800G Pluggable

MULTI-SOURCE AGREEMENT

200G PER LANE FOR FUTURE 800G & 1.6T MODULES

PROMOTERS:

Accelink AOI Baidu 百度

FUJITSU H3C Hisense CIG

HUAWEI SUMITOMO ELECTRIC YAMAICHI ELECTRONICS

Tencent 腾讯 TTL 中国泰尔实验室 China Telecommunication Technology Labs Inphi Think fast.

LUXSHARE ICT LUMENTUM

CONTRIBUTORS:

II-VI Anritsu envision:ensure Centec COMMSCOPE

CREDO we connect eoptolink FIT FiberHome

HGTECH INNOLIGHT KEYSIGHT TECHNOLOGIES

Linktel TECHNOLOGIES HISILICON SEMTECH Sitrus 柏科微电子

NeoPhotonics Tektronix USCONEC VIAVI

MARVELL source PHOTONICS SONT

SiFotonics MAXLINEAR multiLane ZTE

Contents

1. Data center evolution beyond 400G

After many years of standardization and an evolution towards the compact QSFP-DD form factor, 400G modules based on 4x100G line technology and 8x50G AUI began ramping in 2020 for lead hyper scale customers. Meanwhile 100G modules are dominating the overall market predominantly using native 4x25G. These days, the market seems ripe for a technological evolution beyond 400G.

Since the establishment of the 800G Pluggable MSA in 2019, the work on beyond 400G components and standardization has picked up pace and 800G data center architectures are actively discussed in the industry. As shown in Figure 1, different network architectures are needed for AI-focused or classical hyper scale data centers. While the hyper-scale multi-tier data center network employs oversubscription due to the nature of the short packet traffic, AI networks may resemble circuit switches without any oversubscription due to the larger data load exchanged between servers. The connections between top-of-the rack switch (TOR) and leaf switch are in the range of 50m to 60m and can be addressed by 50m (VR) and 100m (SR) interfaces, while leaf to spine and spine to DCI typically use 500m (DR) and 2km (FR) modules.

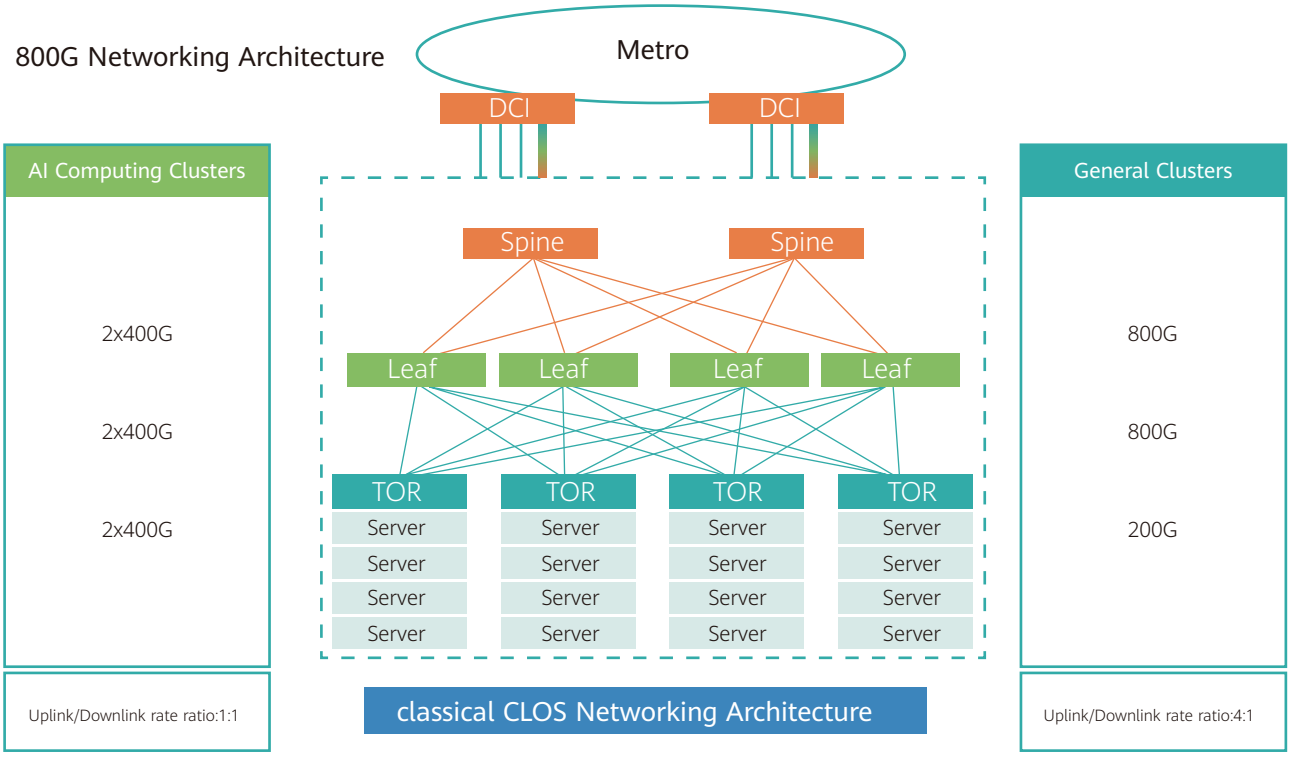


Figure 1 – Data center evolution towards 800G [Ref.1]

Ref. 1: 'Data Center 800G Interconnection Applications and Demands' from Baidu, 800G Pluggable MSA Regular Meeting in Oct. 2019.

The first 800G QSFP-DD and OSFP modules are being sampled in 2021, as shown in Figure 2 (left), and in itself constitute a densification of the standardized 400G interfaces. Here, the power decrease vs. 400G is mainly achieved using native 100G on host and line side and a more advanced processing node for the modem ASICs. Figure 2 shows the technological evolution of Serializer/Deserializer (SerDes) electrical input and output technology. As it is shown, 100G SerDes will start ramping in the market in 2021 and likely become the dominant high-speed interface by 2023. 100G is also expected to become a long living node given the complexity around 200G SerDes design and the implication of performance limitations of electrical 200G on the target architectures and use cases.

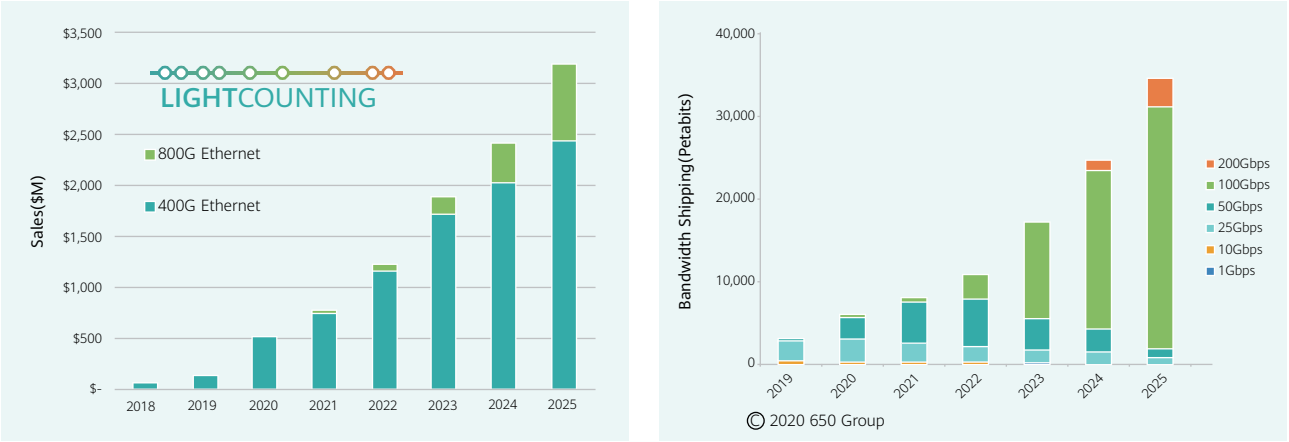


Figure 2 – 800G projections vs. SerDes deployments

The MSA has therefore focused on the definition of cost effective high speed optical interfaces assuming 100G SerDes and published the specifications on low cost 800G (8x100G) PSM8 for 100m and above in 2020. This interface specification will address the 100m reach class, next to multi-mode fiber (MMF) and VCSEL-based modules, and provide the operators a future-proof fiber plant towards 1.6Tb/s and beyond.

For 500m, a reuse of IEEE's DR4 standards and existing 100G single lane optics makes the most economical sense. 800G DR8/2xDR4 with native 100G is optimally suited to leverage 100G SerDes and address 100G fan out necessary in some applications. Introducing a potential 800G DR4 with 4x200G will only succeed when 200G single lane connectivity to the server becomes necessary and 200G SerDes becomes a trend in the market place, and thus is currently not a focus. Moreover, the timeline of 200G SerDes adoption is currently unclear and may shift out further compared to Figure 2 (right), which is likely given the complexity around that technology.

For the 800G 2km FR use case, CWDM4 with 200G/lane optical technology can provide a more cost optimized connectivity compared with 8x100G for higher data center tiers. In 2021, the first 200G/lane PAM4 industry specification was released with the 800G (4x200G) CWDM4 for 2km by the MSA. The specification assumes the future dominant 100G SerDes with 8x100G AUI and includes an additional forward error correction (FEC) in the PHY, leveraging the KP4 FEC from 400GbE clients on the host side and thus achieving full backwards compatibility.

2. 800G Pluggable MSA and related SDOs

The MSA was the first industry alliance focusing on 800G interoperability specifications when founded in 2019. Since then activities have begun in other standards developing organizations (SDO), such as IEEE and OIF. After the conclusion of the bandwidth assessment phase, IEEE has kicked off the study group on beyond 400G in 2021 targeting the definition of the standardization objectives within this year, but doesn't make any decisions on target technology or specifications. After the conclusion of the study group, a task force will be created to define the standards, which typically takes around 4 years. As shown in Figure 3, the completion of 800GbE IEEE standards is expected for the end of 2025, which would be well after the introduction of first, pre-standard modules, in the market, justifying the activities of the MSA. So far, IEEE has decided on objectives focusing on 8x100G using MMF for up to 100m, 8x100G PSM8 for up to 500m, 4x200G CWDM4 for up to 2km and 1.6T PSM8 amongst others.

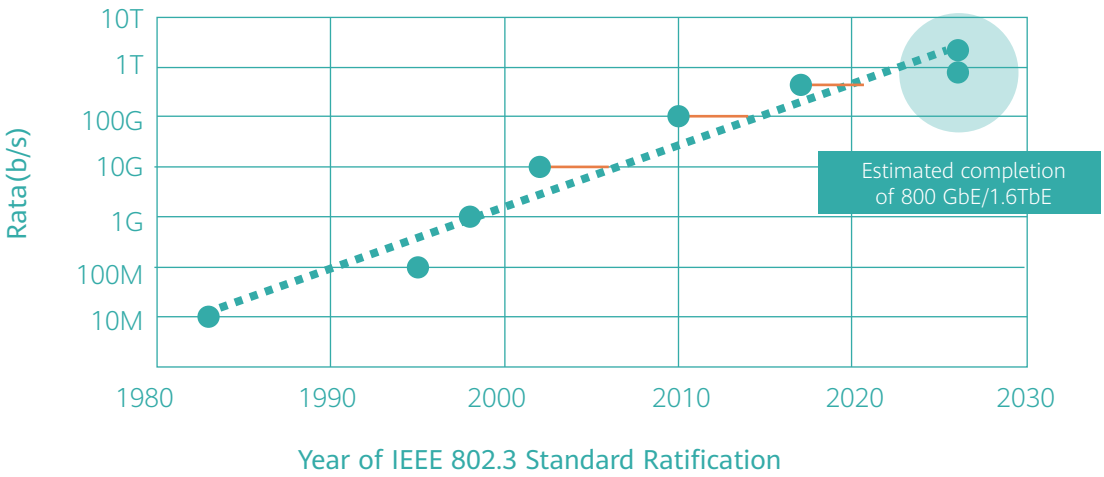


Figure 3 – IEEE standardization roadmap [Source IEEE B400G study group]

800G standardization was also started in 2020 in OIF as an evolution of the 400ZR project, focusing on 2km-10km LR and 80-120km ZR applications using coherent optics. Here, the LR application code targets a data center interconnect (DCI) campus use case. Moreover, OIF has started the standardization of 200G SerDes and historically electrical and optical interconnects used the same modulation formats for short reach interfaces with a sole FEC in the host device, which however might be subject to change. Here, the industry discusses error correction variants with a lower power FEC in the host and extender FECs in optical PHYs.

Overall, it is expected for 200G/lane technology to become the foundation of future 800GbE (4x200G) and 1.6TbE (8x200G) short reach interconnect standards and the work of the MSA lays the foundation for future development and standardization activities. For 1.6TbE, the physical form factor question will drive a lot of the development and standardization discussions in the industry.

3. Progress on 800G PSM8 100m

In order to offer data center operators scalable and future-proof interconnects, the MSA developed 100G/lane PSM8 specifications targeting 100m and above. Using the same fiber plant, the operators can cover 100m and 500m scenarios with a clear evolution path towards 1.6Tb/s PSM8 and beyond. The version 1.0 draft was published in 2020. Since then, transmitter components became available for this application.

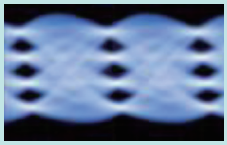
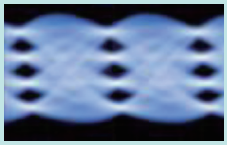
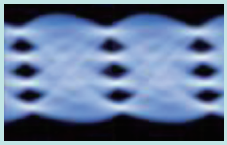
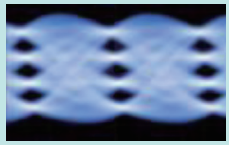
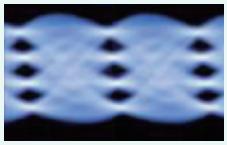
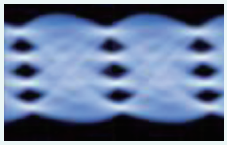
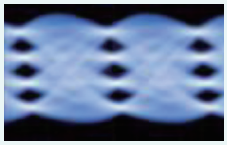
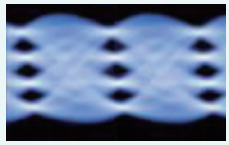
	25℃	50℃	70℃	85℃
53Gbaud-PAM4, SSPRQ in BTB				
After 2 km (λ=1310nm) LUMENTUM				
LD bias	50mA	60mA	70mA	85mA
ER	3.0dB	3.0dB	3.0dB	3.0dB

Figure 4 – Lumentum 100G PAM4 DML Characteristics: Eye Diagram vs. temperature and reach

Figure 4 shows eye diagrams of a Lumentum 100G PAM4 DML tested on carrier with an RF probe. The measurements were conducted using a pulse pattern generator with 53Gbaud (106Gb/s) and an RF amplifier driving the DML, showing clear eye openings in back-to-back transmission and after 2km for temperatures ranging from 25℃ to 85℃ with a relaxed extinction ratio (ER) of 3.0dB.

Figure 5 shows a DML from a second MSA member, AOI. Here, interpolated receiver-sided eye diagram results for a 100G transmission with a TDECQ penalty of 2.54dB using a 5-taps feed forward equalizer (FFE) filter, less than the specification of 4.5dB, and tested on a chip level.

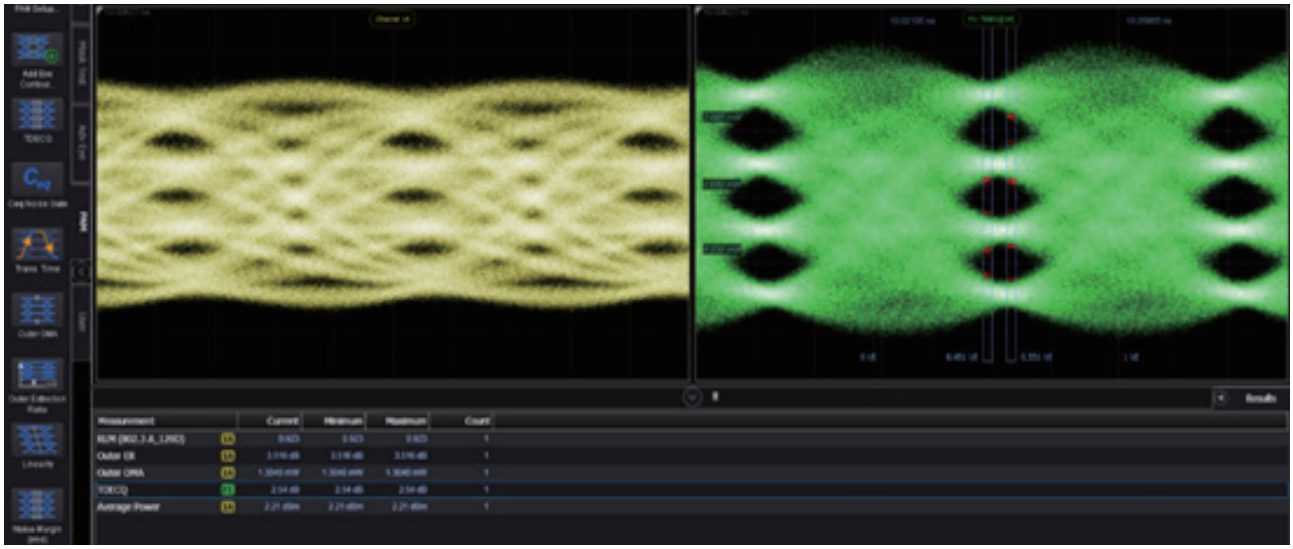
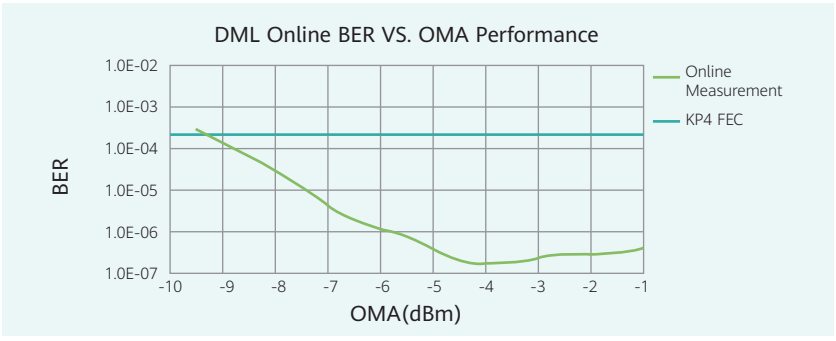


Figure 5 – AOI 100G DML PAM4 receiver-sided eye diagram without equalization (left) and with a 5-taps FFE (right)

At CIOE 2020 in Shenzhen, Huawei showed a live feasibility demonstration of a 100G/lane DML-based module. The vendors for this module demo are shown in Table 1. The BER performance of the module level demo is shown in Figure 6 and it was well below the KP4 FEC limit. The TDECQ for the eye-diagram is 3.0 dB. Thus, the MSA is enabling a new ecosystem for 800G PSM8 modules for reach classes 100m and above.

Table 1 – The vendors participating in the DML based module demo.

Sponsor	Part Name & Pics	Function	Key Features
Vendor within MSA	oDSP Chip	106Gbps per lane PAM4 capability	• High performance CDR • Enhanced equalization • Low power dissipation
LUMENTUM	DML Chip	50G/100G PAM4 capability	• High bandwidth • High ER • Wider operation Temperature • High reliability
USCONEC	• 100m 16-core single mode fiber patch cord • MTP-16 SM APC connector • MTP-LC breakout cable	Parallel SMF fiber link	• High performance MTP-16 connector • LC connector with push-pull tab



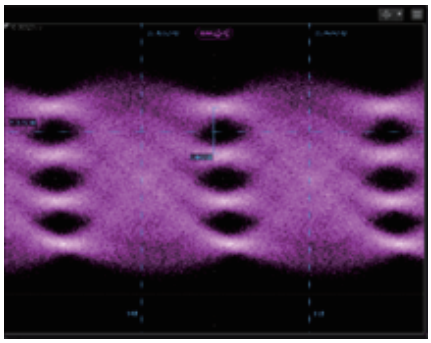


Figure 6 – Huawei module level PSM8 test (left). 100G PAM4 eye diagram at -2dBm OMA (right).

4. 800G FR4 components & specification

4.1 Technology alternatives

400G leaf to spine data center interconnection are addressed either with parallel single mode fiber (PSM) interconnects (DR, 500m) or using 4 lambdas with 20nm coarse wavelength division multiplexing (CWDM4, 2km). For 800G, several technological variants are discussed in the industry for the 2km use case as shown in Table 2. Similarly to the 800G DR8, a densification of 400G FR4 is a technologically more straight forward solution leveraging 100G/lane optics and relying on more advanced CMOS processing nodes to reduce the power consumption of the DSP modem in order to fit into a QSFP-DD/OSFP pluggable. However, in contrary to the DR solution, which requires its inherent feature of the 8x100G break out, the optical lane technology for FR4, which is transmitted over duplex fiber, can be chosen more freely in order to find a more cost optimized solution. The 2x400G FR4 suffers from the higher cost of an increased number of components and therefore cannot compete with 4 lane solutions in the long run.

Table 2 – 800G 2km technological variants

800G options	2x400G CWDM4	CWDM4	Coherent
# Lasers	8	4	2
Laser requirements	DML/EML	EML	Tunable, narrow linewidth, >13dBm
Driver & modulators	8	4	4
PD/TIAs	8 (single-ended PDs)	4 (single-ended PDs)	4 (balanced PDs)
Component bandwidth	>25GHz	>50GHz	>50GHz
FEC limit	2.0E-4	2.0E-3	TBD (higher than IMDD)
Backwards compatible	Yes	Yes	No
Fiber pairs	2	1	1
Power consumption	16-18W	12-14W	20-24W
Cost	$$	$	$$$

Coherent 800G pluggables are being standardized in OIF for transmission above 2km and there has been an ongoing discussion in the industry for the last several years around the applicability of coherent optics inside of the data center. Historically, coherent optics have been introduced for the first time in long haul networks for 40Gb/s and 100Gb/s interfaces due to the inherent limitation of intensity modulated direct detection (IMDD) for larger values of chromatic dispersion (CD) and polarization mode dispersion (PMD). Since then, coherent optics have conquered metro networks and with the arrival of 400G modems are finding a new domain with the 400ZR interconnect standard for 80-120km amplified data center interconnects. The cost difference between coherent and IMDD at 400G is still substantial due to the more sophisticated design of coherent optics and a much lower volume overall. As shown in Table 2, the more sophisticated laser

and generally stronger required forward error correction (FEC) leads to a higher power consumption and prohibitive latency for data center networks, and precludes the use of coherent optics inside of the data center for 2km and below.

4.2 Modulation choice

Higher level modulation in Ethernet was introduced for the first time for 50G per lane signaling using PAM4. For optical transmission, 50G PAM4 is used in 200G (4x50G) optics as well as for 400G LR8. With the transition to 100G, the industry converged on maintaining the same modulation scheme as for 50G, leveraging the design and testing methodology of PAM4. On the host side, electrical 100G PAM4 faced more serious design challenges for use cases such as backplane transmission. The main question for 200G per signaling is whether PAM4 is still feasible or alternative modulation formats needs to be considered, incl. PAM6, PAM8 or DMT.

Higher order modulation formats such as PAM8 or even PAM16 are usually very limited due to its high error floor and so far never have been practical choices in optical standards. Digital multi-tone (DMT) in principle could offer a more efficient use of the spectrum, however it comes at the cost of higher DSP power at the transmitter and receiver due to the need for frequency domain processing and is generally compromised for short reach interconnections due to its high peak-to-average power ratio (PAPR), which severely limits the link budgets for passive transmission. In standardization discussions on 200G SerDes, PAM4 and PAM6 are being discussed as the most likely modulation choices. Historically, the host side modulation choice was also used on the line side, leveraging the host FEC.

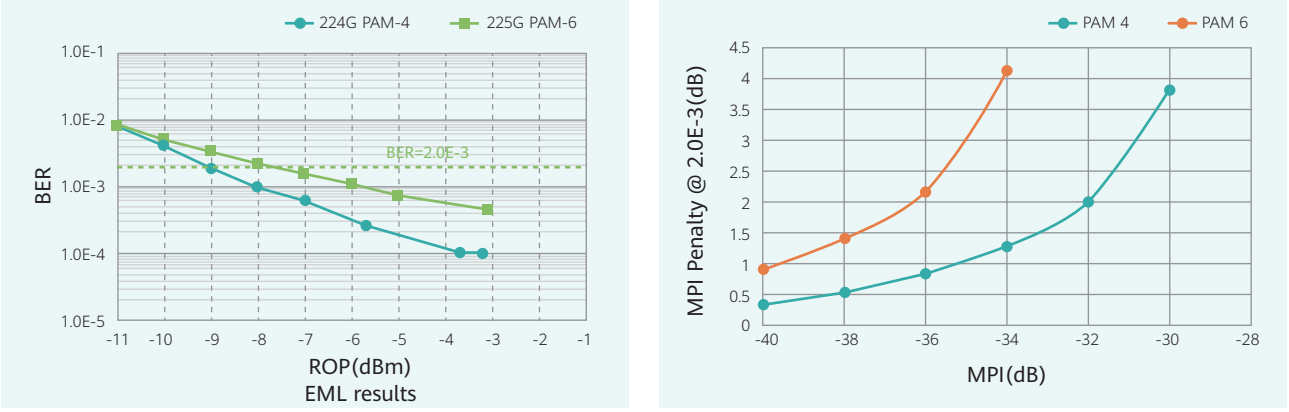


Figure 7 – PAM4 vs. PAM6 performance benchmarking (left). MPI penalty vs. Modulation Format (right)

As shown in Figure 7, the MSA benchmarked PAM4 and PAM6 on advanced IMDD prototypes. It was demonstrated that PAM4 can achieve better overall link budget and lower error floor, which relaxed the burden of FEC dimensioning and can allow for a low power and low latency solution. Moreover, the multipath interference (MPI) tolerance of PAM4 is much higher than PAM6, thus enabling the crucial double and triple links for the FR reach class. For a triple link configuration, common for FR interfaces, an MPI

penalty of -35dB worst case has to be assumed. In this case, the MPI penalty of PAM6 reaches 3dB. The extra 1 dB link budget significantly limits the application scenario of PAM6. Our findings are consistent with the discussions in OIF on 800LR/ZR, where similarly to the 800G Pluggable MSA, a four level signaling format 16QAM (based on four level electrical modulation equivalent to PAM4), will likely be used for optical transmission ranges of 2km and above. Thus, PAM4 is the optimal choice in optical domain, given the performance of optoelectronics devices, ASICs and allocation penalties. This overall alignment between the 2 SDOs might also impact the specification of 200G SerDes in order to achieve an end-to-end consistency of 200G signaling.

4.3 800G FR4 specifications

The use of IMDD is always more cost advantageous than coherent optics, if the transmission is not limited by fundamental transmission effects CD and PMD. CWDM4 4x200G transmission is projected to be the most cost effective technology for the 2km application space. Figure 8 (left) shows measured S21 curves of 112Gbaud transmitter and receiver component prototypes. The high bandwidth PD and EML are provided by Sumitomo. As shown in the figure, the EML bandwidth is well above 50GHz, with RIN< -150 dB/Hz. Thanks to the waveguide design, the PD shows a bandwidth larger than 60GHz with chip responsivity larger than 0.7 A/W. For the RFICs, which are under development, the preliminary results demonstrate a bandwidth larger than 50GHz, with THD< 3%. Optoelectronic devices for 200G/lane have been demonstrated, and will be further improved for even better performance in the coming future.

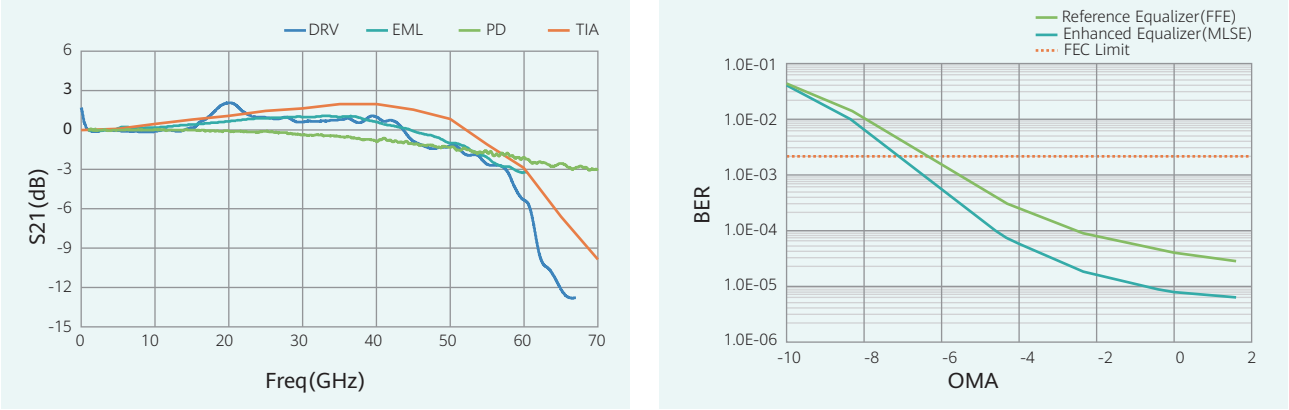


Figure 8 – Measured S21 of 112Gbaud components (left), Verification of 224Gb/s PAM4 with FFE and MLSE (right)

Figure 8 (right) shows an evaluation of 224Gb/s PAM4 transmission performance using a reference linear feed FFE with 21 taps and an enhanced, non-standard, variant including an additional maximum likelihood sequence estimation (MLSE). Based on the numerical and experimental verification of 224Gb/s per lane PAM4, we propose certain baseline parameters for a potential interoperability specification. Compared to 400G FR4, a larger dispersion penalty is anticipated at higher baud rates of 800G FR4, with the limiting performance at 1330nm of the CWDM4 grid. The Transmitter and Dispersion Eye Closure Quaternary (TDECQ) for PAM4 is proposed to be increased from 3.4dB to 3.9dB, as shown in Table 3.

Table 3 – Major interoperability specifications for 224Gb/s PAM4

	Description	Unit	Spec
Tx	Outer OMA / lane (min)	dBm	0.2
	TDECQ (max)	dB	3.9
	Launch power outer OMA - TDECQ	dBm	-1.2
	Extinction ratio (min)	dB	3.5
	RIN-OMA (max)	dB/Hz	-139
Rx	Sensitivity outer OMA	dBm	Fig. 8
	Stressed Rx sens. outer OMA / lane (max)	dBm	-2.1
	Stressed eye closure (SECQ)	dB	3.9

Meanwhile, FFE with increased tap numbers in TDECQ is verified to be necessary by both experimental and simulation results. The system is also more limited by relative intensity noise (RIN) requiring a more stringent specification at -139dB/Hz. Here, a 0.5 dB increase of Tx optical power is proposed to compensate the DGD penalty in 224G/lane scenario, which has come to agreement within MSA. On the other hand, considering the FEC enhancement, the receiver sensitivity is kept the same as that in the 400G FR4 case. The corresponding receiver mask is shown in Fig. 9. The full specification is available on the 800G Pluggable MSA website.

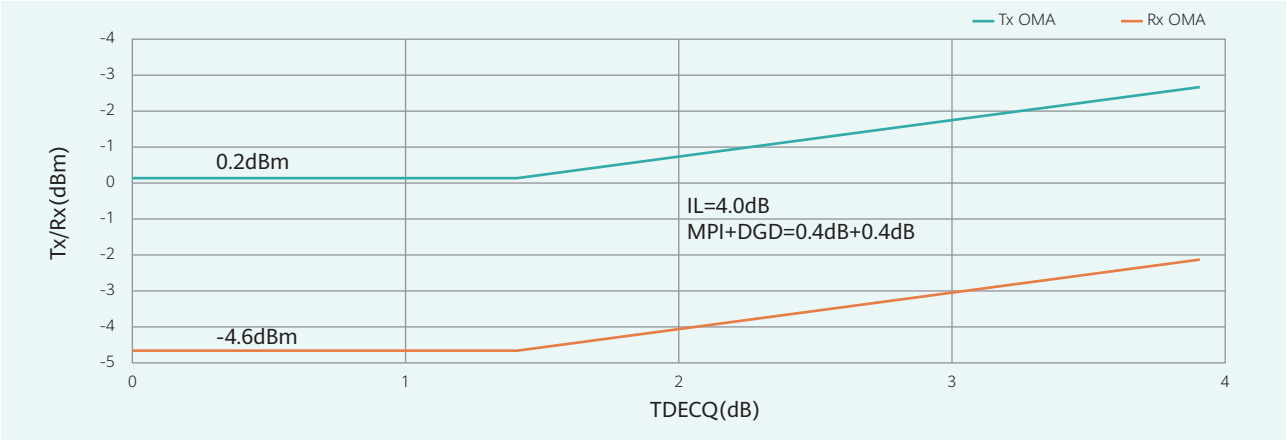


Figure 9 – Receiver sensitivity mask 800G FR4

5. Concatenated 800G FEC

5.1 FECs in optical standards

In the field of optical communications, the spatially-coupled technique is a widely used scheme to construct strong FEC codes, like staircase codes, which was invented by the research team of Professor Frank R. Kschischang from University of Toronto in 2011, as shown on Figure 10. Staircase codes, a class of spatially-coupled product-like FEC codes, are built from Bose-Chaudhuri-Hocquengham (BCH) component codes, can provide binary-symmetric channel (BSC) capacity-approaching performance at high code rates. Specifically, an ITU-T G.709 compatible 6.69% overhead staircase code shows a net coding gain of 9.41 dB at a bit error rate of 1.0E-15, or 0.56 dB from the BSC capacity, and has been adopted by many standards such as IEEE 802.3, G. 709.2, and 400G ZR.

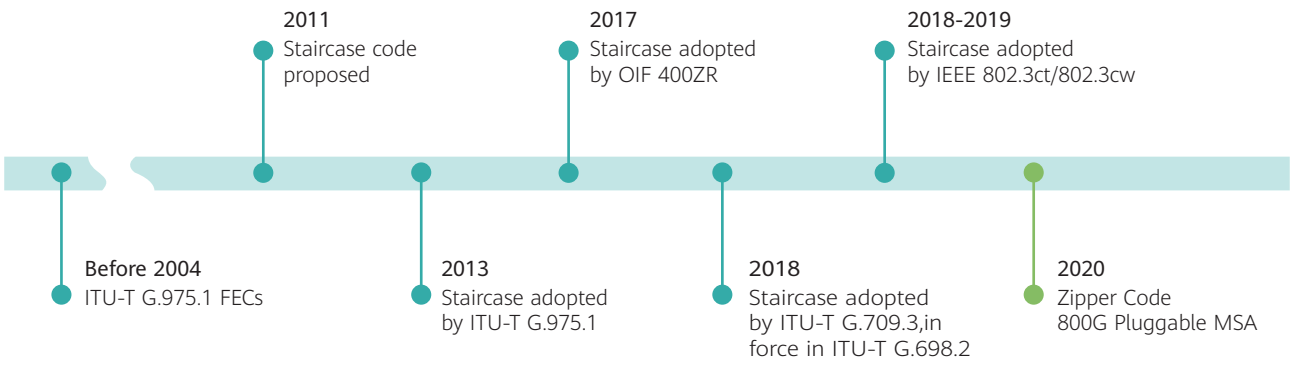


Figure 10 – Forward error correction codes in standards

The zipper code (Z-FEC), which was proposed by the same group recently [1], is an upgrade and generalization of the staircase code. Zipper codes provide a more flexible design framework with three major elements: zipping pair, interleaver mapping and component code. The zipping pair is formed by a virtual buffer and a real buffer. The real and virtual buffers are filled by the data of the current frame and mapped from previous frames, respectively. Parity bits are generated from frames of both halves. The interleaver mapping describes a scheme of how the data in the previous frames are mapped to the virtual buffer. The component code is usually a simple algebraic code, such as BCH code. Each component code word is split into two equal portions, in which one half is drawn from the real buffer and another half is drawn from the virtual buffer. In the zipper code, the buffer of bits actually transmitted (real buffer) is "zipped together" with a buffer of interleaved copies of previously transmitted bits (virtual buffer), constrained to form code words of some constituent code.

5.2 Zipper code for 800G FR4

The zipper code can be used as a powerful code in concatenation with low-complexity hard- or soft-decision codes, providing low-complexity encoding in latency-sensitive applications. A hard-decision zipper code is more favored for 800G-FR4, with high burst error tolerating capability and a compatibility with a low-power hard-output MLSE. Using an iterative decoding algorithm, hard-decision zipper code can meet the performance requirement (2.0E-3) of 800G-FR4 with low power consumption and latency.

An alternative solution are stand-alone soft-decision algebraic codes. For a possible soft-decision solution, a more power-consuming soft-output equalizer, rather than MLSE, is needed, when the performance of FFE-based receiver is not adequate. This soft information contains burst impairments and colored noise, which will seriously degrade the error correcting capability of simple soft-decision algebraic codes. Thus, large interleavers are required to decorrelate the burst errors, which also introduce additional latency. Therefore, soft FEC is not preferred for future data center standards.

There are two options to process the legacy host KP4 FEC - concatenation or termination. In the concatenated scheme, the Reed Solomon (RS) code in the host acts as the outer code, which combines with the zipper code in optical module to form an overall concatenated code as shown in Figure 11 (left). The zipper code matches the bit error rate of optical link and the decoding threshold of the RS code. The concatenated scheme has latency and power-consumption advantages by avoiding decoding and re-encoding processing of the RS code in optical module. Figure 11 (right) shows the concatenated scheme performance in BSC, which reveals sufficient margin for 2.0E-3 threshold [2]. On the other hand, in the terminated scheme a decoding process of the RS code is necessary to remove the parity bits and to recover the original data stream. This decoding will introduce extra latency and power consumption of at least 50~100 ns and ~200 mW in optical module, without even considering the cost of alignment lock, deskew, lane reorder, etc.

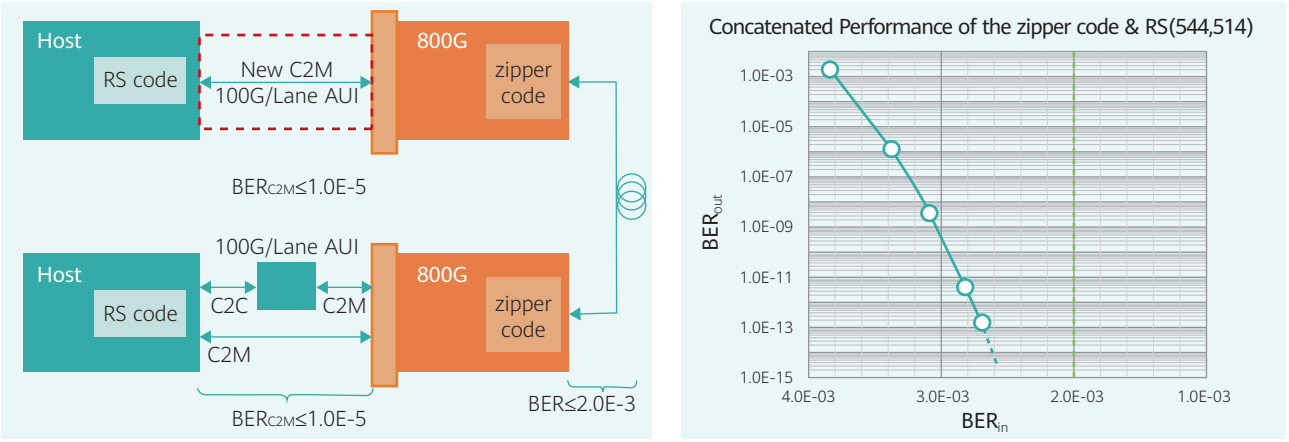


Figure 11 – Concatenated FEC design with the Zipper FEC (left). Performance of the concatenated code (right)

The targeted reach for 800G FR is from 500m to 2km. Correspondingly, the propagation delay varies approximately from 2.5µs to 10µs. A rule of thumb is that in latency-critical applications, the latency of the whole transceiver should not exceed 10% of the propagation delay thus resulting in ~200ns total latency for FEC. The concatenated scheme is therefore a more favored solution. Here, the optical module can directly apply the inner zipper encoder to the received synchronous data streams from C2M lanes, since the C2M lanes from the host will share one clock as similar with the practice of 400G-FR4 (with 8x56G C2M lanes and 4x100G optics).

6. Outlook

800Gb/s datacom connectivity will be driven by the adoption of 100G C2M interface and the transition to higher capacity switches with 51.2Tb/s. The 800G Pluggable MSA has evaluated and proposed several concepts for 200Gb/s per lane optical signaling to enable 800Gb/s FR4 pluggable solutions based on 100G SerDes. Based on the 100G C2M interface with legacy KP4 FEC, the proposal includes an concatenated inner zipper code in optical module to improve overall performance with very low latency. The suggested optical specifications were verified numerically and experimentally and have shown to support EML based transmitters. With IEEE evaluating future standards, 200Gb/s optical PHY specifications are likely to become the foundation of future 800GbE, 1.6TbE and related single lane 200GbE standards. Table 4 shows a possible future evolution of 800G FR. After the first design variant based on fundamental 100G SerDes aligned with the 51.2T switch node, the steps beyond could see a transition towards native 200G/lane signaling and potentially just a retimer for architectures with a shorter distance from host to the optical module, for applications such as on-board optics (OBO) or near package optics (NPO).

Table4 – Generations of 800G FR4 optics

	Electrical i/f	Implementation	Attributes
Gen 1	8x100G-PAM4	Gearbox	Time to market for 51.2T
Gen 2	4x200G-PAM4	DSP	Aligned with 102.4T Switch
Gen 3	4x200G-PAM4	Re-timer	Lower Power

[1] A. Y. Sukmadji, U. Martínez-Peñas and F. R. Kschischang, "Zipper codes: Spatially-coupled product-like codes with iterative algebraic decoding", CWIT 2019.
[2] Yu Tian et al., "800Gb/s-FR4 specification and interoperability analysis", OFC 2021

800G Pluggable

MULTI-SOURCE AGREEMENT

About Us

The 800G Pluggable MSA group was formed in September 5, 2019 and promotes a joint industry exchange and collaboration between data center operators and vendors of infrastructure equipment, optical modules, optoelectronic chips, and connectors.

It focuses on the data center network interconnection scenario, targeting to determine the optimal interconnect architecture, define interface specifications of the 800G pluggable optical modules, build the ecosystem, and guide healthy development of the industry.

- Chairman: Wang Chen, CTTL
- Secretary: Zhang Hua, Hisense
- Spokesperson: Maxim Kuschnerov, Huawei

Please visit our website for more info:
www.800Gmsa.com