multiple work from industry has been published extending the viability of EML-based approaches to 400G by using differential-drive techniques, with Broadcom [16], Mitsubishi [17], and Huawei [18] demonstrating beyond 160 GBd PAM modulations. InP Mach-Zehnder modulators (MZM) with bandwidth beyond 100 GHz are also demonstrated by NTT [19] as candidates for 400G PAM4.

2.2   Silicon Photonics (SiPh) modulator

Silicon Photonics has finally win its position among the mainstream optical transceiver technology during the migration from 100G/lane to 200G/lane, with its industry share projected further expanding. It is natural for the industry to expect the product solutions based on Silicon Photonic evolving to the next signaling rate. However, the biggest hurdle sits on Silicon photonic's path towards 400G/lane is the achievable bandwidth of the modulator. The bandwidth barrier is increasingly treated as a system co-optimization problem across silicon photonic modulator, its high-speed driver and the equalization offered by the DSP. A SiPh MZM of beyond 90GHz bandwidth was demonstrated by AMF in [20] leveraging the on-chip equalizer. Heterogeneous integration combines the benefit of high electro-optic response of the organic material with the mature photonic fabrication platform and device library of silicon photonic, and brings highly energy efficient and high bandwidth modulators viable for 400Gb/s signaling [21]. IMEC reported a 110GHz bandwidth GeSi electroabsorption modulator (EAM) capable of 400Gb/s IMDD signal transmission, more importantly featuring a 300mm SiPh fabrication platform for commercial readiness [22]. More exotic approach such as polarization multiplexing can be used to double the spectral density of an IM-DD link, however its practicality remains in question due to challenges in de-multiplexing the two rotated polarization signals and its lack of interoperability with existing architecture.

2.3   Thin Film Lithium Niobate (TFLN) modulator

TFLN is increasingly viewed as a major enabler for high-bandwidth modulation with emerging demonstrations beyond 110GHz. Optical industry showed a vibrant effort to cultivate the ecosystem for commercial readiness, spanning across IDM startups, foundry services for future fabless operation, and veteran optical component corporations. For system architects, the significance is that TFLN is being positioned not only as a high-performance modulator technology, but also as a manufacturable platform that can be integrated into higher-density photonic stacks, aligning with packaging and density requirements at 400G/lane.

2.4 High-speed Photodiodes

Silicon waveguide photodetectors gain increasing interest owing to its homogeneous integration with the transmitter, thus making the optical transceiver another viable chiplet for the semiconductor industry's modular portfolio. Unlike the challenge in modulator, the progress towards 400G-capable PD has been very encouraging, in fact the competition has shifted from bandwidth race to performance optimization with a focus on responsivity and dark current. Recent demonstrations from TSMC showed record performance, with receiver bandwidth around 110 GHz, responsivity around 0.9 A/W and dark current below 20 nA.

The availability of 100GHz+ bandwidth optoelectronic components greatly expedite the convergence of optical signaling modulation format to PAM4. However optical signaling doesn't stand alone, instead it is tightly associated with electrical signaling. While the first wave of the newer and faster signaling rate typically is implemented with optical interconnect interfaced with 2 lanes of the older and slower electrical interconnect, the most power efficient and long-lasting implementation always put optical and electrical interface at the same signaling rate. Unlike optical interconnect where SMF fiber link has close-to-zero bandwidth restrictions, electrical interconnect is largely constrained by the bandwidth of the passive electrical channel. This brings forth the next section in this study report, status of the electrical channel technology. The progress and the development trajectory of passive channel lay down the boundary of the exploration space of electrical signaling.

3 Electrical Channel Technology Status

3.1 Channel Characteristics and Materials

Electrical signaling for 400G/lane is strongly dependent on the passive link. While active analog frontends and DSP techniques continue to improve, the system-level feasibility is often dominated by end-to-end channel behavior driven by printed circuit board (PCB) materials and stack-ups, copper design, connectors, cabling, and packaging transitions. This passive ecosystem is therefore the central bottleneck when progressing toward 400G/lane electrical interconnection.
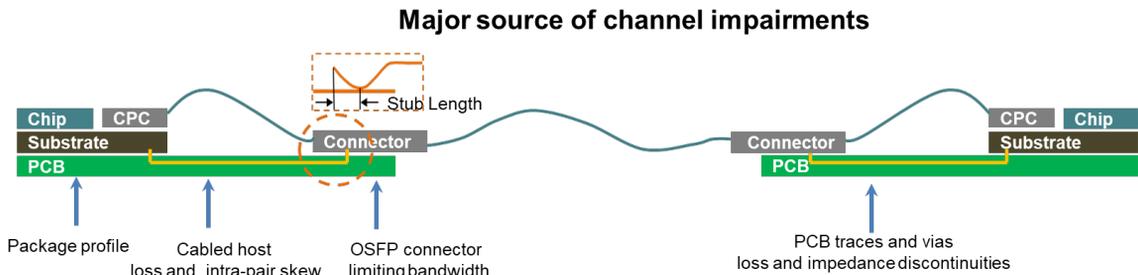

Figure 1: Key stakeholders determining the SI of an electrical channel.

Doubling the electrical data rates imposes a $\sim 2\times$ increase in Nyquist frequency, exponentially exacerbating channel loss, distortion, and noise sensitivity. At 400G/lane, the amplitude of the signal component will decrease due to the increased conductive and dielectric losses of the chip package, PCB, and connectors as Nyquist frequencies push toward 112 GHz (for PAM4) or 90 GHz (for PAM6). Aside from the rapidly increasing insertion loss, link penalty due to passive link impairments become non-negligible if not critical. Due to multiple transitions points between medium and increased I/O density, the noise component due to reflections and crosstalk will increase and become harder to mitigate. Moreover, intra-pair skew can amplify jitter and inter-symbol interference ultimately degrading the eye diagram. For 400G/lane, the receiver skew tolerance window narrows significantly to 1.8 ps for PAM4 and 2.36 ps for PAM6, allocated across the package, PCB, and cable. A comparison of intra-pair skew tolerance of different signaling rate is shown in Table 1, for simplicity only KP4 FEC overhead is considered.

Table 1: Skew tolerance requirement of different signaling rate

Data rate (Gb/s)	106.25	212.5	425		
Modulation	PAM4	PAM4	PAM4	PAM6	PAM8
Time Per UI (ps)	18.8	9.4	4.7	5.9	7.0
Skew tolerance assuming +/- 0.4UI	7.52	3.76	1.88	2.36	2.8

Pluggable modules must overcome critical signal integrity barriers of the front panel connector to survive in 400G per lane. The current golden finger type of connector suffer from the high-frequency roll-off caused by mating stubs and the stringent sensitivity of return loss at Nyquist frequencies. Reducing the stub length to 0.5 mm could increase the effective bandwidth over 80 GHz which is not yet sufficient for PAM4 or PAM6 modulation and comes with the cost of manufacturability.

Meanwhile, emerging interconnect technologies and ongoing advancements in connector design offer promising pathways to overcome these bandwidth limitations. The industry has tried to push the channel bandwidth beyond 100 GHz, and shown evidences to support PAM4 signaling for 400G/lane. A co-packaged copper (CPC) channel with two-phase optimization, board-level followed by connector-level, has been conducted by Luxshare-Tech. As depicted in Figure 2, the insertion loss measured is ∼40 dB at 110 GHz. Besides, Huawei has also demonstrated near-packaged copper (NPC) channels which can be found in [18]. One of these channels shows an insertion loss of 41.2 dB at 106.25 GHz, see Figure 3.
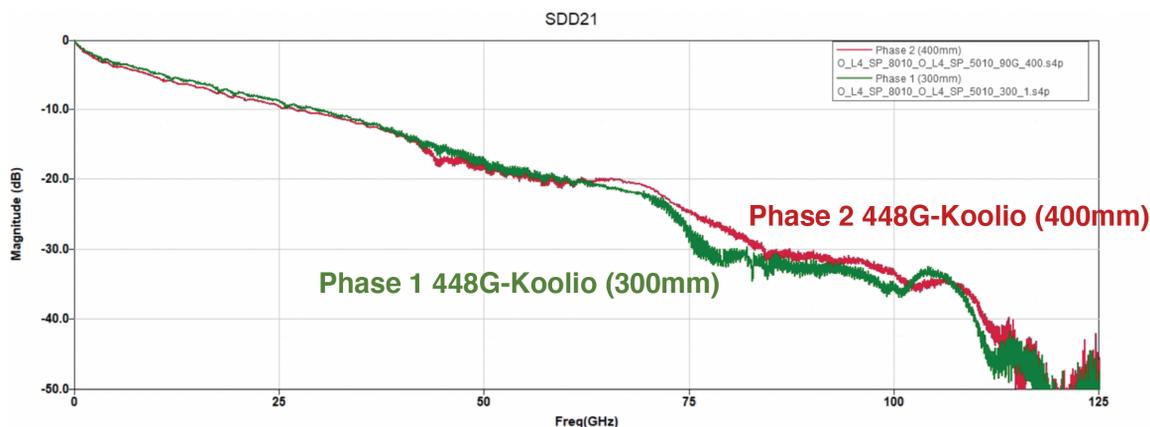

Figure 2: Insertion loss measurement results by Luxshare-Tech with two-phase optimization. Courtesy of Luxshare-Tech.
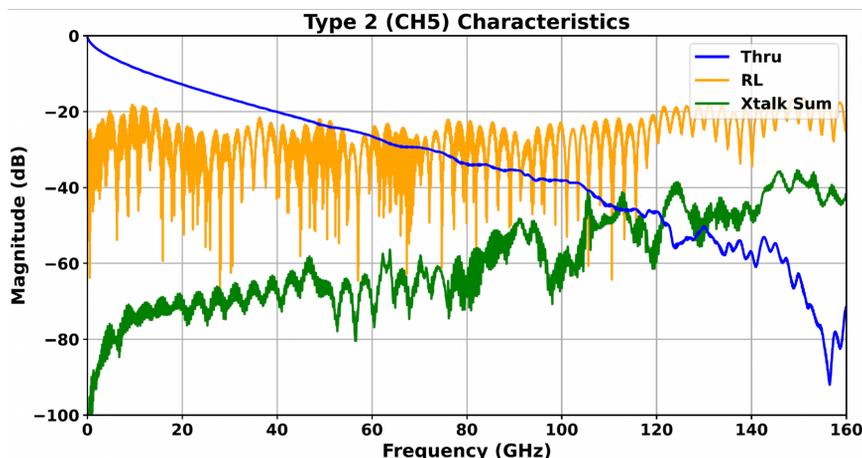

Figure 3: Channel characteristics provided by Huawei. Courtesy of Huawei.

Highly shielded connector interfaces and simplified structures with fewer transitions mitigate crosstalk (by ~20 dB) and impedance mismatches [24]. To combat the "fiber weave effect" (where P and N traces encounter different dielectric constants), designers use spread-glass fabrics (e.g., 1067 or 1078) and route traces at an angle (panel rotation) to the PCB weave [25]. For flyover and DAC applications, manufacturers use co-extrusion techniques to ensure signal conductors are perfectly centered and coupled, achieving "hyper low skew" typically below 1.75 ps/m [26].

3.2 SERDES Technology Status

While 200G/lane introduced strong DSP, 400G/lane mandates even more sophisticated equalizer architecture, enhanced AFE driving capabilities and high-precision CDR to combat extreme high-frequency roll-off. Longer, more complex feed-forward equalizers (FFEs) and decision feedback equalizers (DFEs) are implemented, often with maximum likelihood sequence detection (MLSD) or Bahl-Cocke-Jelinek-Raviv (BCJR) algorithms to unravel severe inter-symbol interference (ISI).

PAM6 increases the complexity of the DSP architecture compared to PAM4, as PAM6 require two times more comparators in unrolled DFE or transitions in MLSE/BCJR. 400G/lane also requires ultra-low jitter CDR designs. The industry is moving toward high-frequency, low-phase-noise phase-locked loops (PLLs) and sophisticated timing recovery algorithms to handle the eye closure caused by multi-level modulations like PAM6.

Furthermore, TX DAC and RX ADC may need an extra bit for PAM6 for achieving the same coefficient resolution as for PAM4. To overcome channel loss, the AFE must deliver higher output swings without introducing non-linearities. This requires 3 nm or 2 nm CMOS processes to reduce the parasitic capacitance of the transistor and provide the necessary analog bandwidth (f_T/f_{max}) to more than 400 GHz (ideally 3.5 to 4 times the Nyquist frequency to allow for sufficient analog headroom, gain, and feedback loop stability). These processes also enable the integration of massive DSP logic and high-speed ADC/DAC arrays—running at sampling rates up to 256 GSa/s—within a manageable thermal and area envelope.
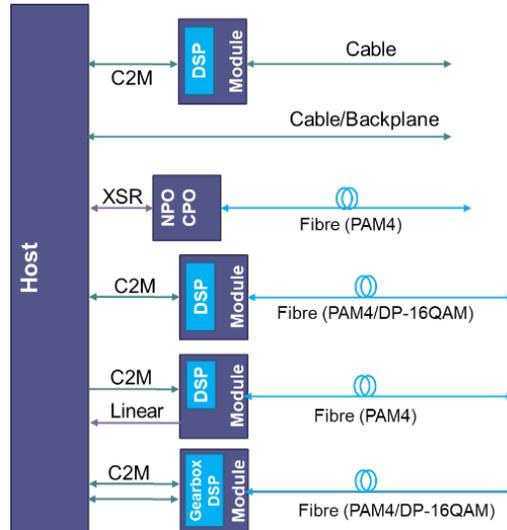
4 Ethernet PHY Architectures

Figure 4: Ethernet PHY architecture options showing Host-to-Module connections with various interface types: C2M (Chip-to-Module), XSR (Extra Short Reach), and Linear configurations.

The signal path from the host system to the transmission medium is established through a chain of functional blocks. The chain begins at the Host, which not only originates the data but also encodes it using Reed-Solomon Forward Error Correction (RS-FEC) to protect against transmission errors. This signal traverses an Electrical Interface acting as the physical bridge between the host

ASIC and the transceiver. In standard implementations, the Module encapsulates a Digital Signal Processor (DSP) to perform signal conditioning and retiming. Finally, the signal is converted for the transmission medium, such as copper cable, backplane, or optical fiber.

From this baseline, several variations emerge to address power, latency, and reach constraints. Linear Pluggable Optics (LPO) attempt to eliminate the DSP entirely by passing analog signals directly from the host. While efficient at lower rates, LPO approaches face practical limits at 200G per lane, where the degraded SNR of 400G PAM4 and passive-link bandwidth demands make purely linear front-panel implementations increasingly difficult. Linear Receive Optics (LRO) offer an intermediate asymmetric structure with linear transmission in one direction and DSP in the other, though this introduces co-design complexities between electrical and optical interfaces. Near-Packaged and Co-Packaged Optics (NPO/CPO) address signal integrity challenges by shortening the electrical channel through Extra Short Reach (XSR) interfaces. This architecture structurally relies on PAM4 electrical signaling to align with the optical interface and achieve system-level density targets. The Gearbox configuration serves as a critical transition tool, bridging two 200G electrical lanes to match a single 400G optical lane.

In conclusion, the 400G era prioritizes high-speed electrical PAMn links, while fiber optical interconnects remain the primary vehicle for 400G PAM4. This divergence underscores that DSP solutions remain the default for reach and NPO/CPO offers a path for density, while linear front-panel approaches face increasing resistance from 400G physical constraints.

5   FEC Considerations for Electrical Links

Due to the uncertainty of whether the industry can break through in terms of the electrical channel bandwidth, both PAM4 and PAM6 are possible modulation candidates for 400G/lane currently. The RS(544,514), also known as KP4, has been adopted as the forward error correction (FEC) scheme for PAM4 signaling from 25G/lane to the recent 200G/lane in IEEE Ethernet standards [27–29]. A natural question is whether KP4 can continue to fulfill the requirement of error performance for 400G/lane with PAM4 or PAM6 signaling. If the answer is negative, considering the backward compatibility, concatenated FEC schemes with KP4 being the outer code are likely to be potential solutions.

For PAM4 with KP4 only, the signaling rate is 212.5 GBd which already requires the Nyquist frequency of 106.25 GHz. Concatenated FEC for electrical PAM4 signaling becomes unacceptable as the inner code leads to a higher signaling rate, and thus further challenges the bandwidth limitation. For PAM6, the signaling rate is 170 GBd with KP4 only, requiring the Nyquist frequency of 85 GHz. Thus, an inner code with small overhead is still possible. For instance, an additional inner code with 6.67% overhead increases the Nyquist frequency to 90.75 GHz. On the other hand, compared to PAM4, PAM6 has a 3 dB signal-to-noise ratio (SNR) degradation for the same bit error rate without FEC due to the narrower spacing between adjacent symbol levels. The inherent SNR degradation for PAM6 further necessitates an inner code.

In the following, FEC architectures will be discussed for PAM4 and PAM6. Inner FEC design for PAM6 will be demonstrated in detail. In addition, the error performance of different FEC schemes is evaluated by the block error ratio (BLER) method introduced in IEEE P802.3dj [29], and is shown at the end of this section.

5.1   FEC consideration for PAM4 over Electrical Links

As argued aforementioned, only KP4 is considered as the FEC scheme in this case. Both electrical and optical links have PAM4 signaling. The unified modulation facilitates all existing optical packaging technologies, including retimed pluggable module, CPO, NPO, LPO, and LRO. Architectures supporting different optical packaging technologies are depicted in Figure 5. Note that there may be an inner code within the retimed pluggable module for optical reach between 500 m and 2 km.
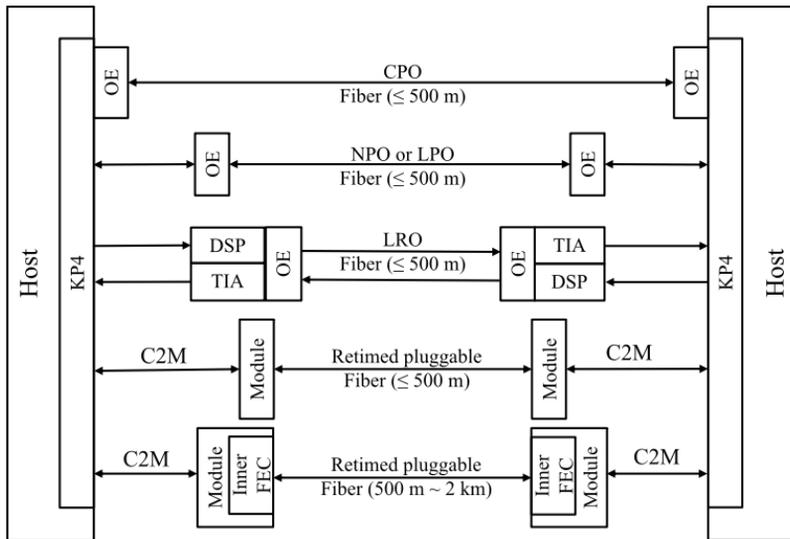
Figure 5: FEC architecture supporting PAM4 over both electrical and optical links.

5.2 FEC consideration for PAM6 over Electrical Links

PAM6 constellation is not as trivial as PAM4 as the number six is not a power of two. Existing popular methods for mapping bits to PAM6 symbols exploit a two-dimensional constellation with six levels in each dimension. In total, there are 36 two-dimensional constellation points. By ignoring four specified points, the remaining 32 points allow the mapping from five bits to a pair of one-dimensional PAM6 symbols, i.e., 2.5 bits per symbol. Depend on which four points are ignored, there are two commonly discussed two-dimensional constellations, cross QAM-32 and framed-cross QAM-32 [30, 31].
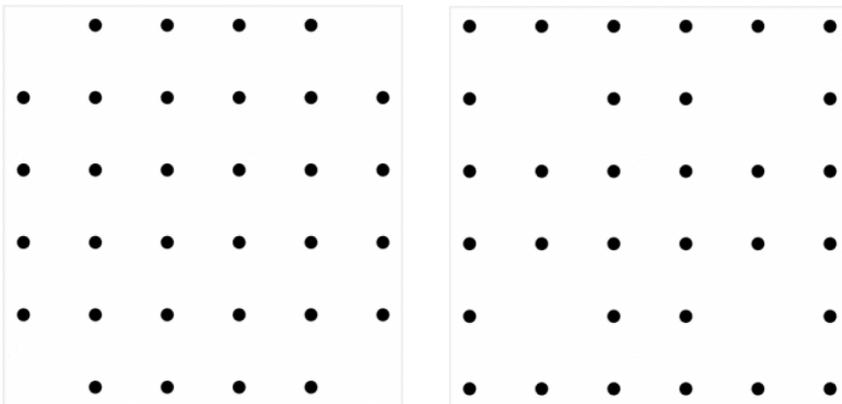

Figure 6: Cross QAM-32 constellation (left) and framed-cross QAM-32 constellation (right).

While PAM6 signaling is over the electrical links, the optical links still utilize PAM4 modulation. The mismatch between different links will prohibit the use of CPO, NPO, LPO, and LRO technologies, leaving the industry with retimed pluggable module as the only option. FEC Architectures with KP4 only supporting retimed pluggable module are the same as those depicted in Figure 5. Concatenated FEC Architectures supporting retimed pluggable module are depicted in Figure 7 where FEC1 is the inner code for electrical links and FEC2 is the inner code for optical links with reach between 500 m and 2 km. When the FEC2 is not needed, the inner FEC1 shall be terminated within the optical modules for partially handling possible errors over the electrical links. When the FEC2 is needed, the inner FEC1 shall be terminated within the optical modules before FEC2 encoding.
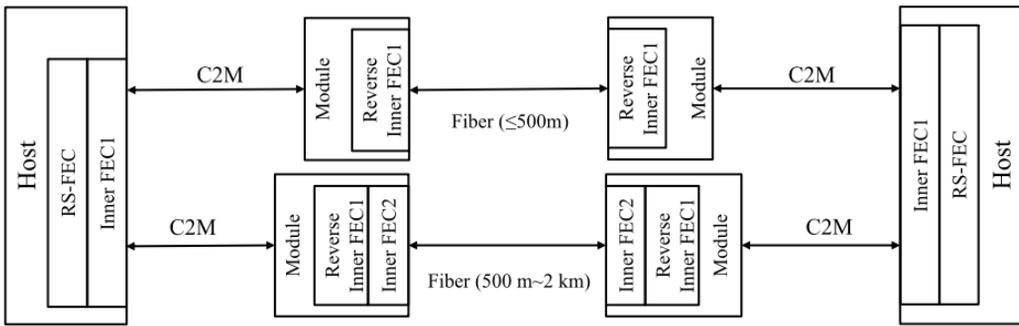
9

Figure 7: FEC architecture supporting PAM6 over electrical links and PAM4 over optical links.

As discussed above, inner codes with low overheads are preferred for PAM6 signaling over electrical links. The simplest way to protecting information bits is to directly encode all the bits with a code. For example, the extended BCH(128,120) code generates eight parity bits based on 120 information bits with a 6.67% overhead. A codeword is obtained by appending the eight parity bits to the 120 information bits. The extended BCH(128,120) code can correct one erroneous bit within a codeword using hard-decision decoding. For the same length of a codeword, reducing the number of information bits (equivalently increasing the number of parity bits) enhances the error correction capability. However, in this manner, the code overhead inevitably gets enlarged. It is desired to improve the error correction capability while keeping the overhead unchanged. Set partitioning (SP) is a promising technique to achieve this goal, which allows a code to encode less bits with the same number of parity bits. The concept of SP dates from [32], and SP-based inner code design has been shown in [33, 34]. In the following, the cross QAM-32 constellation is exemplified to demonstrate the code design idea based on SP.

According to SP, the constellation points can be divided into two subsets or four subsets, as shown in Figure 8. The colors of constellation points denote to which subset they belong. For simplicity, let SPx denote that x subsets are considered. For SP2, one partition bit is associated with each constellation point to indicate the subset information. An inner code based on SP2 encodes $4n + k$ information bits with an (n, k) code. In detail, the first $5k$ information bits correspond to k constellation points. Then, the associated k partition bits are encoded by the (n, k) code to obtain $n - k$ parity bits. Note that there are $4(n - k)$ remaining information bits. Each parity bit is further viewed as a partition bit to append one bit to every four remaining information bits, such that the five bits correspond to the subset indicated by the parity bit. The inner code based on SP2 is denoted by SP2$(5n, 4n + k)$. Differently from the direct encoding, SP2$(5n, 4n + k)$ only encodes partition bits derived from the information bits. If an error in the partition bits is corrected by the implicit (n, k) code, the demodulation decision should also be corrected and should be done within the right constellation subset. A graphical illustration of SP2$(5n, 4n + k)$ encoding is provided in Figure 9.
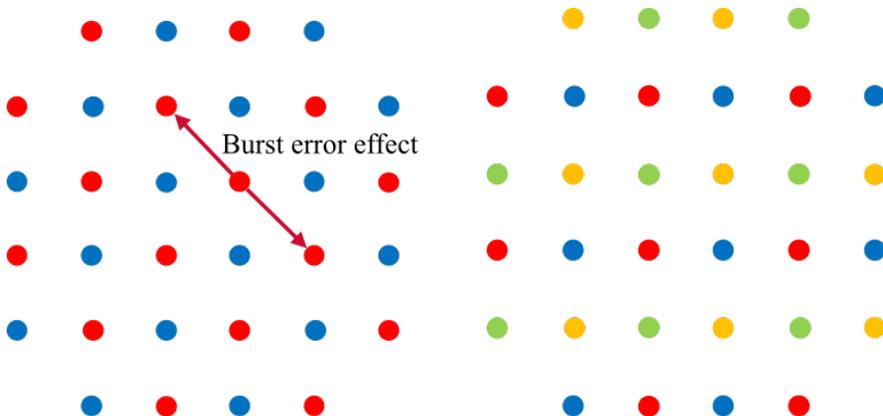

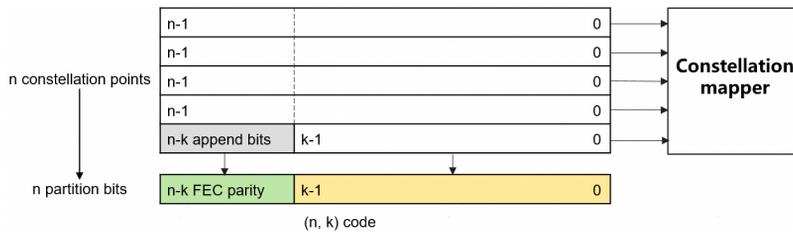Figure 8: Two subsets (left) and four subsets (right) of cross QAM-32 constellation.

Figure 9: Illustration of SP2($5n, 4n + k$) encoding.

Due to the existence of intersymbol interference (ISI) over electrical links, equalization techniques, such as DFE, MLSE, and BCJR algorithm, are usually adopted to handle the ISI. However, these equalization techniques cause burst errors over modulation symbols. With the burst error effect, difference between the transmitted symbol sequence and that after equalization is a zigzag sequence with alternating -1 and +1 [35]. As a result, for PAM6, a two-dimensional constellation point will be corrupted to its adjacent point on the diagonal. However, the burst error effect happens in the same subset for SP2, see Figure 8 (left). The implicit (n, k) code cannot detect such errors as they cause no flips on the partition bits. Motivated by this issue, the inner code based on SP4 is considered. As shown in Figure 8 (right), adjacent constellation points on the diagonal are separated into different subsets for SP4. Each constellation point has two associated partition bits. If a $(2n, 2k)$ code is employed to encode the partition bits, the final inner code should be denoted by SP4($5n, 3n + 2k$). The encoding of SP4($5n, 3n + 2k$) can be easily extended from SP2($5n, 4n + k$), and thus omitted here. A graphical illustration of SP4($5n, 3n + 2k$) encoding is provided in Figure 10.
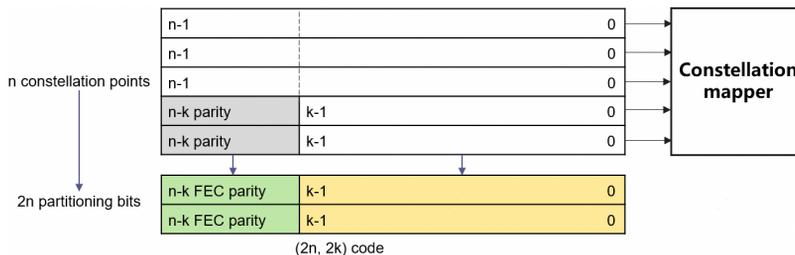

Figure 10: Illustration of SP4($5n, 3n + 2k$) encoding.

The error performance of different FEC schemes is provided in Table 2 using one of Huawei channels [36]. Bin i of the KP4 error histogram is related to the BLER method and indicates the probabilities of a RS codeword with i symbol errors.

Table 2: Error performance of different FEC schemes

Modulation		PAM4		PAM6	
FEC		KP4	KP4	KP4+SP4(180,170)	KP4+eBCH(128,120)
Data rate [Gbps]		425	425	450	453.75
Insertion Loss [dB]		41.3	33.1	33.7	33.6
ICN@bump [mV]		0.85	0.37	0.42	0.43
Alpha		0.98	0.46	0.51	0.53
BER (MLSE HD)		2.97e-7	1.47e-6	3.62e-6	5.48e-6
BER (after inner FEC HD)		-	-	7.36e-7	2.37e-7
BER (after inner FEC SD)		-	-	<1.25e-8	<1.25e-8
KP4 error histogram	Bin 1	1.62e-3	5.54e-3	2.07e-3	7.25e-4
	Bin 2	-	-	-	8.06e-5
	Bin 3	-	-	-	-
	Bin 4	-	-	-	-

6 Comparative Analysis: PAM4 vs PAM6 Electrical

Given the FEC anchor point and implementation landscape, two primary paths are being considered for the 400G electrical interface: continuing with PAM4 at higher baud rates, or transitioning to PAM6 at lower baud rates.

6.1 PAM4 Electrical Interface

The PAM4 approach leverages the logic architecture that the industry has developed over the past four years and maximizes reuse of existing IP and infrastructure. Under this approach, the inner FEC, referred to as FEC1, could reuse the scheme currently employed in 200G/L optics for reaches over 500 m. This same inner code can also be applied to high-loss electrical channels, providing a unified solution across use cases.
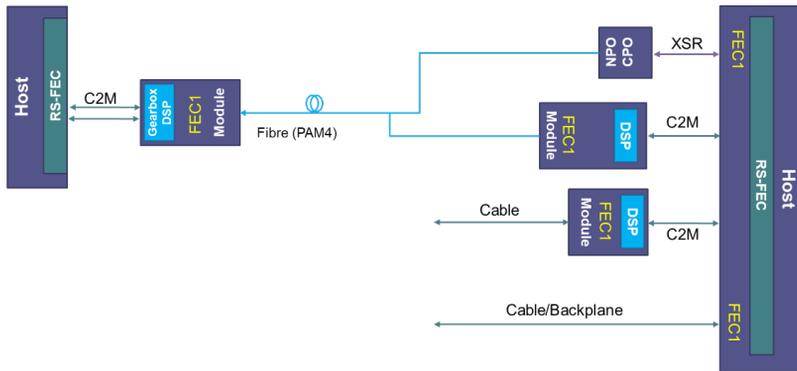

Figure 11: PAM4 electrical interface architecture with FEC1 inner coding. Shows Host with RS-FEC connecting via C2M to Gearbox DSP modules and various module configurations including NPO/CPO (XSR), DSP modules with C2M, and direct Cable/Backplane connections.

The primary engineering challenge lies in advancing cable, connector, and packaging technology to ensure that passive channel bandwidth remains sufficient for the increased baud rates required by PAM4. Importantly, the passive-link constraint is not static: connector, cable, and packaging improvements through 2025 can deliver a clean response to approximately 110 GHz, directly alleviating the feasibility concern for high-baud PAM4. This trajectory narrows the gap to the bandwidth levels often cited for 400G/lane PAM4 implementations, on the order of ~112 GHz, even though difficult use cases such as high-loss backplanes remain challenging.

6.2 PAM6 Electrical Interface

The PAM6 approach reduces the required baud rate for a given data rate and can appear attractive as a response to bandwidth limitations. This lower Nyquist frequency, approximately 85–90 GHz, eases the burden on passive channels and may extend the reach over copper links where high-frequency loss is the dominant constraint.

However, PAM6 introduces trade-offs that must be managed. The reduced spacing between voltage levels makes the signal more vulnerable to noise and link impairments. To compensate while preserving the mandatory KP4 outer FEC anchor, a stronger inner FEC, referred to as FEC2, becomes necessary. This additional coding layer translates into system considerations, including potential increases in latency, power consumption, and silicon die area. Furthermore, architectural alignment with NPO/CPO, which currently favors PAM4, becomes a point of divergence that requires careful system partitioning.
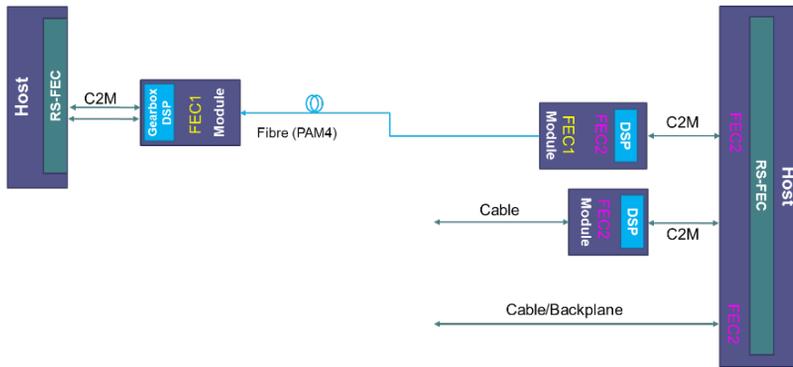
Figure 12: PAM6 electrical interface architecture with FEC2 inner coding. Shows Host with RS-FEC and FEC2 layer connecting via C2M to Gearbox DSP modules and various module configurations including NPO/CPO (XSR), DSP modules, and direct Cable/Backplane connections.

6.3 Comparison

As established above, standalone KP4 provides insufficient coding gain to close the link budget for 400G per lane signaling, making a concatenated inner FEC a mandatory architectural component. Among candidate options, Hamming, SP4, and BCH are frequently evaluated.

Hamming(128,120) benefits from computational efficiency and lightweight implementation, making it a preferred choice for PAM4 systems where the SNR margin is manageable. For PAM6, an inherent modulation penalty is often characterized as an SNR loss of multiple dB, which typically leads to the consideration of more complex inner codes such as SP4(180,170) or BCH(128,120). These stronger codes provide additional gain but involve steeper costs in latency, power consumption, and silicon die area.

Table 3 summarizes key requirements for possible FEC configurations and modulation formats.

Table 3: FEC Configurations and Modulation Formats Comparison

Modulation Format	PAM4		PAM6		
Outer FEC	KP4				
Inner FEC	—	Hamming(128,120)[a]	—	SP4(180,170)	BCH(128,120)[a]
Inner FEC OH [%]	0	6.67	0	5.88	6.67
Data rate [Gbps]	425	453.75	425	450	453.75
Signaling rate [GBaud]	212.5	226.875	170	180	181.5
Nyquist frequency [GHz]	106.25	113.4375	85	90	90.75
Input SNR[b] @1e-13 [dB]	17.53	14.80	20.45	18.41	17.84

[a]Pad insertion may be needed for an integer multiple of common reference clock.

[b]Input SNR values are based on AWGN performance.

While PAM6 reduces the Nyquist frequency, this bandwidth benefit is counterbalanced by the modulation's SNR penalty and the resulting need for heavier inner FEC. The choice between PAM4 and PAM6 thus becomes a system-level optimization problem: balancing the bandwidth relief of PAM6 against the implementation overhead of stronger FEC, versus managing the signal integrity demands of PAM4 to leverage a lighter coding stack.

7 Conclusion

Optical signaling is progressing with high confidence toward 400G per lane, with broad convergence on PAM4 and with multiple viable component and integration approaches under active development. The dominant uncertainty is electrical: passive high-speed components and channels remain the key limiter when progressing toward 400G/lane electrical interconnection. Bandwidth progress through 2025 is encouraging for PAM4-style high-baud approaches, but difficult loss and impair-

ment environments continue to require careful channel engineering and, in many cases, architectural mitigation.

With RS-FEC (KP4) fixed as the end-to-end outer code for interoperability, the FEC stack makes the PAM4/PAM6 trade-offs explicit. While PAM6 offers a path to reduce bandwidth requirements, it typically necessitates stronger inner FEC, introducing considerations around latency, power, and die area. In contrast, PAM4 maintains continuity with existing architectures and efficiency targets but demands high-performance passive links.

For AI infrastructure applications where power efficiency and latency are paramount, the final architectural direction will be determined by the trajectory of passive link development. The critical gating factor is whether the ecosystem can deliver connectors, cabling, and packaging capable of sustaining bandwidths exceeding 112 GHz. If this performance threshold is met, the industry is likely to favor the continuity and efficiency of PAM4; otherwise, the channel constraints will necessitate the complexity and latency trade-offs of PAM6.

References

[1] Z. Jiang *et al.*, "MegaScale: Scaling Large Language Model Training to More Than 10,000 GPUs," in *Proc. USENIX NSDI*, 2024, pp. 745–760.

[2] I. Lamaakal *et al.*, "Tiny Language Models for Automation and Control: Overview, Potential Applications, and Future Research Directions," *Sensors*, vol. 25, no. 5, p. 1318, 2025.

[3] X. Zhuang *et al.*, "What is next for LLMs? Pushing the boundaries of next-gen AI computing hardware with photonic chips," *PubMed Central*, 2026.

[4] K. Qian *et al.*, "Alibaba HPN: A Data Center Network for Large Language Model Training," in *Proc. ACM SIGCOMM*, 2024, pp. 691–706.

[5] LessWrong Community, "Estimates of GPU or equivalent resources of large AI players for 2024/5," 2024. [Online]. Available: https://www.lesswrong.com/posts/bdQhzQsHjNrQp7cNS/estimates-of-gpu-or-equivalent-resources-of-large-ai-players

[6] Epoch AI, "AI training cluster sizes increased by more than 20× since 2016," 2024. [Online]. Available: https://epoch.ai/data-insights/training-cluster-size

[7] Z. Wen *et al.*, "ALIBI: Breaking the GPU Memory Wall for LLM Training," *arXiv preprint arXiv:2410.03861*, 2024.

[8] TechInsights, "Broadcom Ships Tomahawk 5, Industry's Highest Bandwidth Switch Chip," 2024. [Online]. Available: https://www.techinsights.com/blog/tomahawk-5-switches-512tbps

[9] IET Research, "Co-packaged datacenter optics: Opportunities and challenges," *IET Optoelectronics*, 2024.

[10] Upsite Technologies, "Data Center Cable and Airflow Management," 2024. [Online]. Available: https://www.upsite.com/blog/data-center-cable-management-and-airflow-management/

[11] ENERGY STAR, "Manage Airflow for Cooling Efficiency," 2024. [Online]. Available: https://www.energystar.gov/products/data_center_equipment/16-more-ways-cut-energy-waste-data-center/manage-airflow-cooling-efficiency

[12] Optical Internetworking Forum, "OIF Charts Path to 448G/Lane Interconnects," 2025. [Online]. Available: https://convergedigest.com/oif-charts-path-to-448g-lane-interconnects/

[13] TechInsights, "Broadcom Tomahawk 6: 102.4 Tbps Ethernet Switch for AI Data Centers," 2024. [Online]. Available: https://www.techinsights.com/blog/broadcom-tomahawk-6-1024-tbps-ethernet-switch-ai-data-centers

[14] Next Platform, "The AI Datacenter Is Ravenous For 102.4 Tb/sec Ethernet Switch ASICs," Jun. 2025. [Online]. Available: https://www.nextplatform.com/2025/06/03/the-ai-datacenter-is-ravenous-for-102-4-tb-sec-ethernet/

[15] Synopsys/Wisdom Interface, "Designing for 448G: Modulation, DSP, and Channel Trade-offs," 2025. [Online]. Available: https://www.wisdominterface.com/wp-content/uploads/2025/09/designing-for-448g-modulation-dsp-wp.pdf

[16] P. Bhasker *et al.*, "413 Gbits/s PAM-6 O-band CWDM Electroabsorption Modulated Lasers for 400G per lane IM-DD Applications," in *Proc. Optical Fiber Communication Conf. (OFC)*, 2025, pp. 1–3.

[17] S. Okuda *et al.*, "High-speed 340 Gbps PAM4 and 450 Gbps PAM6 Operations of Narrow High-Mesa EML," in *Proc. Optical Fiber Communication Conf. (OFC)*, 2025, pp. 1–3.

[18] X. Chen *et al.*, "540Gbps IMDD Transmission over 30km SMF using 110GHz Bandwidth InP EML," in *Proc. Optical Fiber Communication Conf. (OFC)*, 2025, pp. 1–3.

[19] Y. Ogiso *et al.*, "Uncooled O-band InP MZ Modulator PIC for 3.2 Tb/s (400 Gb/s/lane) Pluggable Transceiver," in *Proc. Optical Fiber Communication Conf. (OFC)*, 2025, pp. 1–3.

[20] H. Wang *et al.*, "90 GHz Silicon Mach-Zehnder Modulator with Integrated Equalizer for 1.6 Tbps (200G/λ) IMDD Transceivers," in *Proc. Optical Fiber Communication Conf. (OFC)*, 2025, paper M2K.1.

[21] L. E. Johnson *et al.*, "Multi-Channel Silicon-Organic Hybrid PICs for 200G/λ and 400G/λ PAM4 Transmission," *arXiv preprint*, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2509.24825

[22] C. Bruynsteen *et al.*, "110 GHz GeSi Electroabsorption Modulator on a 300mm SiPh Platform Enabling High-Density 400G/lane IM/DD Links," in *Proc. European Conf. on Optical Communication (ECOC)*, 2025, pp. 1–4.

[23] S. Phadke *et al.*, "IEEE 802.3 E4AI Study Group - Electrical Interface Considerations," 2025. [Online]. Available: https://www.ieee802.org/3/ad_hoc/E4AI/public/25_0624/phadke_e4ai_01_250624.pdf

[24] TE Connectivity, "400G AI Workshop - Connector Technologies for High-Speed Signaling," SNIA SFF Workshop, Jan. 2025.

[25] Intel, "AN-528: PCB Dielectric Material Selection and Fiber Weave Effect on High-Speed Channel Routing," 2025. [Online]. Available: https://www.intel.com/content/www/us/en/content-details/654621/an-528-pcb-dielectric-material-selection-and-fiber-weave-effect-on-high-speed-channel-routing.html

[26] Samtec, "Hyper Low Skew Twinax Cable for 224 Gbps Applications," *Signal Integrity Journal*, 2024. [Online]. Available: https://www.signalintegrityjournal.com/articles/3949-samtec-announces-hyper-low-skew-twinax-cable-for-224-gbps

[27] IEEE Standards Association, "IEEE Standard for Ethernet," *IEEE Std 802.3-2022 (Revision of IEEE Std 802.3-2018)*, pp. 1–7023, 2022.

[28] IEEE Standards Association, "IEEE Standard for Ethernet Amendment 9: Media Access Control Parameters for 800 Gb/s and Physical Layers and Management Parameters for 400 Gb/s and 800 Gb/s Operation," *IEEE Std P802.3df*, 2024.

[29] IEEE P802.3dj Task Force, "IEEE P802.3dj 200 Gb/s, 400 Gb/s, 800 Gb/s, and 1.6 Tb/s Ethernet Amendment: Physical Layers and Management Parameters for 200 Gb/s and 400 Gb/s Electrical Interfaces," Draft 3.0, 2025. [Online]. Available: https://www.ieee802.org/3/dj/private/8023dj_D3p0.pdf

[30] T. Prinz *et al.*, "Comparison of PAM-6 modulations for short-reach fiber-optic links with intensity modulation and direct detection," in *Proc. Eur. Conf. Opt. Commun. (ECOC)*, 2022, pp. 1–3.

[31] T. Prinz *et al.*, "PAM-6 Coded Modulation for IM/DD Channels with a Peak-Power Constraint," in *Proc. 11th Int. Symp. on Topics in Coding (ISTC)*, 2021, pp. 1–5. [Online]. Available: https://doi.org/10.1109/ISTC49272.2021.9594181

[32] U. Wachsmann, R. F. H. Fischer, and J. B. Huber, "Multilevel codes: Theoretical concepts and practical design rules," *IEEE Trans. Inf. Theory*, vol. 45, no. 5, pp. 1361–1391, Jul. 1999.

[33] C. Liu, "Performance Analysis at 400+Gbps Over Next-Generation VSR Channels," *Ethernet Alliance Technology Exploration Forum*, 2024.

[34] X. He *et al.*, "FEC for 448G: Can Today's FEC Survive Tomorrow's Modulation?" *Ethernet Alliance Technology Exploration Forum*, 2025.

[35] H. Shakiba, "Error Propagation Analysis of MLSE," *IEEE Std 802.3dj*, 2023. [Online]. Available: https://ieee802.org/3/dj/public/adhoc/electrical/23_0420/shakiba_3dj_elec_02_230420.pdf

[36] X. He *et al.*, "A Preliminary Study of Modulation and FEC for 400 Gb/s per Lane NPC Channels," *IEEE 802.3 E4AI Ad Hoc*, Oct. 2025. [Online]. Available: https://www.ieee802.org/3/ad_hoc/E4AI/public/25_1023/he_e4ai_01_251023.pdf