800GPluggable

MULTI-SOURCE AGREEMENT

ENABLING THE NEXT GENERATION OF CLOUD & AI USING 800GB/S OPTICAL MODULES

PROMOTERS:

Contents

1. Cloud Expansion Sets Pace for Optical Modules

Cloud computing and storage have taken over as the technological backbone to a majority of our modern business applications providing infrastructure, platform, software or virtually anything as a service, and to personal appliances covering phones, laptops and various smart devices. Unlike wireless infrastructure and standards like LTE and 5G, where the standardization and technology are ahead of the actual applications, providing a "build it and they will come" business model, the rapid and all-encompassing expansion of cloud applications and services vigorously pushes the development of high-tech electronics and optics, which often seem to run behind the pace set by the end users. The exponential resource growth of artificial intelligence applications and the inherent need for high bandwidth for the transport of big data puts a further strain on data center architectures and the underlying interconnects. Thus, the deployments of the AI cloud, are gaining momentum.

Cloud applications, AR/VR, AI, and 5G application generate more and more traffic. The explosive growth of traffic requires higher bandwidth. As shown in Figure 1, global interconnection bandwidth capacity will grow at a 48% CAGR in 2017 - 2021.

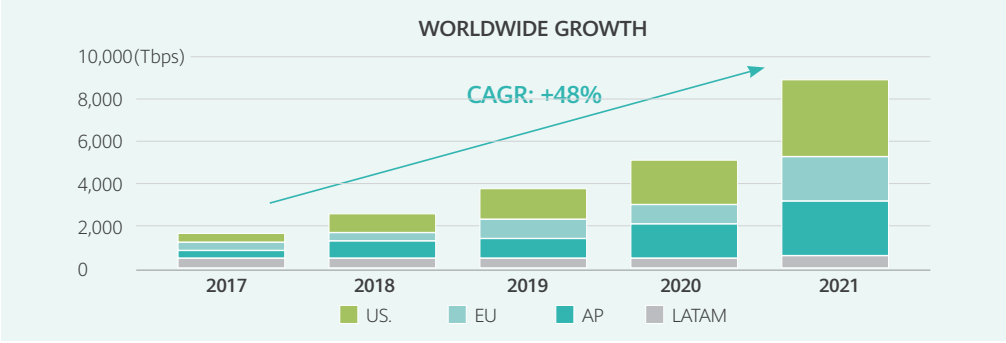

Figure 1 –Global Interconnection Index (Source: Equinix)

As shown in Figure 2, market analysts are projecting a first adoption of 400G Datacom modules in 2020 with a larger adoption of 2x400G/800G modules in 2022-23.

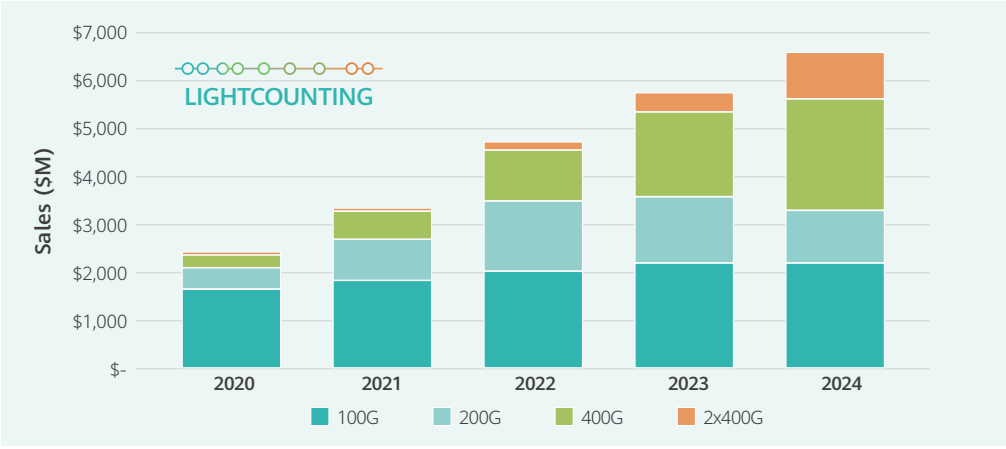

Figure 2 – Projection of the market revenue for datacom modules (Source: Light Counting)

"Our LightCounting Forecast model indicates that operators of Cloud datacenters will need to deploy 800G optics by 2023-2024 to keep up with the growth of data traffic," stated founder and CEO of LightCounting Market Research, Vladimir Kozlov, PhD. "Most of 800G will be still pluggable transceivers, but we expect to see some implementation of co-packaged optics as well."

Data center cloud architectures are being paced by the capacity scaling of switching ASICs, which is doubling approximately every two years, unfazed by the talk about the end of Moore's Law. Today, 12.8Tb/s Ethernet switching chips are being commercially deployed with first chip design firms already prototyping 25.6Tb/s silicon for deployment next year, as shown in Figure 3. This puts further pressure onto the densification of optical interconnects, which do not scale at the speed of CMOS due to the lack of a common design methodology across the various components and a common large scale process.

In the past few years, the rapid expansion of cloud services was fueled by the rapid adoption and price erosion of 100G short reach optical modules based on direct detection technology and non-return to zero (NRZ) modules. After the beginning of the 400GbE Bandwidth Assessment activity in IEEE in March 2011, initial deployment of 400G optical modules is finally starting in 2020 with a stronger ramp projected for 2021, as shown in Figure 2. In fact, in the initial use cases, 400G modules will be mainly used to transport 4x100G over 500m in DR4 application and 2x200G FR4 optics over 2km, not making use of the 400GbE MAC. At the same time, it seems unlikely that IEEE would soon standardize the next generation of optics, such as 800GbE, meaning that the standardization of higher density optics for the transport of 8x100GbE or 2x400GbE for the 25.6Tb/s and 51.2Tb/s switching generations would be well behind actual market timeline requirements of 2021-22. This raises the need for 800G industry interoperability outside of the established standard bodies.

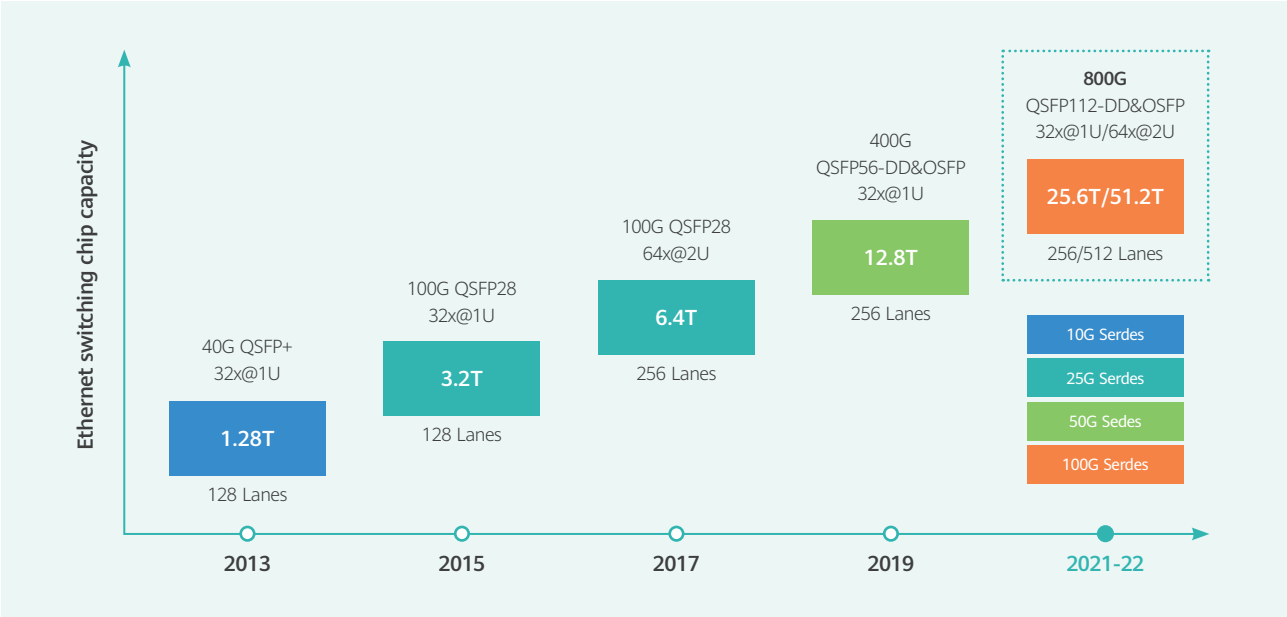

Figure 3 – Historical evolution of data center switching chip capacity

2. Data Center Architectures

The hyper scale data center market is quite fragmented with respect to the used data center architectures or the demand for pluggable optics. Data centers for operators with a larger external customer base offering XaaS are more likely dominated by north-south client-to-server traffic and could have more smaller geographical clusters. On the other hand, operators with a large internal demand for cloud computing and storage see more east-west traffic between servers and could operate their data centers as huge clusters with a higher radix. Even in case of similar use cases, the operators can deploy individual flavors of network architecture or have a subjective preference to a certain interconnects solution such as PSM4 or CWDM4 or other cost-down variants of thereof, such as 100G CWDM4-OCP.

One can derive at least two main types of typical data centers architectures. Figure 4 shows the common abstraction of a hyper-scale data center and its optical interconnect roadmap. In general, these architectures are larger, have a certain convergence from layer to layer, e.g. 3:1, and rely on coherent ZR interconnects at the Spine layer. An important boundary constraint for 800G networking in this case is that 200G interconnects, albeit not serial, are used at the server to TOR layer, whereas the TOR-leaf/spine layer would typically rely on PSM4 4x200G in a fan-out configuration.

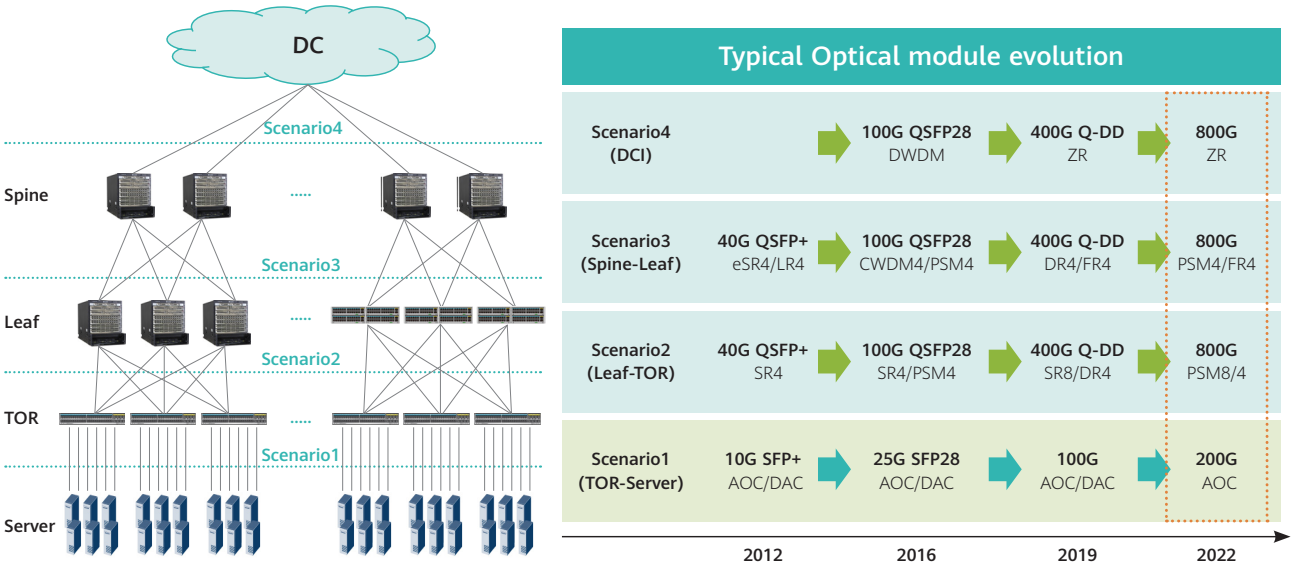

Figure 4 – Typical hyper scale data center interconnect roadmap

For the typical hyper scale data center network (DCN), deploying 200G servers will require an 800G fabric. It's a traffic convergence network, which depends on the balance between service requirements and Capex optimization. Table 1 shows the detailed reach requirements depending on the DCN layer.

Table 1 – Detailed requirements of the typical hyper scale DCN

Scenario	Server to TOR	TOR to Leaf	Leaf to Spine	DCI
Bandwidth	200G	800G	800G	800G
Distance	4m within rack; 20m cross-rack	≥70m 100m is preferred	500m/2km	80km-120km
Module size	QSFP-DD/OSFP	QSFP-DD/OSFP	QSFP-DD/OSFP	QSFP-DD/OSFP

Figure 5 shows the data center network architecture of an AI cluster, with less layers than the hyper scale network due to the fact that it lacks any convergence between the layers. The design of an AI cloud implies different traffic flows with much larger big data flows and less frequent switching.

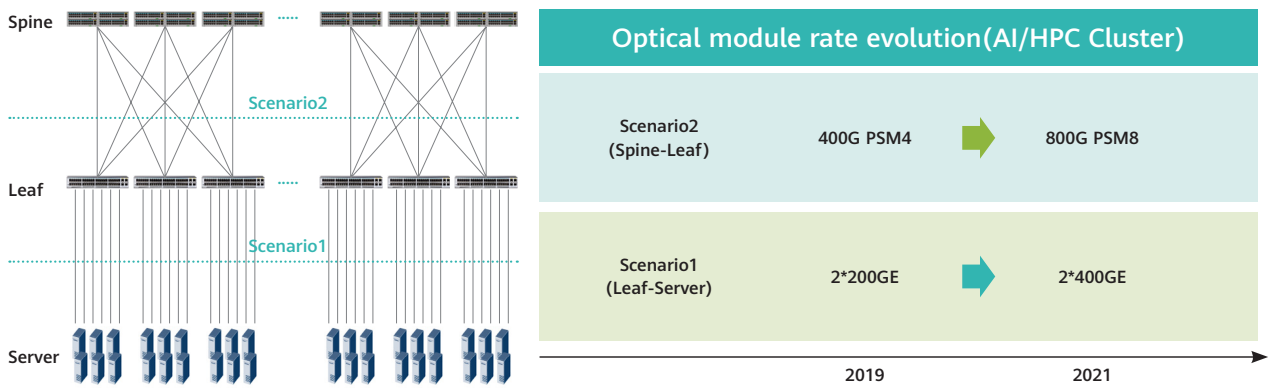

Figure 5 – AI/HPC optical interconnect roadmap

For the AI/HPC cluster DCN, deploying 400G servers will require an 800G fabric. This DCN doesn't have any traffic convergence, with faster deployment than in the case of Figure 4. Table 2 shows the detailed requirements.

Table 2 – Detailed requirements of the AI/HPC cluster DCN

Scenario	Server to Leaf	Leaf to Spine
Bandwidth	400G	800G
Distance	4m within rack; 20m cross-rack	500m
Module size	QSFP-DD/OSFP	QSFP-DD/OSFP
Latency	92ns (IEEE PMA layer)	92ns (IEEE PMA layer)

Not explicitly shown, but also relevant, are DC networks for smaller cloud or enterprises, where the downstream to the server is decoupled from the fan-out rates of the Leaf-Spine layer and typically has slower server interconnect speeds.

3. 8x100G Solution for SR Scenario

3.1 800G SR scenario requirement analysis

For the class of 100m, the industry is facing the basic limitations of VCSEL signaling at speeds of 100G/lane. Here, multi-mode technology will likely allow for reaches of 30-50m, thus only partially covering the SR class, which is primarily employed by Chinese hyper scale data center operators. The MSA targets the development of a low-cost 8x100G module for SR applications, covering the sweet spot of 60-100m, as shown in Figure 6. Particularly, the MSA is intended to specify a lower cost transmitter technology with the potential to leverage sub-linear cost scaling with a high degree of integration. Such a module would allow for an early time-to-market dense 800G solution. A low cost 800G SR8 could also support the potential trends of an increasing switch radix and decreasing server count-per-rack, which may combine to favor middle-of-the-rack (MoR) and end-of-the-rack (EoR) or top-of-the-rack (ToR) architectures, by providing a low cost serial 100G server interconnect. As shown in Figure 6, the MSA will define a lower cost PMD for single mode fiber interconnects based on 100G PAM4. Due to the low latency requirements of SR applications, KP4 forward error correction (FEC) will be used end-to-end with a simple clock recovery and data equalization unit in the module. Finally, the MSA will specify a connector for the PSM8 modules which enables a fan-out to 8x100G.

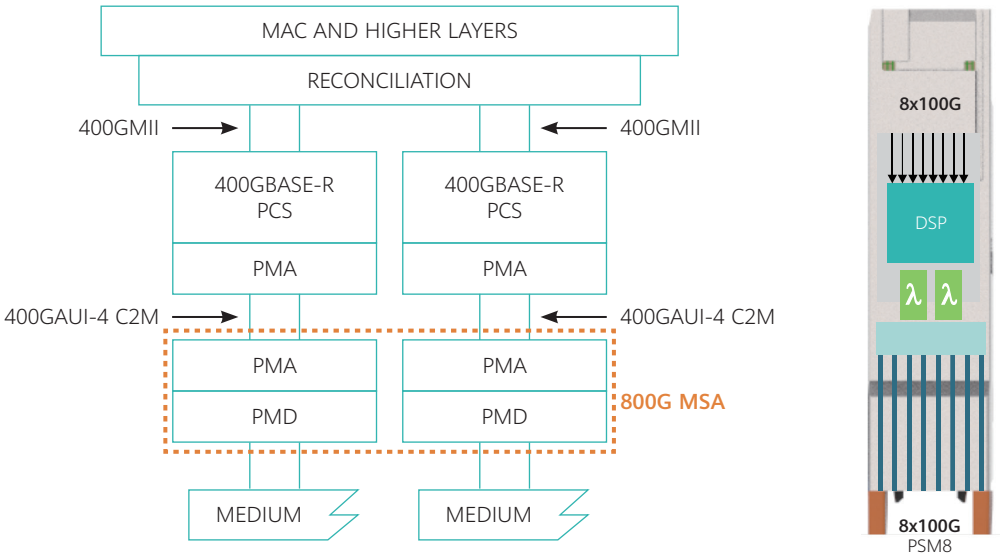

Figure 6– 800G SR8 block diagrams

3.2 Technical Feasibility of 8x100G solutions

As mentioned above, signaling rate up to 100G per lane may limit the evolvement of multi-mode fiber (MMF) based solution from 400G-SR8 to 800G-SR8. Based on the theoretical model used in IEEE, we can reckon that the transmission distance that MMF can support is no more than 50m as the baud rate up to 50G (See Table 3). The limitation factors are from the limited bandwidth of VCSEL and the modal dispersion of MMF. With the optimization in devices, fiber medium as well as enhanced DSP algorithms, 100m MMF transmission may be realized at the cost of higher expense, higher latency, and larger power consumption. Hence, in 800G Pluggable MSA, we recommend that the 800G-SR8 scenario is taken over by SMF based solution.

Table 3 – Fiber channel bandwidth and transmission distance of MMF reckoned by the theoretical model used in IEEE

Bit rate	Signal Type	Fiber Type	Fiber channel bandwidth (GHz•km)	Transmission Distance (m)	IEEE standards
50Gbps	PAM4	OM4	2.301	100m	50G-SR, 100G-SR2 200G-SR4, 400G-SR8
50Gbps	PAM4	OM3	1.541	70m	
100Gbps	PAM4	OM4/OM5	2.301/2.377	50m	Defined now
100Gbps	PAM4	OM3	1.541	35m	-

In order to guarantee the advantages on the cost and power consumption of the SMF based solution, reasonable PMD standard requirements are indispensable in 800G-SR8. The PMD requirements to be defined should ensure that 1) diverse transmitter techniques, such as DML, EML, and silicon photonics (SiPh), can be applied in such scenario; 2) all the potential of the components can be released adequately to achieve the targeting link performance; 3) key parameters in PMD layers should be relaxed as much as possible, in the context of maintaining a reliable link performance. According to these three principles, we will conduct some brief investigations and discussions as follows.

The power budget of the SMF based 800G-SR8 solution would be quite similar with that defined in IEEE 400G-SR8. The only issue to be determined is the insertion loss of new defined PSM8 SMF connectors. It means that the power budget in SR scenario can be achieved without a hitch based on currently mature optical and electronic components and DSP ASICs used in 400GE optical interconnection. Therefore, apart from specifying the connector for the PSM8 modules, the key issue for the definition of PMD parameters in 800 SR8 scenario is to find out the suitable optical modulation amplitude (OMA), extinction ratio (ER), transmitter dispersion eye closure quaternary (TDECQ) of the transmitter and sensitivity of receiver. In order to set these parameters into the suitable position, the bit error ration (BER) performance of the diverse transmitters is investigated and assessed.

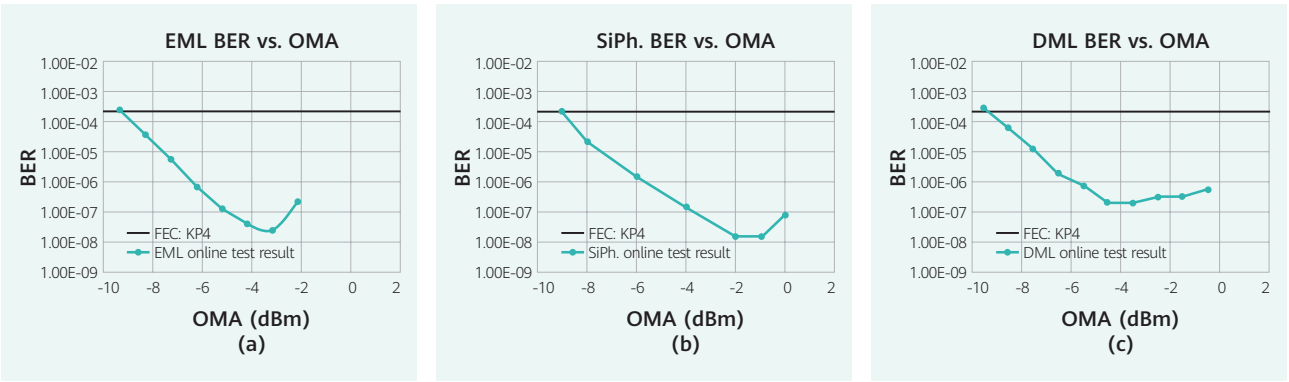

Figure 7 – (a) EML BER vs. OMA results based on commercial available 400G DSP ASICs; (b) Silicon Photonics BER vs. OMA results based on commercial available 400G DSP ASICs, (c) DML BER vs. OMA results based on commercial available 400G DSP ASICs

Figure 7 shows three BER vs OMA curves of 100Gbps PAM4 signal, which correspond to different transmitter technologies respectively, as online results and obtained using commercial 400G DSP ASICs. Actually, the BER performances of EML and SiPh for 100G per lane illustrated in Figure 7 (a) and (b) are well-known results since these two solutions have been extensively discussed in the past few years. Considering relatively low launching optical power of SiPh transmitter and good enough sensitivity of all three solutions, the minimum OMA requirement in 800G-SR8 is recommended to be relaxed appropriately.

The BER performance of the DML in Figure 7 (c) shows that the OMA sensitivity in this case is comparable with that in the case of EML or SiPh, even though the commercial DML used in here has relatively lower bandwidth than EML and SiPh. This result implies that the commercial DSP ASICs used in practice have much stronger equalization ability than the reference receiver IEEE defined in 400GE, and thus it can support the transmitter with comparatively low bandwidth to achieve the targeting power budget 800G-SR required. In order to release the potential of the DSP unit adequately for 800G SR8 PMD, reference receiver for compliance test (i.e. TDECQ) requires to be re-defined to match the practical equalization ability of commercial DSPs, i.e. more taps numbers than currently defined 5 taps are desired. Meanwhile, considering the relatively low sensitivity requirement in SR scenario and restriction of the power consumption of the 800G module, a low complexity DSP mode is recommended in future modules.Another key parameter is ER that is related to the power consumption directly. A lower ER is favored as long as it does not impact the reliability of the link. Based on the above analysis, we believe that a low cost and power consumption SMF-based solution is feasible and promising in 800G-SR8 scenario.

4. 4x200G Solution for FR Scenario

4.1 800G FR scenario requirement analysis

200G per lane PAM4 technology is the next major technological step for optical intensity modulated, direct detection interconnects and will be the foundation for a 4-lane 800G connectivity, as well as an essential building block for future 1.6Tb/s interconnects. As shown in Figure 8, the MSA will define the full PMD and partial PMA layers including a new low power, low latency FEC as a wrapper on top of the KP4 FEC of the 112G electrical input signals, in order to improve the net coding gain (NCG) of the modem. One of the key goals of this industry alliance will be the development of new wide bandwidth electrical and optical analog components for the transmitter and receiver assemblies including digital-to-analog and analog-to-digital (AD/DA) converters. In order to achieve the aggressive power envelop targets of pluggable modules, the DSP chips will be designed in CMOS process with lower nm node and employ low power signal processing algorithms to achieve equalization of the channel.

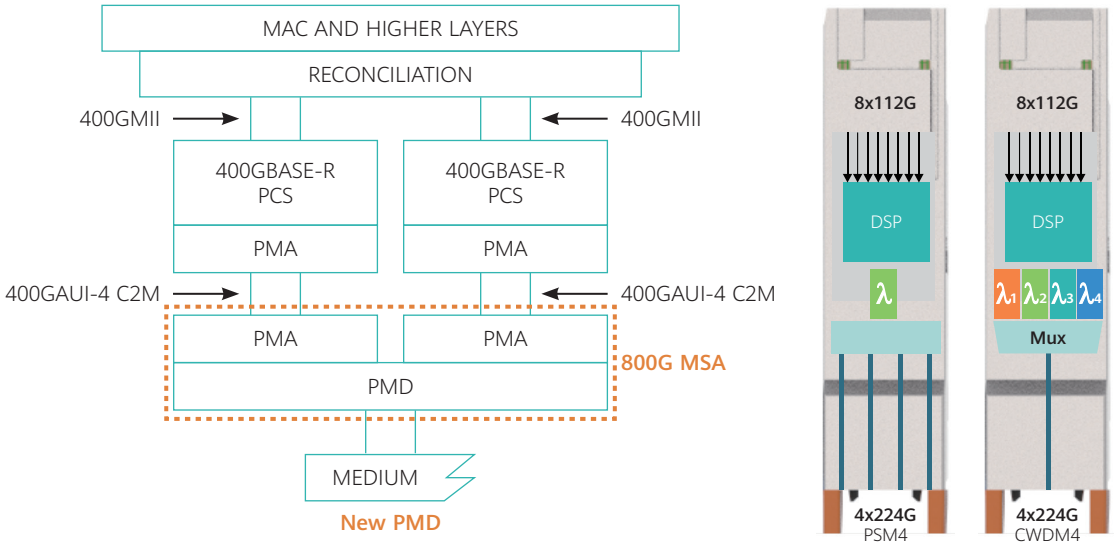

Figure 8– 800G FR4/PSM4 block diagrams

4.2 Technical feasibility of 4x200G solutions

Considering that a temperature controller (TEC) is required in LAN-WDM, which is not desired in 200G/lane scenarios, the power budget will be analyzed based on CWDM4. Link insertion loss, multipath interference (MPI), differential group delay (DGD), and transmitter dispersion penalty (TDP) are the contributions to the link power budget. According to the model released in IEEE standards, MPI and DGD penalty is calculated as listed in Table 4. In view of the increased baud rate of 200G per lane, the dispersion penalty is expected to be larger than that in 100G per lane. A reasonable suggestion for transmitter dispersion penalty (TDP) is 3.9 dB. Hence, taking into account the margin for receiver aging and coupling loss, as well as the typical launching optical power value of the transmitter, we think the receiver sensitivity required should be around -5dBm.

Table 4 – Power budget analysis of 800G-FR4

Description	Simulation value
Link insertion loss	4 dB
MPI penalty	0.4 dB
DGD penalty	0.4 dB
TDP	3.9 dB

Since SNR deteriorates about 3 dB compared with 100G/lane as the baud rate doubles, it is expectable that a stronger FEC is necessary to maintain the reasonable receiver sensitivity (~-5dBm) and margin of error floor. Therefore, as mentioned above, on the top of the KP4, an additional low power, low latency FEC as a wrapper will be carried out in the optical module. The threshold value of the new FEC is determined according to the link performance and power budget requirement.

Link performance of 200G/pane is presented using simulation and experiment. The parameters of the devices adopted in the link are listed in Table 5. The experimental result shows that the receiver sensitivity can reach the target value while the new FEC's threshold is set to 2E-3 as depicted in Figure 9 (a). However, in this experiment, maximum likelihood sequence estimation (MLSE) was required to compensate the excessive inter-symbol-interference induced by channel bandwidth limitation. The dash line in Figure 9 (a) shows the simulation based on the model in which the measured parameters of the devices used in the experiment are adopted. Together with experimental results, simulations show that the system is limited by the bandwidth of components, such as AD/DA, driver and E/O modulators. Considering that high bandwidth components are expected to be available in the years to come, simulation results by using the same system model but with expanded bandwidth is illustrated in Figure 9 (b). It shows the receiver sensitivity of @ 2E-3 can meet the above-mentioned requirement with only FFE equalization in the DSP unit, which is in accordance with the theoretical expectation.

Table 5– Parameters comparison between simulation and experiment

Description	Value in Experimental Setup	Enhanced value in Simulation
3 dB Bandwidth of TOSA	42 GHz	56 GHz
3 dB Bandwidth of ROSA	56 GHz*	56 GHz
3 dB Bandwidth of AD/DA	37 GHz	50 GHz
Effective Number of Bits (ENOB)	4.5	4.5
RIN-OMA	~-137 dB/Hz	-137 dB/Hz
Chromatic Dispersion	7 ps/nm	7 ps/nm
TIA Noise	15 pA/sqrt(Hz)	15 pA/sqrt(Hz)
PD Responsivity	0.5 A/W	0.5 A/W

* ROSA in experiment setup is a high-bandwidth instrument.

Based on the above analysis, TDECQ is still suggested to be followed in compliance testing in the 800G-FR4 scenario. However, FFE tap numbers of the reference receiver adopted in TDECQ measurement is anticipated to be increased to a reasonable value and needs to be further discussed. Additionally, it should be noted that if the ability of future device targeting 100Gbaud underperforms our expectation, more complicated algorithms (e.g. MLSE) may be used in FR4 scenarios, which implies that a new compliance metrology must be developed.

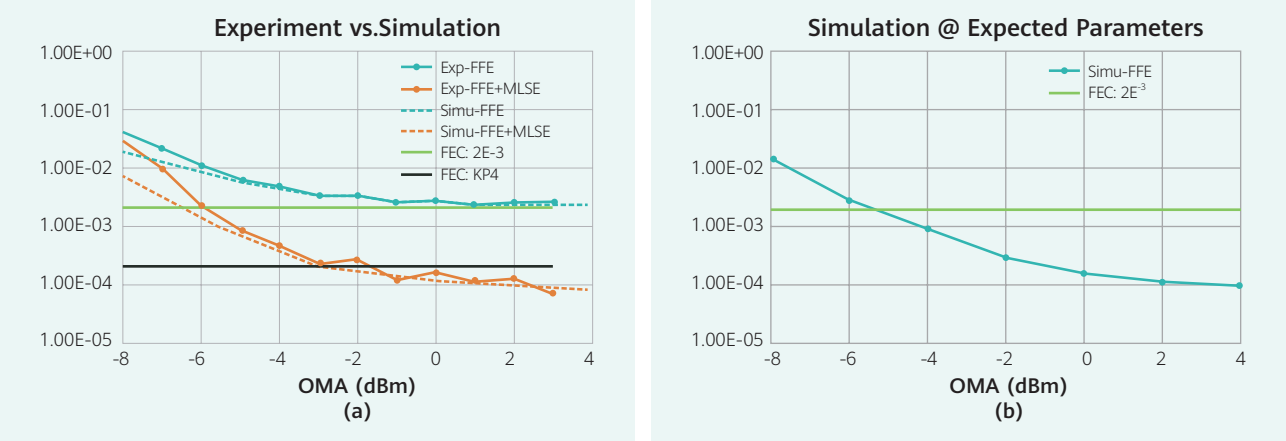

Figure 9 – (a) 200G/lane experiment and simulation results match well with each other; (b) 200G/lane simulation result: FFE equalization can meet the requirement of power budget when component bandwidth in link is improved.

4.3 Packaging for 4x200G solution

For the 4x200G module, the packaging for both the transmitter and receiver needs to be reconsidered to ensure signal integrity within the range under the Nyquist frequency point (56GHz). Two possible solutions for the transmitter are illustrated in Figure 10. Solution A is a traditional approach where the modulator driver (DRV) is close to the modulator. In contrast, in Solution B, DRV in flip-chip design is co-packaged with the DSP unit to optimize the signal integrity on the RF transmission line. Both of these two solutions can be realized by the state-of-art technology. Preliminary simulations show that Solution B can achieve good results and can ensure a bandwidth larger than 56GHz. The ripples on the S21 curve of Solution A are due to the reflection on DRV input and can be optimized through the matching design of the DRV. Eventually, it is expected that the overall performance of Solution A can be further improved.

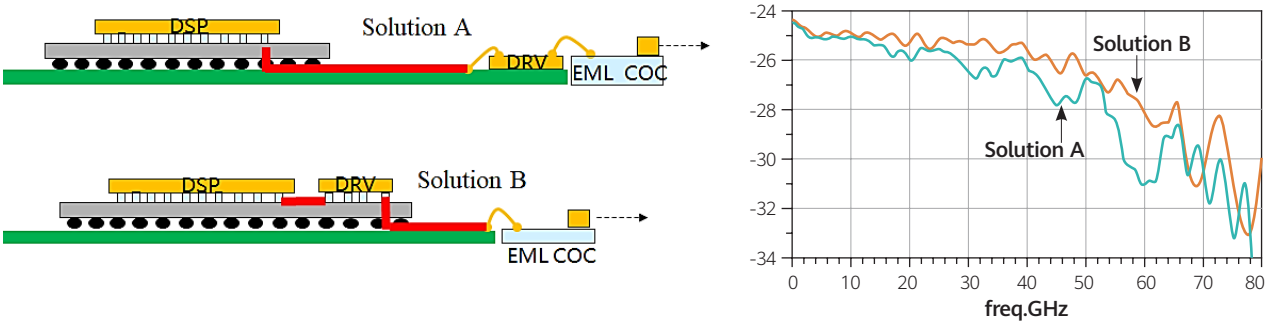

Figure 10 – Two packaging solutions for the transmitter. The S21 simulation puts the RF line (marked in red), the wire-bonding and modulator into consideration, and the bandwidth@-3dB of the EML COC is 60GHz.

At the receive side, the high bandwidth photodiode (PD) with less parasitic capacitance and the high bandwidth trans-impedance amplifier (TIA) are needed to ensure the bandwidth performance of the receiver. There is no obstacle to realizing these components by the state-of-art semiconductor technology. As far as we know, some stakeholders in industry already put much effort in developing these components that are desired to be available in 1~2 years. On the other hand, the connection between PD and TIA is also critical. The parasitic effect in the connection always degrades the performance and thus should be carefully analyzed and optimized.

4.4 Forward error correction (FEC) code for 200G per lane

A stronger FEC with a threshold performance of 2E-3 is required to achieve the sensitivity requirement of 200G PAM receiver. Figure 11 illustrates a comparison between terminated scheme and concatenated scheme. In the first option, KP4 is terminated and replaced with a new FEC with larger overhead. Termination has advantages on NCG and overhead. In the second option, a concatenated scheme keeps KP4 as the outer code and combines it with a new inner code. Concatenation has advantages on latency and power consumption and is more suitable in 800G-FR4 application scenario.

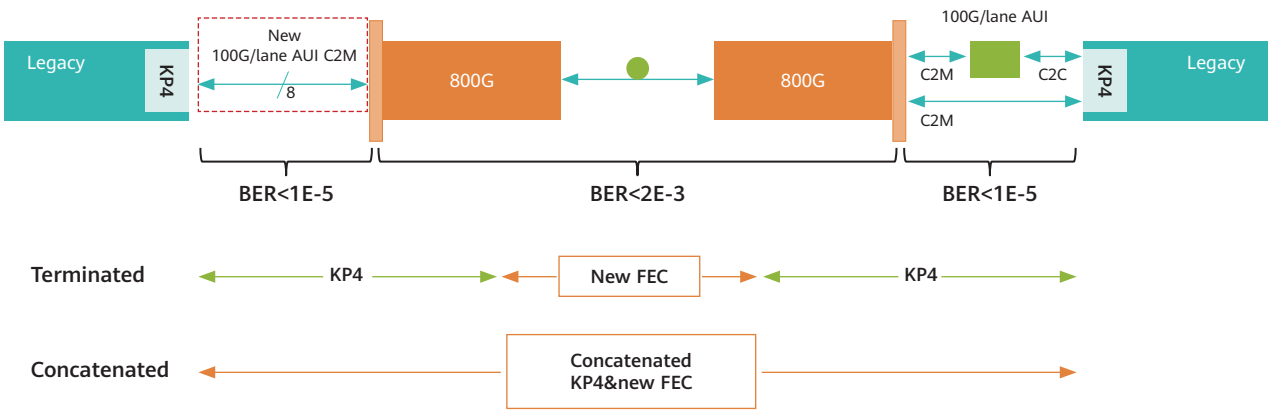

Figure 11 – 800G FEC: Terminated FEC scheme vs Concatenated FEC scheme

Serial concatenation of KP4 and an algebraic code shown in Figure 12 is a straight forward solution to achieve 2E-3 BER threshold performance, as well as to minimize the power consumption and end-to-end latency, since KP4 is not terminated. Noise with bit error rate p_e < 1E-5 introduced in C2M electrical interface is transparent to PMA. The overall performance of the concatenated scheme will not be deteriorated by p_e since p_e is much lower than the decoding threshold of KP4. Hamming codes with single error correcting capability and BCH codes with double error correcting capability are good candidates for the algebraic code in this concatenated scheme. The overhead of these two inner code candidates is ~6%. With a simple soft-in hard-out (SIHO) Chase decoding algorithm of about 64 test patterns, both Hamming and BCH codes can achieve BER threshold better than 2E-3. The symbol distribution defined in 400GBASE-R is inherently an interleaver thus can serve as interleaver (π_e). Interleaver (π_o) with ~10k bit latency is good enough to decorrelate the noise introduced in the optical medium.

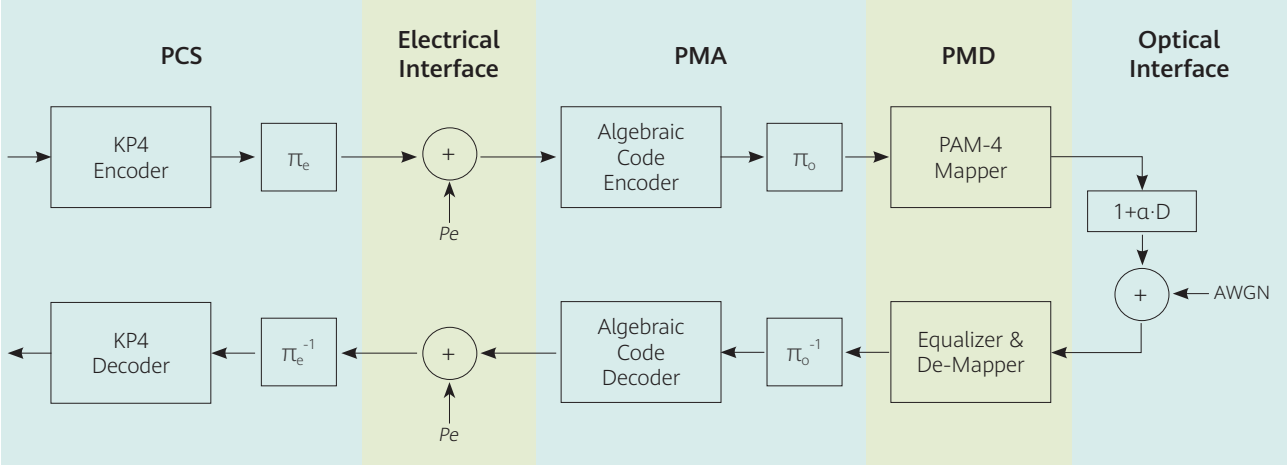

Figure 12– Block diagram of concatenated FEC scheme of KP4 and algebraic code

5. Possible solutions for 800G DR scenario

As shown in Table 6, there are four possible routes in 800 DR scenario. First, 800G SR8 solutions defined in 800G MSA could be defined to extend the reach to 500m. Since parallel fiber solutions require more lanes of fiber, the fiber cost up to 500m is the main concern in this scenario. Second, the 2x400G CWDM4 solution utilizes the available FR4 solution with doubling the pairs of transmitters and receivers. This solution seems to be the balance between fiber resource and technical maturity. However, the power consumption and module cost are its main limitations. Third, there is a possibility that next generation 200G/lane solution may cover this scenario. This solution is believed to be the lowest cost and power consumption with only 4 pairs of transmitter and receivers. As for the available time of this solution, it still requires the feasibility demonstration and industrial maturity considerations. In summary, several solutions are discussed for the DR use case. The MSA will keep track on the technical development, and give suggestions on this application in future.

Table 6 – The possible solutions for 800G DR scenario

Solutions	# of Tx/Rx pair	# of Fibers	# of Wavelength
PSM8 (100G per lane)	8	16	1
2xCWDM4 (100G per lane)	8	4	4
PSM4 (200G per lane)	4	8	1
CWDM4 (200G per lane)	4	2	4

6. Summary and Outlook

In summary, two scenarios, i.e. 800G-SR8 and 800G-FR4, will be defined first in the 800G Pluggable MSA. In the SR8 scenario, to accommodate more technologies into consideration and thus obtain a competitive SMF-based solution, we consider to adjust some key parameters in PMD layer. Therefore, OMA and ER would be relaxed for power consumption, and the reference receiver used in TDECQ measurement would be redefined. We also demonstrated the technical feasibility of 200G/lane optical transmission for 800G-FR4 applications. The experiments and simulations show that a low power, low latency FEC sub-layer should be added into the optical module to achieve the targeting power budget. The details of this new FEC will be presented in the 800G-FR4 standard specifications as so to guarantee the interoperability. Meanwhile, the bandwidth improvement of the components and packaging design optimization are the other two issues that yet require thorough investigations.

The 800G Pluggable MSA targets to release first specifications in Q4/2020, with several subcomponents targeted in the MSA already being prototyped and first 800G modules expected to sample in 2021. With the 400GbE generation ready to be rolled out in the market, 800G pluggable modules will leverage this new eco-system and offer higher density and cost-optimized 100G/lane and 200G/lane interconnects for the next generation of 25.6T and 51.2T switches.

Looking beyond 800G towards 1.6T, the industry begins to see the possible limitations of pluggable modules. SerDes for C2M interconnects is unlikely to scale to 200G/lane using classical PCBs, which might require bringing analog electronics and optics closer to the switching ASIC. But whether the path is leading to co-packaged optics, on-board optics or an evolution of pluggables, we believe that 200G/lane interconnects defined in this MSA, will be an essential building block of the 800G and 1.6T interconnect generations.

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