OIF
PLL INTEROP DEMO
ECOC 2024

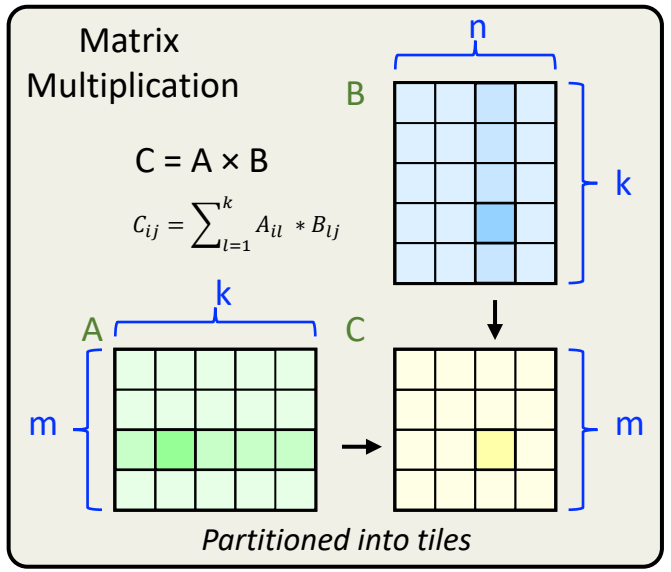
Reining in Power Consumption Trends for Next Generation Optical Networking for AI Compute

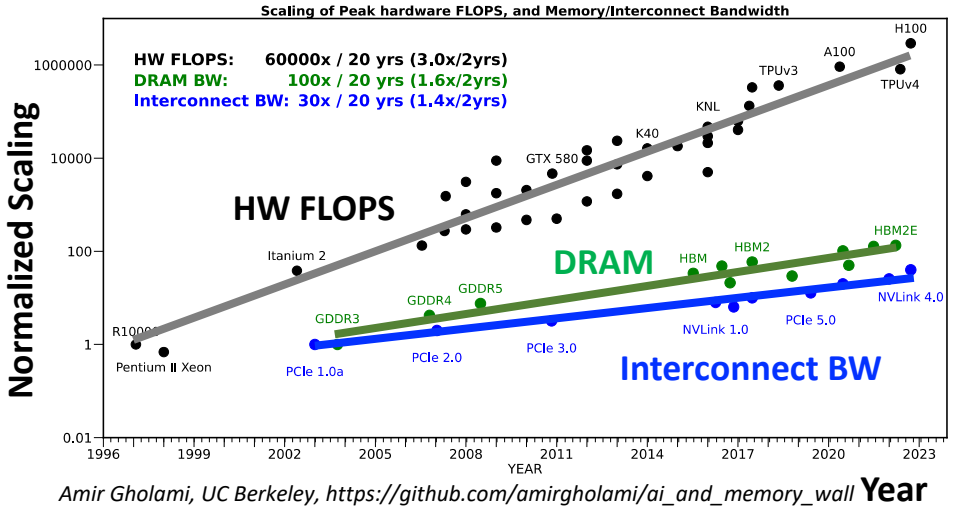
Jeff Hutchins / Ranovus
OIF PLL EEI Vice Chair

Market Focus - ECOC 2024
Tuesday 24 September

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

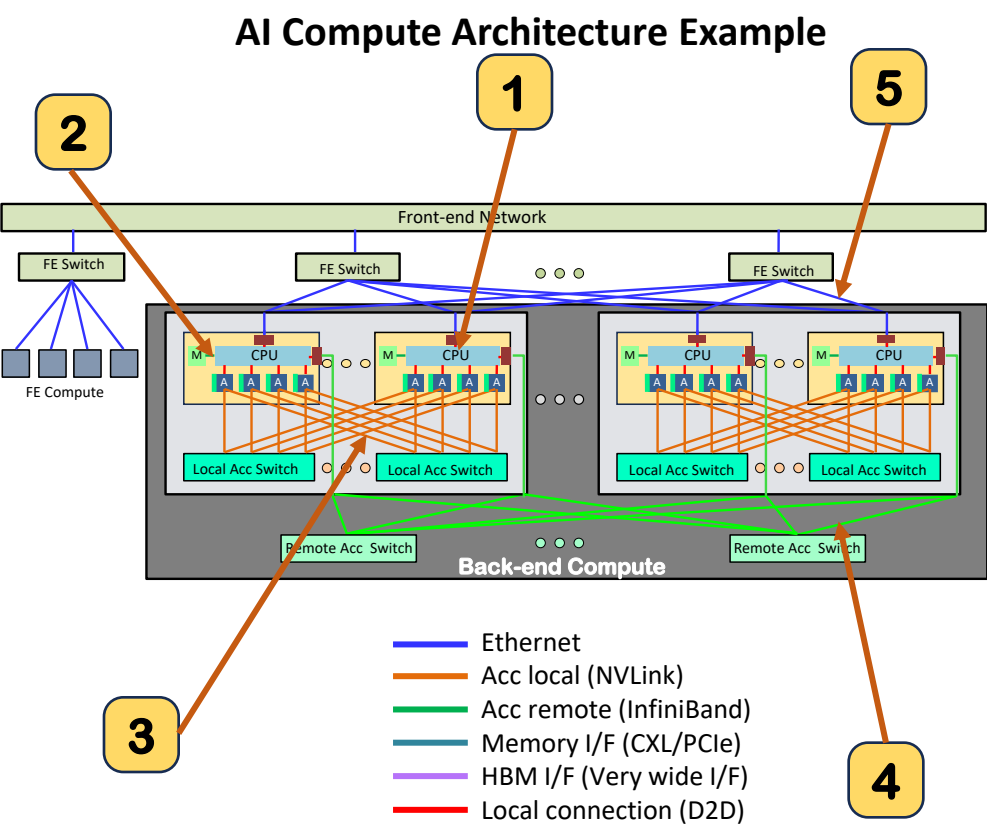




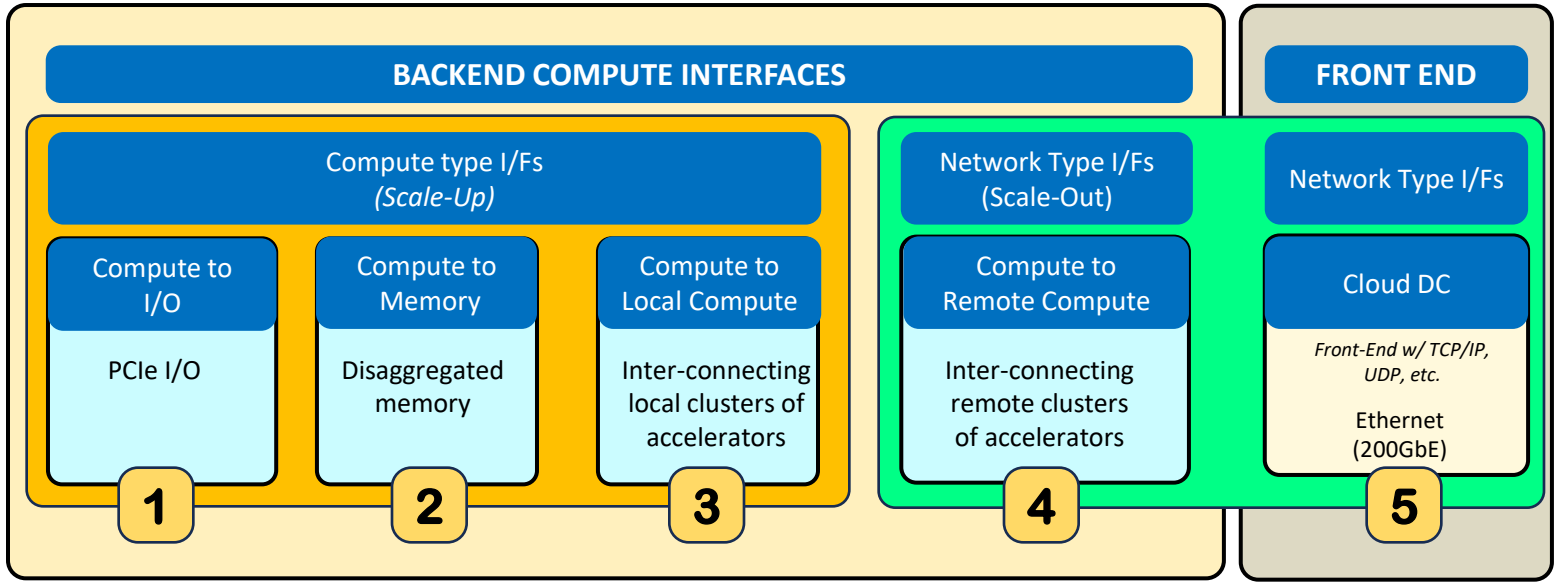
Amir Gholami, UC Berkeley, https://github.com/amirgholami/ai_and_memory_wall **Year**

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

OIF projects address next gen AI compute interfaces



The OIF has initiated several projects to address these needs



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.)

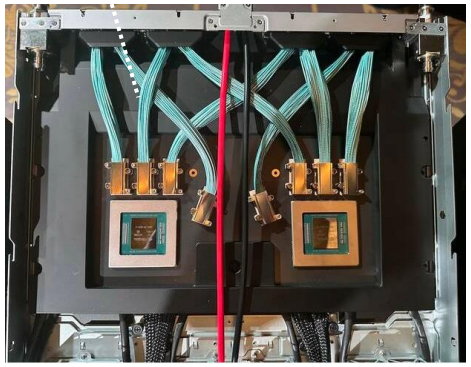
RTLR Project (Retimed Tx, Linear Rx)

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

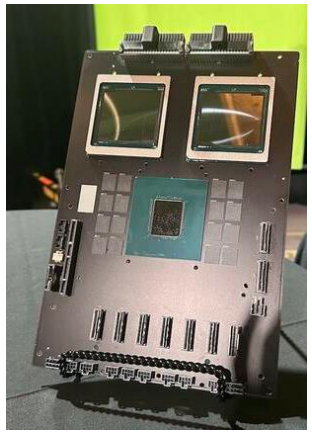
What do these local accelerator links look like?

Scale-up Cables 72 dp per GPU

NVLINK Switch Card



Accelerator GB200 Card



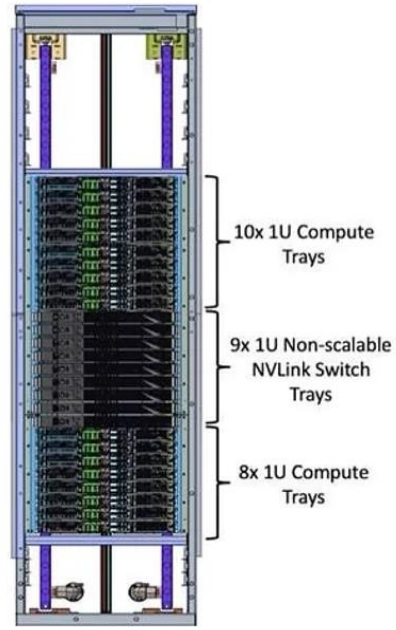
BW/NVL72 :
7.2Tbs per B200 GPU

- 72 dp @ 200G (Tx+Rx)
- 5184 dp per rack

Future GPU:

- 51.2Tbs per GPU, 72 GPUs
- 512 dp @ 200G (Tx+Rx)
- 36864 dp per rack

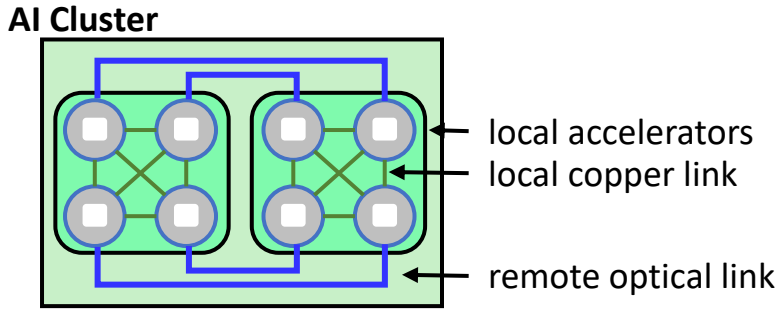
5184 dp x 200G copper cables

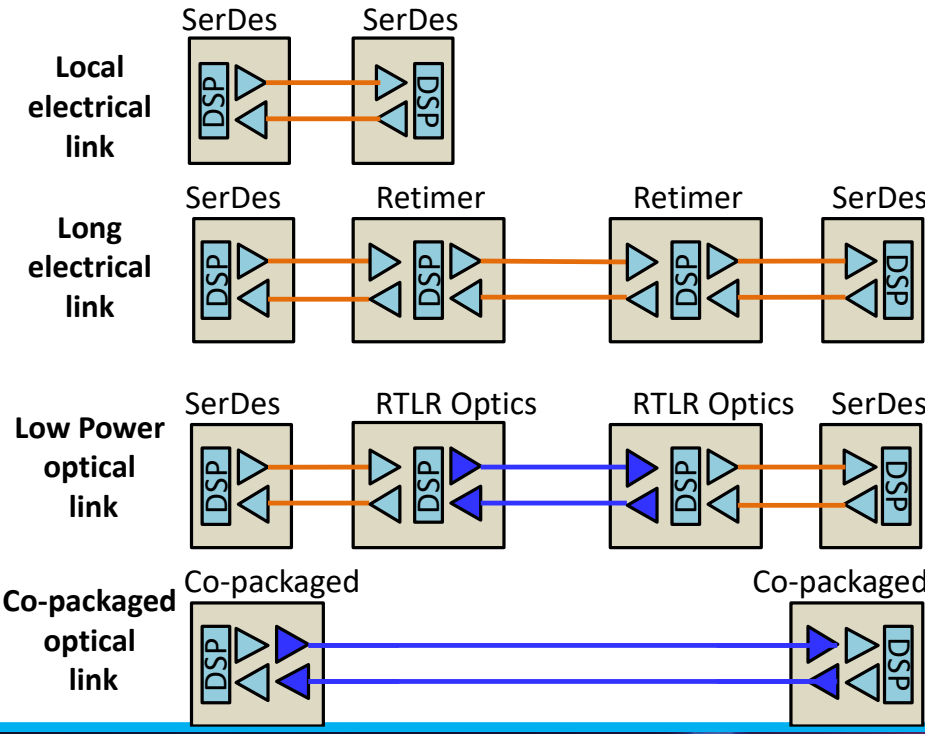


Next gen systems will likely need to at least double interconnect bandwidth with longer reach and interconnect a larger number of accelerators

What approaches can we use?

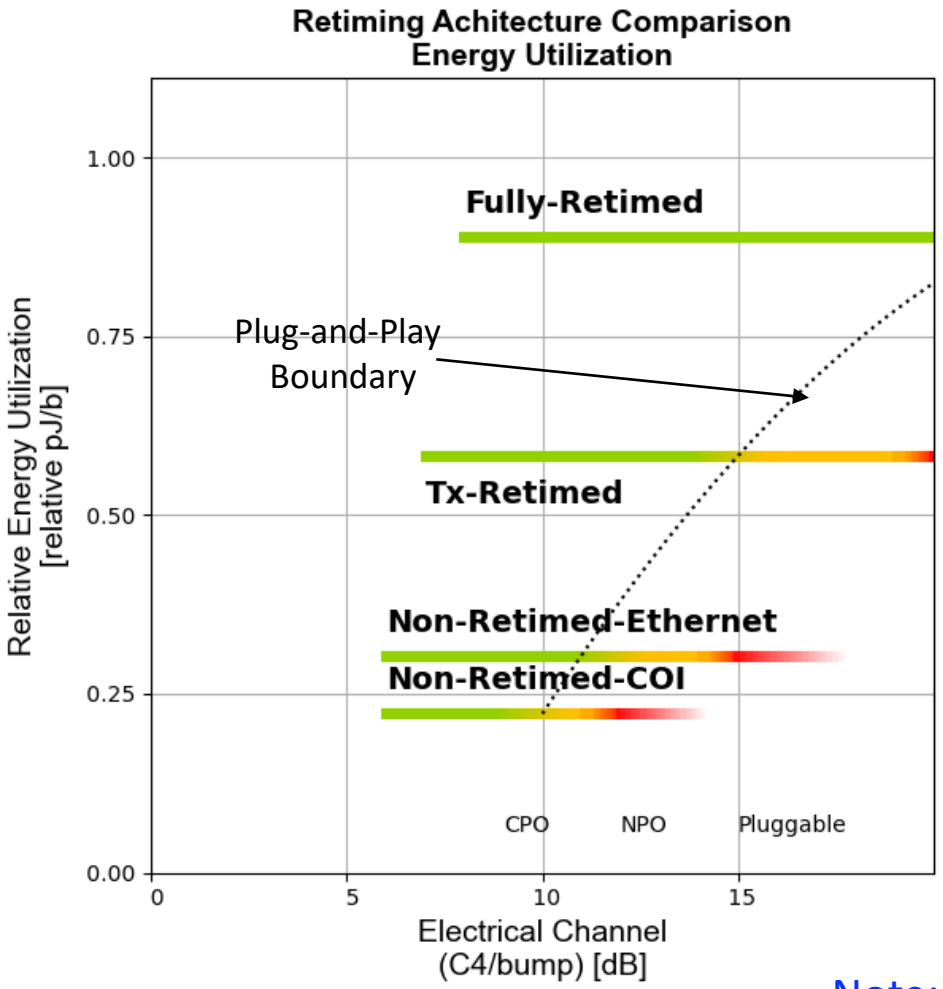
- What is needed?
  - Larger local clusters interconnected with short links
  - Energy efficient, high-speed, low latency, dense interconnects that can scale
- Copper Links
  - Copper is ideal for local connections
  - As the data rates increase, pure electrical link reach becomes shorter
  - Reach can 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

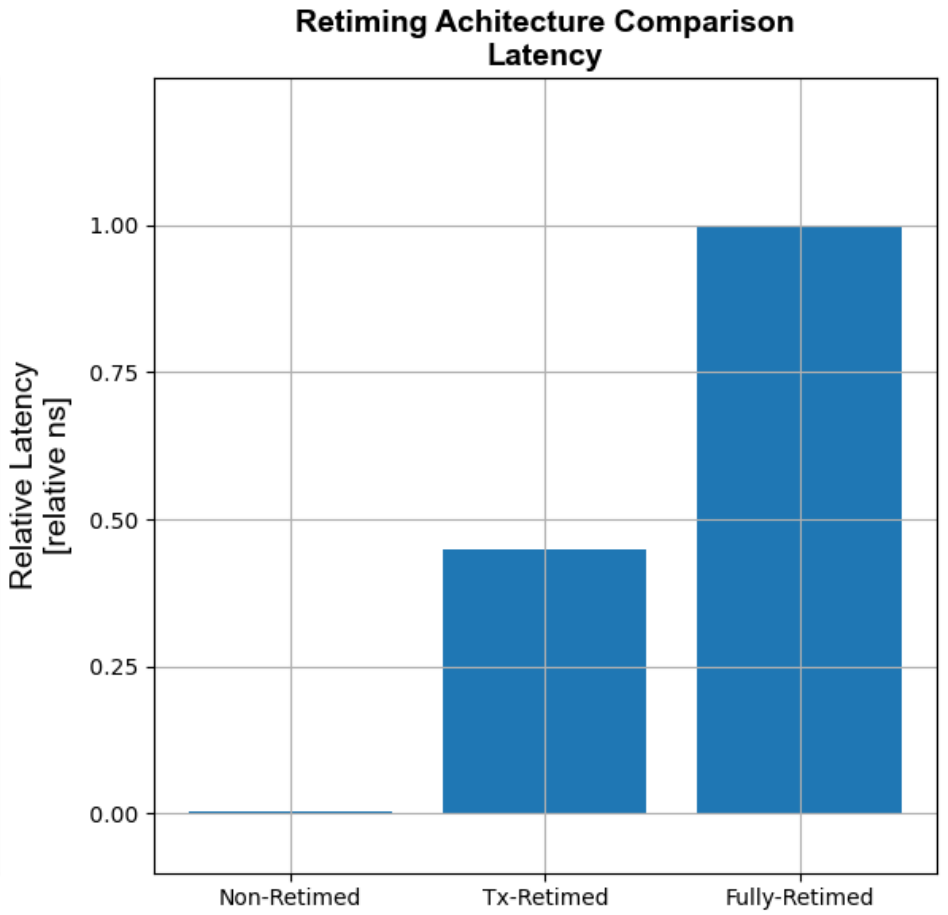




AI scale-up will drive adoption of energy efficient, low latency, & cost-effective interconnections (electrical and optical) to support large AI models

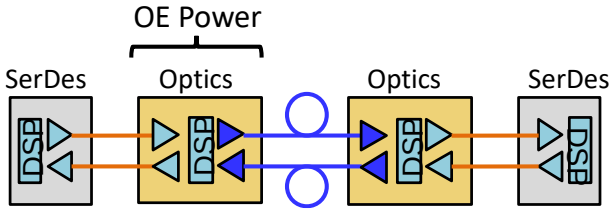
Comparison of EEI optical interface options for 200G





Caveats:

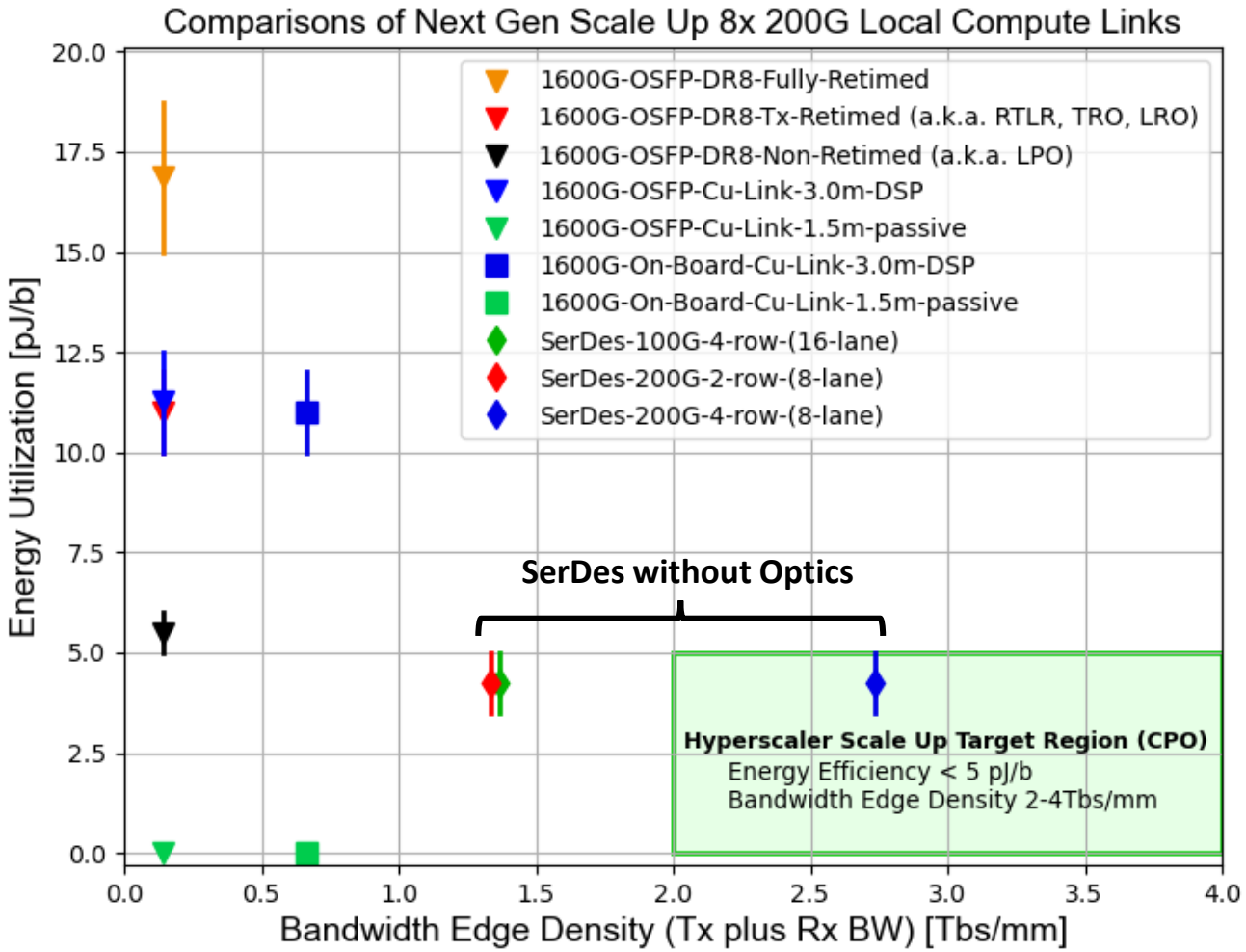
1) Energy utilization consists of optics power only



2) Ethernet links assume TH5 class SerDes
3) Compute Optics (COI) assume low latency FEC
4) Plug-and-play assumes no SerDes tuning, fully compliant

Note: This chart remains a work in progress

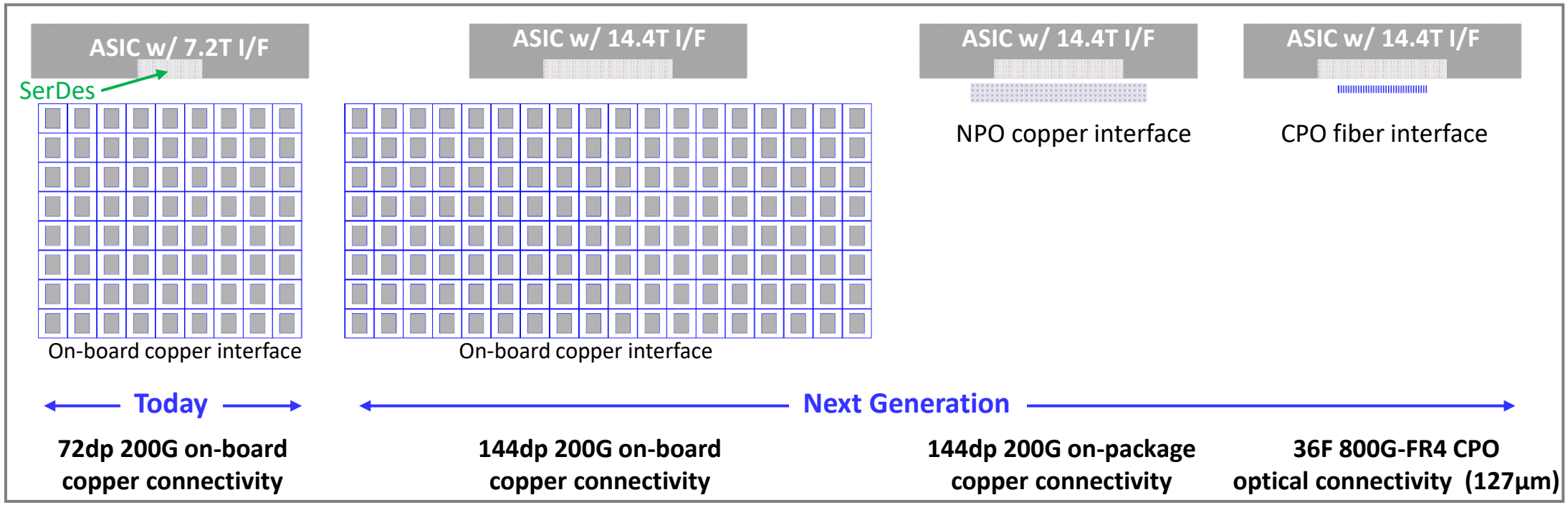
Density Comparisons for 200G Solutions



OIF received hyperscaler requirements for AI compute and the consolidated feedback is published on the OIF website

Currently deployed solutions are outside the target region

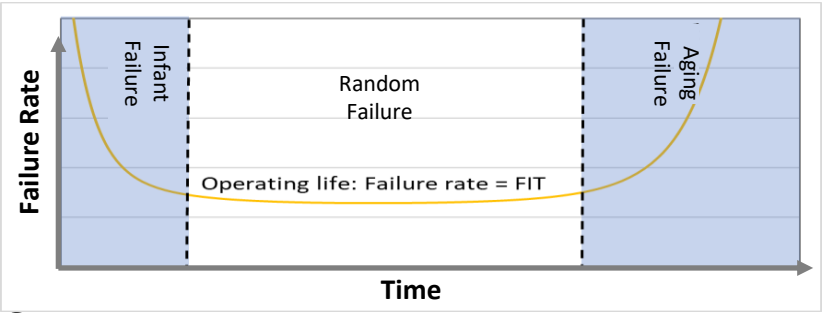
Some alternatives for doubling interface capacity



- **Could transition from 200G to 400G electrical links**
  - Transitioning to 400G electrical links expected to reduce reach and adversely impact power and latency
- **Or could use 200G signaling with two electrical lanes per link**
  - Needs more connector area
- **Or could transition to optical links with CPO**

Hardware reliability for local accelerator interconnect

- AI compute saves intermediate results to enable restart in the event of failure
  - Reducing the failure rate of an interconnect solution is a key objectiv
- Highly integrated IC solutions can be more reliable
  - For example, SR-332 estimates an IC with 20k gates: ~3.7 FIT
  - However, when split into two ICs of 10k gates each, SR-332 predicts ~7.0 FIT, nearly 2x for the same function
  - Optical modules can be an order of magnitude greater
- Lasers are a special case!
  - Lasers have a best-in-class random failure rate of ~ 1 FIT and is a significant contributor for optical links
  - Either replaceability or redundancy is required



OIF Public Webinar: Co-Packaged Optics: Why, What and How, 14 Oct 2020

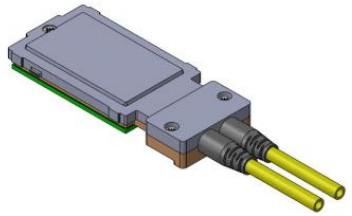
AI compute is adversely impacted by failures Highly integrated solutions are an important step to providing the reliability performance for local optical interconnect of AI Compute

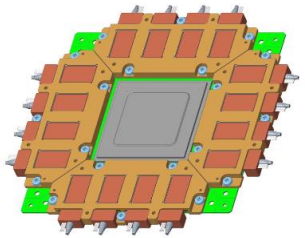
What is the OIF doing?

OIF's Co-Packaging Projects

✓ Co-packaging Framework Project

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



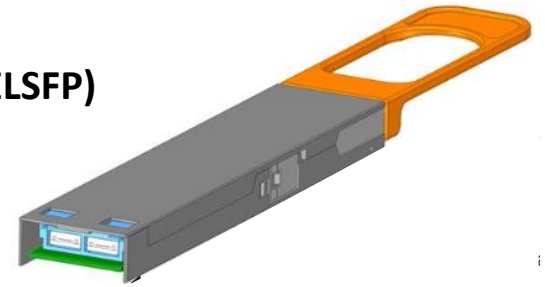


✓ 3.2T Co-packaged Optical Engine

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.)

OIF | ACCELERATING MARKET ADOPTION OF OPTICAL NETWORKING TECHNOLOGIES
PROJECT HIGHLIGHTS 2024

150+ Member Companies → Identifies Industry Needs and Gaps → Develops Implementation Agreements (specifications) → Performs Interoperability Demonstrations

OPTICAL

Multi-vendor Interoperability in Client Form Factors

1600ZR+
- <1000km multispan Coherent DWDM

1600ZR, 800ZR, 400ZR
- >80km Coherent DWDM

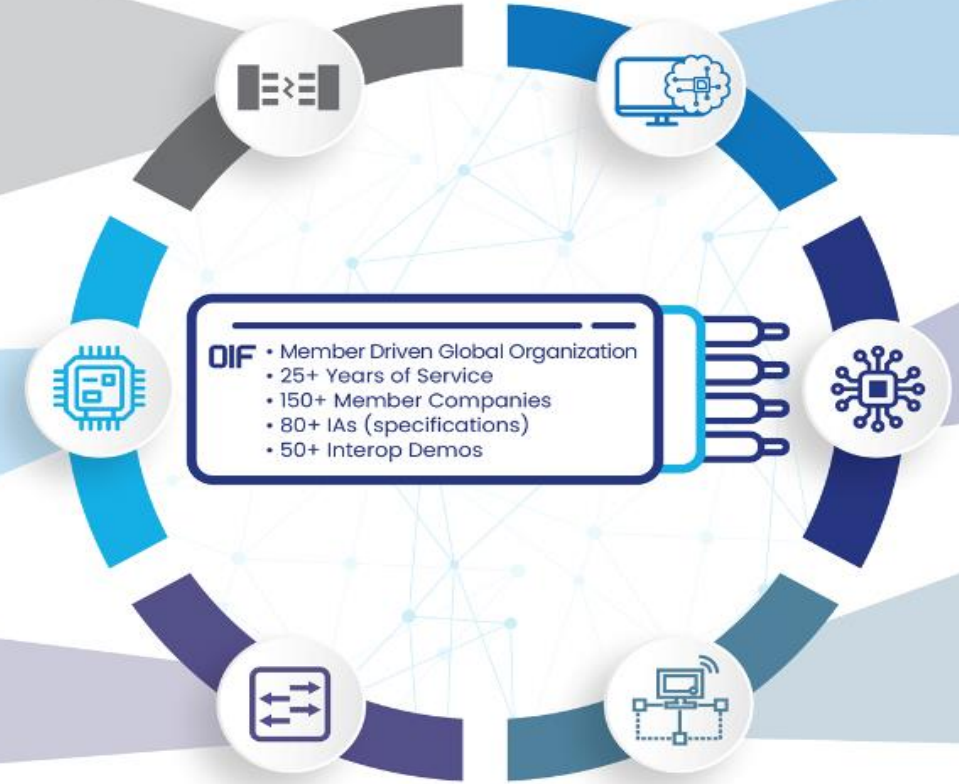
800LR
- <10km Coherent Point-to-Point

ENERGY EFFICIENT INTERFACES

- **Next Generation Low Latency Interfaces for AI/MI & Data Centers**
- **Co-Packaged Modules**
- **External Laser Sources**

PROTOCOL

FlexE
- More Efficient
- Agile Networking



MANAGEMENT

Common Management Interface Specification (CMIS) and Coherent CMIS
- Common
- Flexible
- Extendable

ELECTRICAL

Common Electrical I/O (CEI)
- High-Speed Building Blocks
- 224G, 112G, 56G, 28G
- Protocol Agnostic Link Training

NETWORKING

Transport SDN APIs
- Automation, Programmability

Enhanced Network Operations
- Artificial Intelligence
- Digital Twin
- Awareness Between Application and Optical Layers

Thank you

www.oiforum.com