Next Generation CEI-224G Framework

OIF-FD-CEI-224G-01.0
February 7, 2022

Abstract:

As the OIF looks forward to the higher data rates and/or higher throughput that will be required for the next generation of systems based on 224 Gbps per lane, a consensus has been reached that new specifications and technologies will be required. This framework document represents the efforts of the OIF to identify the hardware interconnection application spaces where the communications and computer industries might benefit from interconnection definitions or "Implementation Agreements" (IA). The objective of this white paper is to identify key technical challenges for next generation systems, define electrical interconnection applications and discuss some of the interoperability test challenges so that the OIF and other industry standards bodies will have a common language, as well as understanding of the development projects that are required for the next generation data rate systems.

Next Generation CEI-224G Framework Contributors:

Document Editor:

Cathy Liu	Broadcom Inc.

Document Co-Editors:

Nathan Tracy	TE Connectivity
Ed Frlan	Semtech
Mike Li	Intel
Gary Nicholl	Cisco
John Calvin	Keysight
Richard Mellitz	Samtec

Contributors:

Karl Bois	TE Connectivity

The OIF is an international nonprofit organization with over 100 member companies, including the world's leading carriers and vendors. Being an industry group uniting representatives of the data and optical worlds, OIF's purpose is to accelerate the deployment of interoperable, cost-effective and robust optical internetworks and their associated technologies. Optical internetworks are data networks composed of routers and data switches interconnected by optical networking elements.

With the goal of promoting worldwide compatibility of optical internetworking products, the OIF actively supports and extends the work of national and international standards bodies. Formal liaisons have been established with CFP-MSA, COBO, EA, ETSI NFV, IEEE 802.3, IETF, INCITS T11, ITU SG-15, MEF, and ONF.

For additional information contact:
OIF
5177 Brandin Ct, Fremont, CA 94538
510-492-4040 Φ info@oiforum.com

www.oiforum.com

OIF

CONTENTS

List of Tables

List of Figures

Glossary†

ADC: An analog-to-digital converter is a system that converts an analog signal into a digital signal.

Application Spaces: Portions of equipment or network architecture that could benefit from having a defined set of interconnection parameters.

ASIC: An application-specific integrated circuit is an integrated circuit (IC) customized for a particular use, rather than intended for general-purpose use.

Backplane: A group of electrical connections used as a backbone to connect several printed circuit boards together to make up a switch, computing or storage system.

BCH: Bose–Chaudhuri–Hocquenghem forward error correction (FEC) codes form a class of cyclic error-correcting codes that are constructed using finite fields.

BER: Bit Error Ratio is the number of bit errors divided by the total number of transferred bits during a studied time interval.

CDR: Clock and data recovery, a component that re-establishes the timing of a signal that may have degraded due to impairments on a transmission line, the retimed signal is now able to continue further to its destination.

CEI: Common Electrical IO, an OIF Implementation Agreement containing clauses defining electrical interface specifications.

CPO: Co-packaged optics.

CTLE: Continuous time linear equalizer.

DER: Detector error ratio.

DFE: Decision feedback equalizer. An equalizer by adding a filtered version of previous symbol estimates to the original filter output.

DSP: Digital signal processing.

Faceplate: A plate, cover, or bezel on the front of a device.

FEC: Forward error correction gives a receiver the ability to correct errors without needing a reverse channel to request retransmission of data.

FFE: Feed forward equalizer.

FPGA: A field-programmable gate array is an integrated circuit designed to be configured by a customer or a designer after manufacturing – hence the term field-programmable.

Gbps: Gigabits per second. The throughput or data rate of a port or piece of equipment. Gbps is $1x10^9$ bits per second.

GBd: The baud rate is the number of electrical transitions per second, also called symbol rate. Giga Baud is $1x10^9$ symbols per second.

IA: Implementation Agreements, what the OIF names their defined interface specifications.

IC: Integrated Circuit

I/O: Input Output, a common name for describing a port or ports on equipment

ISI: Intersymbol interference.

LDPC: A low-density parity-check code is a linear error correcting code, a method of transmitting a message over a noisy transmission channel. An LDPC is constructed using a sparse Tanner graph. LDPC codes are capacity-approaching codes.

LR: Long reach. CEI LR specifies backplane/midplane and copper cable electrical interfaces.

MCM: Multi chip module, a specialized electronic package where multiple integrated circuits (ICs), semiconductor dies or other discrete components are packaged onto a unifying substrate, facilitating their use as a single component (as though a larger IC).

Mid-board optics: an optical transceiver that is mounted on a PCBA away from the PCBA edge, close to a switch ASIC to reduce the amount of PCBA trace loss between an ASIC and the optical transceiver. This is in contrast to the common practice today of locating optical transceivers at the PCBA edge.

Midplane: Some backplanes are constructed with slots for connecting to devices on both sides, and are referred to as midplanes.

MLSE: Maximum likelihood sequence estimation is a mathematical algorithm to extract useful data out of a noisy data stream.

MR: Medium reach. CEI MR specifies chip-to-chip electrical interface.

NG: Next generation.

NRZ (PAM2): Non return to zero, a binary code in which 1s are represented by one significant condition (usually a positive voltage) and 0s are represented by some other significant condition (usually a negative voltage), with no other neutral or rest condition.

NPO: Near-package optics.

OE: Optical engine.

O-to-E and **E-to-O**: Optical to electrical interface and Electrical to optical interface, a component that converts an optical signal to an electrical signal or vice versa.

OFDM: Orthogonal frequency duplex modulation, a method of encoding digital data on multiple sub carrier frequencies

PAM: Pulse amplitude modulation, a form of signal modulation where the message information is encoded in the amplitude of a series of signal pulses. For optical links it refers to intensity modulation.

PAM4: Pulse amplitude modulation-4 is a two-bit modulation that takes two bits at a time and maps the signal amplitude to one of four possible levels.

PCBA: Printed circuit board (PCB) assembly, an assembly of electrical components built on a rigid glass-reinforced epoxy based board.

Repeater: A low-latency electronic device that receives a signal and retransmits it. Repeaters are used to extend transmissions so that the signal can cover longer distance. Besides signal equalization, clock and data recovery (CDR) function could be also added to remove jitter from received signal effectively.

RS: Reed Solomon FEC coding, it is a type of block code. Block codes work on fixed-size blocks (packets) of bits or symbols of predetermined size. It can detect and correct multiple random and burst errors.

SerDes: A Serializer/Deserializer is a pair of functional blocks commonly used in high speed communications to transfer data over a relatively low number of lanes.

SNDR: Signal-to-noise-and-distortion ratio is a measurement of the purity of a signal.

SNR: Signal-to-noise ratio.

Tbps: Terabits per second. The throughput or data rate of a port or piece of equipment. Tbps is $1x10^{12}$ bits per second

TME: Test and measurement equipment.

Twinax copper cable: A type of copper cable similar to coaxial cable, but with two inner conductors instead of one.

VSR: Very short reach. CEI VSR specifies chip-to-module electrical interface.

XSR: Extra short reach. CEI XSR specifies die-to-optical engine (D2OE) and die-to-die (D2D) electrical interface.

† Some definitions include content from www.wikipedia.com

1 Executive Summary

In the past the OIF has supported the communications industry by generating implementation agreements such as CEI-56G and CEI-112G that have been shared openly with other industry standards bodies. These implementation agreements (IAs) have defined the parameters and required performance levels necessary to support the development of cost- and power- effective broad industry ecosystems. As the OIF anticipates the next generation of higher electrical data rates at 224 Gbps per lane, it is becoming apparent that new technological solutions will be required at many levels within future communication systems. The objective of this CEI-224G framework document is to identify and define the hardware application spaces that could possibly benefit from future CEI-224G OIF Implementation Agreements across the multiple levels of hardware. Identifying and defining these application spaces will allow the OIF and others in the industry to have a common language, or understanding, as decisions are made to initiate new development projects.

The technical challenges of next generation data rate systems are discussed, as well as test and interoperability issues that will need to be addressed for the various interconnection applications. Although some technical options are mentioned, it is not the scope of this document to define specific technical solutions for these applications or the priority with which the application spaces should be addressed.

As in the past, it is critical that the industry maintain interoperable interfaces for application spaces to enable cost effective component, subsystem, and system development and deployment. This will ensure interoperable fiber, connectors, electrical interfaces, etc. Identification of the critical application interconnections is the first step to meeting this requirement. The goal of this document is to build consensus across the industry on the applications spaces and motivate the initiation of collaborative discussions that are required to generate a broadly agreed set of project developments and objectives.

2 Introduction

2.1 Purpose

The OIF Next Generation Electrical Interconnect Framework identifies application spaces for next generation systems and identifies areas for future work by the OIF and other standards bodies. The scope of this document explores next generation electrical interconnection interfaces at 224 Gbps per lane.

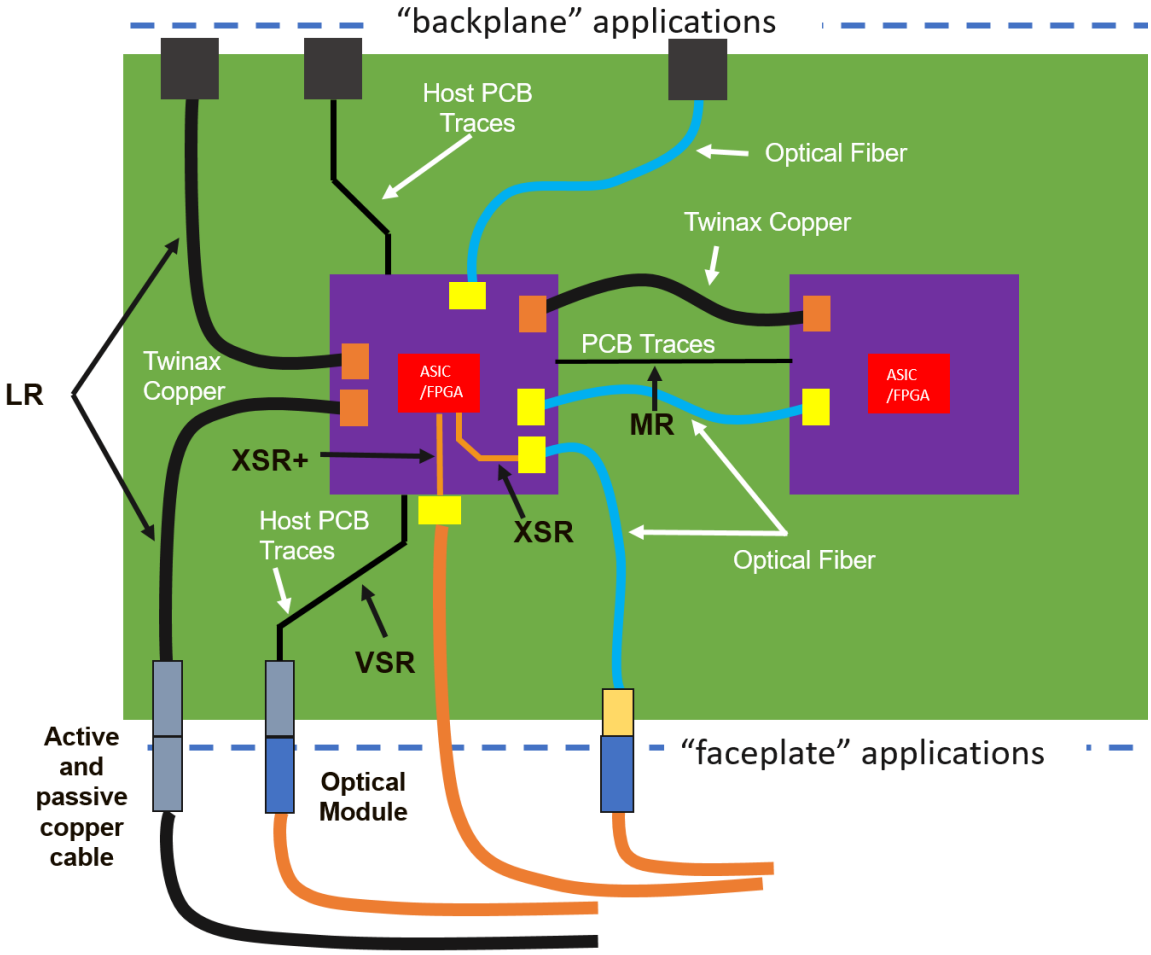


Figure 1 Interconnect Application Spaces

As shown in the Figure 1, interconnection interfaces in a typical system are needed for die-to-die, die-to-OE (optical engine), chip-to-module, chip-to-chip within a PCBA (printed circuit board assembly), between two PCBAs over a backplane/midplane or a copper cable, or even between two chassis. These interfaces may be unidirectional or bi-directional, optical or electrical, and may support a range of data rates.

For each application space, the IAs that follow from this framework should identify requirements to support interoperability across the various application spaces for optical and electrical links. They may include, but are not limited to:

- Relative cost considerations
- Electrical link reach and loss budgets
- Power consumption (including pJ/bit power efficiency)
- Channel requirements and general characteristics
- Signal levels and target BER
- Link Latency
- Test and measurement methodologies

The Framework Document may recommend a number of follow-on sub-projects to address interoperability for specific application spaces.

2.2 Motivation

With the growth of artificial intelligence (AI) computing and 5G/6G data traffic next generation systems are being driven by the need to handle the increasing volume of data traffic. Recently Ethernet IEEE 802.3, OTN ITU-T and OIF CEI have kicked off their efforts targeting signaling beyond 112 Gbps per lane. So what comes next beyond 112 Gbps for electrical interfaces over copper (Cu) channels? Will it be 200+ Gbps?

To meet the next-generation system bandwidth requirement, Ethernet is aiming at aggregate interconnects and modules with 800GbE and up to 1.6TbE capacity and next generation switches targeting 100T capacity. Other industry organizations' roadmaps are also facing needs for these higher data rates as well. A temporary solution could be increasing the number of electrical lanes. However, this on its own will not be sufficient to meet bandwidth growth and will not provide a long-term solution for the best integration density and cost. The next generation systems are constrained by limits on power consumption, by limits on the size of a system, and by the need to provide a cost effective solution. Higher capacity lanes would address these limitations. Therefore, 200+ Gbps per lane is critical to scale the device capacity and efficiency of networking and compute bandwidth.

2.3 Challenges and possible solution space

The historical evolution of OIF CEI long reach (LR) projects is listed in Table 1. OIF has developed IAs for different electrical interfaces for 11 Gbps (CEI-11G), 25 Gbps & 28 Gbps (CEI-25/28G), 56 Gbps (CEI-56G), and 112 Gbps (CEI-112G) which are at its disposal. Recently the work group is starting on 224 Gbps (CEI-224G) project to meet higher date rate needs.

Table 1 Historical evolution of CEI-LR projects

OIF CEI projects	CEI-25G	CEI-56G	CEI-112G	CEI-224G
Timeline	2011-2014	2014-2018	2018-2021	2021-
Ethernet rate	100G	50/100/200G	100/200/400G	200/400/800/1600G
Switch capacity	3.2T	12.5T	25T/50T	50T/100T
Per-lane data rate	25 Gbps	56 Gbps	112 Gbps	224 Gbps
Modulation	NRZ	PAM4	PAM4	TBD
Insertion loss	25dB at 12.5GHz	30dB at 14GHz	28dB at 28GHz	TBD
Reach objectives	5m copper cable	3m copper cable	2m copper cable	1m coper cable
Pre-FEC BER target	1e-15	1e-4	1e-4	TBD
SerDes architecture	Analog	Analog/DSP	Analog/DSP	TBD

However, copper interconnects are severely bandwidth limited and it is increasingly difficult to achieve the same link distances using higher signaling rates. The predominant next generation interconnect challenges to overcome are presented in Figure 2 in a solution space diagram, and are discussed in greater detail in subsequent sub-sections.

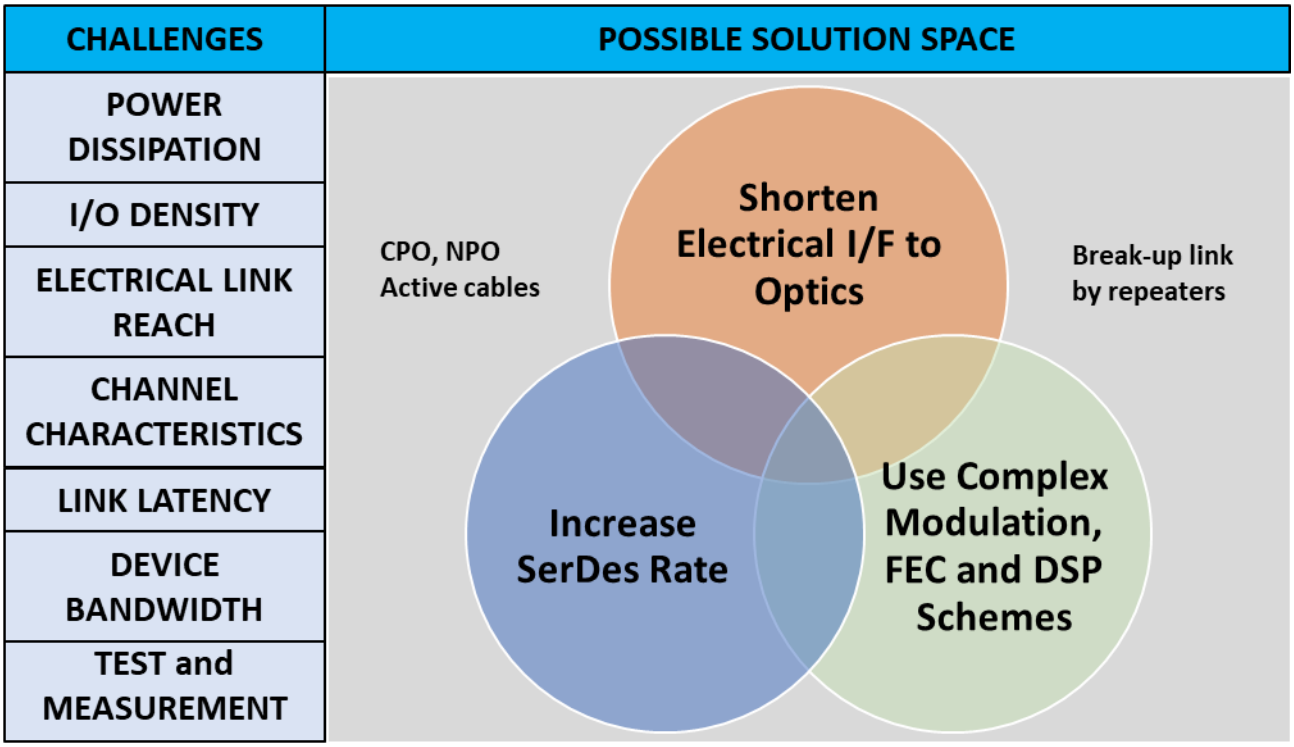


Figure 2 Next generation interconnect challenges

2.3.1 Challenges of cost, power and electrical link reach

It is sometimes useful to look back as to where we have come from and see what lessons we can learn and hopefully apply going forward. Over the past 12 years there has been relentless advancement within the industry resulting in an 80 times in overall system bandwidth. This has been primarily driven by several factors, including the development of seven generations of switching silicon (80x increase in bandwidth), four generations of SerDes speed increments (from 10G to 100G) and four increments in switch radix (from 64 SerDes per chip to 512 SerDes per chip) as shown in Figure 3. The current state of the art switching silicon has a capacity of 51T, based on 512 SerDes with each running at 112G. Scaling switch capacity beyond 51T will require higher speed electrical interfaces beyond 112G.

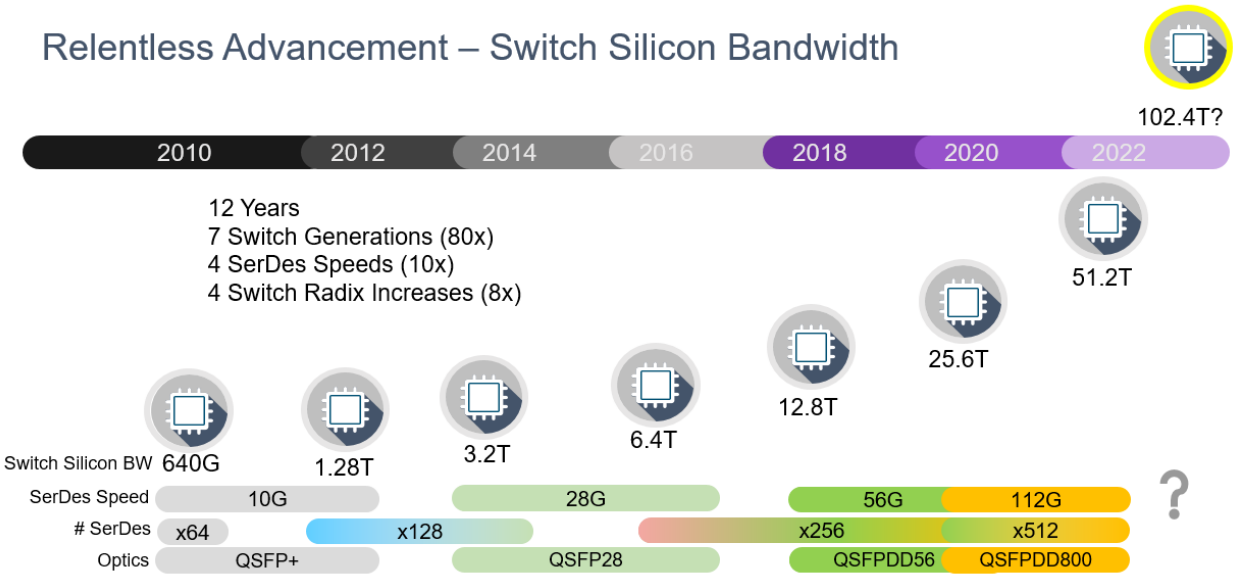


Figure 3 Relentless advancement – switch silicon bandwidth

However, as we think about doubling the bandwidth again the more fundamental question is how to manage the endlessly increasing power required by systems and networks? As Rakesh Chopra (Cisco) commented at a recent OIF webinar Cu (See you) beyond 112 Gbps "Power is everything, limiting what can be built, what customers can deploy and what our planet can sustain". Rakesh closed his talk with the following call to arms "We are at an inflection point in the industry where the pace of bandwidth growth and innovation isn't slowing down, and power is growing at an unsustainable rate. Therefore, it is a moral imperative that the industry solve this power challenge, ultimately driving a new set of innovations to bring these next-gen technologies to the industry."

Power is rapidly becoming the limiting factor for next generation networks and systems. It is limiting what equipment vendors can build and it is limiting what end users can deploy in their data centers. Data centers used to be limited by the size of the equipment that they would deploy. This is probably not the case anymore. Today the fundamental limitation for data centers is around the power per rack, the power per row and the power per building. From an equipment vendor perspective, it is no longer all about how small you can make your equipment, it's more about how power efficient you can make your equipment. In fact, we may see an expansion of

equipment size in order to drive that efficiency vector moving forward. There is also a business driver for reducing power as well, in that every watt data center operators consume in networking is a watt that they cannot consume in servers (which are what ultimately generate revenues).

However, looking at Figure 4 it is not all bad news. Over the past 12 years where the overall system bandwidth has increased by 80 times, the total system power has only increased 22 times. During this period the power per bit (per second) has actually been falling, so the power efficiency is getting better. The challenge is that it is not sufficient, and we need to look at ways of changing the slope of bandwidth growth and power efficiency.

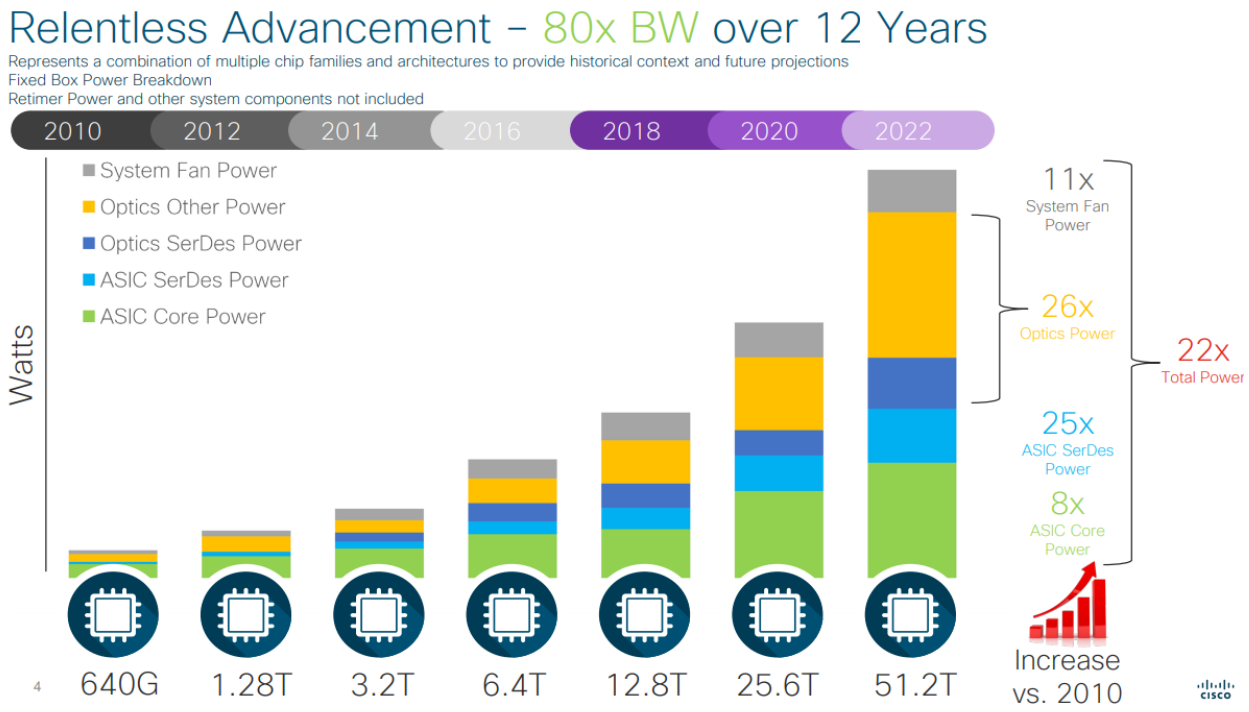


Figure 4 Relentless advancement – 80x BW over 12 years

Another observation from Figure 4 is that the SerDes power (both on the host and in any pluggable optics module) is increasing at a faster rate than any of the other components of the total system power. As we start to develop the next generation of higher speed electrical interfaces, it is more important than ever to focus on ways of minimizing the SerDes power.

From a system design perspective there are several approaches to minimizing SerDes power (either directly or indirectly) and some of these are illustrated in Figure 5.

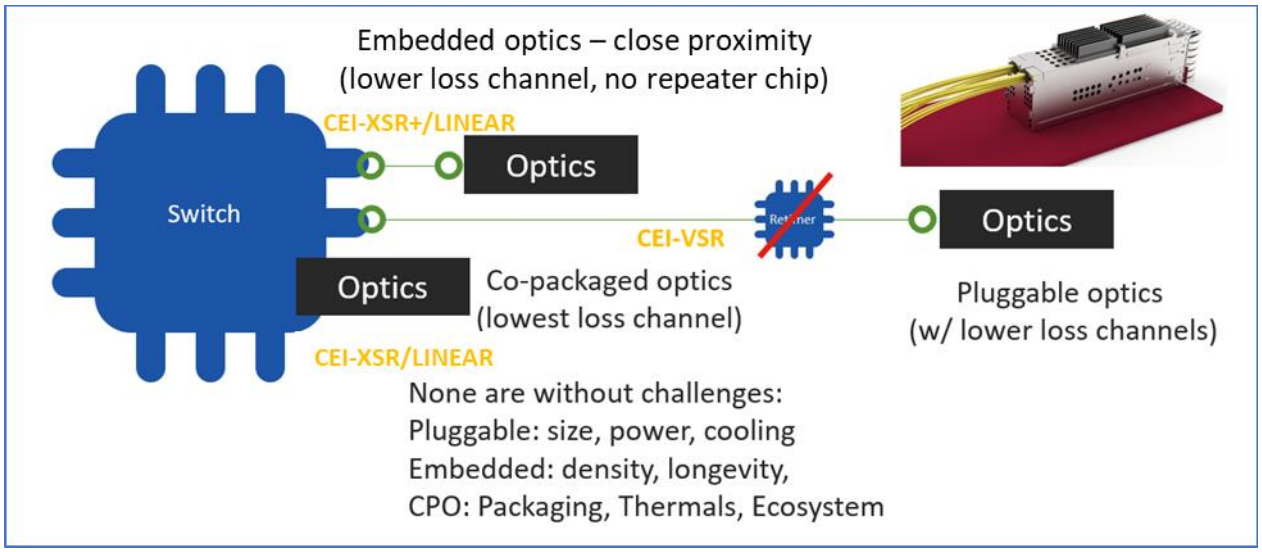


Figure 5 Approaches to minimizing SerDes power

One approach to reduce overall SerDes power is to reduce the number of SerDes. If the channel loss between the switch chip and the pluggable optics module on the faceplate can be lowered (potentially by using cabled hosts rather than PCB traces for the interconnect), then repeater chips that are often required between switch and optics (especially on longer traces) can be eliminated. Eliminating a repeater chip halves the number of SerDes per switch port (from four to two), with a potentially significant savings in power. In fact, it may be preferable to have a slightly higher power SerDes in the switch and the optics module, if this eliminates the need for a repeater.

Another approach to eliminate repeaters is to move the optics away from the faceplate and closer to the switch chip. This is sometimes called embedded optics or near package optics. This approach has the added advantage of a shorter and therefore lower loss electrical channel, with the potential for further power savings.

Continuing down the path of moving the optics closer to the switch chip, the final solution is to mount the optics on the same package as the switch itself. This is called co-packaged optics. This results in the shortest and lowest loss electrical channel between the switch and the optics, and therefore potentially offers the lowest power solution (lowest power per bit).

None of the approaches listed above is easy. They all represent a significant departure from how systems are typically built today. They each come with their own set of engineering, manufacturing and deployment challenges. At this point it is not clear which of these approaches (if any) will become the mainstream dominant solution. The industry will likely see deployment of all these different approaches (at least for the foreseeable future). It is therefore important that the CEI-224G project consider solutions that enable all these approaches.

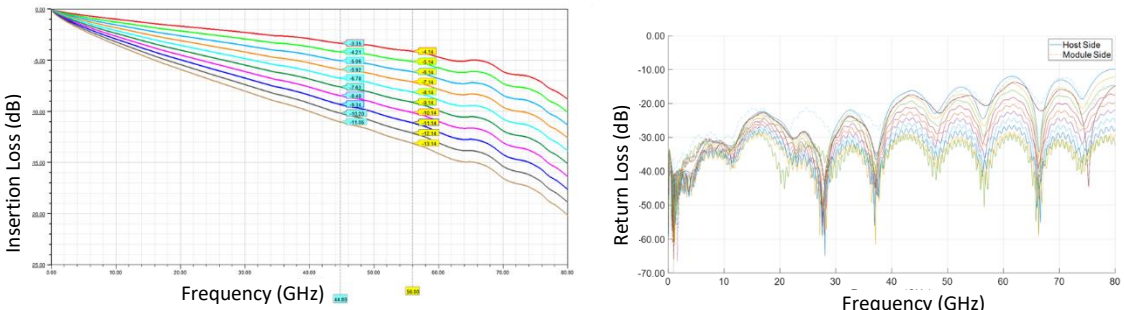
In summary, power is going to be one of the most critical considerations for the CEI-224G project and is likely to drive new power-focused system and network architectures, with a corresponding new set of next-generation electrical interface requirements.

2.3.2 Challenges of channel requirements and characteristics

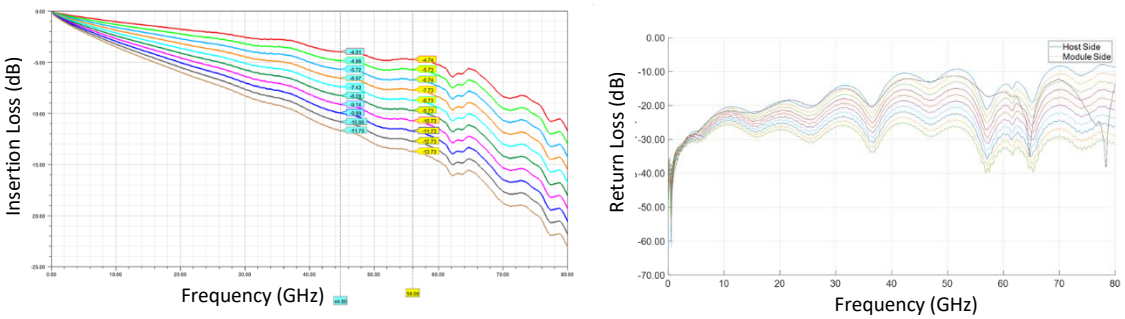
As one considers the opportunity for doubling the electrical data rates, a number of practical considerations emerge. Within comparable passive interconnect technology, increasing the operating data rate leads to degradation in the Signal-to-Noise Ratio (SNR). The amplitude of the signal component will decrease due to the increased conductive and dielectric losses of the medium. Due to shorter unit intervals at a higher rate, combined with inevitable mechanical feature transitions in the channel, the noise component due to reflections and crosstalk will increase and be harder to mitigate. Moreover, the accompanying need for increased I/O density further exacerbates the concern for cross talk due to port proximity. Finally, power consumption has become a critical parameter for network operators. Hence, channel definitions will need to carefully balance the usage of repeaters and high-performance equalization schemes vs. latency and power consumption. Bounds on the power envelope and active component maximum feature set will drive improved channel performance. Where will all of this come from?

When considering physical reach of an electrical channel, current materials that make up the medium significantly contribute to the total signal attenuation. This is agnostic of applications from short reaches such as die-to-die electrical links in a multi-chip module or an optical engine, to the other end of a spectrum when considering longer reaches such as backplane and copper cable applications. The chip substrate and printed circuit board (PCB) materials (dielectric and conductor) impact the total loss. Advancements are being made in these materials to support higher data rates, but still other media will likely need to be considered such as twinax cable for internal applications such as “cabled hosts” to preserve the channel reach. In addition, connectors will need to be improved for optimized high speed performance. Other ideas to be considered are flexibility on impedance for various components such as connectors and chip packages where it provides an improvement for the overall channel assembly. One of the effects that became clear during 112 Gbps IA developments was the impact that return loss could have in low loss channels. Based on this learning, 224 Gbps applications (and channel metrics) will need to more heavily weight the impact of factors beyond insertion loss. Specific VSR channel examples are illustrated in Figure 6. The tuning of a connector solution (including host and module attach points defined by MSA) to different system reference impedance yields very different channel performance (see oif2021.444).

(a)



(a)



(b)

Figure 6 224 Gbps Very Short Reach (VSR) channel simulation for different system reference impedance optimization; namely (a) 90 and (b) 100 Ohm differential system reference impedance.

Another key factor to consider when discussing channel requirements is the modulation solution(s) that will be targeted for 224 Gbps. A modest increase in modulation complexity can reduce the signaling baud rate and significantly impact the quality and cost of the printed circuit board material that is required to support the channel reach or loss. On the other hand, the increased modulation level will also decrease the SNR for the end-to-end system. Moreover, this added modulation complexity may bring an engineering challenge to the chip design, but once done, may have a small incremental impact on the combined die functionality and cost. The cost for advanced channel design/acquisition cost, and cost for advanced SerDes design/acquisition cost need to be both considered and balanced. The same thought process must be considered before it is dismissed regarding modulation solutions for different reaches for example. While the benefit of having a common modulation solution across both optical and copper applications is obvious, again, the incremental cost of a 'dual mode' modulator where modulation 1 is optimized for lower loss chip to optical channels and modulation 2 is optimized for higher loss chip to chip copper channels, may be minor compared to the overall deployment costs when the volume of copper channels (exotic PCB material vs good PCB material) is considered.

As industry considers IO port and circuit density increases driven by aggregate bandwidth roadmaps, great care will be required to ensure crosstalk is well managed in high density channel applications. Careful design optimization of connector and silicon package footprints as well as

printed circuit board lithography processes will become critical to mitigate the harmful efforts of these circuit transitions. In addition, increasing port densities can negatively impact channel reaches due to the need for isolation of traces. One key architecture to be evaluated for potential positive benefits is the termination of cabled host twinax cable solutions that directly terminate to chip packages rather than terminating to the host printed circuit boards near the chip package. This minimizes all the parasitic effects of the connector to host and host to package transitions as well as reducing the package effects.

Power consumption of a channel has become one of the highest priorities for network operators and is a critical consideration. Preservation of passive copper channels, where possible, enables the power reduction by not requiring short optical link, as long as the electrical transmitter and receiver can be realized in a power efficient manner. These silicon and channel tradeoffs must be carefully considered and debated to ensure the end user gets the best possible power efficiency. The industry has been very clear on the need to carefully design systems for power efficiency in light of the forecasted power envelope growth. As an example, whereas repeaters might have been sought to mitigate channel reach limitation, their use may not fit in the system power envelope. Therefore, the margin needs to be accommodated by a cleaner passive channel.

2.3.3 Challenges of material characteristics, properties, fabrication and modeling

New or improved manufacturing or materials for PCBs, cables, connectors, packages are likely. Measurements demonstrated in section 2.3.5 show a 25% improvement in package routing which may be required for 224 Gbps per lane. Scalable models created from this (Figure 7) show a comparison to 1 inch of the OIF 112 Gbps per lane LR reference package routing model.

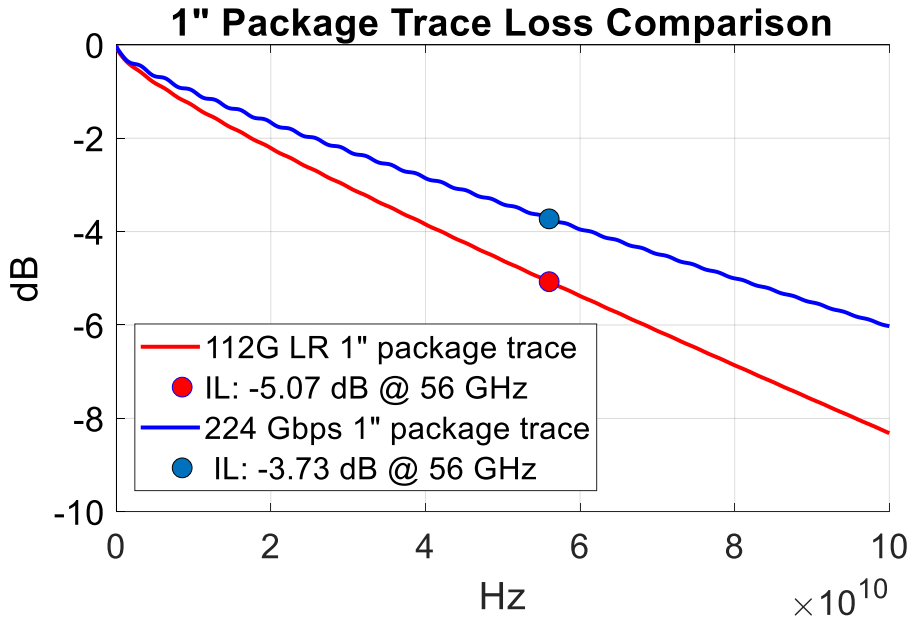


Figure 7 Package Trace Improvement

Insertion loss approximating 1.3 dB per inch at 56 GHz are detailed by Amphenol in the OSFP200GEL MSA. The aforementioned package and PCB models converted to IEEE 802.3 Annex 93A transmission line parameters in Table 2 may be used for a 224 Gbps link simulation starting point.

Table 2 Starting Point for 224 Gbps Package and PCB Trace models (IEEE 802.3 Annex 93A)

Parameter	Setting	Units
board_tl_gamma0_a1_a2	[0 0.000567732 2.90358e-05] †	
board_tl_tau	0.0058	ns/mm
board_Z_c	100	Ohm
package_tl_gamma0_a1_a2	[0 0.000644085 0.00018018]	
package_tl_tau	0.0057	ns/mm
package_Z_c	[87.5 87.5 ; 92.5 92.5]	Ohm

† Insertion loss approximating 1.3 dB/in at 56 GHz

Closer attention to details of electromagnetic model management is required as unit intervals approach 9 ps. Some of these include skew and non-transverse fields. The effect is that excess P/N skew and a transmission path coupling to itself can result in larger than expected pulse response precursor impairments. Figure 8 depicts a 9 ps unit pulse at the end of a 15 dB (at 56 GHz) channel with and without 4 ps intrapair skew. This type of phenomena is exacerbated by a shrinking UI and introduces renewed considerations for material, fabrication, modeling, and SerDes design.

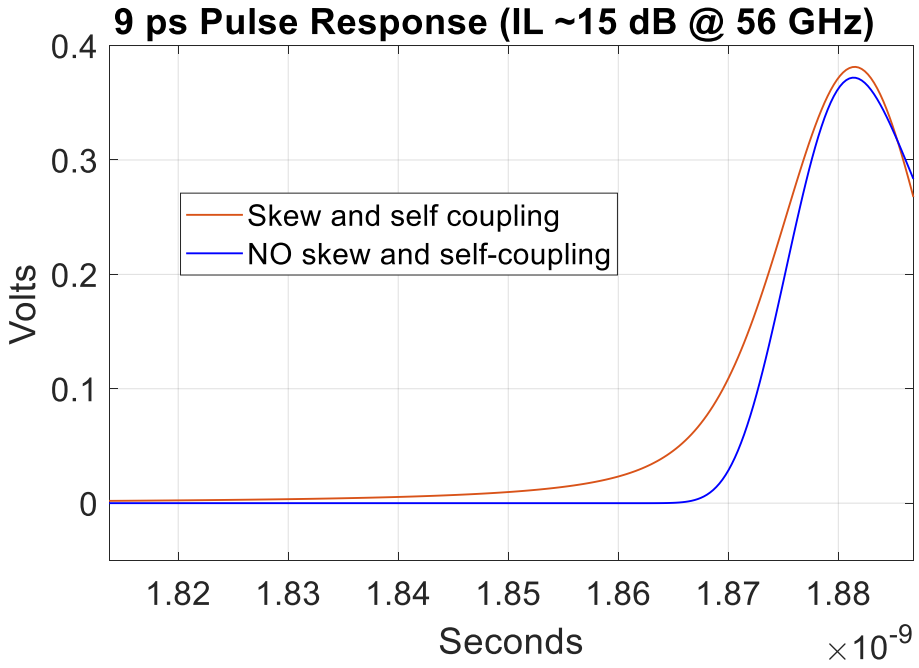


Figure 8 Pulse Response Illustration of Pre-cursor Challenge

2.3.4 Challenges of modulation, equalization, target DER, and FEC/latency

If we follow the SerDes technology evolution of doubling the data rate per lane every 2-3 years, the next generation I/O data rate will be 224 Gbps. In CEI-112G implementation agreements (IA), PAM4 signaling can drive die-to-die/optical engine (OE), chip-to module (VSR), chip-to-chip (MR) and even chassis-to-chassis interfaces (LR) through substrate traces within a package, PCB traces over a board, or copper cables at 112 Gbps. In order for a next generation electrical interconnect to achieve 224 Gbps per lane, the leading options requiring consideration are:

Doubling of the PAM4 signaling rate: The current 112 Gbps per lane electrical interfaces predominantly use the 4-level pulse-amplitude modulation (PAM) scheme PAM4 at a signaling rate of 56 GBd, 4-level signaling with 2 bits of information per symbol. Maintaining a PAM4 modulation, a data rate of 224 Gbps requires a baud rate of 112 GBd.

An increase of the PAM constellation size: To totally avoid increasing the baud rate to 112 GBd for a 224 Gbps data rate, a higher level modulation scheme such as PAM16, 16-level signaling with 4 bits per symbol, would be needed. Alternatively, potentially more feasible PAM6 and PAM8 could achieve 224 Gbps by decreasing the baud rate to 90 GBd and 75 GBd respectively.

Higher complexity modulation schemes: There do exist various Multicarrier Modulation schemes such as Orthogonal Frequency Division Multiplexing (OFDM) and Discrete Multitone Modulation (DMT) which could ease system bandwidth requirements but at the expense of more complex solutions than that offered by a PAMN system.

Use of fiber optic cables: With signaling rate increasing, electrical channels such as PCB traces or copper cables both have bandwidth limitations over certain reach distances. Alternatively, optical fiber cables can be used to transmit high bandwidth data over long distances. There are new emerging technologies such as silicon photonic, chiplet and co-packaging optics (CPO) where the majority of the channels exist in the optical domain.

Even with the most advanced printed circuit board or cable technology, the insertion loss would become difficult to conquer for the option of doubling the signaling rate to 112 GBd. High bandwidth also presents challenges for the design of packages, connectors, passive channels, and transceivers.

To overcome the bandwidth limitation, higher modulation levels could be considered. A higher level PAM modulation offers a larger unit interval and a reduced bandwidth requirement. However, the SNR requirements to achieve a certain error rate depends on the modulation scheme chosen. Table 3 lists key parameters for different PAM schemes, such as the number of bits per symbol, signaling rate, unit interval, fundamental frequency, SNR penalty, and jitter tolerance. We can see that increasing modulation levels can help reduce the bandwidth and the sampling rate, but at the cost of higher SNR penalties and sensitivity to noise and jitter.

Table 3 Key parameters for different PAM schemes

Data rate, Gbps	112	224					
Number of PAM levels	**4**	**4**	**5**	**6**	**7**	**8**	**16**
Bits per symbol *	2	2	2.25	2.5	2.75	3	4
Signaling rate, GBd	56.25	112.5	100	90	81.82	75	56.25
Unit interval, ps	17.78	8.89	10	11.11	12.22	13.33	17.78
Fundamental frequency, GHz	28.125	56.25	50	45	40.91	37.5	28.125
Required SNR at slicer, dB **	20.42	20.42	22.43	24.04	25.40	26.56	32.55
SNR penalty, dB	0	0	2.01	3.62	4.97	6.14	12.12

* Assumes an efficient mapping of bits to PAM symbols.
** For BER = 1e−6.

Undoubtedly, a common modulation scheme between optical and electrical links provides interface with lower power/cost advantages. If the optical link uses a certain PAM modulation scheme (say PAM4) it may also be of benefit for the electrical links to support the same modulation scheme, without the need of using gear-box which will add power/area/cost to the module. Then the optical engine just needs to convert signals linearly across the O-to-E & E-to-O interfaces. Furthermore, backward compatibility to legacy NRZ and PAM4 signaling formats needs to be considered as well. Hence, we should consider the tradeoffs between a common modulation vs. differentiated modulations in the total implementation complexity, power and cost. For the LR interface over backplane and copper cables, it may be worth to investigate advanced modulation formats beyond PAM4 to maximize the copper reach at the expense of modulation scheme commonality.

If PAM4 or even higher order modulation schemes have to be used to squeeze the signal bandwidth to match the harsh frequency response of the channel, advanced signal processing and advanced forward error correction (FEC) schemes would be considered for CEI-224G electrical links. As the processing node improves, advanced digital signal processing (DSP) after an analog-to-digital converter (ADC) can provide cost-effective and strong equalization for lossy and noisy channels. As shown in Figure 9, multi-tap feed forward equalization (FFE) and decision feedback equalization (DFE) can work together to cancel inter-symbol interference (ISI) caused by channel insertion loss and return loss. Advanced detection schemes such as maximum likelihood sequence estimation (MLSE) and its variants may be used to further improve the equalization capability of the receiver without significantly enhancing noise and crosstalk from aggressor channels. Furthermore, advanced DSP and FEC would be combined to provide joint optimization for a system design. Unlike an analog based receiver, a DSP receiver could provide non-binary soft information to the FEC decoder to enable soft decision decoding and achieve higher coding gain.

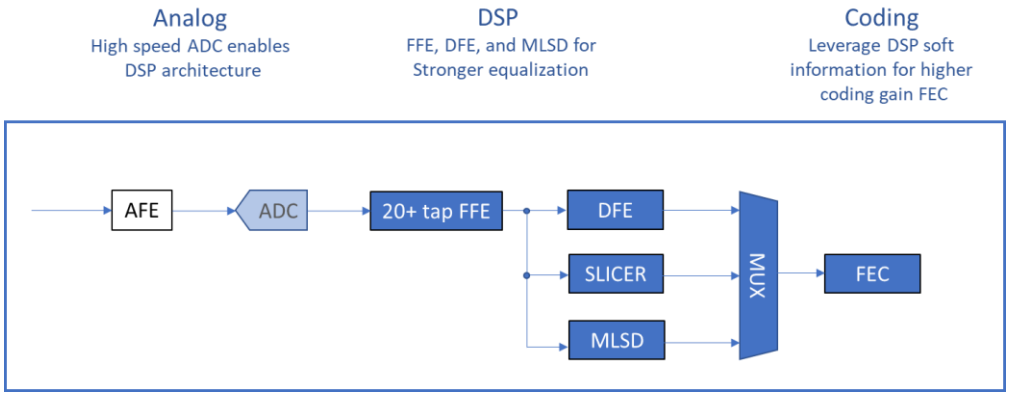


Figure 9 Block diagram of a DSP receiver

CEI-224G interfaces specifically addressing optical applications (such as XSR, NPO and VSR) must carefully consider the tradeoff between equalization complexity and channel performance in order to maintain a reasonable power envelope for OEs and NPO/VSR modules. Powerful DSP approaches (which are feasible for higher loss interfaces) may not be suitable for such interfaces and should be amenable to simpler analog and or reduced complexity DSP architectures.

Before looking into different FEC options, let's review the existing error correction architectures. Because of the SNR penalty of PAM4 over NRZ, FEC is assumed to be used in the system to achieve corrected BER of 1e-15 or better. There are two major types of FEC architectures in a multi-part link system as shown in Figure 10: shared FEC and terminated FEC. For a shared FEC architecture, chip-to-chip and/or chip-to-module electrical links and the optical link share a single FEC encoder and decoder, while in a terminated FEC architecture each electrical link and optical link have their own encoder and decoder. We will assume both FEC architectures are possible in the next generation system. It is worth pointing out that in 100GE/200GE/400GE a single shared FEC is used.

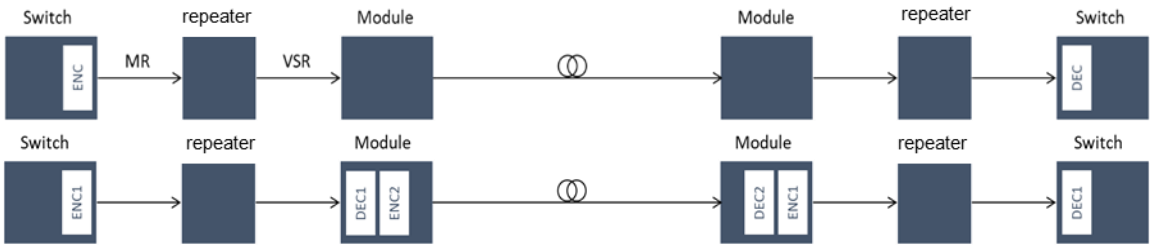


Figure 10 Shared FEC architecture (top) and Terminated FEC architecture (bottom)

The required BER target on the electrical link is normally more stringent than the optical link in a shared FEC architecture since the bulk of the coding gain is assigned to the toughest part of the channel (the optical link), and a relatively smaller coding gain is allocated to the electrical links. To allow simpler and lower latency FEC codes than Ethernet's RS (544, 514, 15) (a.k.a. KP4 FEC) to be used in the system OIF CEI-112G implementation agreements (IA) specify following BER targets:

- 1e-9, 1e-8 and 1e-6 for CEI-112G-XSR
- 1e-6 for CEI-112G-VSR
- 1e-6 for CEI-112G-MR
- 1e-4 for CEI-112G-LR

To achieve more robust system performance a dedicated FEC for electrical links could be used. By doing this, we can relax the BER target from current CEI-112G IAs to one or two orders of magnitude higher, say 1e-5 or 1e-4 for VSR and MR interfaces. Relaxing BER from 1e-6 to lower targets such as 1e-5 or 1e-4 can significantly improve the SerDes SNR margins by 0.96 dB and 2.19 dB, respectively. However, terminating the FEC at each segment of electrical and optical interfaces will introduce more encoding/decoding complexity and latency to the overall link.

Alternatively, concatenated coding can improve both optical link and electrical link performance without terminating the FEC at each segment of the link. As shown in Figure 11, an outer code at the host side and an inner code at the line side can be concatenated to enhance the overall coding gain. The outer code could be legacy RS (544, 514, 15) FEC with a simple Hamming or BCH code as the inner code. The inner code encoder and decoder could be disabled or bypassed for backward compatibility if extra coding gain is not needed.

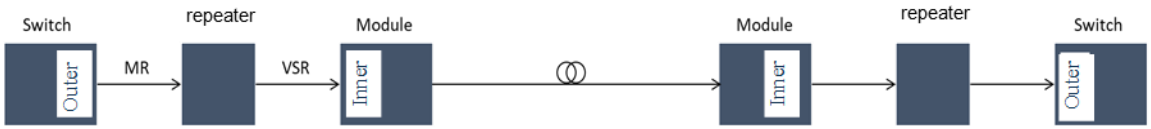


Figure 11 Concatenated FEC architecture for a multi-part link system

Besides terminated and concatenated architectures with KP4 FEC, there are many other advanced FEC options such as longer RS codes, product codes, and low density parity check (LDPC) codes that could be considered to obtain further coding gains. Furthermore, a soft decision decoding algorithm leveraging soft information provided by a DSP receiver could provide 2 dB to 3 dB more coding gain than hard decision decoding.

However, stronger FEC for the electrical link to achieve better coding gains normally come at the cost of coding overhead, coding latency and coding complexity. The trade-off of performance and cost for different applications needs to be carefully considered, especially for CEI-224G projects with low latency or/and low power application requirements.

2.3.5 Challenges of test and measurement

At this early time in the 224 Gbps stage of development the test and measurement equipment (TME) industry is not offering metrology grade hardware-based clock recovery systems that can function at these speeds. If early silicon designers can facilitate a correlated reference clock out of their test silicon, then a sampling architecture is the preferred avenue to address these early measurement needs. If such a correlated reference clock is not available, the only avenue forward is to leverage a real-time architecture. Areal-time oscilloscope solves the need for

physical clock recovery by leveraging deep acquisition memory and over-sampling to an extent that permits a digital signal processing approach to clock recovery and advanced PAM4 measurements.

While bandwidth requirements for 224 Gbps are still a matter of future standards development a simple extension of the logic behind 40 GHz requirements for 112G would suggest 80 GHz (3 dB) following a 4th order Bessel Thomson trajectory to ~120 GHz would be reasonable. Silicon packaging will also play a key part in finalizing decisions around modulation format and required measurement bandwidth expectations.

Preliminary GL102 and GZ41 material microstrip transmission models based on 35 mm traces suggest a strong reliance on traditional equalization methods to recover a relatively high BER signal relative to similar equalization methods applied at 112G.

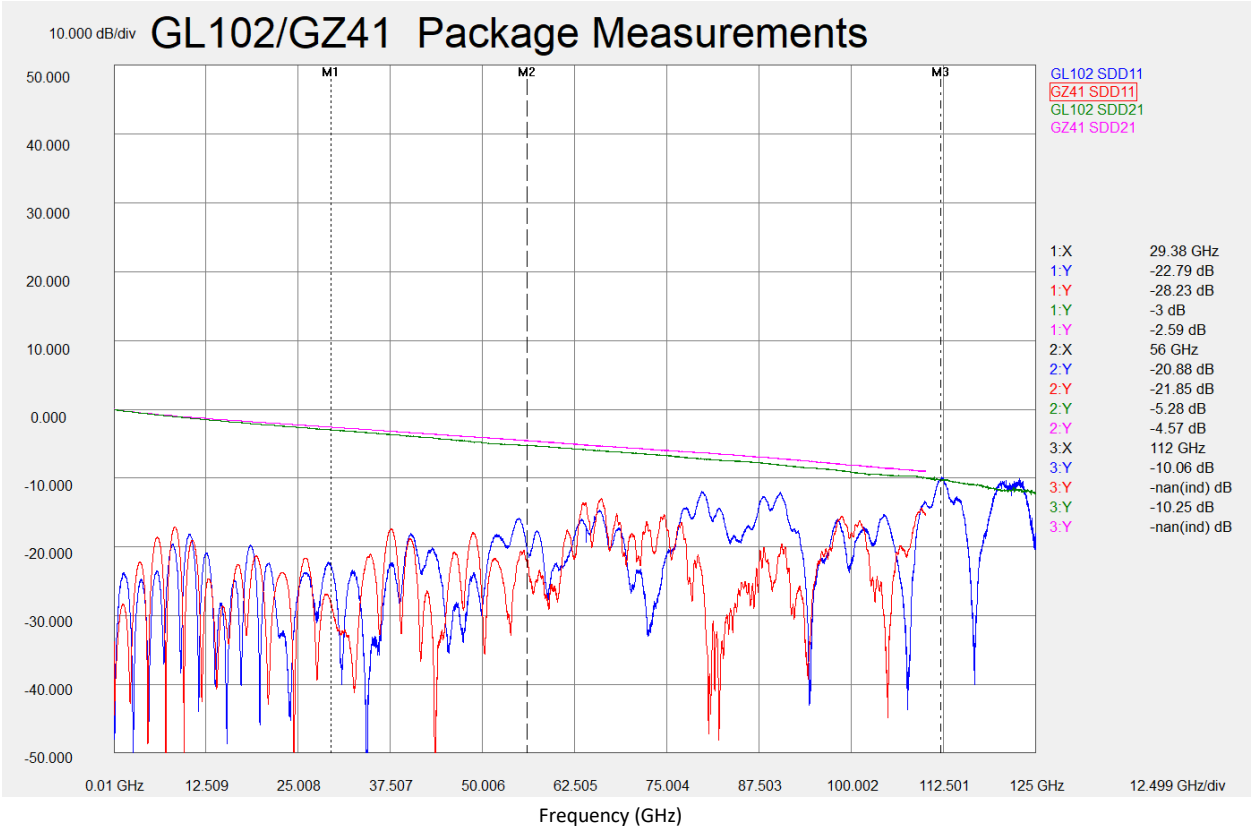


Figure 12 GL102/GZ41 (differential insertion/return loss) Package measurements

A 112 GBd PAM4 precision stimulus system incident to either of the illustrated packages above results in the following measurements.

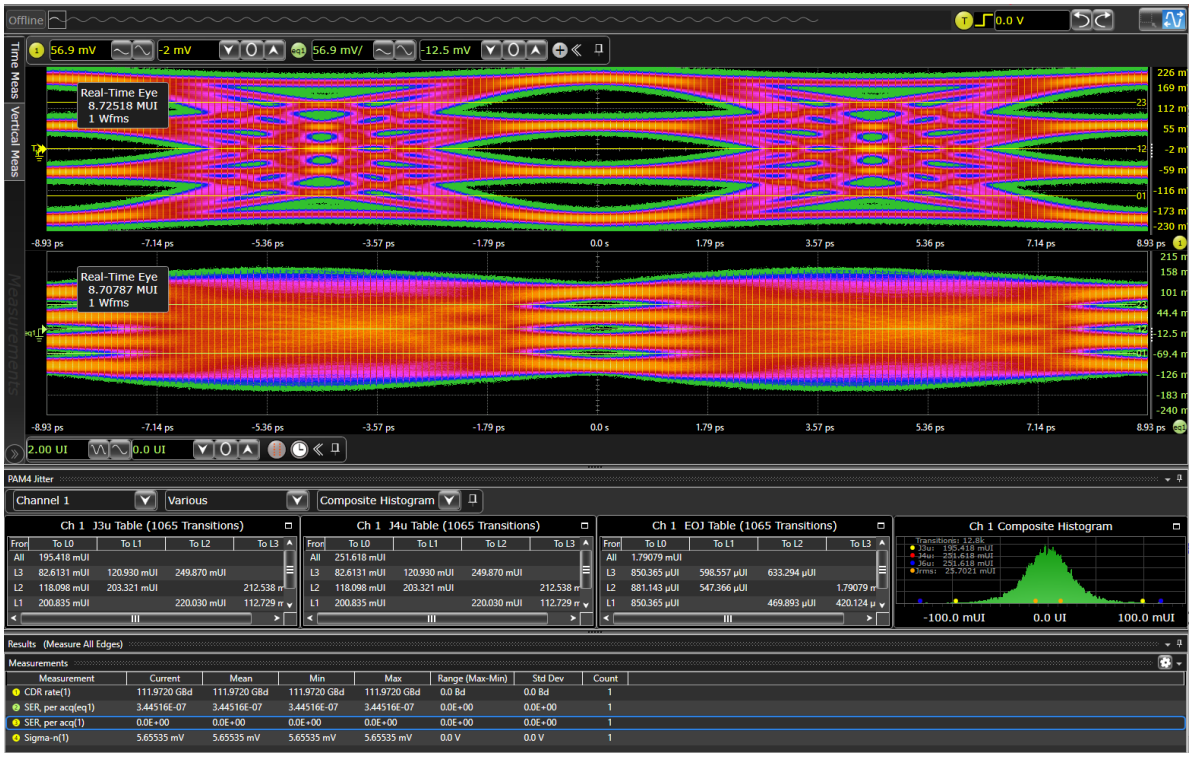


Figure 13 Real-Time instrument measurement of a 112 GBd PAM4 precision stimulus system incident to the illustrated packages

An analysis of this series of measurements at effectively the "TP0" point proposed by the measured 200G package model offer the following insights.

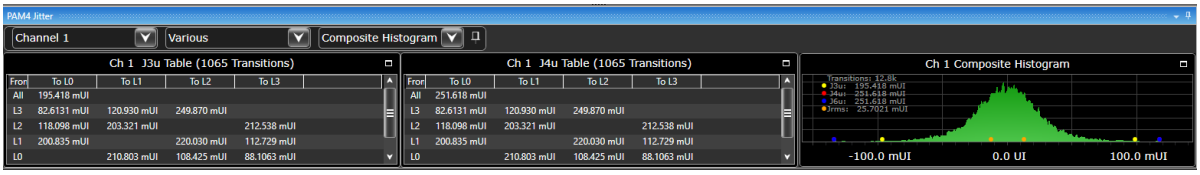
Measurement	Current
1 CDR rate(1)	111.9720 GBd
2 SER, per acq(eq1)	3.44516E-07
3 SER, per acq(1)	0.0E+00
4 Sigma-n(1)	5.65535 mV

Measurement 1: The instrument has locked onto the 112 GBd data rate with a Digital Signal Processed Clock Recovery and has effectively provided stable first or second order tracking.

Measurement 2: After a nominal CTLE and FFE tuned operation (no DFE yet) a symbol error rate of 3.4E-7 is observed.

Measurement 3: Within the boundaries of 8E6 UI's there are no incident symbol errors in the supplied 112 GBd PAM4 signal.

Measurement 4: Observed noise levels (with no instrument noise removal) when integrated out to 130 GHz tend to the high side of nominally 6 mV. SNDR will be significantly compromised by this unless some creative noise reduction methods are leveraged.



J3u and J4u are higher than expected limits in this example setup. A significant part of this is the stimulus system and slew rate limiting occurring at these speeds. If noise de-embed was applied at the instrument level, these jitter values of ~195 mUI should drop about 20-30% to ~150 mUI. This may well suggest the need for some nominal level of equalization to be performed in concert with the 12-edge jitter calculations, which is normally not the case. In this example the loss compensation was simply for a 5 dB package model to reach a TP0 level of experimental results.

Similar jitter results are also observed after a 3X serialized version of this package model is applied for a net loss at Fbaud of ~15 dB. The quality of the signal is clearly heavily impaired at this point and it's likely that a high tap count DFE would be needed along with deeper FFE's (Figure 14) to fully recover a 1E-4 raw BER signal after a 15 dB channel.

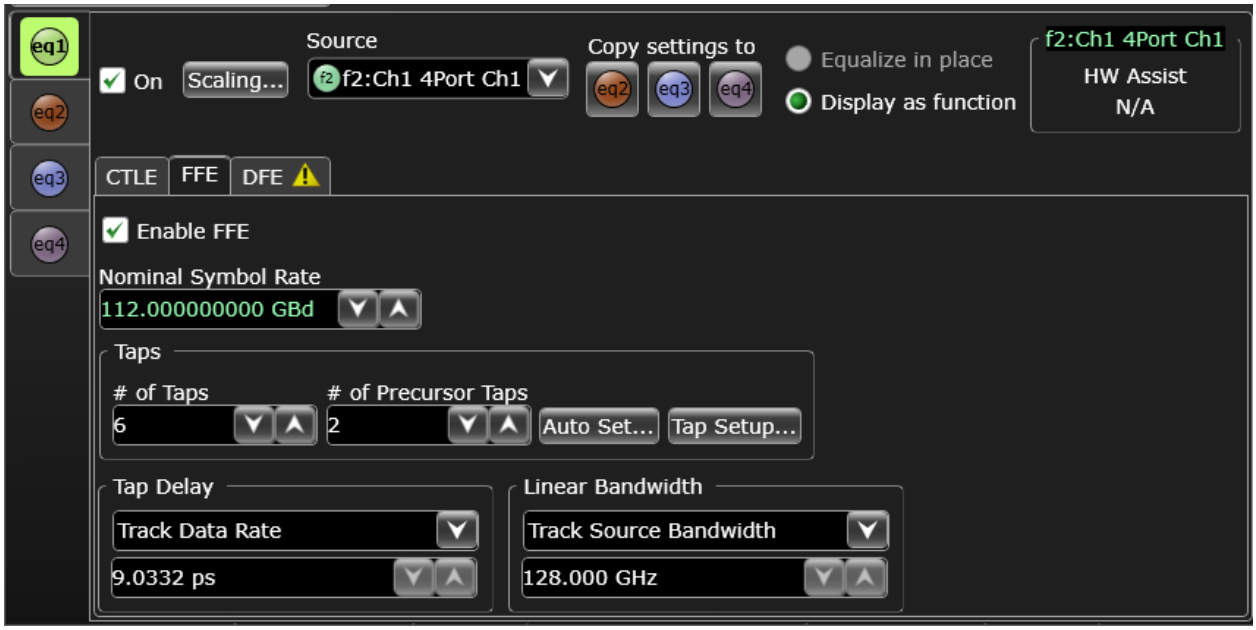


Figure 14 Equalization configuration for 112 GBd PAM4

An Equivalent-Time instrument analysis of this same precision 112 GBd PAM4 stimulus will depend on careful consideration of an explicit forwarded clock to use in the interim while a physical 112 GBd CDR system is released for this speed.

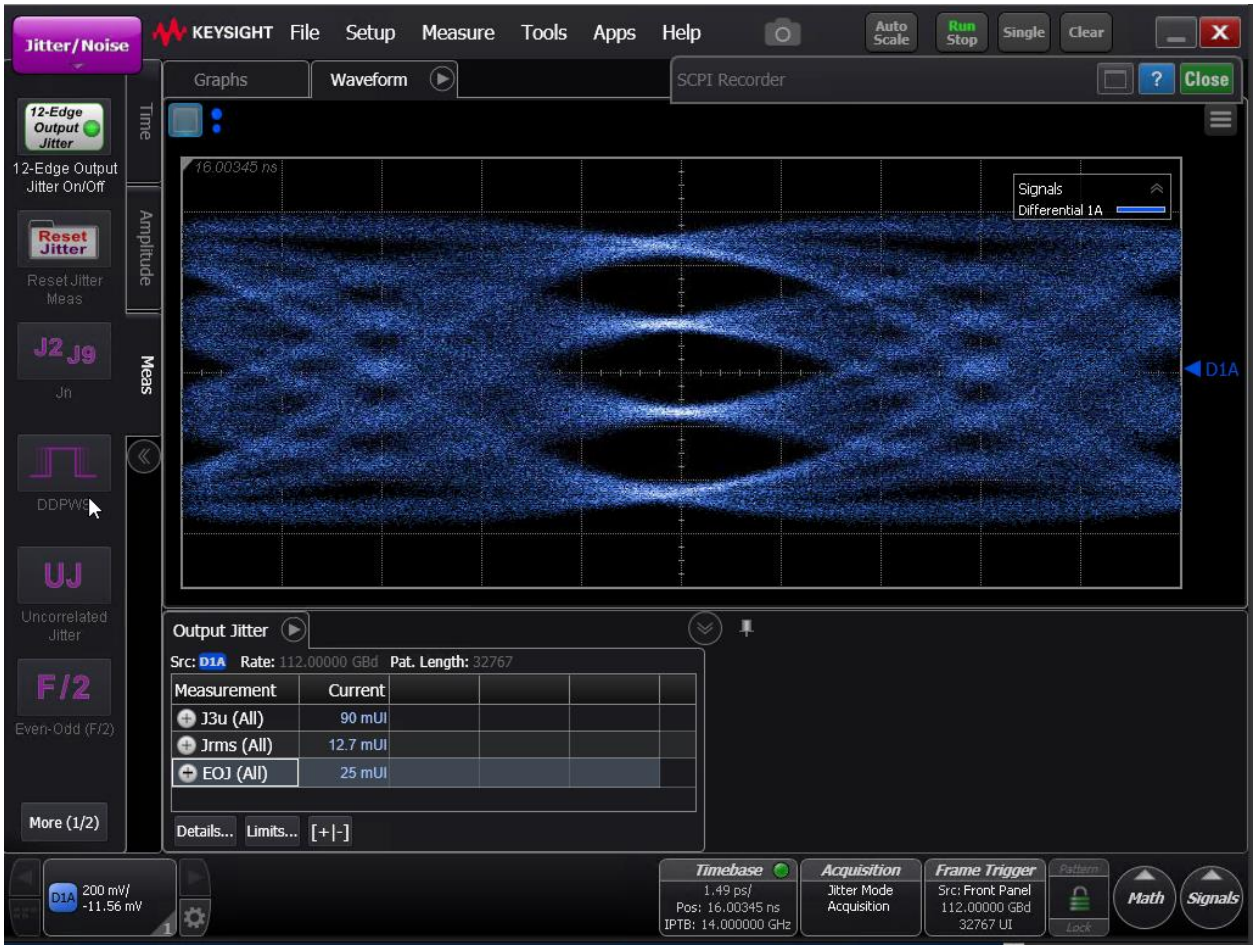


Figure 15 Equivalent-Time instrument measurement of a 112 GBd PAM4 precision stimulus

In this configuration, we see the practical realization of 12-edge jitter values inline with existing 56 GBd PAM4 jitter specs, normalized on a per-UI basis. In Figure 15, a 90 mUI J3u value offers validation that under carefully controlled forwarded clock conditions there is satisfactory margin to evaluate 112 GBd PAM4 signals today.

While instrumentation capability will be advancing over the course of early 224 Gbps technology introduction, it's important to recognize that today, all the tools for both signal stimulus and acquisition/analysis exist in a form to advance early standards and spec development.

2.4 Summary

As time proceeds, ICs will become faster and denser. To cope with the issue of interconnect capacity and density of future systems, 224 Gbps per lane is critical to scale the switch capacity and efficiency of networking bandwidth.

The implementation of CEI-224G interconnect technology poses several challenges especially in relation to: bounded power dissipation, limited I/O density, high channel and device bandwidth,

low latency, and optimal electrical link reach. Highlighted were the side-effects of some solutions, which are a result of the complex inter-dependencies of: higher integration, increasing SerDes rate, complex modulation, FEC and DSP schemes, chip break-out and routing, signal conditioning, thermal & power issues, package footprint, etc.

In conclusion, further study is required to decide on a solution for each of the challenges identified, in order to achieve a cost effective CEI-224G interconnect solution that satisfies the power density pJ/bit/dB requirement.

The CEI-224G Interconnect Framework explores the interconnect needs for next generation systems and identifies applications for possible work at the OIF or other standards bodies to address the industry's next generation needs.

The purpose of this document is to foster communications between optical/electrical interworking technology users & providers, which comprises an ecosystem of: end users, system and equipment vendors, cable/connector vendors, and silicon vendors. Also, this document is to serve as a "Statement of Understanding" between optical/electrical interworking technology users and providers, for achieving coordinated solutions for NG Interconnects.

3 Interconnect Applications

The next generation (NG) interconnect application spaces mentioned in section 2 can be broken down into the following applications.

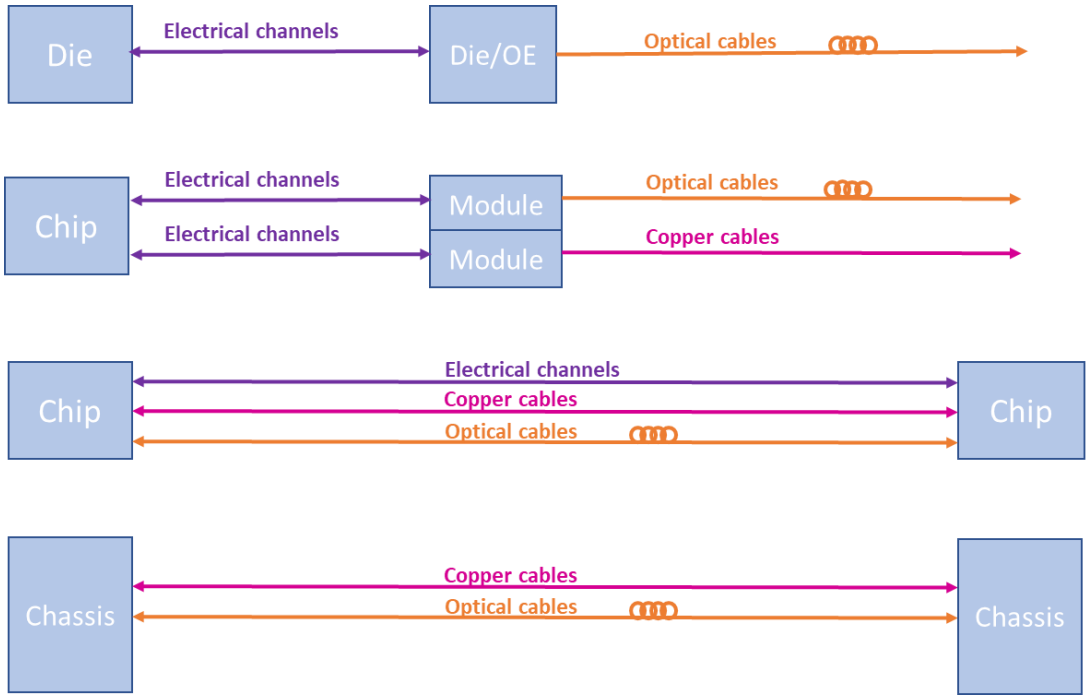


Figure 16 Interconnect Application Spaces

3.1 Die to Die Interconnect Within a Package

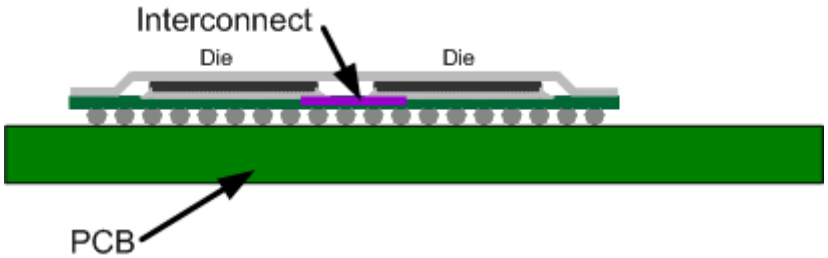


Figure 17 Die to Die within an MCM Interconnect Application Space

It may be necessary to use multiple dies within a multi-chip module (MCM) to achieve the industry's objectives. These co-packaged solutions can communicate with low power since the substrate provides a high quality communication channel.

The communication channel would typically be less than 50 mm package substrate trace. This short electrical link may allow for a much simpler interface and require less power than other

existing standard electrical interfaces. For example, equalization is unlikely to be needed and it may be possible to assume such short links are synchronous (single reference clock going to all chips), removing the need for a frequency tracking CDR.

Future dies may also have direct optical input/output, such that the die to die interconnect would be optical.

3.2 Die to optical engine within a package

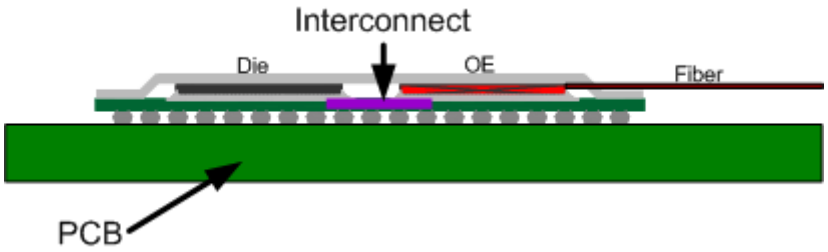


Figure 18 Die to Optical Engine MCM Interconnect Application Space

It may be necessary to use a die and an optical engine within a multi-chip module (MCM) or a co-packaged optic (CPO) to achieve the industry's objectives. These co-packaged solutions can communicate with low power since the substrate provides a high quality communication channel.

The communication channel would typically be less than 50 mm package substrate trace with or without a socket interconnector. This short electrical link may allow for a much simpler interface and require less power than other existing standard electrical interfaces.

If the optical link uses a certain PAM modulation scheme it may also be of benefit for the electrical links to support the same modulation scheme. Then it would be possible that the processing of the modulation scheme would be in the chip (i.e. the optical engine just needs to convert signals linearly across the O-to-E & E-to-O interfaces).

3.3 Chip to Nearby Optical Engine

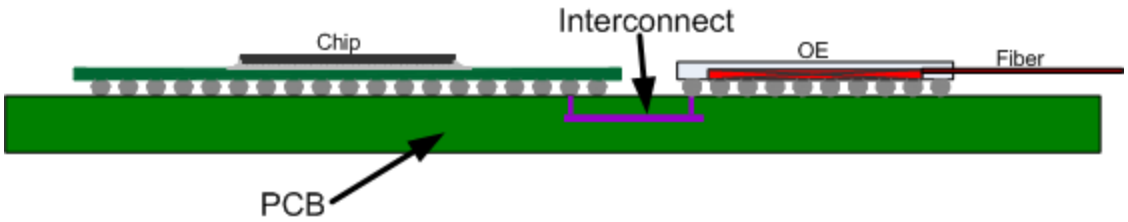


Figure 19 Chip to nearby OE Interconnect Application Space

It may be useful to place an optical interface very close to the host chip (rather than placing the optical device within a host MCM due to heat restrictions of the optical components). In this case, a short electrical link of less than 150 mm PCB trace is anticipated. Although this type of link will

require more power than a link within a multi-chip module, the short reach of this channel would still imply power could be saved. Again the same modulation format may be appropriate for such links for backward compatibility and if the optical side uses a certain format.

3.4 Chip to Module

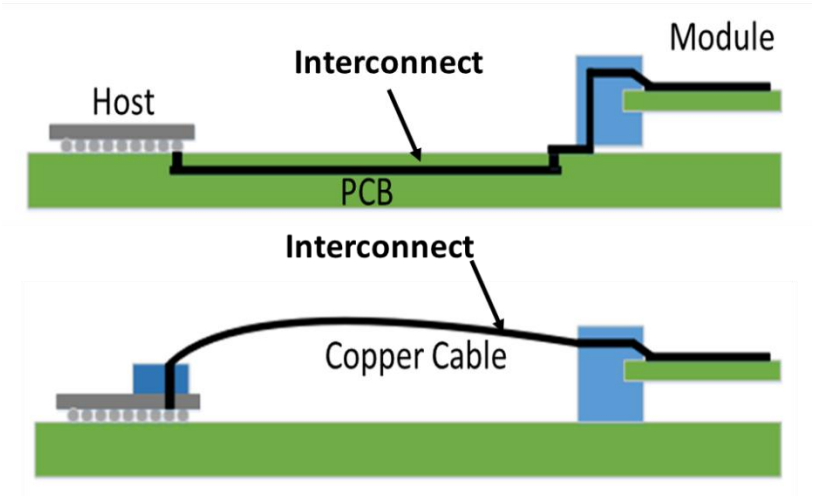


Figure 20 Chip to Module Interconnect Application Space

It is common in modern communication systems to support pluggable modules at the front faceplate of the equipment. The electrical link used to connect these pluggable modules can extend to beyond 200 mm of host PCB trace plus a connector and minimum 20 mm module PCB trace. Besides traditional PCB-based hosts (top), new cabled-host (bottom) implementations could be considered to provide more port flexibility and link budget as shown in Figure 20. At higher data rates placing retiming devices inside the pluggable module provides support for longer host traces but the inclusion of complex equalization features can overburden the limited power budgets of the pluggable module. Alternatively, a linear chip to optical engine (OE) interface could be considered to enable low power, low cost, small form factor 224G serial optical modules in CPO, NPO and VSR applications.

Again the same modulation format may be appropriate for such links for backward compatibility and if the optical side uses a certain format. However, advanced modulation formats beyond PAM4 or/and advanced FEC and equalization features are all possible solutions for the chip to module interconnect.

3.5 Chip to Chip within PCBA

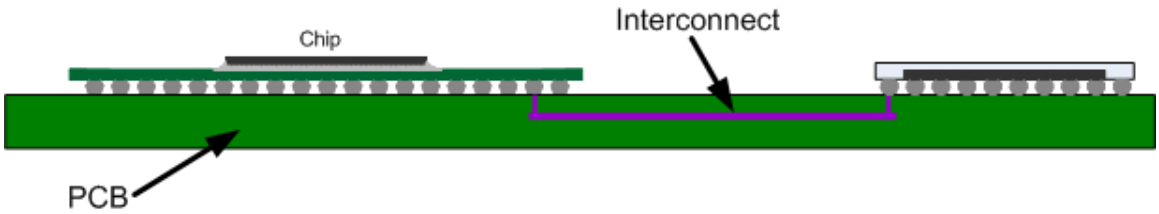


Figure 21 Chip to Chip within PCBA Interconnect Application Space

An interconnection interface may be needed between two chips on the same PCBA or on a daughter card or shorter mid-plane. By definition, this interface is relatively short ranging from 1 cm to perhaps 50 cm PCB trace and up to one connector.

Most chip to chip environments can save power if one can assume that both chips use the same power sources and the same reference clock, so that the signal noise sources are reduced in comparison to systems where the devices at each end of a channel are fully independent.

This interface would conventionally be electrical. It would, however, also be possible to use a combination of electrical and optical interfaces or even optical waveguides within the PCBA.

Advanced modulation formats PAM4 and beyond, FEC and equalization features are all possible solutions for the chip to module interconnect.

FEC may be a requirement to meet the BER – however the choice of the FEC must be considered carefully to address coding overhead, coding latency and power consumption concerns.

3.6 PCBA to PCBA across a Backplane/Midplane or a copper cable

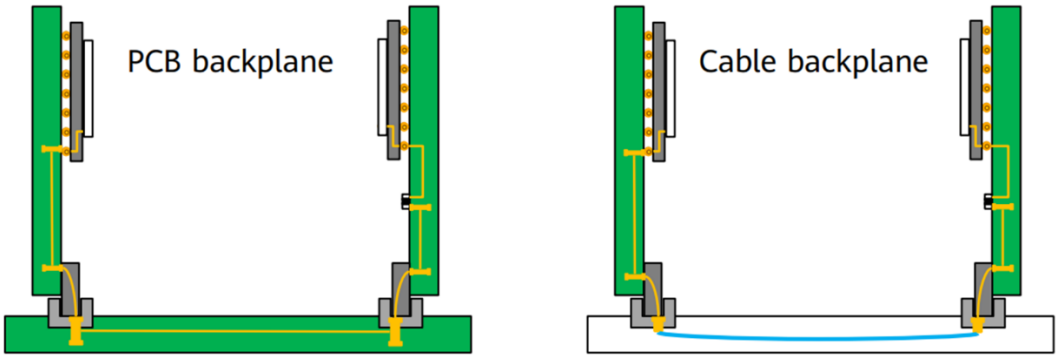


Figure 22 PCBA to PCBA across a Backplane or a Copper Cable Interconnect Application Space

This interface communicates between two cards across a backplane/midplane or a copper cable within a chassis and is up to 1 m combination of backplane and copper cable with up to 2 connectors.

This interface would conventionally be electrical. Due to the longer length channel, these interfaces would resemble the OIF's CEI type solutions. It would however also be possible to use an active copper cable, a combination of electrical and optical interfaces or even optical waveguides within the PCBA.

In addition, it may be appropriate to use advanced modulation formats PAM4 and beyond in the link allowing for increased throughput density at the same baud rate.

FEC may be a requirement to meet the BER – however the choice of the FEC must be considered carefully to address coding overhead, coding latency and power consumption concerns.

3.7 Chassis to Chassis within a Rack

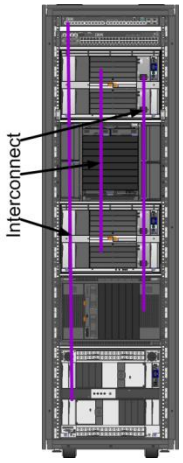


Figure 23 Chassis to Chassis within the Same Rack Interconnect Application Space

This interface ranges up to 3 m and could be either optical or active copper cables. Wider interfaces (i.e. optical multi fiber cable or parallel pair copper cables) could be analyzed for this application.

In addition, it may be appropriate to use advanced modulation formats PAM4 and beyond in the link allowing for increased throughput density at the same baud rate.

FEC may be a requirement to meet the BER – however the choice of the FEC must be considered carefully to address coding overhead, coding latency and power consumption concerns.

3.8 Rack to Rack side-by-side

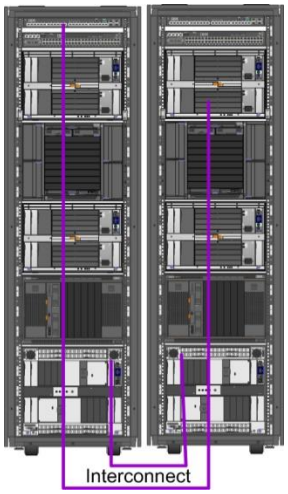


Figure 24 Rack to Rack side-by-side Interconnect Application Space

This interface ranges from 3 m to 10 m. Wider interfaces (i.e. optical multi fiber cable or parallel pair active copper cables) could be analyzed for this application.

It may be appropriate to use advanced modulation formats PAM4 and beyond in the link allowing for increased throughput density at the same baud rate.

FEC may be a requirement to meet the BER – however the choice of the FEC must be considered carefully to address coding overhead, coding latency and power consumption concerns.

Although there are links that are longer than 10 m, these are considered outside the scope of this document.

3.9 Longer links

Although there are links that are longer than rack to rack side-by-side, these are considered optical interfaces and outside the scope of this document.

3.10 Interconnect Application Summary

Table 4 Interconnect Applications

Intra Interconnect Application	Distance Up To	Types of interfaces
Die to Die in a Package	~50 mm	Electrical
Die to Optical Engine in a Package	~50 mm	Electrical
Chip to nearby optical Engine	~150 mm	Electrical
Chip to pluggable module	~200 mm	Electrical
Chip to chip within PCBA	50 cm	Electrical or Optical
PCBA to PCBA across a backplane/midplane/cable	~1 m	Electrical or Optical
Chassis to Chassis within a rack	~3 m	Electrical or Optical
Rack to Rack side-by-side	~10 m	Electrical or Optical

4 Points of Interoperability

The Optical Internetworking Forum (OIF) promotes the development and deployment of interoperable networking solutions and services through the creation of Implementation Agreements (IAs) for optical networking products. It is therefore important for any next generation interconnects to consider the interoperability points to be defined in the agreement. The IA must also develop realistic measurement techniques for the defined interoperability test points.

A next generation interconnect may be either electrical or optical. The possible interoperability points are shown in Figure 25.

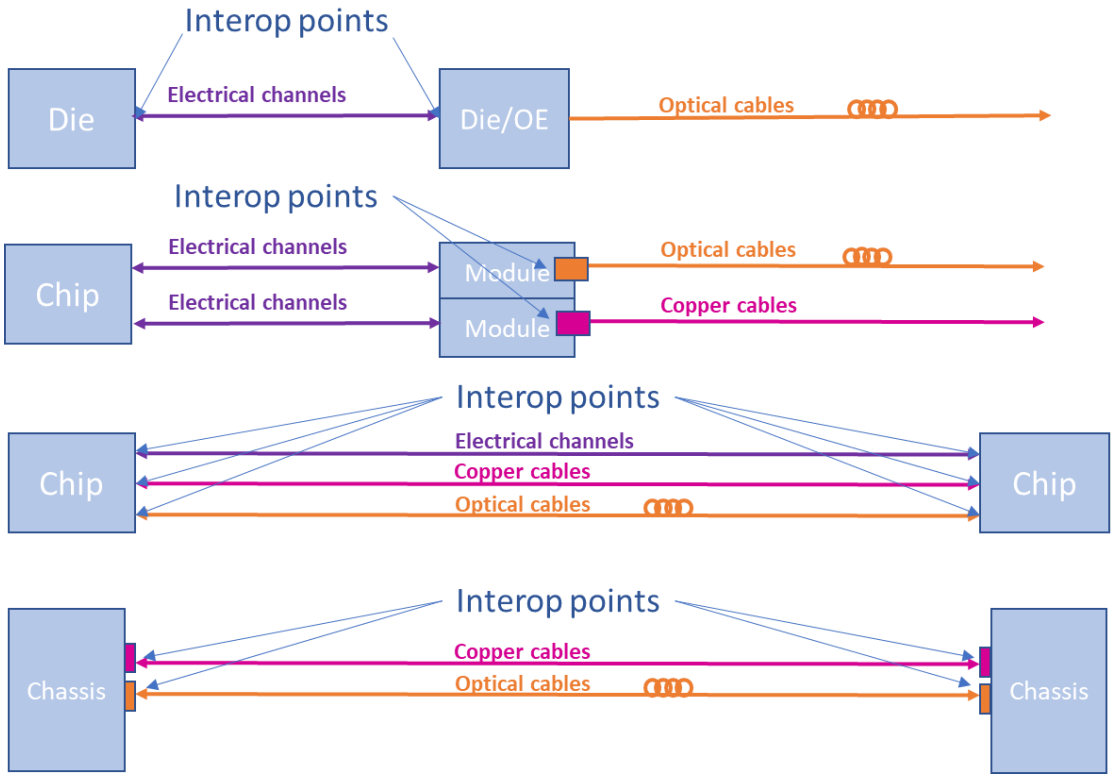


Figure 25 Interconnect Application Space Showing Points of Interoperability

The chip to chip interoperability points are best defined at the ball of the IC or packaged device. This allows chip makers to design directly to the specification and avoids the confusion of defining a load channel, which may not represent the real life system interconnect. The challenge with this method is the verification of compliance at a point that is not measureable in a real system.

The chip to module interconnect contains a separable connector at the faceplate of the host equipment. This provides a natural point to test for interoperability. Hence, the chip to module interoperability points are best defined at the host connector interface. This allows both the host

and module designers to verify their designs directly against the specification. The challenge is to specify a signal in the middle of a connector. A reference test board with the mating connector is required to provide measurement points that can be used by test equipment. The chip to module interoperability points can be measured at the end of a host or module compliance board. The compliance board method can be extended to higher signaling rates, noting that higher losses of both product channels and compliance board traces are to be expected.

5 Opportunities for Future Work

Section 3 identifies the many applications where next generation systems might benefit from an identified interconnect definition or "Implementation Agreement (IA)". All of these "interconnections" are possible areas for new projects within the OIF or other standards bodies. Section 2 identifies many additional specific areas that might be investigated for future OIF activities. Possible topics for future investigation include advanced system architectures, advanced modulations, lower loss interconnection systems, and so on.

6 Relation to Other Standards

These projects will potentially benefit from liaison activities with study groups and task forces and industry organizations including:

- ITU SG
- IEEE 802.3
- Fibre Channel
- InfiniBand
- Gen-Z
- JEDEC
- PCI-SIG
- CXL
- CCIX
- OpenCAPI

7 Summary

Service providers, network customers and data center operators have clearly communicated that higher data rates like 224 Gbps are required for client links to support higher data rates on the backbone networks. These next generation data rates need to be implemented while also addressing challenges associated with power dissipation, density, performance, reach and cost. In addition, compatibility with legacy data rates and networks will be required in many applications. These goals can be achieved by having consensus amongst a broad cross section of component, subsystem, and system suppliers to leverage new technologies that drive signaling, architecture, and integration developments. As has been demonstrated in the past, most recently at 112 Gbps, the OIF is proposing to play a key role in coordinating industry activity to identify and develop critical technical solutions that will enable next generation data rates to be cost effectively deployed in the development of next generation equipment and networks.