Energy Efficient Interfaces

OIF Interoperability Demo
ECOC 2024

Copyright © 2024 OIF

Energy Efficient Interfaces @ ECOC 2024

- Energy Efficient Interfaces (EEI)

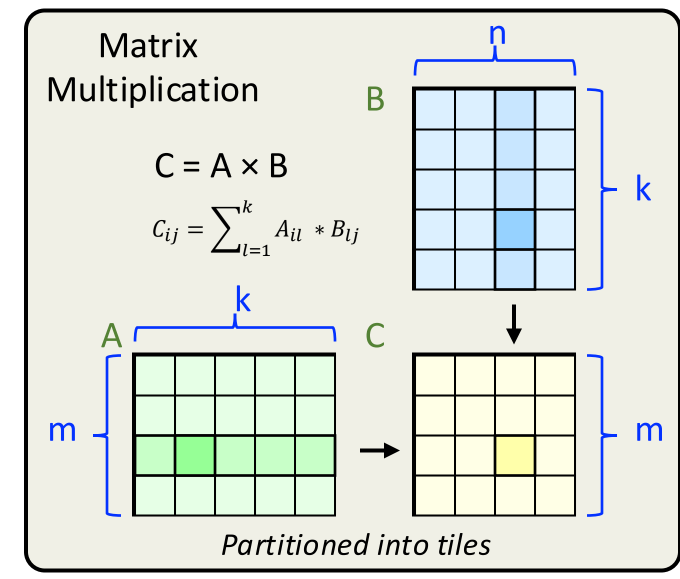
- EEI Interoperability agreements
  - Co-Packaging Framework Document
  - 3.2T Optical Module for Co-Packaging Project
  - ELSFP Project
  - Electrical Interfaces for Co-Packaging

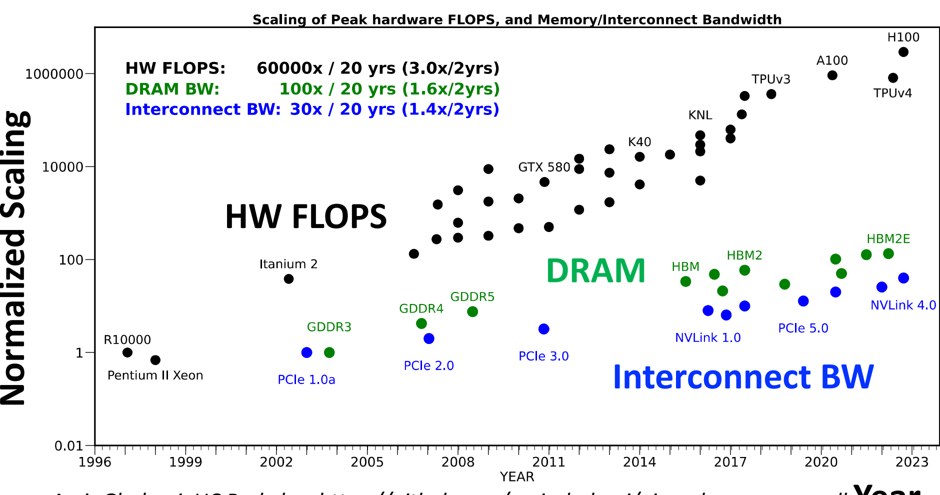
- Interoperability Demonstrations

Copyright @ 2024 OIF

The Challenge!

- AI training utilizes large quantities of matrix multiplication
  - GPUs are designed to accelerate “multiply and add” operations used in AI matrix multiplication
  - Each row in matrix A is paired with every column in matrix B – Lots of computation with lots of parameters!
- Large AI models can partition the computation into smaller chunks
  - Tile computations can be handed off to clusters of local and remote compute accelerators
  - However, the completion of a tile in matrix C must wait for all contributing tiles to complete
- The time to complete the computation depends on:
  - Computation speed
  - Interconnect bandwidth
  - Memory speed

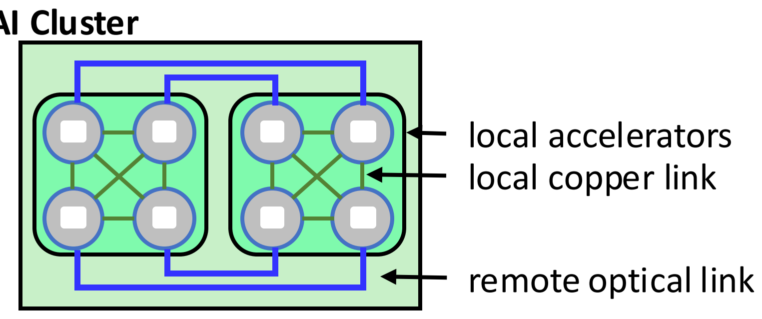


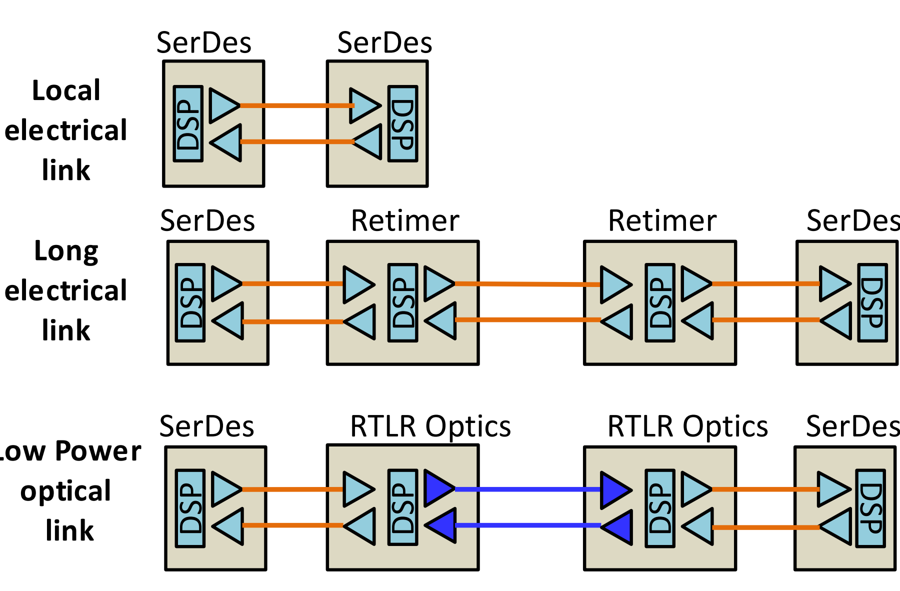


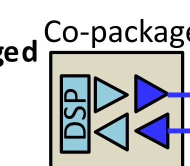
AI scale-up will drive adoption of optical chiplets to achieve lower latency, energy efficient, & cost-effective interconnections to support large AI models

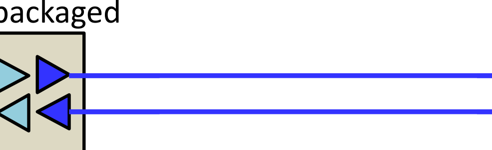
What approaches can we use?

- What is needed?
  - Larger local clusters interconnected with short links
  - Energy efficient, high-speed, low latency, dense interconnects that can scale
- Copper Links
  - Copper is ideal for local connections
  - As the data rates increase, pure electrical link reach becomes shorter
  - Reach can be extended with additional DSP capability and/or with the addition of retimers, trading off additional latency and power
- Optical Links
  - With some addition of E-O power, the electrical signaling can be converted to optical and then travel without needing additional retimers to restore the signal
  - If the electro-optical conversion is co-packaged with the ASIC, additional power and latency can be saved









AI scale-up will drive adoption of optical chiplets to achieve lower latency, energy efficient, & cost-effective interconnections to support large AI models

Copyright @ 2024 OIF

Applications for energy efficient links for AI

AI Compute Architecture Example

Front-end Network

FE Switch / FE Switch / FE Switch

FE Compute

CPU, M, A, Local Acc Switch

Remote Acc Switch / Remote Acc Switch

Back-end Compute

Callouts 1, 2, 3, 4, 5

Legend:
- Ethernet
- Acc local (NVLink)
- Acc remote (InfiniBand)
- Memory I/F (CXL/PCIe)
- HBM I/F (Very wide I/F)
- Local connection (D2D)

The OIF worked with end-users to summarize their application requirements

BACKEND COMPUTE INTERFACES

Compute type I/Fs *(PCIe/CXL like interfaces)*

Network Type I/Fs

	1 Compute to I/O	2 Compute to Memory	3 Compute to Local Compute	4 Compute to Remote Compute
Description	PCIe I/O	Disaggregated memory	Inter-connecting local clusters of accelerators	Inter-connecting remote clusters of accelerators
Interfaces	PCIe	HBM DDR / LPDDR CXL.mem	NVLink UALink xGMI CXL.cache	InfiniBand Ethernet UEC

Requirements (1–4):
- ❑ Low Latency (< 5ns + TOF, EOE & E2E)
- ❑ Energy Efficient (< 5 pJ/b)
- ❑ Dense Connectivity (1 → 2 T/mm)
- ❑ High Degree of Link Accountability

FRONT END

Network Type I/Fs

5 Cloud DC

Front-End w/ TCP/IP, UDP, etc.

Ethernet (200GbE)

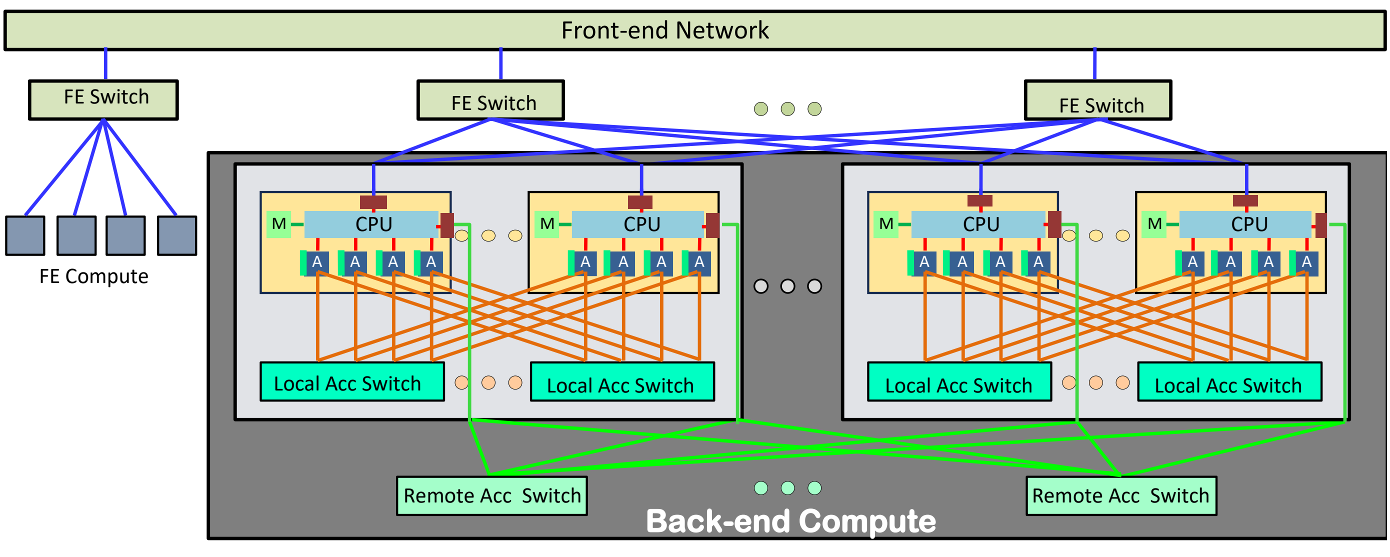
- ❑ Standards based
- ❑ Energy efficient
- ❑ Potentially high density for some

Ethernet (200GbE)

What kinds of Energy Efficient links are needed?

- Energy efficient, high-speed, low latency interconnects
  - PCIE-like
    - Memory
    - Local Accelerator interconnect
    - I/O
  - Low latency Ethernet
    - Remote accelerator interconnect
  - Ethernet

AI Compute Architecture Example



Back-end compute can leverage co-packaged solutions using optical chiplets for dense, energy efficient, low latency interconnections

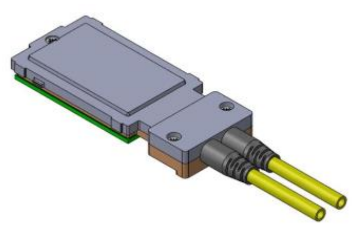
Copyright @ 2024 OIF

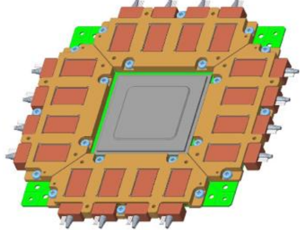
What is the OIF doing?

OIF's Co-Packaging Projects

✓ Co-packaging Framework Project

OIF-Co-Packaging-FD-01.0 – Co-Packaging Framework Document



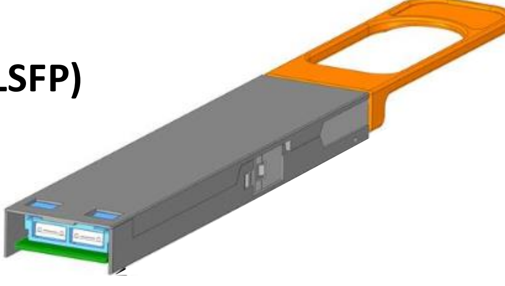


✓ 3.2T Co-packaged Optical Engine

OIF-Co-Packaging-3.2T-Module-01.0 – Implementation Agreement for a 3.2Tb/s Co-Packaged (CPO) Module

✓ External Laser Source (ELSFP)

External Laser Small Form Factor Pluggable (ELSFP) Implementation Agreement (August 2023)



✓ Management Interface for ELSFP

OIF-ELSFP-CMIS-01.0 – Implementation Agreement for External Laser Small Form Factor Pluggable (ELSFP) CMIS

Energy Efficient Interfaces for AI

System Vendor Requirements Document for Energy Efficient Interfaces

- Document the EEI requirements as provided by the end-users for AI/ML optical and electrical links

Energy Efficient Interface Framework

- Study and initiate new standards for dense, low power, low latency links for AI/ML

RTLR Project (Retimed Transmitter, Linear Receiver)

- Address lower latency and low power applications utilizing transmit retimed optical transceivers (e.g. Ethernet, UEC, etc.)

COI Project (Compute Optics Interface)

- Address energy efficient, low latency photonic interfaces for transport of traffic for AI scale-up applications (e.g. PCIe, NVLink, UALink, etc.)

Copyright @ 2024 OIF

OIF Projects Addressing Next Gen AI Compute Interfaces

AI Compute Architecture Example

Front-end Network
FE Switch
FE Switch
FE Switch
FE Compute
CPU
Local Acc Switch
Remote Acc Switch
Back-end Compute

- Ethernet
- Acc local (NVLink)
- Acc remote (InfiniBand)
- Memory I/F (CXL/PCIe)
- HBM I/F (Very wide I/F)
- Local connection (D2D)

BACKEND COMPUTE INTERFACES				FRONT END
Compute type I/Fs *(Scale-Up)*			Network Type I/Fs *(Scale-Out)*	Network Type I/Fs
Compute to I/O	Compute to Memory	Compute to Local Compute	Compute to Remote Compute	Cloud DC
PCIe I/O	Disaggregated memory	Inter-connecting local clusters of accelerators	Inter-connecting remote clusters of accelerators	*Front-End w/ TCP/IP, UDP, etc.* Ethernet (200GbE)
1	2	3	4	5

COI Project (Compute Optics Interface)

- Address energy efficient, low latency photonic interfaces for transport of traffic for AI scale-up applications (e.g. PCIe, NVLink, UALink, etc.)

RTLR Project (Retimed Tx, Linear Rx)

- Address lower latency and low power applications utilizing transmit retimed optical transceivers (e.g. Ethernet, UEC, etc.)

Copyright @ 2024 OIF

Adapted from oif2023.269, Ram Huggahalli, Microsoft

Energy Efficient Interfaces @ ECOC 2024

- Energy Efficient Interfaces (EEI)

- EEI Interoperability agreements
  - Co-Packaging Framework Document
  - 3.2T Optical Module for Co-Packaging Project
  - ELSFP Project
  - Electrical Interfaces for Co-Packaging

- Interoperability Demonstrations

Copyright @ 2024 OIF

Co-Packaging Application Spaces
Framework Project

Data-Center Networking | AI Training / ML | Disaggregation

End Node
Sparse →
Compute End-Point

DRAM
Server Chip
PCIe
NIC
CPO
CPE
HBM
NPU/GPU
Server CPU Package
Local Memory
Compute Fabric

Optical Engine
Electrical Engine

Switching Node
Dense →
Switches

Ethernet Switch ASIC
High Perf. Switch ASIC
Switch
MC
Remote Memory
Ethernet

Low Latency I/F
Ethernet I/F

100 – 800Gb/s	Bandwidth	High (Multi-Tbps per compute node)
µs - ms	Latency	Low (<200ns RT memory latency)
10 – 50 FITs	Reliability	Very High (< 1 FIT)

Source: OIF CPO Webinar, Ram Huggahalli 10-14-2020

Copyright @ 2024 OIF

Co-Packaging Architectures (1)

Framework Project

Co-Packaged using socket for engine

Copper Cables Or Fibers

OE or EE

Engine Substrate

soldered

ASIC

socketed

Front Panel

Co-Packaged Assembly Substrate

Host PCB

Co-Packaged with soldered engine

Copper Cables Or Fibers

OE or EE

Engine Substrate

soldered

ASIC

soldered

Front Panel

Co-Packaged Assembly Substrate

Host PCB

Copyright @ 2024 OIF

Co-Packaging Architectures (2)

Framework Project

Co-Packaged using copper cable assembly

Copper Cable

or

To Front Panel or Backplane

Copper Cables Or Fibers

soldered

ASIC

OE or EE

Co-Packaged Assembly Substrate

Engine Substrate / RDL

Front Panel

Host PCB

Near-Packaged using socketed engine

Copper Cables Or Fibers

ASIC

OE or EE

ASIC Substrate / RDL

Engine Substrate

socketed

socketed

Front Panel

Co-Packaged Assembly Substrate

Host PCB

Copyright @ 2024 OIF

Interfaces Studied for Interoperability

Framework Project

Thermal, Mechanical & Environmental

Copper Cables Or Fibers

Electrical

OE or EE

ASIC

Engine Substrate

soldered

soldered

Co-Packaged Assembly Substrate

MIS

Power

Socket

Optical (Electrical) & RLS

Front Panel

Host PCB

Application Example

- **Switch Generation:** 51.2Tb/s
- **Lane Speed:** 106 Gb/s
- **Interface Architecture:** XSR based AUI, 400G-FR4 PMD
- **Motivation:** System power reduction, ecosystem & operational readiness

Reliability and Repairability

Copyright @ 2024 OIF

Energy Efficient Interfaces @ ECOC 2024

- Energy Efficient Interfaces (EEI)

- EEI Interoperability agreements
  - Co-Packaging Framework Document
  - 3.2T Optical Module for Co-Packaging Project
  - ELSFP Project
  - Electrical Interfaces for Co-Packaging

- Interoperability Demonstrations

Copyright @ 2024 OIF

Example System Attachment

3.2T Optical Module

- 16 x 3.2T Modules = 51.2T Switch Capacity

Co-Packaged Assembly Substrate (Interposer)

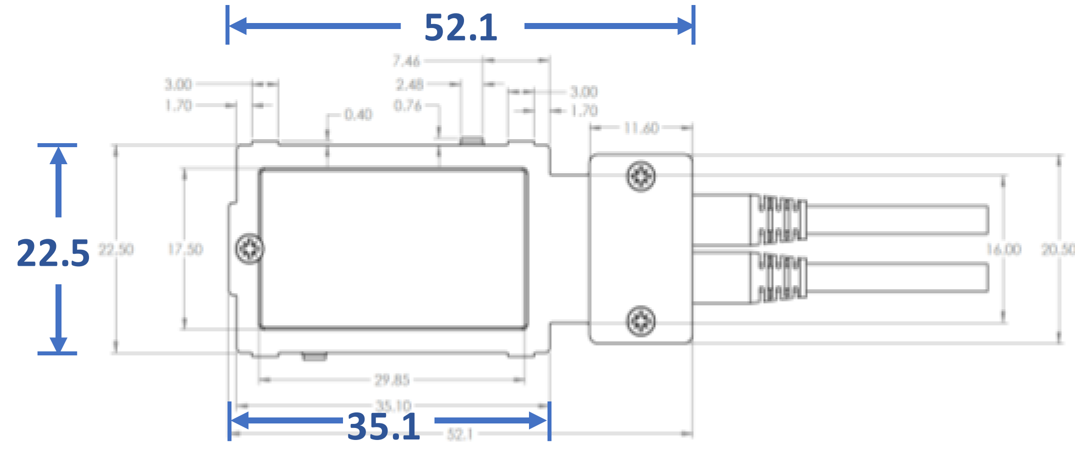
Channel components cross-section

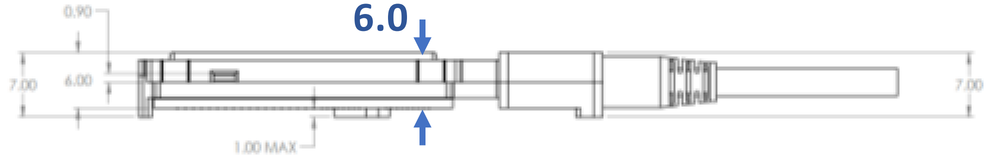
Copyright @ 2024 OIF

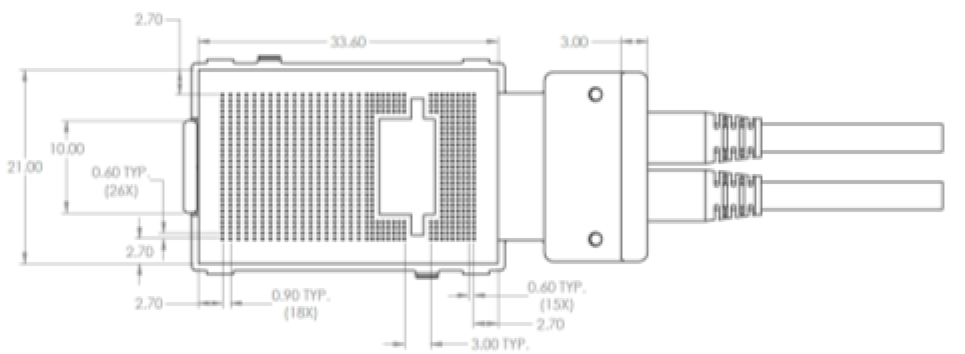
3.2T Module Dimensions

3.2T Optical Module

- 32 x 112G XSR to Standard Optics:
  - 8 x 400G DR4
  - 8 x 400G FR4 (incl. 200G mode)
- Copper Cable Assembly compatible
- Power capability:
  - 56W (Internal Laser option)
  - 48W (External Laser option)







Copyright @ 2024 OIF

LGA Pin Map

3.2T Optical Module

- Supply rails:
  12V, 3.3V, 2.6V, 1.8V, 1.2V, 0.9V, 0.7V
- Comms Electrical: 1.2V SPI
- Comms protocol: CMIS
- 400G and 800G (2x400G) port grouping defined
  - *For low power modes and 2x400G-FR4 cable assignment*

VCC_0P7A
VCC_0P7B
VCC_0P9
VCC_1P2
VCC_1P8
VCC_2P6
VCC_3P3
VCC_12

400G Port
2 x 400G Grouping

Ingress, to Switch

Egress, to Fiber

Top view

RF region
0.9mm x 0.6mm pitch

DC/Low-Speed region
0.6mm x 0.6mm pitch

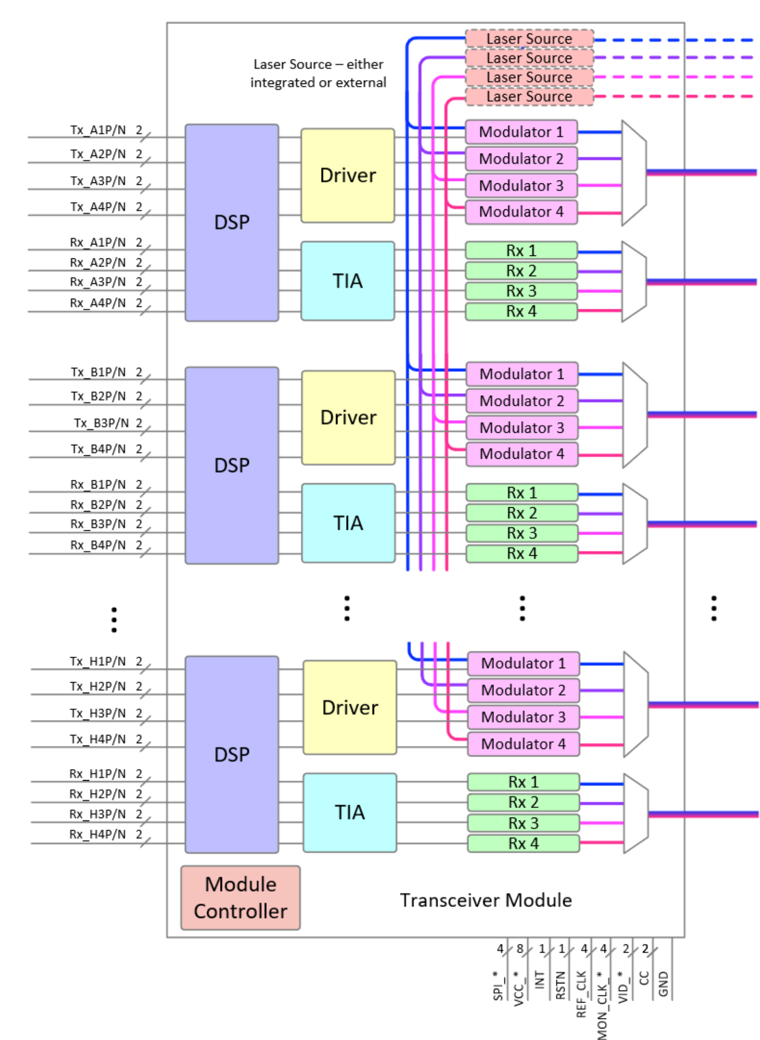
Copyright @ 2024 OIF

3.2T Optical Module Functionality

3.2T Optical Module

- FR Module example ->

- *How does this all fit in?*
  - *3D integration*
  - *Die/functionality integration*
    - *Optics (Laser + Modulator + PD)*
    - *EIC (Driver/TIA/Control)*



Copyright @ 2024 OIF

Energy Efficient Interfaces @ ECOC 2024

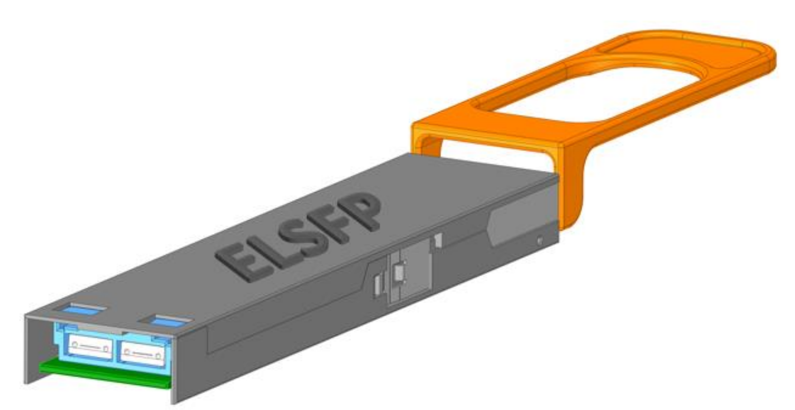
- Energy Efficient Interfaces (EEI)

- EEI Interoperability agreements
  - Co-Packaging Framework Document
  - 3.2T Optical Module for Co-Packaging Project
  - ELSFP Project
  - Electrical Interfaces for Co-Packaging

- Interoperability Demonstrations

Copyright @ 2024 OIF

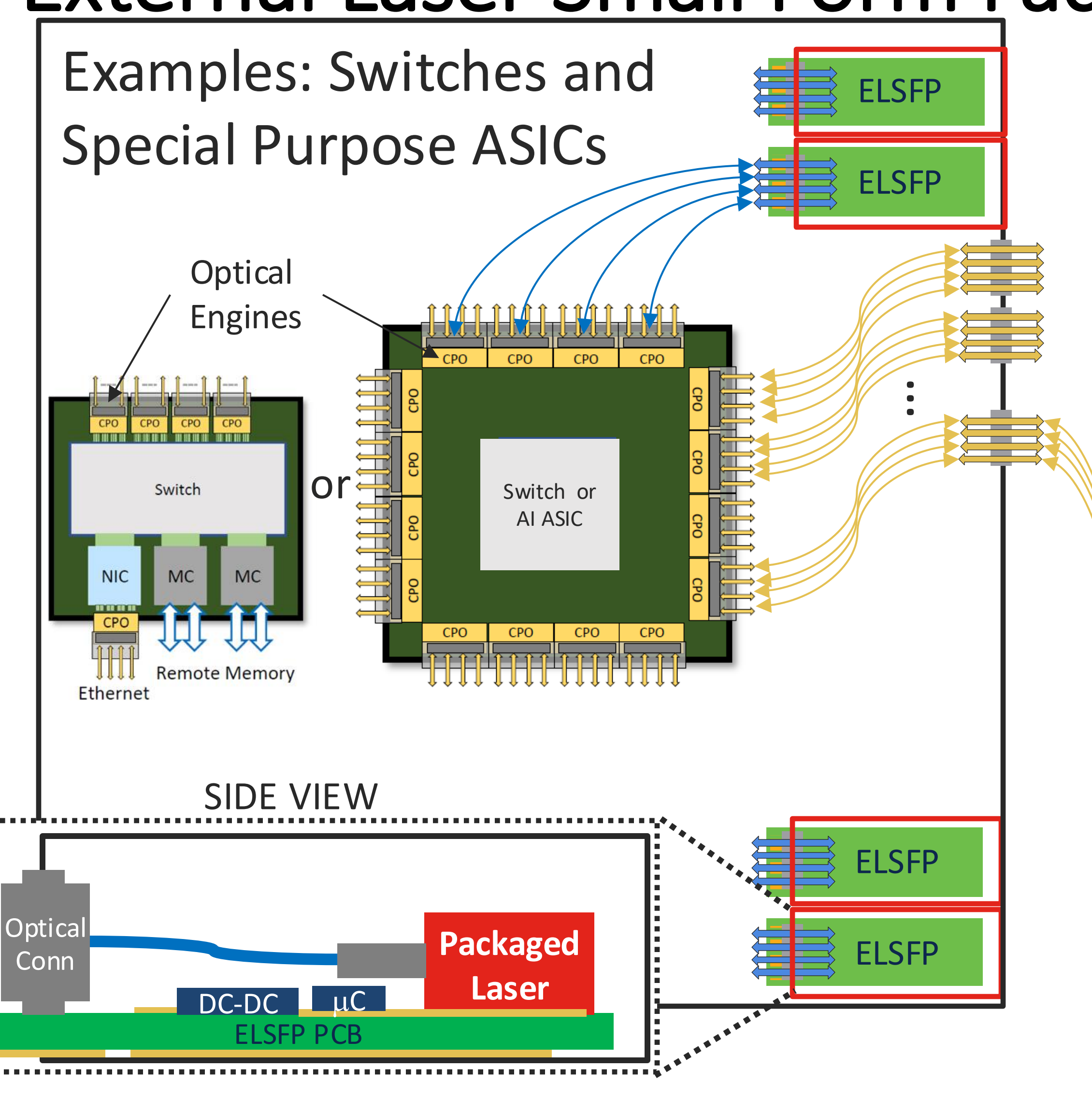
Why ELSFP?
ELSFP Project



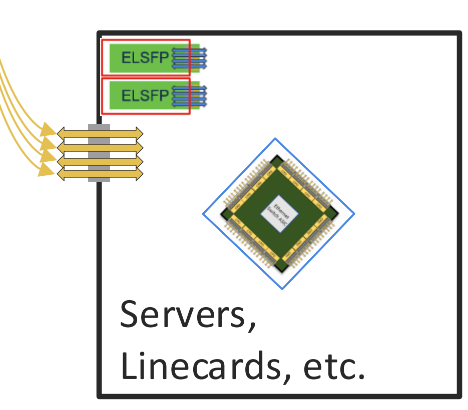
- OIF defining common External Laser Pluggable
- Industry need for co-packaged and near-packaged systems
  - Systems need faceplate density
  - External laser modules need common specification for economies of scale
- Form factor to span multiple system generations
  - Plan for optical & thermal scaling

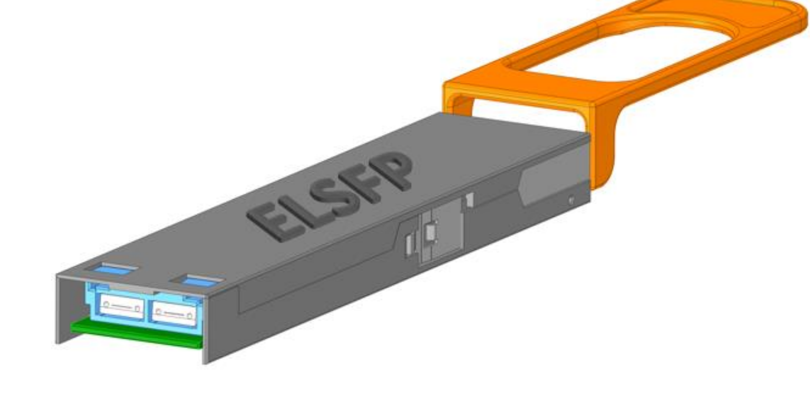
Copyright @ 2024 OIF

External Laser Small Form Factor Pluggable (ELSFP)



- ELSFPs provide CW laser power for optical engines (OEs).
- Decreases thermal power density in the system
- Each large system will likely need multiple (i.e. 8 or 16) ELSFPs
- The light from a given ELSFP can feed more than a single OE.
- A pluggable form factor helps to ensure total system reliability and a “hot swap” replacement if a single laser or ELSFP module fails.
- Eye safety is achieved by a blind mate optical connector internal to the system.





Initial Technical Concept
ELSFP Project

Density
- Blind mate pluggable
- Width similar to OSFP (16 modules wide with standard management I/O)

Commonality
- Industry standard 3.3V Supply
- CMIS (Common Management Interface Specification)

Scaling
- Optical Power Classes
- Thermal Power Classes
- Belly-to-belly configurations
- Riding heat sink for system flexibility
- 2 “MT like” ferrules for future proofing
  - Support for 8 PM fibers per MT
  - Support for multiple OE modules

Copyright @ 2024 OIF

Single Port ELSFP Design

ELSFP Project

Press fit host connector

Riding heat sink

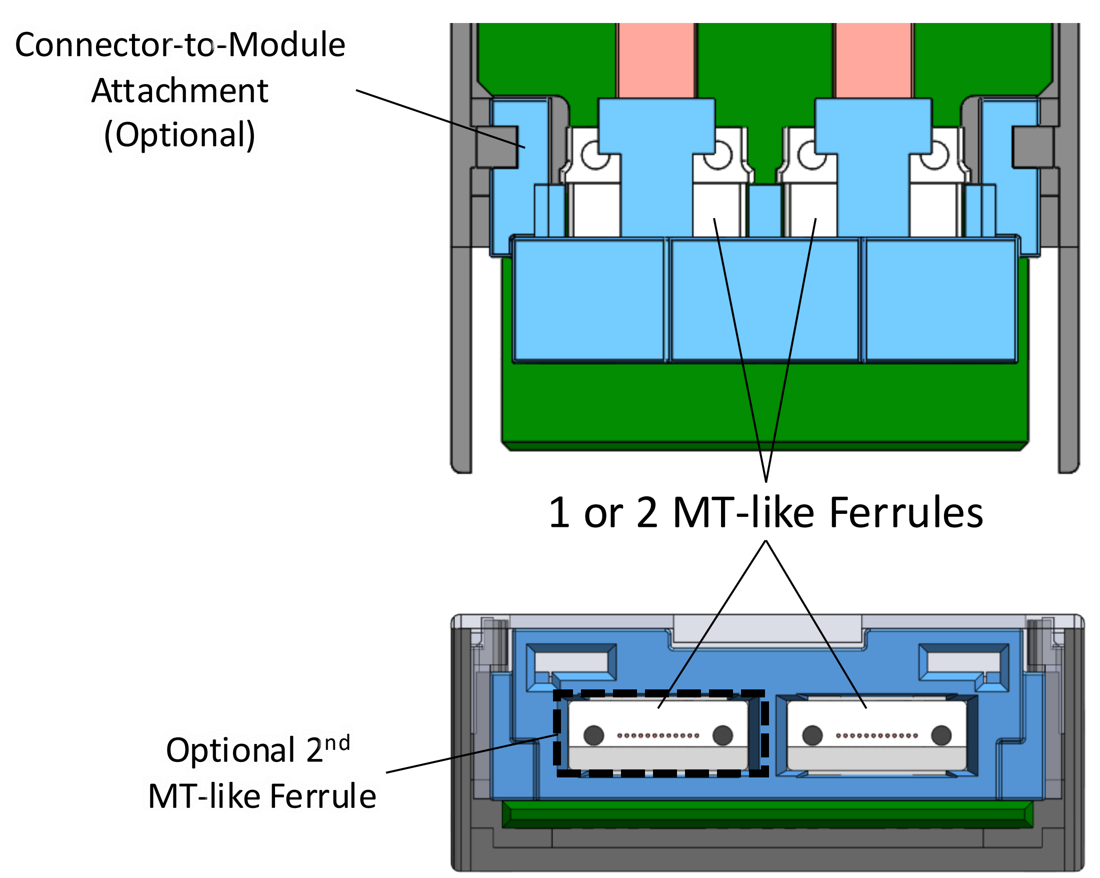
Optical connectors snap into host connector housing

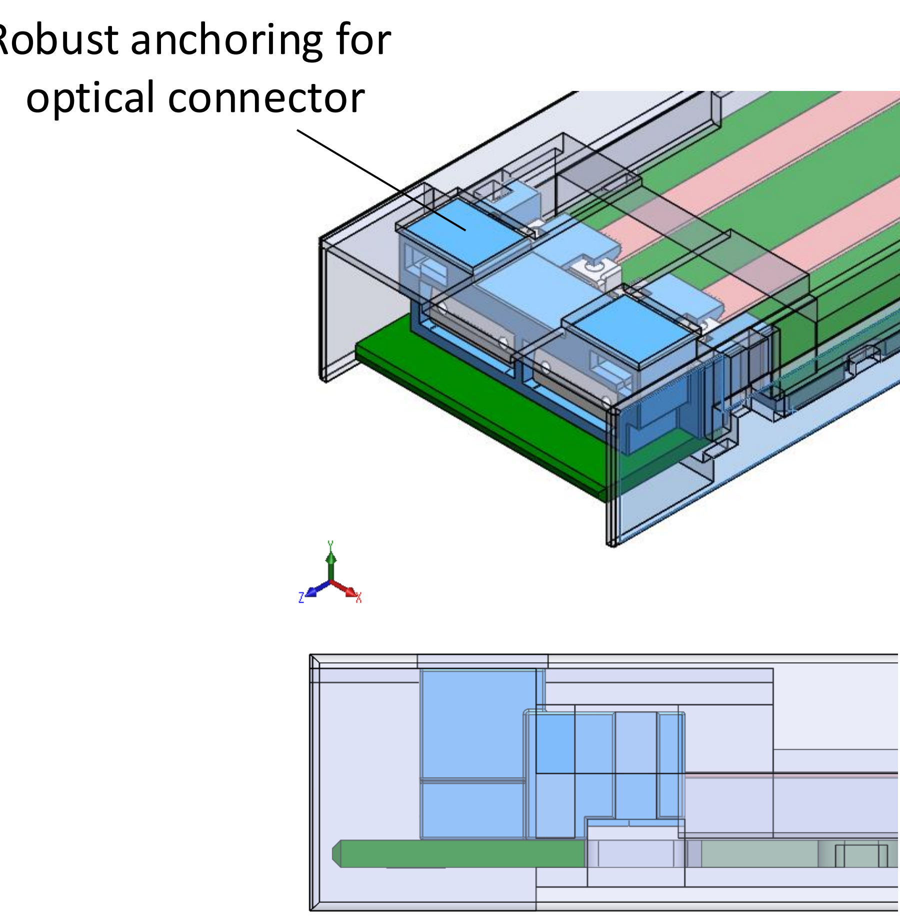
Separately applied connector and cage

OSFP style latch

Shroud providing paddle card and optical connector protection

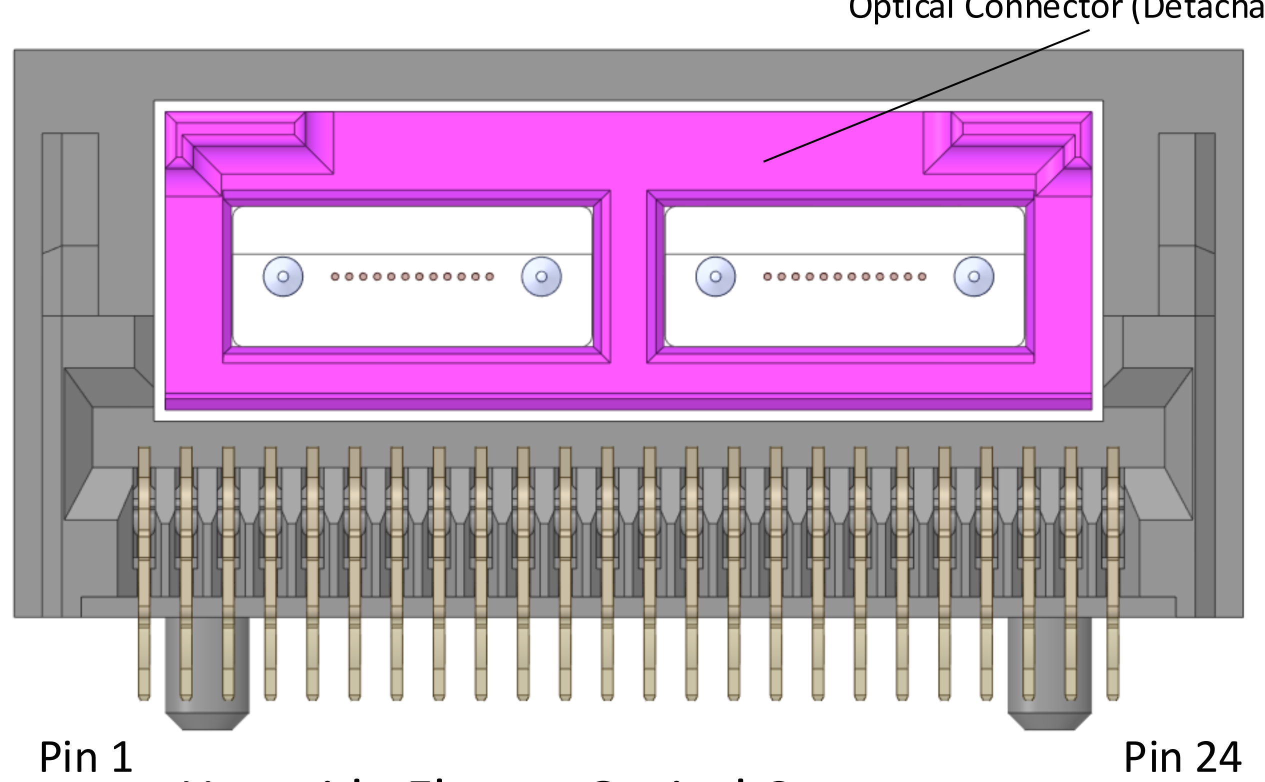
ELSFP Module-Side Optical Connector





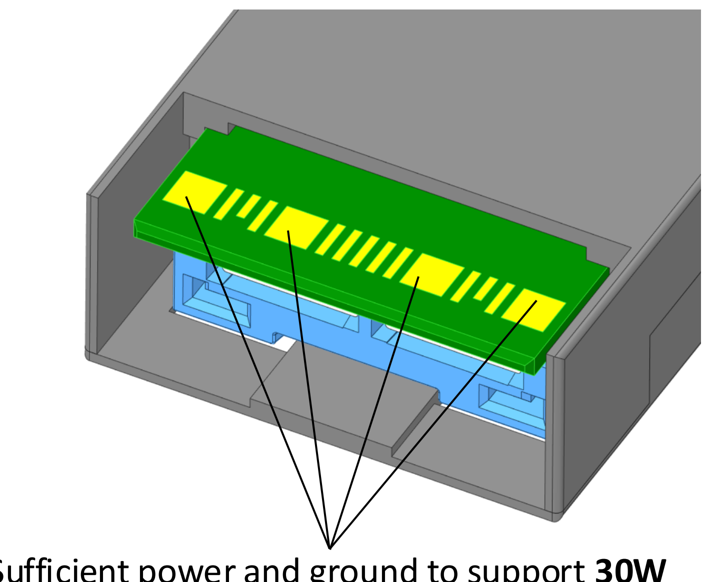
Copyright @ 2024 OIF

ELSFP Electro-Optical Connector



Host side Electro-Optical Connector

Module Bottom side Electrical Contacts



Sufficient power and ground to support **30W**

Additional pins for control/management, laser safety (i.e. presence pin), and spares for future proofing
Optical connector sub-assembly (pink) is separable from the board mounted electrical connector sub assembly

Copyright @ 2024 OIF

ELSFP Optical Power Classes

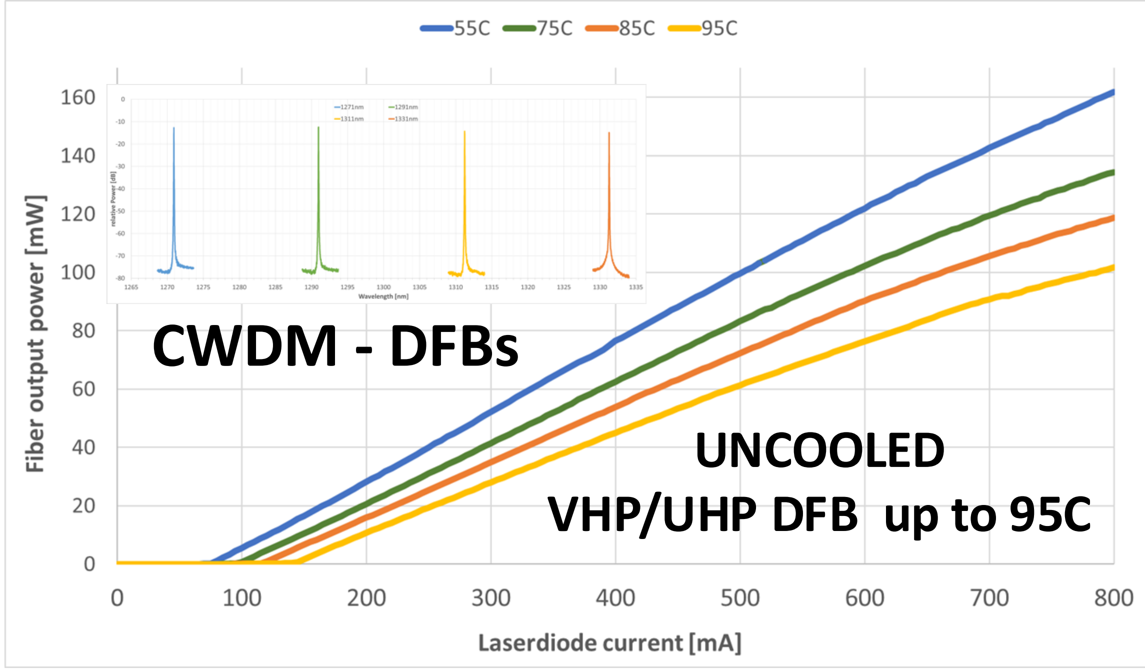
ELSFP Optical Power Classes	Power/λ/Core +/- 1.5dB
Super Low Power - SLP	2dBm
Ultra Low Power - ULP	5dBm
Very Low Power - VLP	8dBm
Low Power - LP	11dBm
Medium Power - MP	14dBm
High Power - HP	17dBm
Very High Power - VHP	20dBm
Ultra High Power - UHP	23dBm
Super High Power - SHP	26dBm

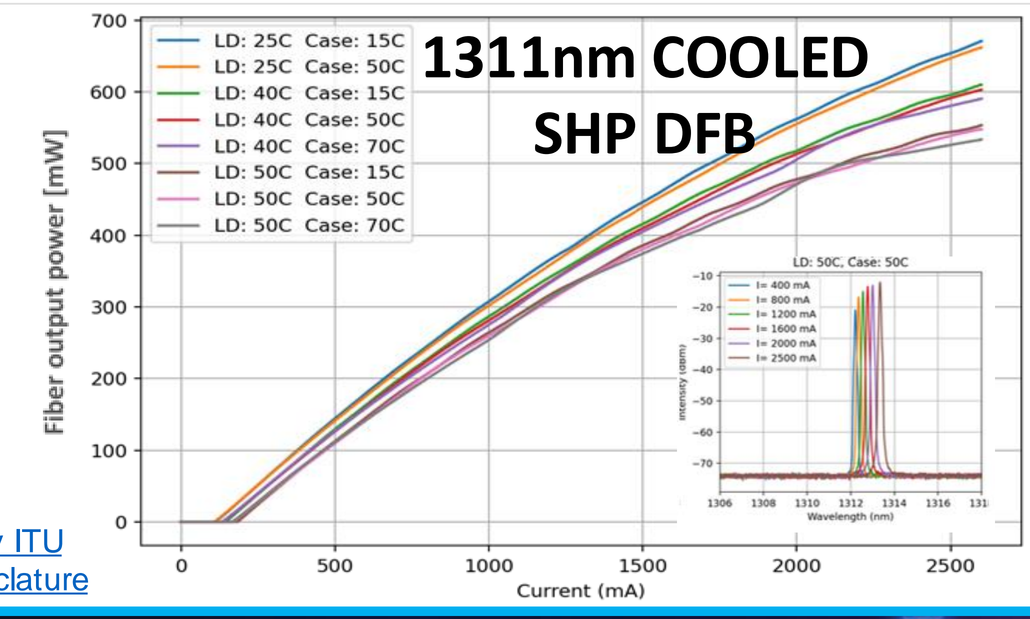
Combs

Single-Channel

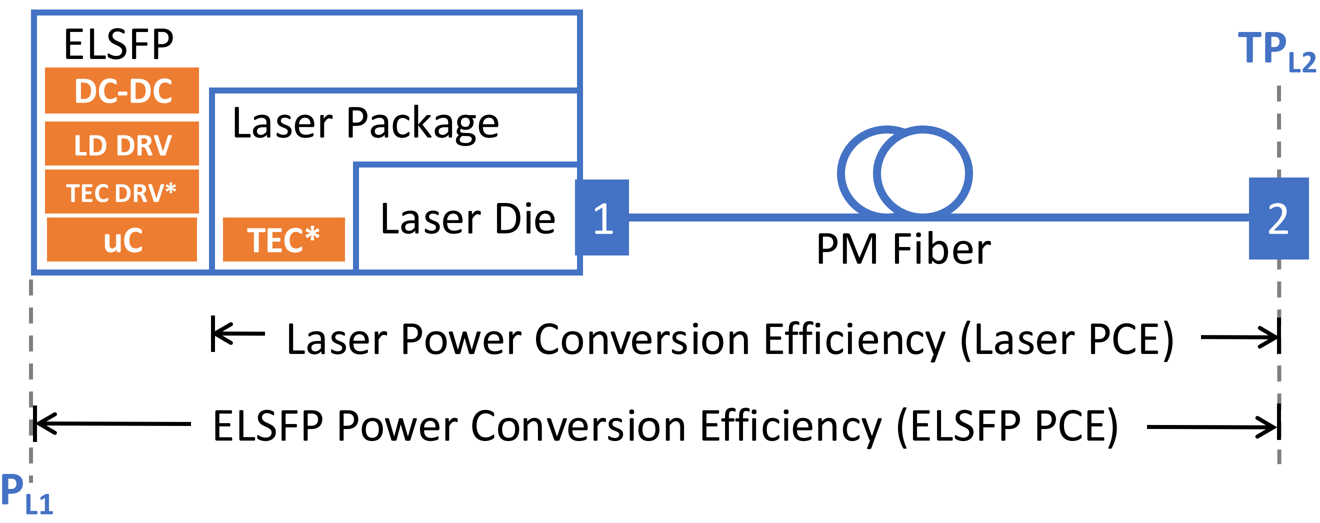
Multi-Channel

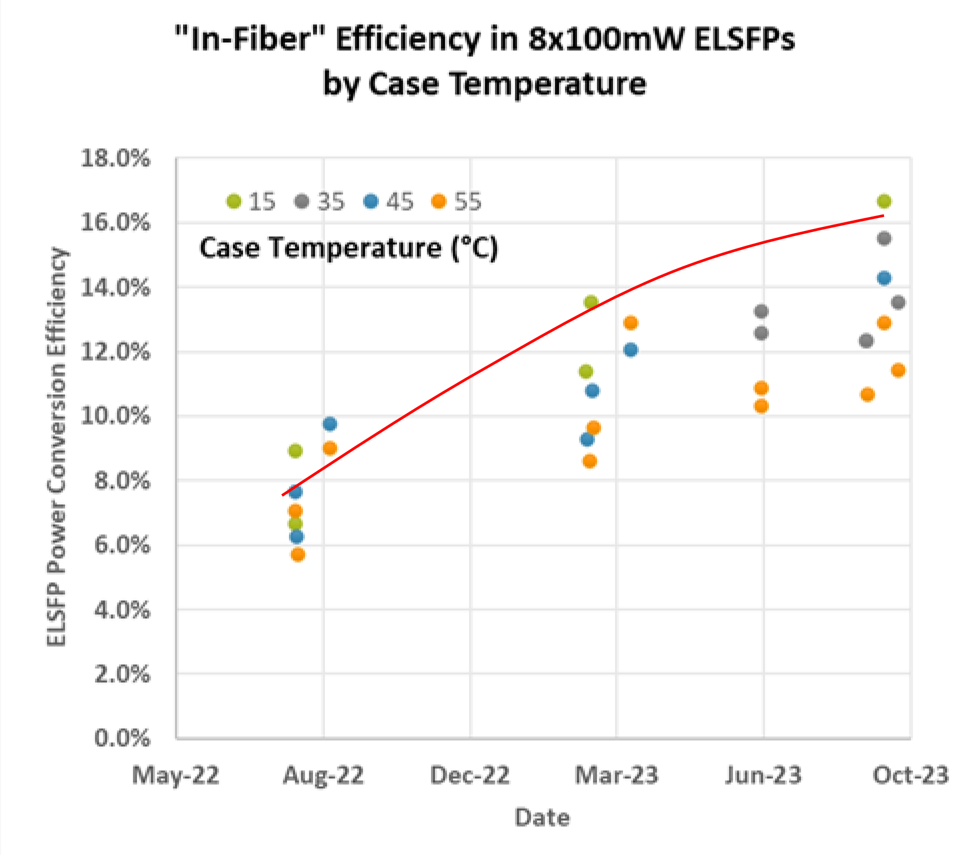
*Naming convention inspired by ITU Radio Frequency Band Nomenclature





ELSFP's eco-system drives innovation





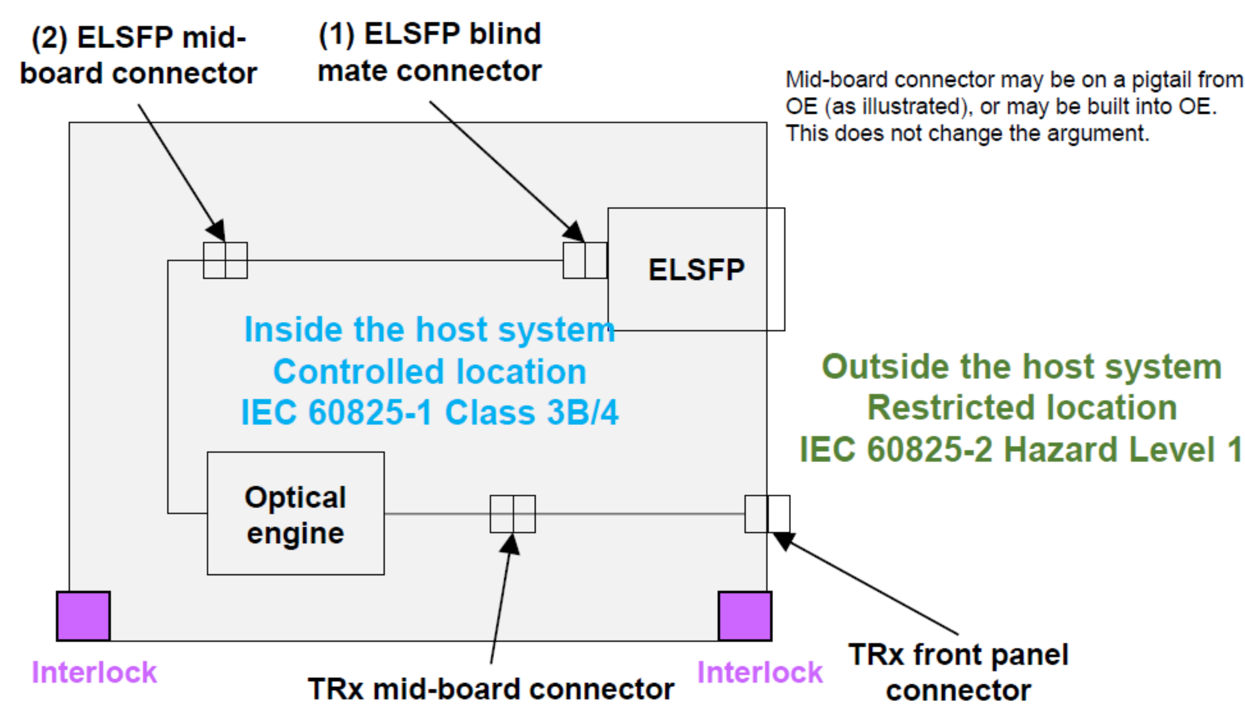
The ELSFP's eco-system continues to innovate and has yielded impressive improvements in energy efficiency (PCE), a key component of next generation energy efficient interfaces

Copyright @ 2024 OIF

Eye Safety

ELSFP's blind mate optical connector paired with a system interlock enables a safer co-packaged system implementation for users.

Similar to EDFAs with powerful CW lasers, Class 3B and 4 lasers can be used inside ELSFP and systems can be deployed in unrestricted locations.



Copyright @ 2024 OIF

Energy Efficient Interfaces @ ECOC 2024

- Energy Efficient Interfaces (EEI)

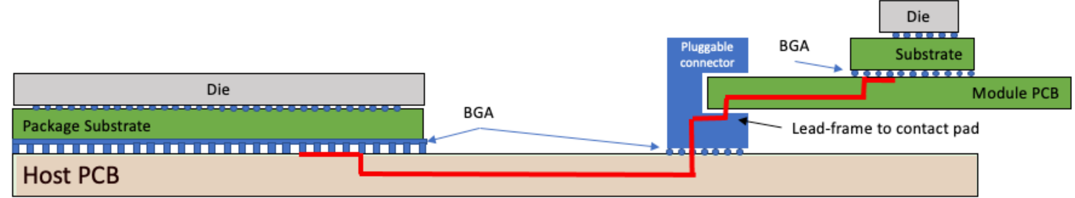
- EEI Interoperability agreements
  - Co-Packaging Framework Document
  - 3.2T Optical Module for Co-Packaging Project
  - ELSFP Project
  - Electrical Interfaces for Co-Packaging

- Interoperability Demonstrations

Copyright @ 2024 OIF

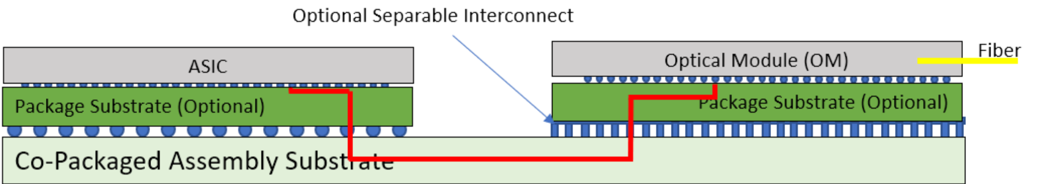
CEI – An Essential Building Block for Co-packaging

Pluggable Module Channel Example Illustration



- Channel loss: 16dB ball to ball (22-24dB bump to bump)
- Typical pluggable connectors: IL of ~1dB with RL of -10dB @26.5GHz

CPO/NPO Channel Example Illustration

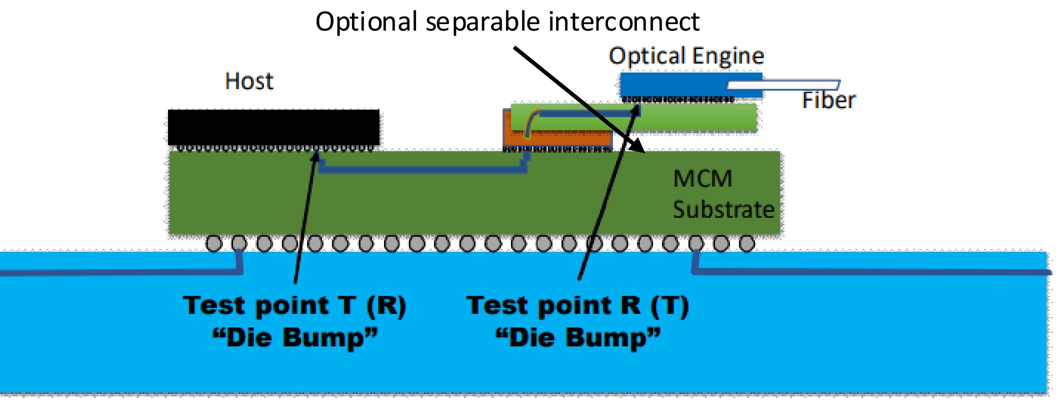


- Channel loss: CPO – 10dB bump to bump; NPO – 13dB bump to bump
- Optional separable interconnect performance example: LGA socket: IL of ~0.05dB with RL of -40dB @26.5GHz (*oif2020.341.01*, *Nathan Tracy*)
- Avoids/reduces major discontinuities.
- Optical modules are not end user pluggable.

- Significant power saving opportunity over VSR to be captured.
- A broad interoperable ecosystem is the key to success and can only be achieved through standardization.

Copyright @ 2024 OIF

CEI-112G-XSR-PAM4 for Co-packaging



Category	IL at Nyquist (Max, dB)	BER (Max)
CAT1	**10**	**1e-6**
CAT2	10	1e-8
CAT3	8	1e-9

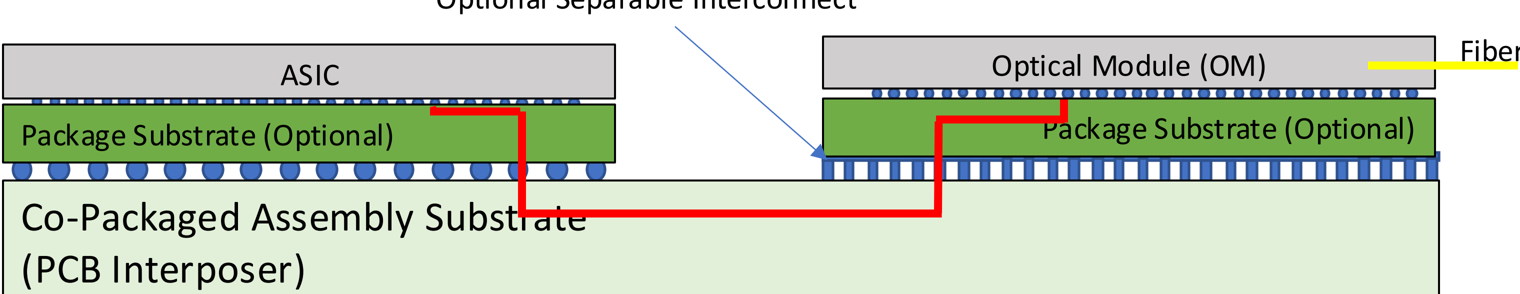
- Baud rates supported: 36 Gsyms/s to 58 Gsyms/s
- Based on loss and jitter budgets between TX and RX using copper signal traces in a SIP(System in a Package) to enable low power consumption
- Three channel categories are defined, allowing optimization for various applications.
- Timeline
  - Project started in April 2018.
  - Draft specification is becoming technically stable. Few pending items to be addressed.

Copyright @ 2024 OIF

CEI-112G-XSR+ -PAM4 for Near Packaging

- The emergence of Near Package Optics (NPO) Architecture
  - Co-packaging requires significant package substrate size increase and technology advancement, which adds risk to goals of availability, cost and multi-vendor support.
  - Instead of a monolithic package approach, Near Packaging relies on advanced PCB technology for dense high-speed routing without significant power penalty.
  - Near Packaging architecture takes advantage of existing technologies and more robustly enables an open ecosystem implementation.
- Additional margin also strengthens a broader supply base for co-packaging implementation and adoption.

- Baud rates supported: 36 Gsyms/s to 58 Gsyms/s
  - Optimize for Ethernet rate @ 106.25Gbps – the key application for CPO/NPO
  - Insertion loss < 13dB @ 26.5625GHz Nyquist bump to bump with up to 1 separable interconnect.
- Enable the lowest practical energy consumption (pJ/b) implementation.
- Leverage specification methodology and other work from existing CEI 112 projects.



Copyright @ 2024 OIF

Energy Efficient Interfaces @ ECOC 2024

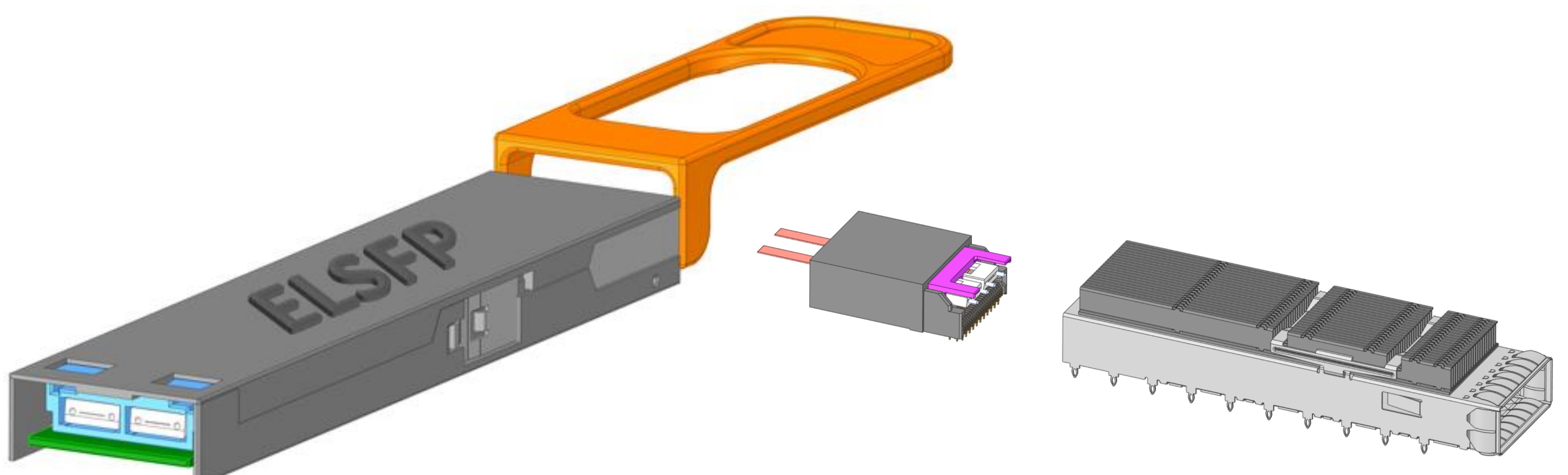
- Energy Efficient Interfaces (EEI)

- EEI Interoperability agreements
  - Co-Packaging Framework Document
  - 3.2T Optical Module for Co-Packaging Project
  - ELSFP Project
  - Electrical Interfaces for Co-Packaging
  - Compute Optics Interface (COI)
  - Retimed Transmitter Linear Receiver (RTLR)

- Interoperability Demonstrations

Copyright @ 2024 OIF

EEI ELSFP - External Laser Small Form Factor



- Lasers: 8 lasers per module (1310nm)
- Output power: 23 dBm (UHP)
- Both cooled and uncooled lasers

PCIe Protocol over Optics I (OS Enumeration)

Aurora A1000

CXL Memory Controller End Point

SMF

Optical PCIe Retimed Modules

CPU Root Complex

PCIe 5 Optical Protocol Link from CPU to CXL Memory Controller

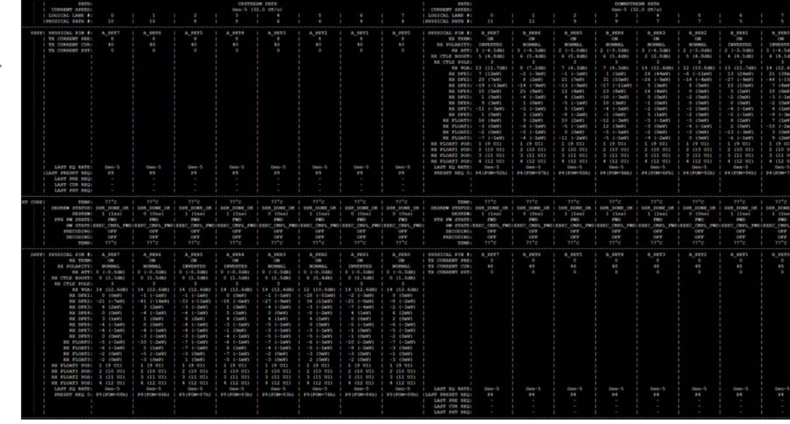
PCIe 5 full protocol link over optics using industry standard CPU and (first public demonstration)

CXL memory controller

- 512GT/s link capacity (32GT/s x 8 x 2)

PATH STATES					
Upstream Path			Downstream Path		
PTH	STATE	SPEED	PTH	STATE	SPEED
04	0x13-->FWD	Gen-5	05	0x13-->FWD	Gen-5
06	0x13-->FWD	Gen-5	07	0x13-->FWD	Gen-5
08	0x13-->FWD	Gen-5	09	0x13-->FWD	Gen-5
10	0x13-->FWD	Gen-5	11	0x13-->FWD	Gen-5

- Link diagnostics



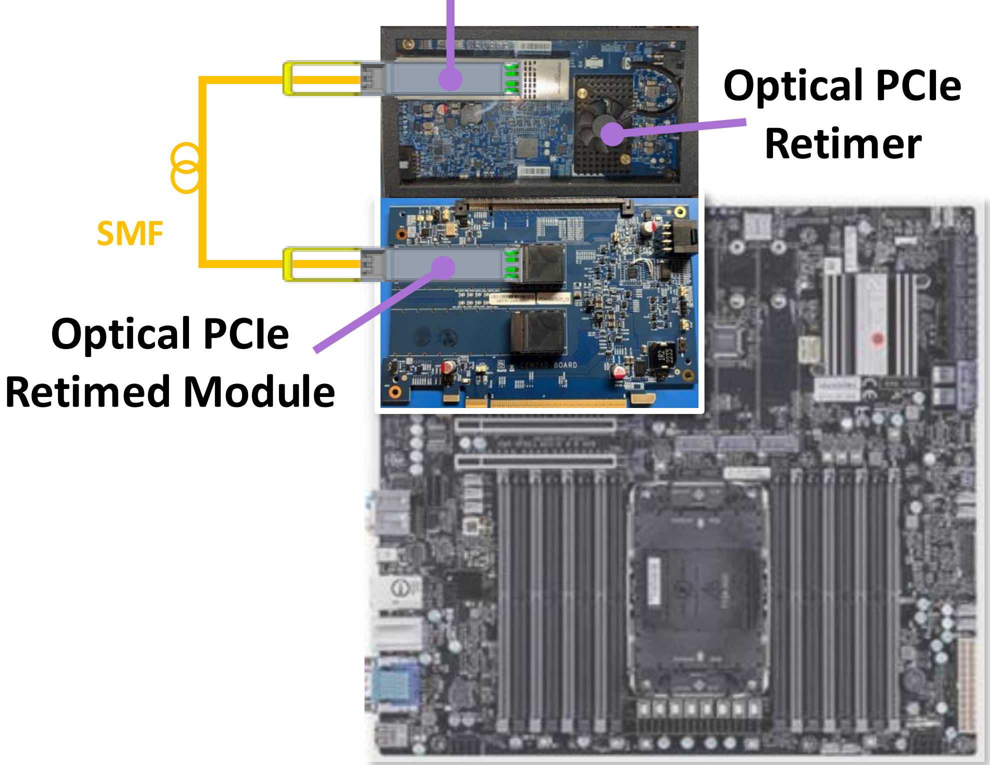
- Link Training and Status State Machine (LTSSM)

- Ubuntu server OS shows enumerated PCIe device driver from the BIOS/OS

```
root@sc-opto1: ~
root@sc-opto1:~# lspci -s 16:00.0 -vvv | grep 'CXL\|LnkSta'
16:00.0 CXL: Device 1dfa:01e2 (rev 01) (prog-if 10 [CXL Memory Device (CXL 2.x)])
        LnkSta: Speed 32GT/s (ok), Width x8 (ok)
        LnkSta2: Current De-emphasis Level: -6dB, EqualizationComplete+ EqualizationPhase1+
    Capabilities: [3d4 v1] Designated Vendor-Specific: Vendor=1e98 ID=0000 Rev=1 Len=56: CXL
        CXLCap: Cache- IO+ Mem+ Mem HW Init+ HDMCount 1 Viral+
        CXLCtl: Cache- IO+ Mem- Cache SF Cov 0 Cache SF Gran 0 Cache Clean- Viral-
        CXLSta: Viral-
root@sc-opto1:~#
```

Copyright @ 2024 OIF

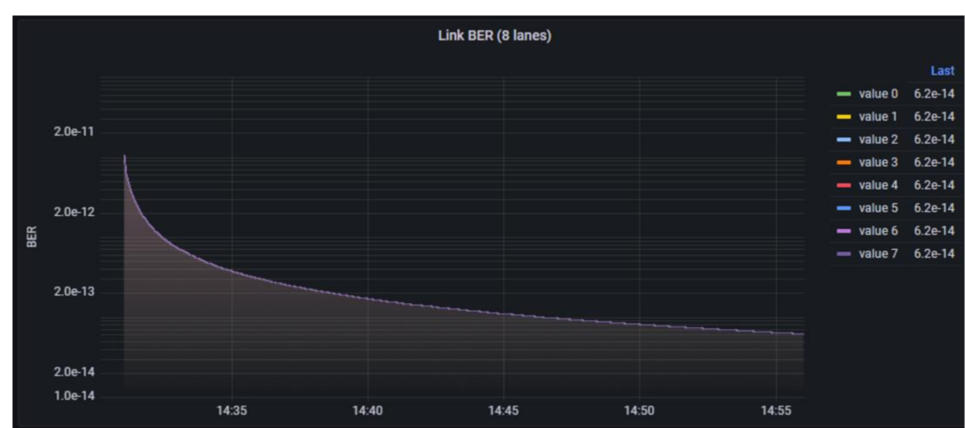
PCIe Protocol over Optics II (Performance)



**PCIe 5 Performance (BER)
Retimed to Unretimed Optics**

Retimed module to unretimed module

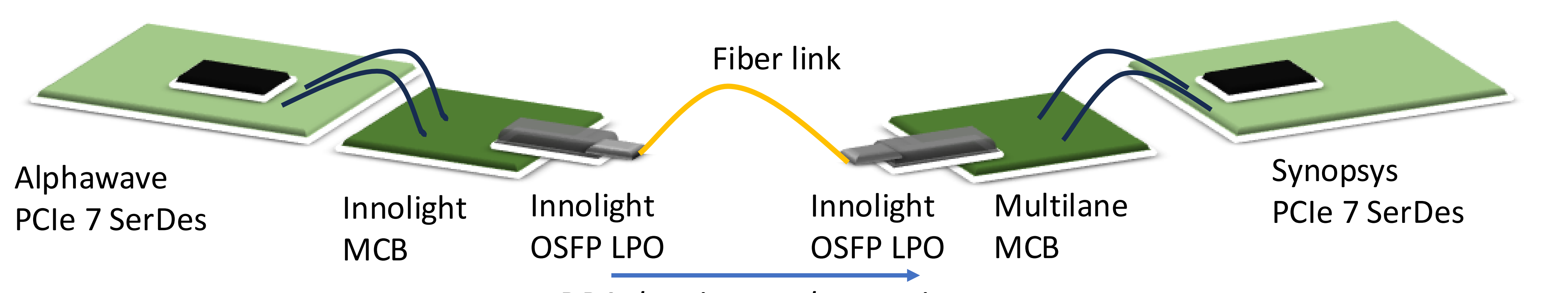
- Integrated Retimed Module (Astera Labs), to
- Unretimed LPO Module (Accelink or Innolight)
- Demonstrating performance <<1e-12
- Running live prbs data stream x 8 bi-directional, error free



BER live graph (BER reducing further vs time captured)

Copyright @ 2024 OIF

PCIe 7.0 Performance Over Optics



DR4 showing one lane active

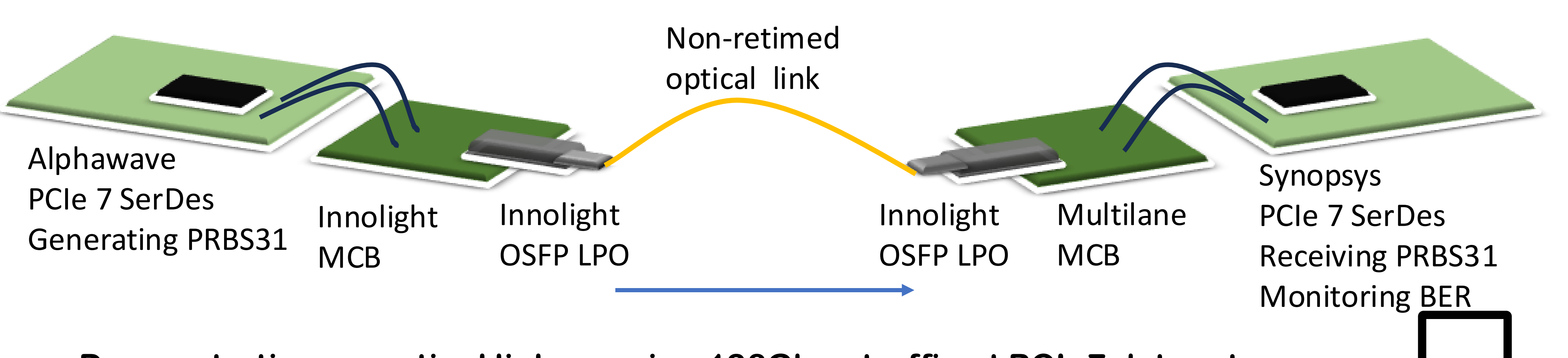
This PCIe 7.0 over Optics demonstration features a test chip silicon TX transmitting PRBS 31 PAM4 at 128Gbps over an optical link. The setup includes an Innolight module compliance board mated with an Innolight linear pluggable optical module (LPO) over a fiber optics channel. This is connected to another Innolight linear pluggable optical module (LPO),which is mated with a Multilane module compliance board. The test chip silicon RX then receives the PRBS 31 signal at 128Gbps. The receiver test chip demonstrates bit error rate (BER) performance and displays the received eye diagram over a linear optical link

Key points:

- Demonstrates a developing eco-system enabling energy efficient link capable of transporting PCIe data rates

Copyright @ 2024 OIF

PCIe 7.0 Performance Over Optics



Demonstrating an optical link carrying 128Gbps traffic at PCIe7 data rates

Data Signal: PCIe7 (128Gbps PRBS31 PAM4)

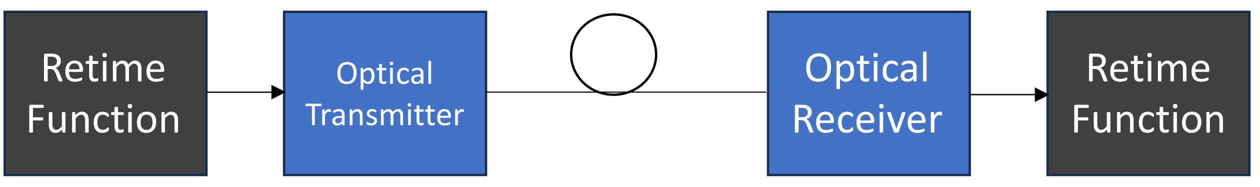
Optical Link: 800GBASE-DR4 OSFP linear modules supporting 128Gbps traffic

Copyright @ 2024 OIF

RTLR (Retimed Tx, Linear Rx)

Fully Retimed Optical Link: Highest Power, Longest Reach

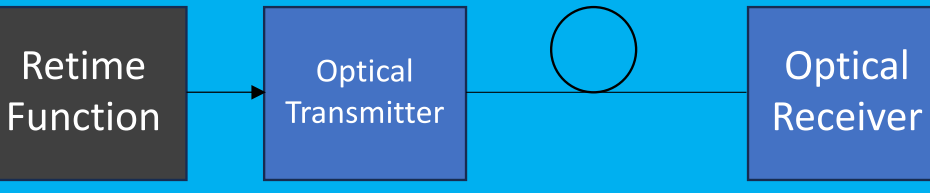
DSP/retimer module OIF-CEI-112G-VSR-PAM4 supports 16 dB channel on egress with some optical output compliance expectation



Ingress path includes DSP/retimer in the module and supports 16 dB channel to Host ASIC

Retimed Transmit Linear Receiver (RTLR) Optical Link: Balance of Reach, Power

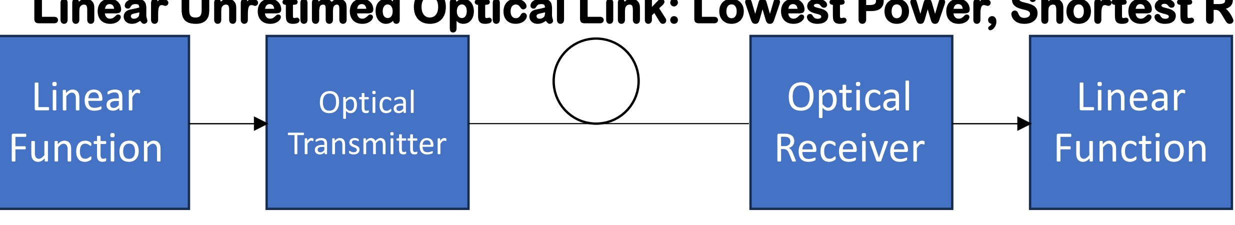
OIF-EEI-112G-RTLR is meant to be identical to above on the egress channel



Ingress path removes the DSP/retimer in the module and uses an enhanced version of OIF CEI-112G-Linear-PAM4 specifications by utilizing host ASIC DSP SerDes capability

Linear Unretimed Optical Link: Lowest Power, Shortest Reach

Egress path removes the DSP/retimer in the module and uses OIF CEI-112G-Linear-PAM4 specifications by utilizing host ASIC DSP SerDes capability



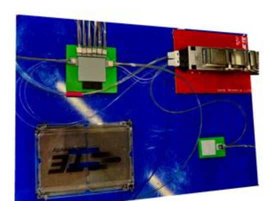
Ingress path removes the DSP/retimer in the module and uses OIF CEI-112G-Linear-PAM4 specifications by utilizing host ASIC DSP SerDes capability

Copyright @ 2024 OIF

Conceptual Demo for AI Compute

Depicting an AI Backend Compute and its various links

Compute Chassis
Showing an array of compute and switch cards interconnected with a variety of optical connectivity options



Next Gen switch card located in Compute Chassis
Highlighting an ASIC with 4T/mm edge bandwidth and Ethernet interfaces on board.
Optical links powered by ELSFP

Accelerator Cards
Variety of next gen PCIe compute cards plugged into a PCIe chassis

AI backend compute employs low latency links to interconnect local accelerators in a cache coherent way. The local links are typically PCIe-like (NVLink, UALink, etc).

Groups of compute clusters are interconnected with lower latency Ethernet / InfiniBand connections

Copyright @ 2024 OIF

ALPHAWAVE
SEMI

AOI

AsteraLabs

CICT
Accelink

INNO LIGHT

LESSENGERS
Transferring Light Through Nanowire

Linktel
TECHNOLOGIES

SENKO
Advanced Components

SUMITOMO
ELECTRIC

O-Net
Technologies

SYNOPSYS

TE
connectivity

USCONEC

OIF
PLL INTEROP DEMO
ECOC 2024

OIF
SCAN ME