OIF CEI-56G Application Note

Common Electrical Interface at 56 Gb/s

Abstract:

The OIF is well along in the development of the CEI-56G suite of high speed interconnect Implementation Agreements (IA)s . The OIF is developing IAs for nine 56G clauses, spanning five reaches and three modulation techniques.

The OIF has a long legacy of developing IAs for serial electrical interfaces, with the CEI family being the principal industry definition for SerDes for the 6 Gb/s, 11 Gb/s and 28 Gb/s generations. Interfaces that leverage the CEI family of specifications have been built in most of the major ASIC flows and FPGA devices. CEI has been adopted or adapted into multiple interfaces by IEEE 802.3, InfiniBand and Fibre Channel.

About the OIF:

The OIF facilitates the development and deployment of interoperable networking solutions and services. Members collaborate to drive Implementation Agreements (IAs) and interoperability demonstrations to accelerate and maximize market adoption of advanced internetworking technologies. OIF work applies to optical and electrical interconnects, optical component and network processing technologies, and to network control and operations including software defined networks and network function virtualization. The OIF actively supports and extends the work of national and international standards bodies. Launched in 1998, the OIF is the only industry group uniting representatives from across the spectrum of networking, including many of the world's leading service providers, system vendors, component manufacturers, software and testing vendors. Information on the OIF can be found at http://www.oiforum.com.

For additional information contact:

The Optical Internetworking Forum, 48377 Fremont Blvd, Suite 117,

Fremont, CA 94538 USA

510-492-4040 info@oiforum.com www.oiforum.com

Table of Contents

Table of Figures

Glossary†

2.5D: Refers to a type of die-to-die integration via a silicon interposer having through-silicon vias (TSVs) connecting its top and bottom metal layers

3D: Refers to a three-dimensional (3D) integrated device in which two or more layers of active electronic components (e.g., integrated circuit dies) are integrated vertically into a single circuit where through-silicon vias (TSVs) are commonly used for die-to-die connection.

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.

BCH: Bose, Ray-Chaudhuri, Hocquenghem forward error correction (FEC) codes are a class of advanced 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.

BGA: Ball Grid Array, a package type

CDR: Clock and data recovery, a block in a receiver that recreates a clock or clocks in a receiver by looking at the statistical edge information in the received data and then uses that clock or clocks to sample the received data.

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

Clock Forwarding: A clock can be forwarded in parallel with the data to allow a transceiver to avoid using CDR in order to save power. This is typically done in USR applications.

EMB: Effective modal bandwidth, see TIA-492AAAD.

ENRZ: Ensemble Non Return to Zero, a multi-wire code in which 3 bits are modulated with the Hadamard Transform onto four wires.

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

FPGA: Field Programmable Gate Array – a reprogrammable logic device that often includes SerDes.

FR4: A grade designation assigned to glass-reinforced epoxy printed circuit boards (PCB).

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

GBd: The baud rate is the actual number of electrical transitions per second, also called symbol rate. One GigaBaud 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: Inter-Symbol Interference, the eye closure caused by the energy remaining on the channel from previous unit intervals, which typically impacts the horizontal eye opening.

ISI-Ratio: The ratio of the largest code eye to the smallest code eye, which is a simple metric to evaluate the impact of ISI on horizontal eye closure for a given code.

LR: Long Reach – long enough to reach a chip located across a backplane

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.

MUX/DEMUX: Multiplex / demultiplex, a multiplexer (or mux) is a device that selects one of several analog or digital input signals and forwards the selected input into a single line, Conversely, a demultiplexer (or demux) is a device taking a single input signal and selecting one of many data-output-lines, which is connected to the single input.

MR: Medium Reach – long enough to reach a chip located across a PCBA including on a daughter-card

NRZ: 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.

PAM: Pulse amplitude modulation, a form of signal modulation where the message information is encoded in the amplitude of a series of signal pulses.

PAM-4: Pulse amplitude modulation-4 is a two-bit modulation that will take two bits at a time and will map 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.

RS: Reed Solomon FEC coding, this 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: Serializer/Deserializer

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

USR: Ultra-Short Reach – just long enough to reach another die within the same package

VSR: Very-Short Reach – long enough to reach a transceiver in a front-panel optical module

XSR: eXtra-Short Reach –long enough to reach a nearby device or mid-board optical module

† Some definitions include content from www.wikipedia.com

The OIF CEI Legacy

The OIF has a long legacy of developing IAs for high speed electrical interfaces. The CEI family has been the principal industry definition for SerDes for the 6 Gb/s, 11 Gb/s and 28 Gb/s generations. The OIF SxI interface was CEI's immediate predecessor and served the 3 Gb/s generation. The OIF SPI/SFI 3 and 4 electrical interfaces served the 800 Mb/s and 1.6 Gb/s generations. This long history of 7 generations and six doublings gives the OIF a unique viewpoint on the needs of the industry for today and tomorrow.

Interfaces that support the CEI family of specifications have been built in most of the major ASIC flows and FPGA devices. CEI has been adopted, adapted, or influenced the development of multiple interfaces including IEEE 802.3, InfiniBand and Fibre Channel, SATA, SAS, RapidIO and HyperTransport interfaces as well as numerous proprietary protocols.

Name	Rate per pair	Year	Adopted, Adapted or Influenced
CEI-56G	56Gb/s	2016	The future is bright
CEI-28G	28 Gb/s	2011	InfiniBand EDR, 32GFC, SATA 3.2, 100GBASE-KR4 and 100GBASE-CR4, CAUI4, SAS-4
CEI-11G	11	2008	InfiniBand QDR, 10GBASE-KR, 10GFC, 16GFC, SAS-3, RapidIO v3
CEI-6G	6	2004	4GFC, 8GFC, InfiniBand DDR, SATA 3.0, SAS-2, RapidIO v2, HyperTransport 3.1

SxI5	3.125	2002-3	Interlaken, FC 2G, InfiniBand SDR, XAUI, 10GBASE-KX4, 10GBASE-CX4, SATA 2.0, SAS-1, RapidIO v1
SPI4, SFI4	1.6	2001-2	SPI-4.2, HyperTransport 1.03
SPI3, SFI3	0.800	2000	(from PL3)

Introduction

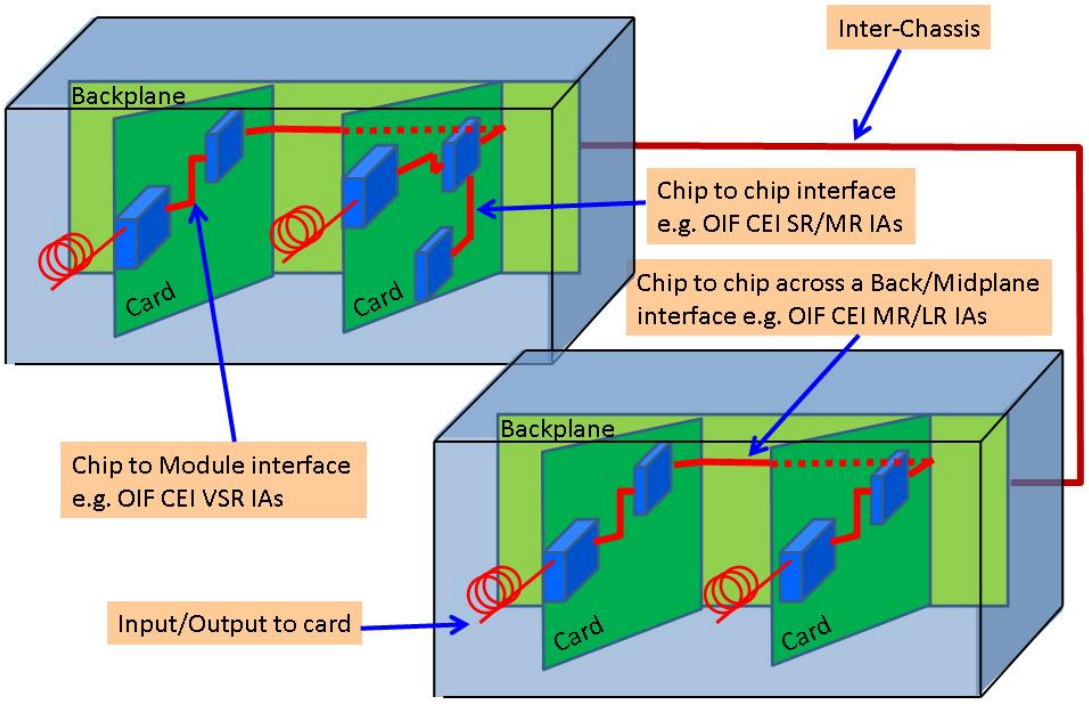

Figure 1 Interconnect Application Spaces

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

Considerations for these links may include:

- Cost
- Link Performance
- Power Consumption
- Channel loss budgets
- Choice of PCB material
- Modulation technique and signal levels
- Number of lanes, channel configuration and general characteristics
- Reference clock jitter
- Forwarded Clock or CDR
- Latency
- Connector performance
- Reliability
- Size
- Operating Temperature

Motivation for CEI-56G

Next generation systems are being driven by the need to handle the increasing volume of data traffic. At the same time, 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.

These needs drive next generation systems to ever increasing communication port densities. The increased density leads to smaller surface areas available to dissipate the heat generated and therefore requires decreased power consumption for a port.

The industry currently has electrical interfaces for 10Gb/s (OIF's CEI-11G), 25Gb/s & 28Gb/s (OIF's CEI-25/28G) in production, and work is coming along on 56Gb/s (OIF's CEI-56G). However, channels produced with copper PCB traces are severely bandwidth limited, and because of this it is increasingly difficult to achieve the same link distances using higher signaling rates.

Improvements in IC integration, which has been relentlessly driven by Moore's Law over past decades, have enabled higher densities of logic gates at escalating higher clock rates to

be used in IC designs. These trends have allowed the industry to deploy increasingly more complex communication systems at each generation to meet the infrastructure needs.

However, as one digs deeper, the future appears to be challenging. Many different technologies need to converge to improve the throughput. These complex ICs have increasing gate counts, but the power dissipation per gate and I/O speed have no longer scaled at the same rate as the gate count. In addition, the numbers of electrical connections (bumps and package pins) are also not scaling at the same rate - leading to a power, capacity and port count gap for the next generation interconnect interfaces.

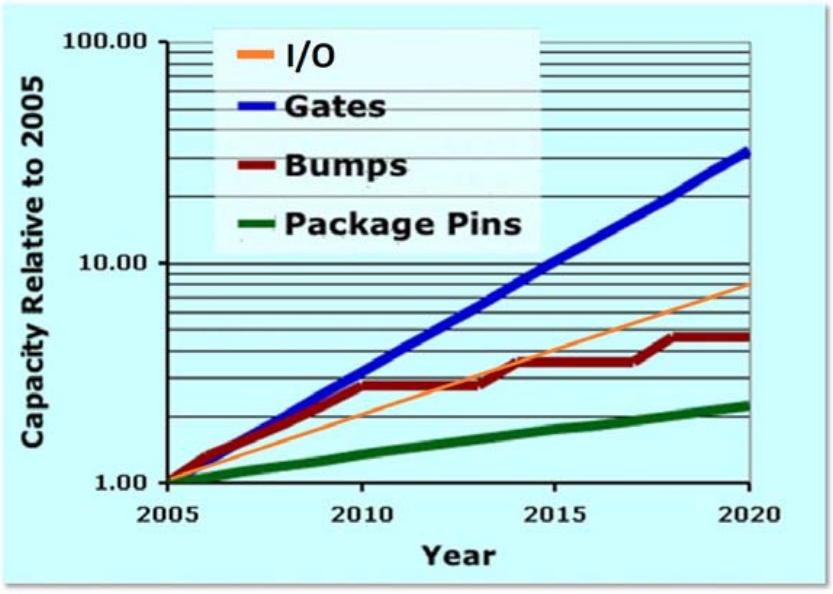

Figure 2 Scaling of Gates, Bumps, Pins and I/O (Ref. Xilinx, I/O Line added)

The predominant interconnect challenges to overcome are presented below in a solution space diagram, and are discussed in greater detail in subsequent sub-sections.

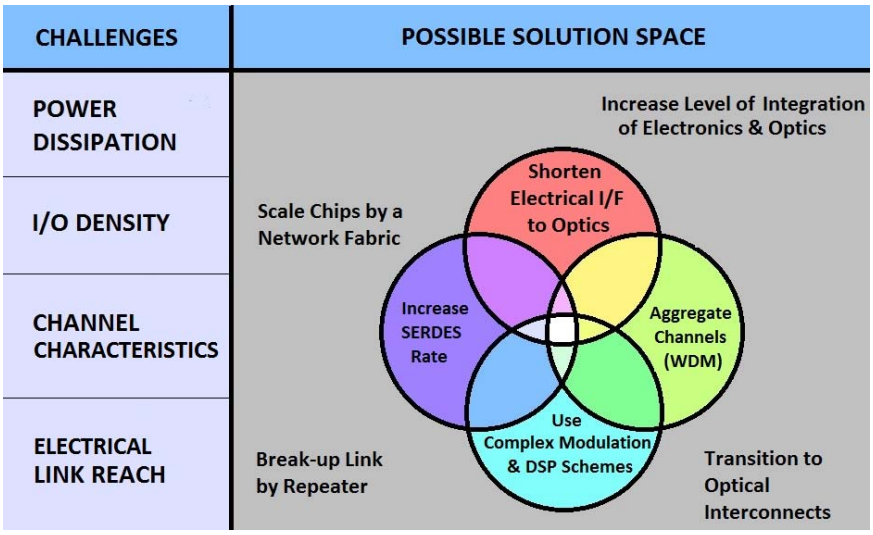

Figure 3 Interconnect Challenges

CEI-56G Challenges

Die Area

SerDes implementing CEI-56G are expected to be larger than SerDes supporting CEI-28G interfaces because of the need for more advanced equalization. Additionally the size of the analog portion of the interfaces do not scale as well as the digital portions.

This area situation has rendered some previous chip architectures unuseable. This has led to a trend of SerDes "out-boarding" where an Ultra-Short Reach (USR) or eXtra Short Reach (XSR) interface is put on the main ASIC and the longer reach SerDes are put in the same package or nearby.

Chip Power Dissipation

SerDes implementing CEI-56G are expected to be higher power than SerDes supporting CEI-28G interfaces because of the need for more advanced equalization. Additionally the power dissipation of the analog portion of the interfaces do not scale as well as the digital portions.

This has furthered the trend mentioned above of SerDes "out-boarding" where an Ultra-Short Reach (USR) interface is put on the main ASIC and the longer reach SerDes are put in the same package.

Package Power Dissipation

When package power is the key constraint, the use of an XSR interface can help move some of the power dissipation outside of the main ASIC and onto other chips nearby on the PCBA.

Chip Power Wiring

Lower voltage power sources are used with smaller technology nodes. The net effect of higher integration of low voltage semiconductors is a significant increase in total device current, which requires high current source power supplies which must be controlled within several mV tolerances, which in turn requires additional power pins per device to accommodate the high electrical currents.

System Power Dissipation – Mid-planes

Because system cooling can escalate beyond physical limits, multiple power saving strategies may be needed.

Since the I/O power is related to the distance that the electrical signals travel for a given channel's properties, reducing the distance that the electrical signal must be driven can reduce power dissipation. One technique that has been used is to use mid-planes.

With this physical architecture, a switch board is mounted horizontally behind the vertically mounted front boards. With this architecture, the electrical links are shorter and only have to traverse one connector.

System Power Dissipation – Mid-board optics

In some cases, it may make sense to integrate the optics close to the ASIC package, thereby minimizing the need to drive electrical signals far. These are sometimes called optical engines and the technique is referred to as using mid-board optics.

I/O Densities on Chips and Connectors

The maximum number of useful I/Os for high speed serial links per device is not only limited by the available package technology itself, but also by the ability to route the device on the PCBA.

In order to maintain signal integrity for a high speed serial link design, it is required to be able to route a differential pair between two package balls when escaping from the inner ball rows of a ball grid array (BGA), and it therefore may become more costly to use packages with a ball pitch below 1.0 mm.

In addition, for every ball row from the edge of the package on which differential pairs have been placed, a separate circuit package layer has to be used. Differential pairs in the outer 4 rows of the BGA requires 4 signal layers on the PCBA, while the 6 outer rows would require 6 layers, and thus the PCB layer stack grows with every inner BGA row to be routed.

Channel Materials and Characteristics

Link applications are characterized by the supported loss budget and signal impairments. Loss is determined by link length, board materials, back-drilling, via type usage, the number of connectors, connector types, and package materials. Increases in signaling rate cause more frequency-dependent loss and thus a lower Signal to Noise Ratio (SNR).

The deviation of the loss from a smooth curve is also important as the ripples in a loss curve are not easily equalizable. This is called Insertion Loss Deviation (ILD).

In green field applications, advanced materials for the channel (PCB and connector) can result in an decreased loss and can thus support either an increased electrical data rate or a decreased amount of required equalization.

Channel Reach

In communication systems usually the front plate area is occupied by pluggable modules for inter-system communications, while the intra-system traffic between the PCBAs is connected over the backplane/midplane. To support reasonable system dimensions, system backplane/midplane connections typically need to bridge distances of up to 70–100 cm. In some more recent applications, alternative channel architectures replace the backplane with direct plug orthogonal (DPO) structures where I/O line cards on the front side of the equipment directly plug into switch fabric linecards on the rear side of the equipment, thus directly connecting the linecards without the loss of the backplane between them

The introduction of repeater devices into the data path can increase the effective reach at the expense of power and board area.

Cables

Twinax, micro-coax and/or flex circuits are also an option to reduce loss and extend reach for the high speed links, particularly in systems with fewer high speed links. In some circumstances, this can allow the main PCBA to be built from low cost material. In another case cable assemblies are now being integrated onto backplane connector and used in lieu of PCB backplanes due to their ability to reduce loss and extend reach. The use of cables brings their own share of challenges.

Latency

CPU-to-CPU and CPU-to-Memory applications of CEI are often intolerant of latency. This often precludes the use of heavyweight FEC in these applications. The characteristic method that is intolerant of latency is the use of credit-based flow control where a receiver grants the transmitter the right to send a certain amount of data. Only a moderate amount of buffering is typically provided in these systems, most often in more costly SRAM. Most of the major CPU to CPU protocols use credit-based flow control.

Summary of challenges

As time proceeds, ICs will become faster and denser. To cope with the issue of interconnect capacity and density of future systems, photonic interconnects will become an even more important connection technology.

The implementation of 56 Gb/s Interconnects technology poses several challenges especially in relation to: die area, chip power dissipation, package power dissipation, system power dissipation, limited I/O density, channel loss, channel reach, and latency. Highlighted were the side-effects of some solutions, which are a result of the complex inter-

dependencies of: higher integration, complex modulation schemes, chip break-out and routing, signal conditioning, thermal & power issues, and package footprint.

Modulation Techniques

Advanced modulation techniques can be used to lower the Nyquist frequency requirement for a link using a bandwidth limited electrical channel. These techniques require a higher SNR for equivalent BER. The provision of channels with more correlations to measure against can also help. Forward error correction (FEC) can increase the effective SNR and therefore the supported loss budget at the expense of higher power dissipation, complexity, and latency. Bandwidth limitations and increases in signaling rate result in impairments that may be compensated by using equalization techniques, which themselves also impact power consumption and complexity.

NRZ

NRZ (Non Return to Zero) is the modulation technique used by most current speed links. It requires a channel with two correlated conductors. It uses two signal levels and conveys one bit per baud. To support 56Gb/s per pair, NRZ runs at a symbol rate of 56 GBaud and has a Nyquist frequency of 28 GHz. It has an ISI-Ratio of one, meaning the ratio of the largest eye to the smallest eye is always the same.

At 56 Gb/s, NRZ is useful in applications where the channel's frequency dependent loss is not excessive. That loss has been found to be excessive in many long reach applications.

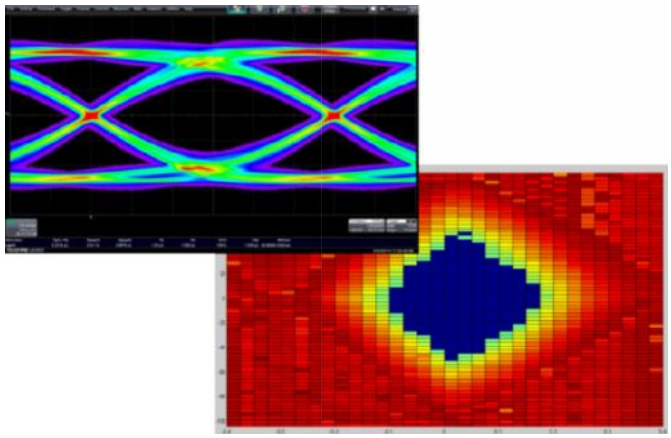

Figure 4 Typical NRZ Eye Diagram and Statistical Eye Plot

ENRZ

Ensemble NRZ (ENRZ) requires a channel with four correlated conductors. Useful channels can be constructed with two correlated loosely coupled pairs. It uses four signal levels and conveys three bits per baud over the four conductors by modulating the three sub-channels collectively using the Hadamard matrix. It To support a 56Gb/s per pair effective rate, ENRZ runs at a symbol rate of

37.3 GBaud and has a Nyquist frequency of 18.6 GHz. It has an ISI-Ratio of one, meaning the ratio of the largest eye to the smallest eye is always the same. ENRZ can often be used without Forward Error Correction, even on some long reach channels.

At a 56 Gb/s effective rate, ENRZ is useful in longer reach applications, particularly when the use of FEC is not desired.

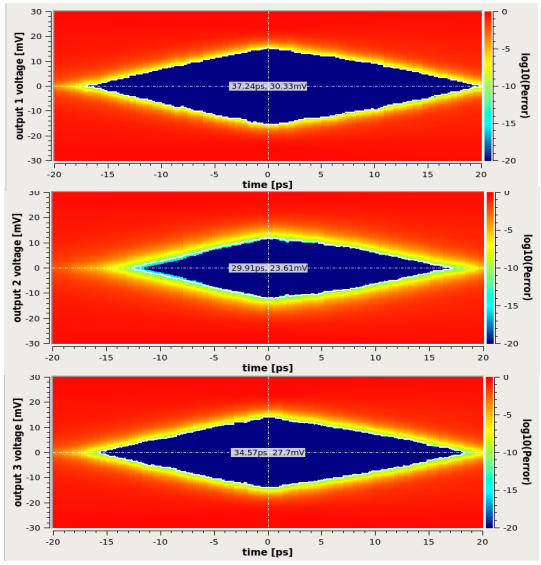

Figure 5 ENRZ Typical Statistical Eyes

PAM-4

PAM-4 requires a channel with two correlated conductors. It uses four signal levels and conveys two bits per baud. To support 56Gb/s per pair, PAM-4 runs at a symbol rate of 28 GBaud and has a Nyquist frequency of 14 GHz. It has an Intersymbol Interference Ratio (ISI-Ratio) of three, meaning the largest code eye is three times as large as the smallest code eye. It has a vertical impairment of 9 dB as compared to NRZ and has a horizontal impairment due to its high ISI-Ratio. PAM-4 is usually paired with equalization and Forward Error Correction.

At 56 Gb/s PAM-4 is useful in applications where the channel's frequency dependent loss would otherwise be excessive.

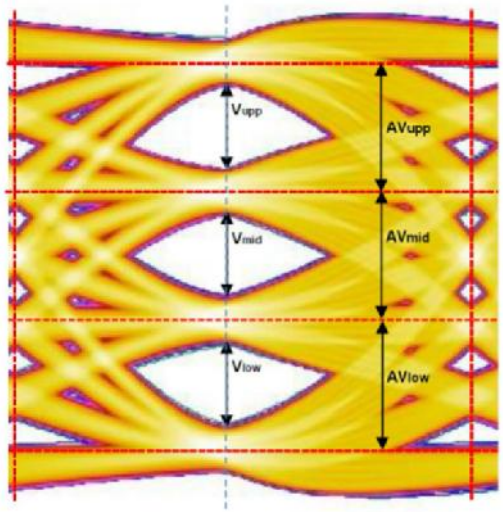

Figure 6 Typical PAM-4 Eyes

Interconnect Reaches and Application Spaces

The five CEI-56G reaches can be summarized in the following table:

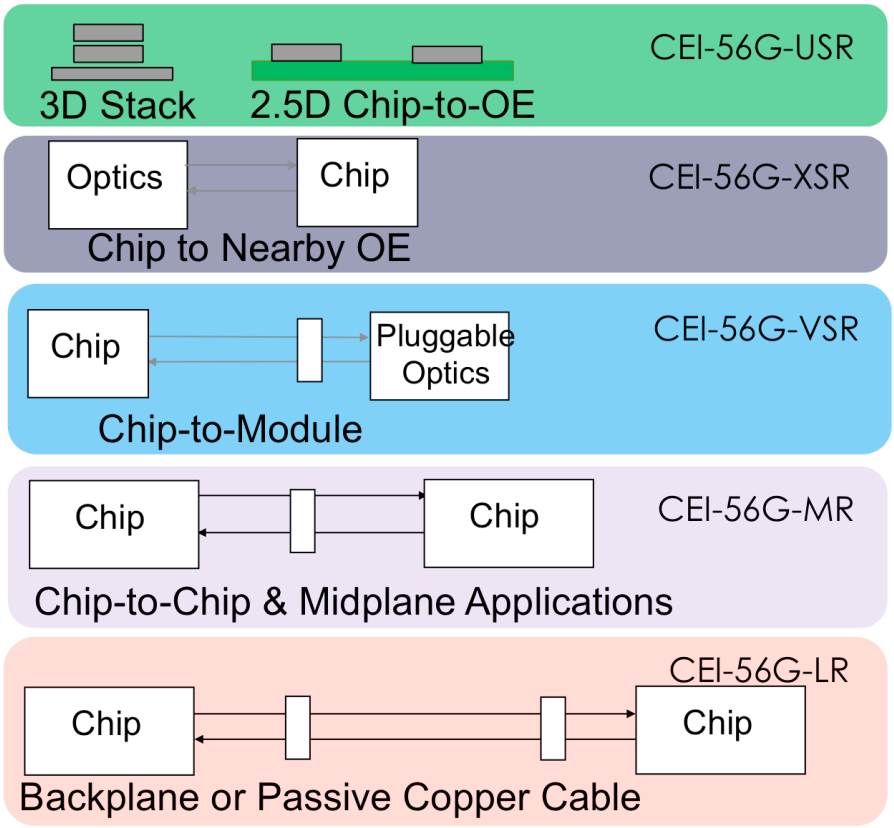

Figure 7 CEI-56G Reaches

The use of these reaches for interconnect application spaces can be broken down as follows.

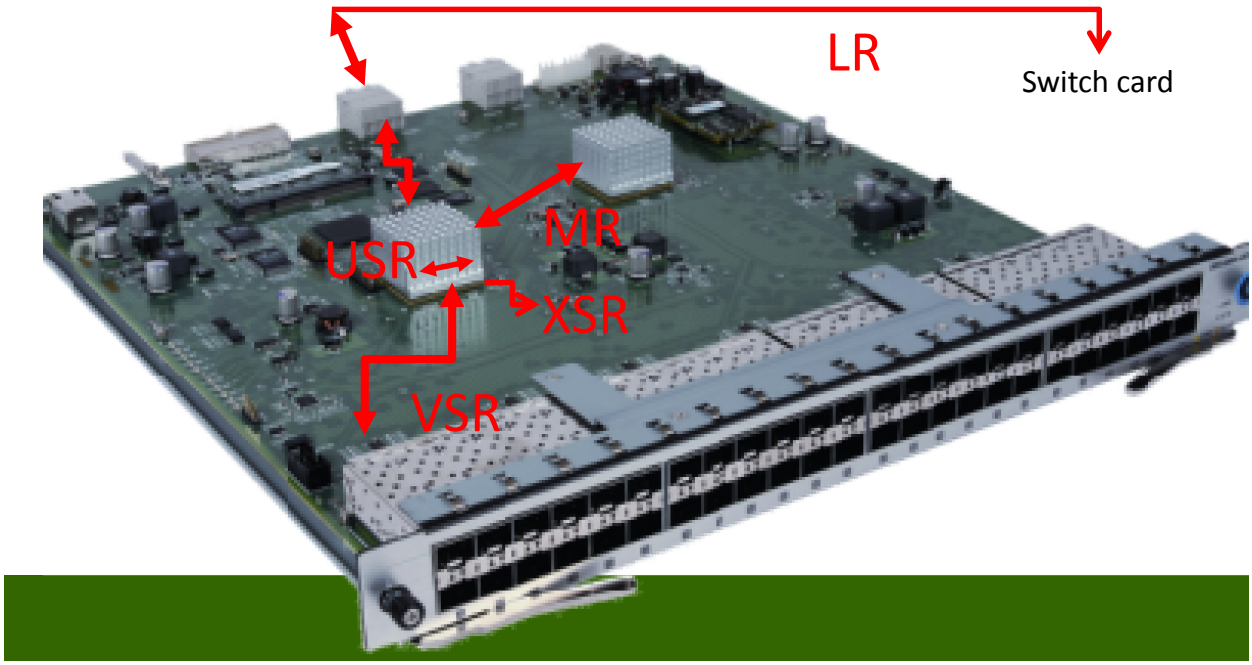

Figure 8 Use of CEI-56G Reaches

USR - Die to Die Interconnect Within A Package

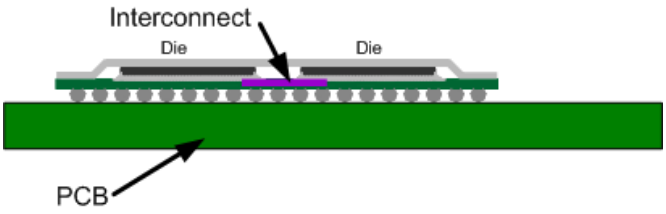

Figure 9 USR - Die to Die

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

The communication channel is less than 10mm. This short electrical link may allow for a much simpler interface and require less power than an existing standard electrical interface. 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.

A typical use for this is to off-board the SerDes from a switch ASIC.

Clock forwarding can be used in the place of CDR for USR links.

USR - Die to Optical Engine within a Package

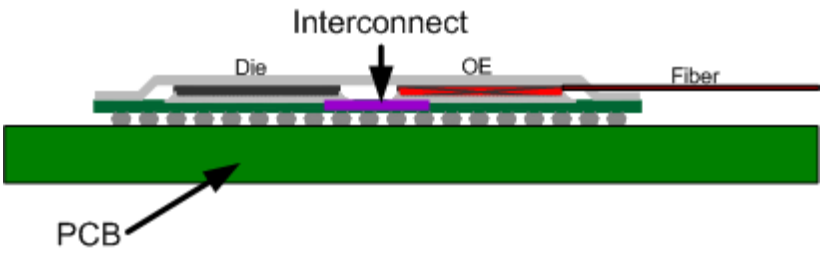

Figure 10 USR - Die to Optical Engine

It may be necessary to use a die and an optical engine 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 10mm. This short electrical link may allow for a much simpler interface and require less power than an existing standard electrical interface.

XSR - Chip to Nearby Optical Engine

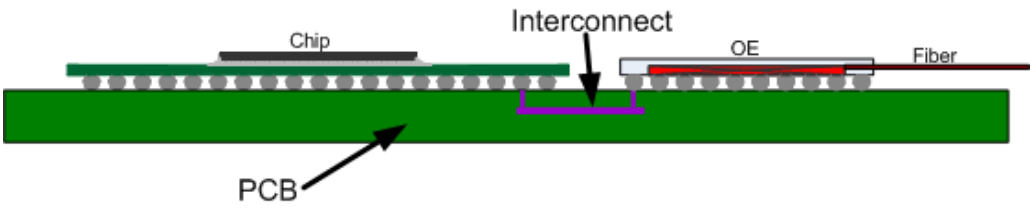

Figure 11 XSR Chip to Optical Engine

It may be useful to place an optical interface very close to the host chip. Some optical devices cannot sit within a host MCM due to heat restrictions of the optical components. In this case, a short electrical link of less than 50mm is anticipated. The short reach of this channel allows power to be saved. These optical modules are often called mid-board optics.

XSR - CPU to CPU

CPUs can be connected via a short connection over high rate SerDes. Some CPUs use a coherency protocol to make the different processors appear to be in the same coherency domain.

It is generally not possible to use heavy-weight Forward Error Correction in CPU applications due to the latency requirements of these interfaces, which typically use credit-based flow control.

XSR – DSP Arrays

DSPs are sometimes connected in arrays to process high rate information such as from a RADAR or a LIDAR. An autonomous vehicle is an example of such an application. Sometimes the interfaces are unidirectional and other times, bidirectional. Both are typically latency sensitive similar to CPUs as even the unidirectional applications often have complex, time sensitive control flows.

XSR - CPU to Memory Stack

CPUs can be connected via a short connection to a memory stack. A classic way to accomplish this is to have a device made with a logic process that forms the base layer device for the actually memory dies.

It is not possible to use heavy-weight Forward Error Correction in this application due to the latency requirements of memory interfaces. CPU threads or state machines are often waiting for the results of the memory access. Memory caches help, but do not eliminate the significant performance hits that CPUs see when latency is added.

VSR - Chip to Module

It is common in modern communication systems to support pluggable modules at the front faceplate of the equipment. This facilitates low cost initial deployment of the equipment if some ports are left unpopulated. A pay as you go policy is then used until the entire faceplate is populated with pluggable modules. The electrical link used to connect these pluggable modules can extend to beyond 30cm. At higher data rates this challenges the ability of the host chip to drive these long trace lengths within the power constraints of large switch chips. 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. Advanced modulation formats (such as PAM or DMT schemes), Forward Error Correction (FEC) and equalization features are all possible solutions for the chip to module interconnect.

One technique that has been used is to engineer the electrical links to have a lower native BER than that of the optical link. In this way, a single FEC can be used to protect a CEI-56G link at each end and an optical link in the middle.

MR - Chip to Chip within a PCBA

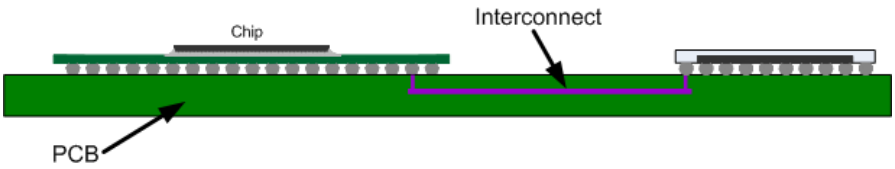

Figure 12 MR - Chip to Chip

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 up to 50cm. This interface could include a single connector.

LR – Chip to Chip across a Backplane/Midplane

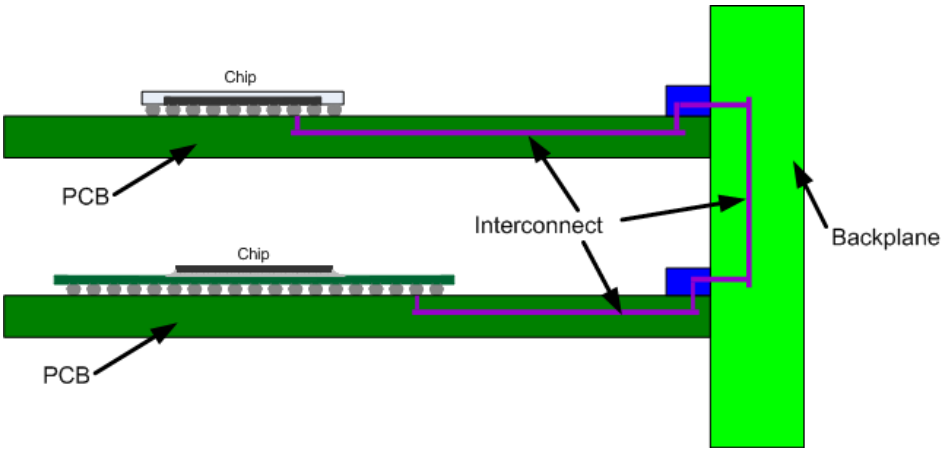

Figure 13 LR Chip to Chip across a backplane

This interface communicates between two cards across a backplane/midplane within a chassis and is less than 1m with up to 2 connectors.

FEC may be a requirement to meet the BER – however the choice of the FEC must be considered carefully to address both latency and power concerns. Possible FEC implementations include RS or BCH.

LR – Chip to Chip across a Cable

While not an explicit goal of the CEI-56G project, cabled versions of CEI-28G have been used to support both backplane and chassis to chassis interconnects.

Interoperability Points

CEI-56G defines three different interoperability points for the various reaches. These are the points at which compliance the the OIF IAs can be determined by measuring the properties of an implementation and comparing those results to the parameters in the IA.

Die Bump

CEI-56G-USR defines interoperability at the die bump of a flip-chip semiconductor device

Package Ball

CEI-56G-XSR, CEI-56G-MR and CEI-56G-LR all defines interoperability at the package ball.

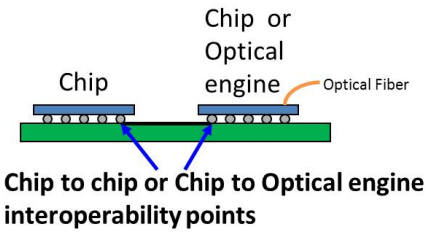

Figure 14 Package Bump Interoperability Points

OIF chose package balls as the interoperability points for these three specifications because the entire channel is typically the responsibility of the same company. XSR does not have a connector and therefore no other interoperability point is available. For MR and LR, which do have connectors, the system designer still owns the entire channel and therefore it does not make sense to constrain how the system designer allocates the link budget.

Module Interconnect

CEI-56G-VSR defines interoperability at the module interconnect.

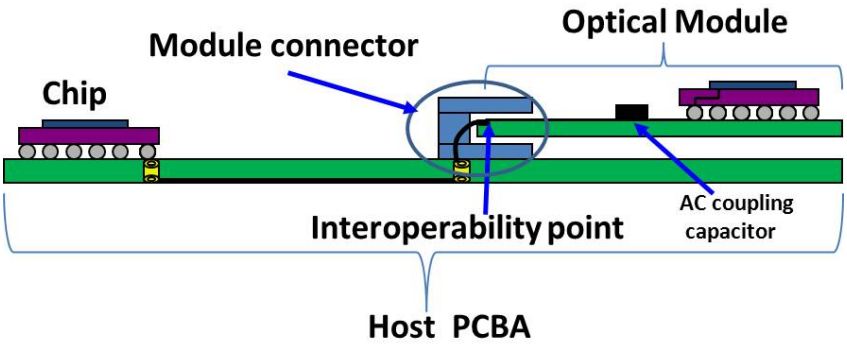

Figure 15 Chip to Module Interoperability Point

Host compliance boards (HCB) and module compliance boards (MCB) are defined to measure compliance.

Summary

The CEI-56G family of clauses defines a set of interfaces that can be used to solve most high speed interconnect needs. Five separate reaches and three separate modulation techniques can support the needs of most applications.