

OPEN
Compute Project



SHORT REACH OPTICAL INTERFACES

1.0.0

Authors:

Bobby Lu, editor, Lightelligence

Bijan Nowroozi, section 1, Lightmatter

Jeff Hutchins, sections 2 & 7, Ranovus

David Piehler and Jian Chen, section 9, Dell

Karen Liu, section 6, Ciena

Kazuya Anazawa, section 4.6.2, NTT

Rang-Chen Yu, section 8, TeraHop

Raymond Law, sections 5 & 10, LOTES

Ron Swartzentruber, sections 1.1 & 11, Appendix A, editor, AMD

Tamal Das, section 4, Intel

Date: August 5, 2026



This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Table of Contents

1 Introduction 5

1.1 Objectives 6

2 Defining Short Reach Optical Interconnect 8

2.1 SROI as a Common Optical Physical Layer 9

3 The Unique Challenges for Optical Scale-up Networks 11

3.1 The need for radix 11

3.2 Interoperability 13

3.3 Power 13

4 The Anatomy of Modern AI Fabric 16

4.1 The Solution: Multi-Dimensional Parallelism 20

4.1.1 The Parallelism Tax: Network-Centric Reality 22

4.2 Scale-Up Domain: Interconnects within a Pod 23

4.3 Scale-out: Interconnecting multiple Pods 28

4.4 Scale-Up and Scale-out: I/O Coherency with Memory Semantics 31

4.5 The Scale-Up Bottleneck: The Epicenter of Interconnect Challenge 33

4.6 Scale-Up Switching 33

4.6.1 Electrical Switching 34

4.6.2 Optical Circuit Switching (OCS) 34

5 Copper Implementation 39

5.1 Data Rate and Insertion Loss Summary 39

5.2 Channel Topology 40

5.3 Optical Channel Topology 48

6 Optical Key Metrics 52

6.1 Targets for Top-Level Metrics 52

6.1.1 Power 53

6.1.2 Cost 54

Date: August 5, 2026

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

6.1.3 Hardware Reliability 55
6.1.4 Latency 56
6.2 Technology Metrics 56
6.2.1 Interconnect Radix 56
6.2.2 Linear and Areal BW Density at XPU 57
6.2.3 Areal Density at Faceplate 58
7 Reliability 60
7.1 Link reliability 60
7.2 Hardware reliability 61
7.2.1 Naively replacing standard pluggable optics for passive copper links 61
7.2.2 Hardware reliability regimes 61
7.2.3 Repair 63
7.2.4 Importance of diagnostics 63
7.2.5 Resiliency 64
7.2.6 The importance of operating temperature 64
7.2.7 Lasers 64
7.2.8 Qualification Regimen 65
8 Scale-up Optical Transceiver Options 66
9 Looking Forward: IMDD Optical Transmission at 400G 73
9.1 Optical impairments 73
9.2 Materials issues 74
9.3 Emerging Alternatives 75
10 Looking Forward: Copper Transmission at 400G 76
10.1 Co-Packaged Copper Topology Length and Insertion Loss at 400G 76
10.2 Summary 77
11 Conclusion 79
12 Glossary 82

Date: August 5, 2026

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

13 References 86
14 License 90
15. About Open Compute Foundation 91
Appendix A. SROI Requirements Summary 92

Date: August 5, 2026

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

1 Introduction

The rapid scaling of AI training systems has created a new class of interconnect challenge. As accelerator clusters expand beyond single-rack deployments into multi-rack Pods spanning up to 30 meters or more, passive copper interconnects reach their physical and economic limits. The scale-up domain, where tightly coupled accelerators exchange memory-semantic traffic at extreme bandwidth and low latency, must transition to optical interconnects.

However, optical technologies developed for traditional datacenter networking are not directly suited to this regime. Short Reach Optical Interconnects (SROI) for scale-up applications operate in a fundamentally different environment from data center optics in several aspects:

- Shorter interconnect distances (up to 30 meters)
- Support high radix
- Cost sensitive environment accustomed to inexpensive copper interconnects
- High reliability expectations (both the hardware and the link reliability) compared to traditional data center optics
- Lower power utilization and lower round trip latencies realized through architectures driving improved energy efficiency

This white paper, developed within the Open Compute Project (OCP), the open systems community where hyperscalers and their ecosystem partners collaborate on hardware design, defines the SROI problem space for scale-up interconnects and establishes system-level recommendations for optical adoption in scale-up fabrics. OCP is uniquely positioned to lead this effort because its members design, build, and operate AI infrastructure at the scale where these challenges are most acute.

In modern AI systems, the scale-up network is the primary performance bottleneck: any bandwidth or latency limitation leaves high-value accelerators underutilized, directly increasing training time and infrastructure cost. Current copper implementations are limited to approximately 1–1.5 meters at lane rates of 200 Gb/s. Next-generation Pods containing 500 to 1,000 or more accelerators will require multi-rack connectivity, with tens of thousands of optical lanes per rack. These lane counts impose cost, power, and reliability constraints significantly more stringent than those of conventional pluggable optics.

Date: August 5, 2026

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

This paper analyzes the system drivers for SROI adoption, compares copper and optical topologies, defines key performance and reliability metrics, evaluates module placement strategies, and examines the optical transmission landscape at 400G per lane. The findings indicate that successful SROI deployment requires convergence toward power efficiency expressed in pJ/b in single digits per link end, cost trending towards passive copper on a per-bit basis evaluated at total system level, and reliability approaching silicon-grade FIT rates. Meeting these targets will require purpose-built, highly integrated optical solutions rather than incremental adaptations of existing datacenter transceivers. The pace of AI infrastructure investment, with new accelerator generations arriving every 12 to 18 months and system design cycles compressing accordingly, means the window for defining these requirements is narrow. This paper aims to provide system-level clarity that the broader standards and supply chain ecosystem needs to respond.

1.1 Objectives

This white paper has three primary objectives.

1. Define the SROI problem space

The scale-up interconnect domain presents a set of requirements that are not adequately captured by existing optical reach classifications. Current IEEE and OIF optical specifications are optimized for channel impairments, namely fiber attenuation, chromatic dispersion, and optical signal-to-noise ratio, which are negligible at the distances relevant to intra-rack and inter-rack AI pod connectivity. Equally, the cost, power, density, and reliability constraints imposed by scale-up fabrics differ materially from those of conventional datacenter optics. This paper defines Short Reach Optical Interconnect as a distinct optical reach regime, characterizes the system environment in which it operates, and establishes why existing specifications are insufficient to address it. The intended audience includes system architects, optical module vendors, semiconductor designers, and standards contributors who require a shared technical foundation.

2. Establish system-level requirements for optical adoption in scale-up fabrics

Optical interconnect cannot be evaluated in isolation. The relevant performance boundaries, namely energy per bit, cost per lane, hardware reliability, and link latency are determined by the system context such as lane count per rack, memory-semantic traffic characteristics, failure sensitivity of tightly coupled compute pods, and the economic

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

baseline set by passive copper. This paper translates those system-level constraints into quantitative targets and technology metrics for optical solutions. It further analyzes copper channel topology to establish the physical reach limits that define when optical interconnect becomes architecturally necessary rather than optional and evaluates module placement strategies across the spectrum from pluggable to co-packaged implementations.

3. Provide the technical foundation for coordinated standards development.

The AI infrastructure ecosystem is advancing faster than the standards and supply chain processes that would ordinarily govern a technology transition of this scale. Accelerator bandwidth is doubling on one- to two-year cycles; formal standards processes operate on longer timelines. Without early convergence on system-level objectives, the risk is that fragmented, protocol-specific optical implementations will prevent the volume aggregation necessary to drive cost reduction and reliability maturation. This paper is intended to give standards organizations including IEEE 802.3, OIF, and JEDEC, as well as multi-source agreement bodies the system-level input required to scope and sequence their work. This document is not a specification. It does not define electrical or optical parameters at the transceiver level, nor does it mandate physical implementation choices. It defines the boundaries within which those specifications must operate to serve the scale-up use case.

Taken together, these objectives reflect OCP's position as the community where the operators of AI infrastructure at the largest scale articulate requirements to the broader ecosystem. The window for establishing this framework is narrow. The conclusions and recommendations in this paper are offered as a structured starting point for that process.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

2 Defining Short Reach Optical Interconnect

The electrical interconnect ecosystem has long recognized that physical reach class determines architecture. JEDEC, IEEE 802.3, and OIF specifications establish channel classes in which reach defines SerDes architecture, equalization strategy, power budget, and packaging model.

At the chip-to-chip scale, extra-short-reach (XSR) and very-short-reach (VSR) electrical interfaces as defined at OIF, operate across millimeters to centimeters with minimal equalization and low power. UCIe™ extends this model to die-to-die interconnect within a package or across an interposer.

At longer electrical reaches, short-reach (SR) and medium-reach (MR) channel classes introduce progressively greater equalization and signal conditioning as insertion loss increases. PCIe™ Gen6 and Gen7 and the CXL™ reflect these reach-dependent tradeoffs.

For optical links, IEEE 802.3 defines objective reach classes such as DR and FR for 500 meters, 2 kilometers, and beyond. These specifications are optimized to address fiber attenuation, chromatic dispersion, and optical signal-to-noise ratio over those distances. Their architectural complexity reflects those channel impairments.

This reach taxonomy exists because each regime is governed by different dominant impairments. Designing for the wrong impairment increases power, cost, and implementation complexity without system benefit. Electrical and optical standards have converged on reach-defined optimization for this reason.

Short Reach Optical Interconnect, as defined in this paper, represents a system-defined optical reach regime operating up to 30 meters, typical of intra-rack and inter-rack connectivity within an AI Pod. At these distances, single-mode fiber insertion loss is minimal and chromatic dispersion is negligible. Multimode fiber at these distances similarly presents no meaningful modal limitation at current and projected lane rates.

In this regime, the optical channel is not the dominant constraint.
The primary design objectives shift to endpoint integration parameters:

- Optical lane density at the package boundary or faceplate
- Energy efficiency per bit at massive aggregate bandwidth
- Cost per lane at rack-scale volumes

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

- Reliability consistent with synchronous, memory-semantic compute fabrics

These objectives differ materially from those that drive medium-reach and long-reach optical specifications. SROI links may leverage adaptations of existing datacenter optics, but the governing optimization objectives differ fundamentally from medium and long-reach designs. It is a reach-defined regime analogous to XSR and VSR in the electrical domain, where integration efficiency rather than channel compensation governs architecture.

Electrical protocol evolution illustrates the same principle. PCIe and CXL extend memory-semantic traffic across longer electrical reaches through retimers and additional equalization, but with increasing impact on power, latency, and reliability. Similar architectural choices apply when these protocols are extended over optical media. Retiming topology, FEC strategy, and signal conditioning approach directly influence power, latency budget, and system-level availability.

Emerging scale-up protocols such as UALink™ and ESUN, which build on IEEE 802.3dj 212.5 Gbps electrical specifications, further reinforce the need for a reach-aligned optical physical layer. The optical interface must preserve protocol latency and error objectives without inheriting unnecessary complexity from longer-reach optical designs.

Understanding SROI as a distinct optical reach regime, rather than as a reduced-distance implementation of DR or FR classes, is central to this white paper. Electrical reach classes evolved over decades. AI infrastructure scaling is compressing that timeline. Accelerator bandwidth is doubling on one- to two-year cycles, and system design windows are correspondingly short. A clear reach taxonomy for short optical links is required to align standards development, supply chain investment, and system architecture decisions.

2.1 SROI as a Common Optical Physical Layer

Scale-up accelerator interconnect is the most immediate driver of SROI, but it is not the only one. Multiple protocol families operating within the rack are approaching copper's practical limits at current and projected lane rates.

These include:

- Scale-up fabrics such as UALink and ESUN
- PCIe Gen6 and Gen7 attachment
- CXL-based memory expansion and pooling
- NVMe™ over PCIe or CXL fabrics

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Each protocol has distinct transaction semantics and topology. All share a common physical constraint: high bandwidth at low latency across centimeter-to-meter distances where copper channel loss and equalization complexity scale unfavorably with baud rate.

Short reach fiber transport is protocol-agnostic. The optical medium carries modulated symbols independent of upper-layer semantics. Protocol differentiation remains in the electrical domain at the endpoints. The optical physical layer must meet defined objectives for lane rate, energy efficiency, density, and reliability, but it need not be protocol-specific.

Defining SROI as a reach-based optical physical layer enables volume aggregation across protocol families. Aggregated volume drives cost reduction and accelerates reliability maturation. Fragmented, protocol-specific optical implementations would have the opposite effect.

This convergence also supports emerging disaggregated architectures. Rack-scale memory pooling, composable I/O, and distributed storage fabrics all benefit from a unified short-reach optical infrastructure. At shorter distances, in-package optical I/O can be viewed as the extreme short tier of the same reach-defined regime, where integration density and energy efficiency dominate.

For OCP, the objective is to define SROI recommendations at a level that supports multi-protocol use while remaining aligned with existing IEEE and OIF electrical specifications. A reach-aligned optical framework reduces ecosystem fragmentation and enables coordinated advancement of optical interfaces for AI infrastructure.

SROI is therefore best understood as a reach-defined optical physical layer serving multiple electrical protocol families within the rack-scale domain.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

3 The Unique Challenges for Optical Scale-up Networks

The requirements for Short Reach Optical Interconnects (SROI) used as scale-up networks differ from the traditional Ethernet front-end networking in several ways.

3.1 The need for radix

While front-end Ethernet favors “fat” pipes, scale-up links prioritize radix and the need to interconnect every accelerator with every switch as shown in Figure 3-1. The “orange” scale-up links provide this full mesh connectivity.

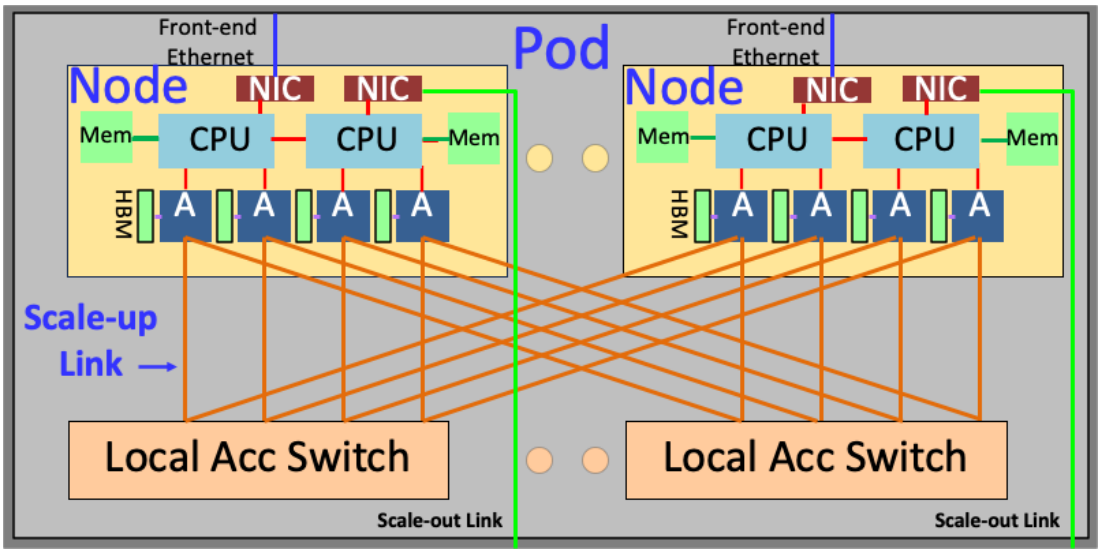


Figure 3-1- - Scale-up mesh connectivity

The radix of a Pod architecture is another important consideration and determines the Pod size. Important design considerations are:

- The feasible I/O bandwidth for the accelerator and for the switch IC
- The aggregate link rate which is needed per I/O port on the accelerator IC and switch IC

For example, the current UALink_200 1.0 supports several different link capacities per UALink station with lanes operating at 212.5Gbps.

- A x1 link with 212.5G link bandwidth over 1 lane

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

- A x2 link with 425G aggregate link bandwidth over 2 lanes
- A x4 link with 850G aggregate link bandwidth over 4 lanes

Full mesh connectivity requires each accelerator to connect to all the switches. The number of accelerators in a Pod is therefore determined by the switch I/O bandwidth and the link bandwidth:

$$\#Accelerators_{POD} = \frac{Bandwidth_{Switch}}{Bandwidth_{Link}}$$

Allocating a small bandwidth to each link results in slow accelerator-to-accelerator communication but supports a large number of accelerators in a Pod. Alternatively, allocating a large link bandwidth constrains the number of accelerators that can be interconnected in a Pod.

The hypothetical example shown in Figure 3-2 shows the accelerator count for various link rates for various switch I/O bandwidths.

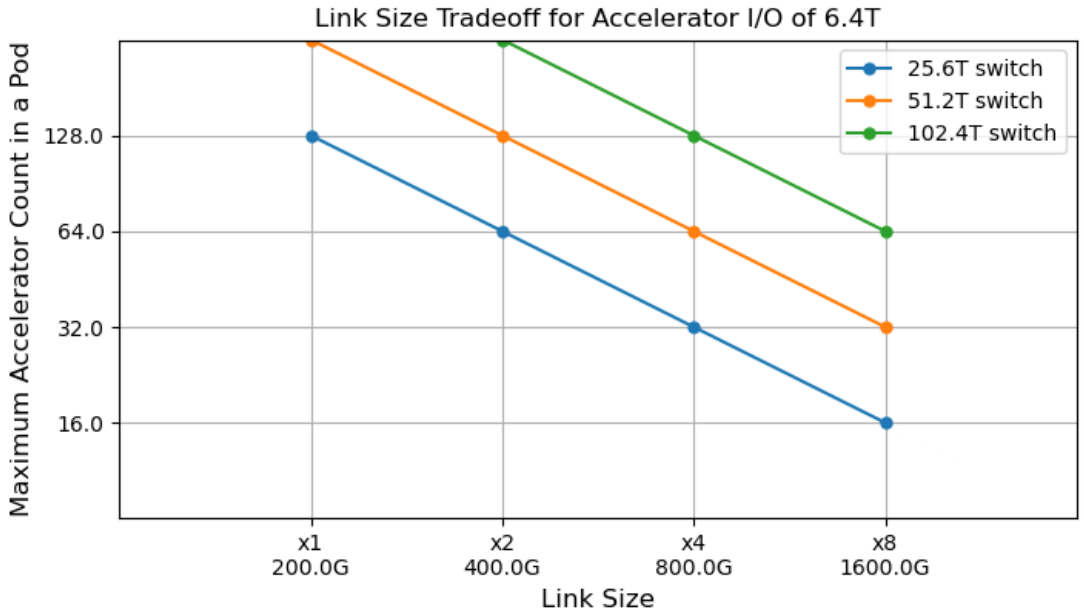


Figure 3-2 - Link size tradeoff against Pod size

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

3.2 Interoperability

The scale-up links need compatible interfaces on either end of an optical link which requires the optical interface on the accelerator to match the optical interface on the switch. Some implementations may prefer an optical standard such as specified by the IEEE 802.3's 800GBASE-FR4 or 800GBASE-DR4 specifications. However, given the number of interfaces and fiber infrastructure required, some implementors may decide that a non-standard optical interface provides better performance and cost (such as power reduction and/or fiber count reduction).

3.3 Power

Current accelerator Pods use passive copper links, but these are for short links on the order of a 1 m. Longer links may employ optics. There are three types of retiming architectures used in optical transceivers: non-retimed, tx-retimed, and fully-retimed. Increasing retiming capability to the optical transceiver increases its power consumption. Electrical channels with lower losses enable the use of lower power interfaces.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

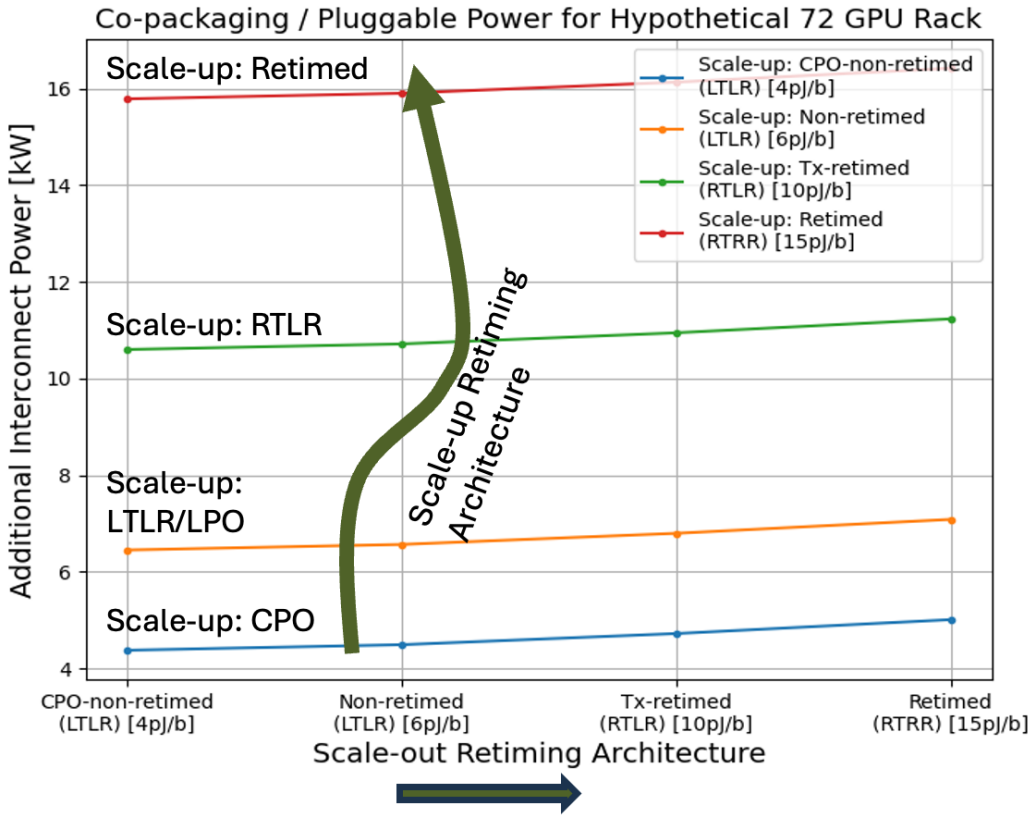


Figure 3-3 - Power consumption example for an accelerator with optical I/O

Figure 3-3 shows an example power consumption estimate for a 72-accelerator system with 7.2 Tb/s I/O bandwidth on each accelerator with each of the links supporting 400G total bandwidth in either direction. Typically, the scale-up bandwidth is a factor of ten greater than the scale-out bandwidth. The vertical axis shows the total interconnect power for such as system in kilowatts. The horizontal axis shows four different interface retiming options for the scale-out interface. There are four curves on the plot with each one representing a different retiming choice for the scale-up interfaces.

Note that in Figure 3-3 the curves are relatively flat from left to right indicating a relatively small increase in power as the scale-out interface changes from non-retimed to fully retimed. However, notice the large increase in power between the scale-up interfaces shown by the four colored lines. The horizontal movement along each curve is modest (the

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

scale-out retiming choice matters less), whereas the vertical gap between curves is large (the scale-up retiming choice dominates total power).

The architect has a choice to prioritize the ease of repair, or lower power consumption, or link robustness.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

4 The Anatomy of Modern AI Fabric

The modern AI landscape is defined by a relentless, exponential growth in computational demand. Research shows that the training compute for frontier models has grown at a staggering rate of $4.5\times$ per year, doubling roughly every 5-6 months, [1]. The architecture of high-performance computing has been fundamentally reshaped by a singular force: the explosive growth of AI models. This trend, driven by the advent of the parallel-friendly Transformer architecture, has enabled an explosion in model size, but it has also fundamentally reshaped data center design. We have moved from general-purpose data centers to highly specialized AI factories. As models have scaled from millions to trillions of parameters, the industry has collided with immense barriers that make distributed supercomputing an absolute necessity.

Table 4-1- Estimated Training Memory Requirement for well-known AI models

Model	Developer	Reported/Estimated Parameters	Training Memory (FP16/BF16)	Training Tokens	Training FLOPs	Training Cluster Configuration
GPT-3Dense [6]	OpenAI	175 B	~1.5 TB	300B	3.14×10^23	10K V100 GPUs
GPT-4MoE [7]	OpenAI	1.8 T	~15 TB	13T	2.3×10^25	25K A100 GPUs
PaLMDense [8]	Google	540B	~4.5 TB	780B	3.4×10^24	6K TPU V4
Llama-3.1Dense [2]	Meta	405 B	~4 TB	15T	3.8×10^25	16K H100 GPUs
Grok-3MoE	xAI	2.7 T	~22 TB	13T	9.0×10^24	[1]200K H100 GPUs
DeepSeek-V3MoE [4]	DeepSeek AI	671 B	~5.6 TB	15T	4.4×10^24	2K H800 GPUs
Pangu-UltraMoE [5]	Huawei	718 B	~6 TB	13T	4.1×10^24	6K Ascend 910B

[1] *xAI's Colossus cluster is planned to expand to 200K, the initial training cluster was widely reported as 100K H100 GPUs.*

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Training a state-of-the-art foundation model is an astronomical undertaking, demanding quintillions of floating-point operations (FLOPs). A model like GPT-3, for instance, is estimated to require on the order of 3.14×10^{23} FLOPs [6]. The impracticality of this task for a single machine is staggering; a top-tier NVIDIA H100 GPU delivers approximately 2×10^{15} FLOPs per second [9], meaning it would take about five years of non-stop computation to complete the task with 100% utilization. This makes it clear that distributing the workload across thousands of accelerators is essential simply to shrink training times from years to a practical timeframe of weeks or months.

Table 4-2 - Well-known AI Accelerators and their Capability, [9], [10], [11], [12]

Parameter	**Unit**	**A100**	**H100**	**B200**	**B300**	**VR200***	**TPUV5p**	**MI300X**	**MI355X**	**Ascend 910B**
Tensor Arch	-	Ampere	Hopper	Blackwell	Blackwell Ultra	Rubin	V5	CDNA 3	CDNA 4	Ascend 910B
BF16 Core dense	TFLOPs	312	1000	2500	2500	-	-	1300	2500	~320-400
BF16 Core Sparse	TFLOPs	624	2000	5000	5000	-	459	2600	5000	-
Power	W	400	700	1200	1400	1800	-	750	1400	310[2]

Even more prohibitive than the compute requirement is the "Memory Wall," a twofold crisis of both capacity and bandwidth. As models grow, the memory required to store their state—the weights, gradients, and optimizer data—far exceeds the capacity of any single AI accelerator. A 1-trillion parameter model can demand over 10 TB of memory, dwarfing the 80-288 GB of on-package High-Bandwidth Memory (HBM) on even the most advanced GPUs. This capacity crisis makes it physically impossible to load a frontier model onto a single device. Furthermore, performance is constrained by the "Bandwidth Wall," as the

[2] Power varies by product implementation and should be verified against the cited Ascend 910B hardware configuration.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

compute cores in an accelerator are so fast they frequently sit idle, waiting for data to be loaded from HBM. This "memory-bound" condition means that an accelerator's real-world performance hits a wall long before its theoretical peak FLOPs can be reached. Table 4-3 gives a glimpse of available HBM memory capacity and bandwidth per accelerator chip in well-known XPUs and these are way beyond memory wall.

Table 4-3 - Available High Bandwidth Memory in well-known Accelerators, [9], [10], [11], [12]

Parameter	**Unit**	**A100**	**H100**	**B200**	**B300***	**VR200***	**TPUV5p**	**MI300X**	**MI355X**	**Ascend 910B**
Spec.	-	HBM2	HBM3	HBM3e	HBM3e	HBM4	HBM2e	HBM3	HBM3e	HBM2e
HBM Stacks	-	5	5	8	8	8	-	8	8	8
Data Rate	Gb/s	2.43	5.2	8	8	6.4	3.27	5.2	8	3.27
Total Memory	GB	40	80	192	288	288	95	192	288	64
Total BW	TB/s	1.52	3.3	8	8	12.8	2.7	5.3	8	3.2

To overcome these challenges, the industry has converged on a new fundamental unit of computing: the Pod and the Node. Unlike general-purpose servers, a Node is a dense, highly specialized system designed for massively parallel computation. A typical high-end node contains a host CPU and system RAM, but its core consists of four to eight powerful AI accelerators, such as GPUs or TPUs as depicted in Figure 4-1. These accelerators are tightly coupled within the chassis by an ultra-high-speed interconnect, creating a potent, self-contained computational unit. Examples of scale-up interconnects include NVLink™, UALink™, and ESUN.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

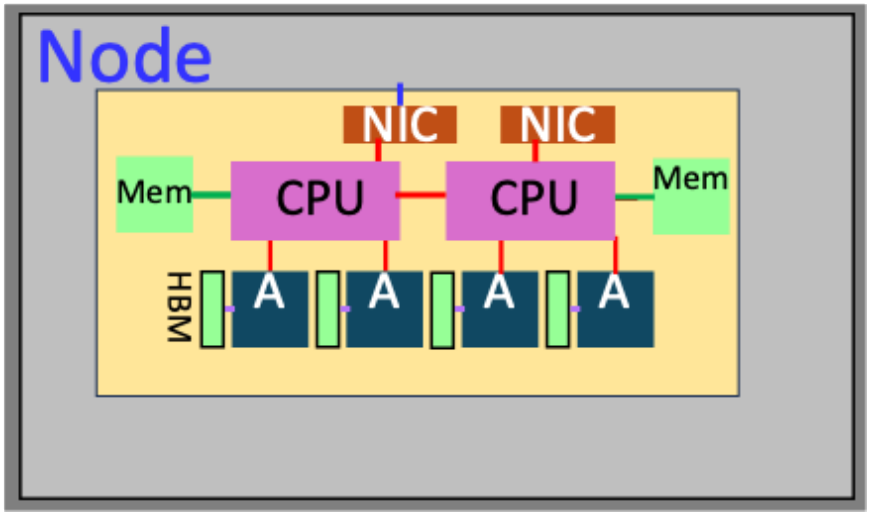


Figure 4-1 - Basic Compute Unit: Node

As model sizes increase, the only viable path forward is to distribute the AI model and its workload across a vast cluster of these nodes. The communication network that links accelerators is no longer a secondary component but a first-order performance bottleneck, making distributed training across thousands of connected accelerators the defining paradigm of modern AI. To train a single, massive model within this paradigm, the workload must be "fractured" across these accelerators. This is achieved through a multi-dimensional approach to parallelism, where modern training frameworks orchestrate a complex interplay of tensor, pipeline, and data parallelism—often augmented with specialized techniques like expert parallelism for Mixture-of-Experts (MoE) models [13], [14].

Configuring this hybrid strategy is a sophisticated optimization problem, requiring architects to meticulously balance computational load, memory constraints, and the ever-present bottleneck of inter-processor communication. The effectiveness of this entire distributed system hinges entirely on the performance of the underlying hardware fabric. Therefore, to fully grasp how these critical communication challenges are being addressed, it is essential to deconstruct the modern AI cluster into its distinct network domains. This analysis will isolate the most critical bottlenecks and establish the context for the advanced short-reach optical interconnects that are vital to the future of AI.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

4.1 The Solution: Multi-Dimensional Parallelism

To overcome the compute and memory walls, a single massive model must be "fractured" across thousands of accelerators using a multi-dimensional approach to parallelism. This orchestration of different strategies places extreme and varied demands on the underlying network infrastructure that connects the accelerators. These specialized networks, or network fabrics, must be architected to handle the unique traffic patterns generated by each parallelism technique. A key characteristic of this traffic is the creation of large, sustained data transfers known as "elephant flows," which can saturate network links and must be managed efficiently. These are characterized as large, sustained, high-bandwidth data transfers between AI accelerators. The baseline for large-scale training involves combining three core strategies:

- Data Parallelism (DP): The simplest approach, where the model is replicated on each accelerator, which processes a different slice of the data. Gradients are synchronized across all accelerators using an All-Reduce communication collective. To overcome the memory limitations of simple replication, Fully Sharded Data Parallelism (FSDP), [14], is now standard. FSDP shards the model parameters, gradients, and optimizer states across the data-parallel group, drastically reducing the per-GPU memory footprint at the cost of more frequent communication (All-Gather and Reduce-Scatter).
- Pipeline Parallelism (PP): This strategy addresses model depth by splitting the layers into sequential stages, with each stage running on a different set of accelerators. Data flows through this "assembly line," allowing for the training of models that are far too deep to fit on a single device.
- Tensor Parallelism (TP): This strategy tackles model width by splitting the mathematical operations within a single layer, [13]. For example, a large weight matrix is split column-wise and row-wise across multiple accelerators, which then work in concert to compute the full operation. This requires frequent, low-latency communication within a tightly coupled group of GPUs.

As model architecture has evolved to handle longer contexts and greater parameter counts more efficiently, new dimensions of parallelism have emerged:

- Expert Parallelism (EP): This is essential for Mixture-of-Experts (MoE) models, used by state-of-the-art models from Llama and DeepSeek. In MoE, only a fraction of the model's "expert" sub-networks are activated for any given token. EP distributes these

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

experts across different accelerators. This creates a challenging, sparse All-to-All communication pattern as the network must dynamically route tokens from their source accelerator to the specific accelerator housing the chosen expert.

- Context Parallelism (CP): The self-attention mechanism's memory requirements scale quadratically with sequence length, creating a bottleneck for long-context models. Context Parallelism, used in the training of Llama-3, addresses this by sharding the activations along the sequence length dimension. This allows models to process extremely long contexts (e.g., 128k tokens or more) by distributing the activation memory load across accelerators.

Table 4-4 - Communication Pattern during Training, [13], [14]

Parallelism Type	Granularity	Communication Pattern	Resulting Network Traffic	Key Bottleneck	Ideal Use Case
DP	Model Replica (Sharded)	All-Gather / Reduce-Scatter	Sustained, high-bandwidth 'Elephant Flows'	Network Bisectional Bandwidth	Scaling training throughput across many nodes.
TP	Tensor Slice	All-Reduce / All-Gather	Sustained, high-bandwidth 'Elephant Flows'	Intra-Node Interconnect Latency	Fitting very large individual layers (high hidden size) into GPU memory.
PP	Layer Group	Point-to-Point	Sustained, high-bandwidth 'Elephant Flows'	Pipeline Bubble / Inter-Stage Bandwidth	Fitting very deep models (many layers) that cannot fit on a single node.
EP	Expert Network	All-to-All	Sustained, high-bandwidth 'Elephant Flows'	All-to-All Latency & Bandwidth	Scaling Mixture-of-Experts (MoE) models with massive parameter counts.

The combination of these strategies results in the 3D or 4D Parallelism used to train today's largest models. Table 4-4 summarizes these patterns and the expectations from the

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

network. For instance, the training of Llama-3 405B reportedly utilized a 4D scheme with degrees of 8 for TP, 16 for PP, 8 for DP, and 16 for CP, [2]. This complex orchestration of different communication patterns—All-Reduce, All-to-All, All-Gather, and point-to-point—places extreme demands on the network.

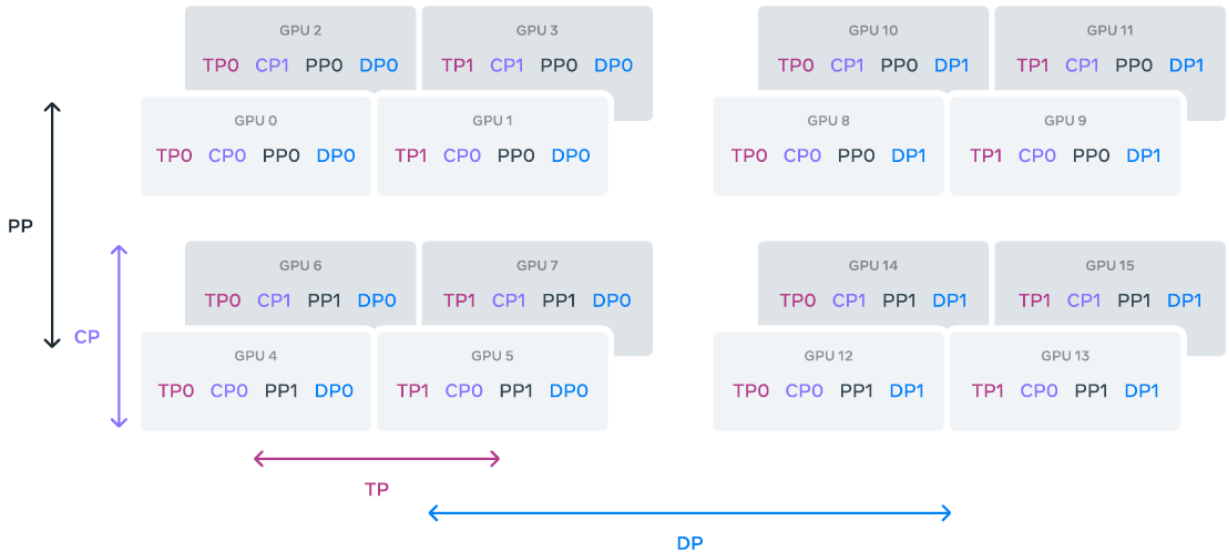


Figure 4-2 - Illustration of 4D parallelism in Llama3

4.1.1 The Parallelism Tax: Network-Centric Reality

This reliance on multi-dimensional parallelism fundamentally transforms the nature of data center traffic. Unlike traditional workloads dominated by "north-south" traffic (client to server), AI factories are defined by a massive, sustained flood of "east-west" traffic between

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

the accelerators themselves. The collective communication patterns described previously manifest as large, synchronized "elephant flows" that require a specialized, lossless, high-bandwidth, and low-latency backend AI fabric. These elephant flows are predominantly RDMA (Remote Direct Memory Access) traffic, arising from the necessity to transfer significant memory chunks—such as model gradients or activations—between XPUs for synchronization. RDMA allows one computer to access the memory of another directly without involving the operating system or CPU of either machine, significantly reducing latency and making it ideal for high-performance inter-XPU communication.

A critical characteristic of these RDMA-based elephant flows is their inherently low entropy. Packet headers (source/destination IP, ports, protocol) often remain constant for the duration of these large transfers. Traditional load balancing mechanisms in switches, which rely on hashing these header fields to distribute traffic across multiple paths, perform poorly with such low-entropy traffic. The hash function consistently maps these flows to the same link, leading to severe network imbalances where some links become heavily congested while others remain underutilized. This inefficiency directly impacts training time and XPU utilization, necessitating more sophisticated network designs. To handle this "parallelism tax," the backend AI fabric has evolved into two distinct but interdependent network domains, defined by their approach to memory access and the specific communication patterns they serve.

4.2 Scale-Up Domain: Interconnects within a Pod

The scale-up fabric is the ultra-high-speed network that connects a group of accelerators organized into multiple nodes into a single, cohesive compute unit called a Pod and provides support for load/store memory semantics. Today the scale-up domain is confined to a rack as shown in Figure 4-3 where the “orange” links represent the backplane’s scale-up connectivity. These “orange” links connect accelerators to a scale-up switch located in the mid-rack region and employ copper links with limited reach (~ 1 m).

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

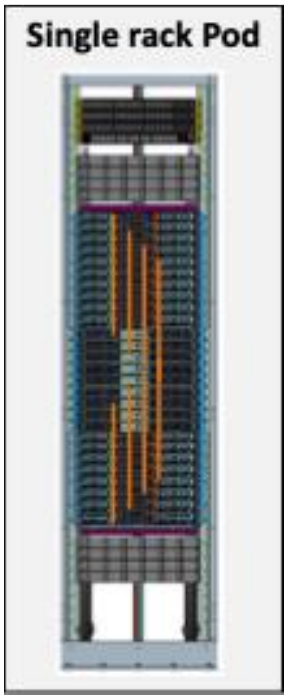


Figure 4-3 - Single Rack Pod

Creating larger Pods will require longer links connecting nearby racks as shown in Figure 4-4 by the “orange” scale-up links. These links will need a reach longer than the current ~1 m used by today’s single rack Pod.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

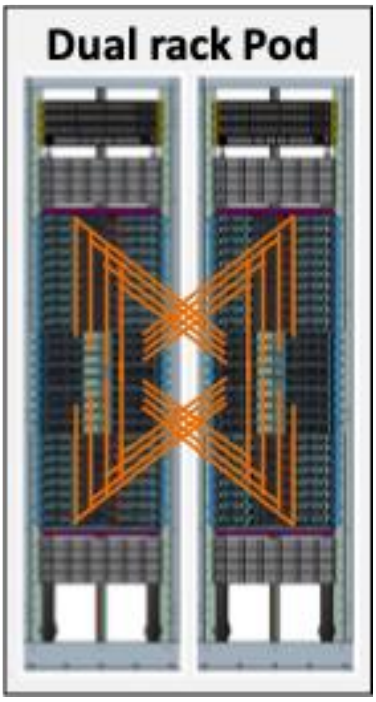


Figure 4-4 - Dual Rack Pod

Figure 4-5 shows a Pod with multiple nodes interconnected with scale-up links. The accelerators (A) in the Pod are interconnected all-to-all by a high-speed, low-latency scale-up connectivity, which allows them to function as a single, coherent computational unit.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

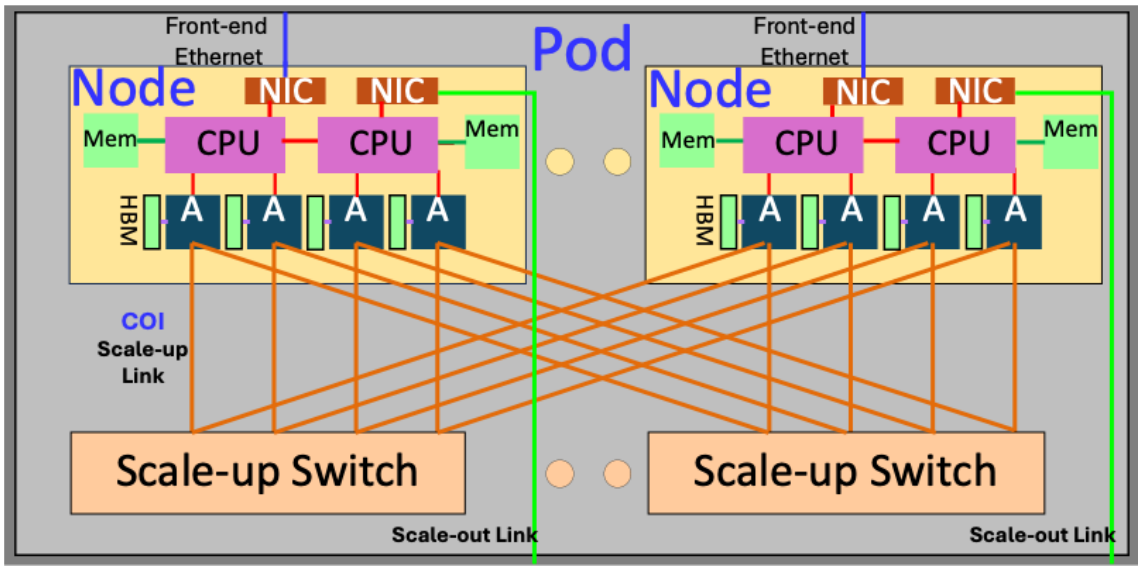


Figure 4-5 - Scaling AI: Pod containing multiple nodes interconnected with scale-up links

A brief survey on scale-up network fabric in well-known AI servers is presented in Table 4-5. Architects employ several physical topologies to build these multi-chassis scale-up fabrics:

- Switch-Based Architectures: The most common approach involves using high-radix, specialized switches to create a non-blocking Clos network. This allows any GPU in the cluster to communicate with any other GPU through a scale up switch (Figure 4-5).
- Direct-Connect Meshes: An alternative is to connect nodes directly to their neighbors without a central switch. A 3D-Torus topology, used in Google's TPU pods, arranges nodes in a three-dimensional grid, which is highly efficient for workloads with strong communication locality.
- Hybrid Topologies: scale-up systems may combine multiple topology types—such as fully connected, switched-fabric, ring, or mesh structures—to balance bandwidth, latency, scalability, and implementation cost. The specific arrangement depends on factors including the number of XPUs, required communication patterns, link bandwidth, and physical constraints.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Regardless of the topology, the defining characteristic of the scale-up fabric is its support for memory-semantic communication, enabled by protocols like NVLink and the open Ultra Accelerator Link (UALink) standard. These allow an accelerator to perform atomic memory I/O coherent operations (loads, stores) directly with the memory of another accelerator, which is essential for fine-grained model parallelism techniques like Tensor and Pipeline Parallelism to operate efficiently across the entire cluster.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Table 4-5 - Scale-Up Specification in well-known Accelerators, [15]

	Parameter							
Example	**Protocol**	**Line Rate**	**Uni BW/XPU**	**#Links /XPU**	**AI Cluster**	**N/W Topology**	**Scale-Up Switch**	**Switch Ports**
Unit	-	Gb/s	GB/s	-		-	-	-
A100	NVLink3	50	300	12	DGX A100	3D-Mesh	NVSwitch2	36
H100	NVLink4	100	450	18	DGX H100	Clos-1	NVSwitch3	64
B200	NVLink5	200	900	18	NVL72	Clos-1	NVSwitch4	72
B300	NVLink5	200	900	18	NVL72	Clos-1	NVSwitch4	72
VR200	NVLink6	400	1800	18	NVL144	Clos-1	NVSwitch5	72
TPUV5p	ICI	-	600	6	TPU POD	3D-Torus	OCS	-
MI300X	Infinity Fabric	32	448	7	MI300X	Dragonfly	Direct Connect	NA
MI355X	Infinity Fabric	38.4	537.6	7	MI355X	Dragonfly	Direct Connect	NA
Ascend 910B	Huawei HCCS	-	350	3	Atlas900	Fat-Tree	CloudEngine	672

4.3 Scale-out: Interconnecting multiple Pods

The scale-out fabric connects multiple independent scale-up nodes or pods together to form a massive, distributed training cluster as shown in Figure 4-6.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

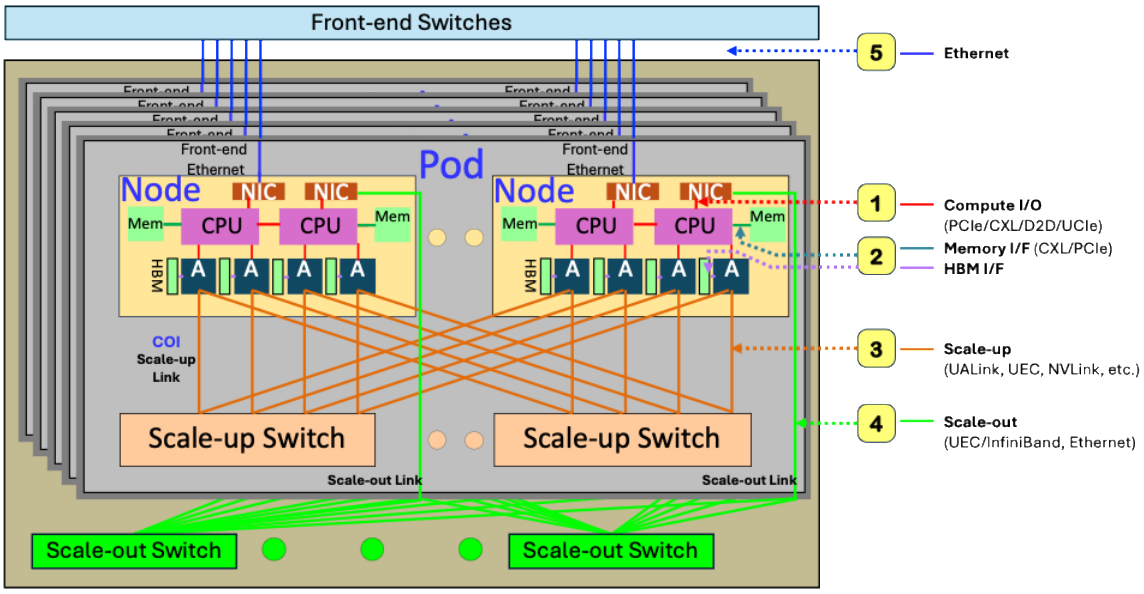


Figure 4-6 – Interconnecting AI Pods

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Table 4-6-Scale-Out Specification in well-known Accelerators

	Parameter				
	Protocol	**Uni. BW/XPU**	**Topology**	**N/W Interface**	**Scale-out Switch**
Example	-	Gb/s	-		-
A100	InfiniBand	200	Fat-Tree	PCIe NIC	InfiniBand Switch
H100	InfiniBand	400	Fat-Tree	PCIe NIC	InfiniBand Switch
B200	InfiniBand	400	Fat-Tree	PCIe NIC	InfiniBand Switch
B300	InfiniBand	800	Fat-Tree	PCIe NIC	InfiniBand Switch
VR200	InfiniBand	1600*	Fat-Tree		InfiniBand Switch
TPUV5p	Ethernet	50	Fat-Tree		Ethernet Switch
MI300X	InfiniBand	400	Fat-Tree	PCIe NIC	InfiniBand Switch
MI355X	InfiniBand	400	Fat-Tree	PCIe NIC	UEC Switch
Ascend 910B	SuperLink	100	Fat-Tree	PCIe NIC	CloudEngine

The scale-out fabric operates on a model of decoupled compute and memory which can scale independently. Table 4-6 depicts how a generic scale out network fabric may look like. Communication between these nodes is not based on direct, atomic memory operations but on explicit message passing. For high performance, this is typically accomplished using RDMA, which allows the network interface card (NIC) of one node to communicate directly with the memory of another node, bypassing the host CPU and minimizing latency. This communication is imperative for the All-Reduce and All-Gather collectives from Data Parallelism and the particularly stressful All-to-All patterns from Expert Parallelism. Research shows that large language model (LLM) training generates sparse communication patterns, meaning full any-to-any connectivity across all GPUs is not always necessary for efficient performance. This insight has led to specialized topologies designed to optimize for

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

these unique traffic patterns, [3]. Technologies like InfiniBand and emerging open standards like Ultra Ethernet, [16], are designed to provide the extreme bandwidth, advanced congestion control, and efficient, RDMA-based message passing required for these scale-out topologies, [3].

4.4 Scale-Up and Scale-out: I/O Coherency with Memory Semantics

The distinction between these network tiers goes beyond mere performance metrics; it reflects a fundamental difference in design philosophy and purpose. Scale-up networking is a form of vertical scaling, focused on making a single system more powerful by adding resources like processors and memory into a tightly coupled domain. In contrast, scale-out networking is a form of horizontal scaling, which involves adding more independent systems to work in parallel.

The scale-up network's goal is to support I/O coherency across multiple physically distinct accelerators. Protocols such as (NVlink, ESUN, and UALink) are designed for this purpose. This is a load/store communication model, ideal for the frequent, small memory read/write operations common in tightly coupled AI training like Tensor and Pipeline Parallelism. The scale-out network, using Ethernet or InfiniBand™ with RDMA, operates on a model of memory decoupling where each node has its own separate memory space. While highly efficient for larger data transfers between clusters (as in Data and Expert Parallelism), it is fundamentally a message-passing system and not as performant for the fine-grained memory operations that characterize the scale-up domain.

The performance of the entire multi-billion-dollar cluster is therefore no longer gated by the accelerators alone, but by the co-optimized performance of both the scale-up and scale-out fabrics. A bottleneck in either domain can leave expensive compute resources idle, making a holistic design approach essential. This philosophical difference manifests in starkly different quantitative requirements, as summarized in Table 4-7.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Table 4-7-Comparative Analysis of AI Network Tiers

Characteristic	Scale-Up Network	Scale-Out Network	Primary DC Network
Primary Application	Tightly coupled XPU-to-XPU communication within a compute domain	Inter-cluster communication, data/pipeline parallelism	General connectivity, storage access
Typical Scale	Small (up to 256 XPUs)	Medium (up to 10,000 nodes)	Very Large (100k - 1M endpoints)
Reach/Distance	< 30 meter	< 2 km	≥2 km
Round-Trip Time	< 1 microsecond	< 10 microseconds	> 100 microseconds
Bandwidth per XPU	~1000GB/s (~8000 Gb/s)	~50GB/s (~400Gb/s)	~100–200 Gb/s (~12.5–25 GB/s), shared node-wide[3]
Dominant Interconnect	Proprietary (e.g., NVLink), Open (e.g., UALink), PCIe	Ethernet, InfiniBand	Ethernet
Communication Semantics	Coherent I/O (Load/Store)	Disaggregated Memory (Message Passing)	Network
Key Performance Driver	Ultra-low latency, massive bandwidth	High throughput, cost-efficiency	Scalability, reliability

The performance difference is immense. On a normalized basis, the nominal scale-up bandwidth of approximately 1,000 GB/s is about 20 times the 50 GB/s scale-out bandwidth and approximately 40–80 times the 100–200 Gb/s primary data-center bandwidth. These figures are indicative; actual comparisons depend on whether bandwidth is measured per XPU, per node, per direction, or as aggregate bidirectional bandwidth. Similarly, the latency targets are drastically different, with scale-up aiming for sub-microsecond RTT limiting the bottleneck during retries. This semantic and performance gap explains why industry efforts

[3] It highlights the "Bandwidth wall" by showing that if one tries to train a model over general DC network, each GPU would only have a tiny fraction of the pipe.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

to advance AI networking are proceeding on two parallel tracks: initiatives like the Ultra Ethernet Consortium (UEC) focus on enhancing Ethernet for the scale-out domain, while separate initiatives like the Ultra Accelerator Link (UALink) and ESUN focus on creating open, memory-semantic fabrics for the scale-up domain.

4.5 The Scale-Up Bottleneck: The Epicenter of Interconnect Challenge

The synthesis of these points leads to a critical conclusion: for the most demanding AI workloads, the scale-up network is the primary performance bottleneck. The efficiency of training massive models is directly gated by the speed at which processors within the compute domain can communicate and synchronize. Any latency or bandwidth constraint in this fabric forces expensive XPUs to sit idle, dramatically reducing XPU utilization (Hardware Flops Utilization, HFU) and increasing the time and cost of training.

The ultimate goal of the scale-up network is to create a "virtual single GPU" from a hardware perspective, allowing a single software application to access the combined memory and compute resources of the entire cluster as if they were a monolithic entity. Achieving this requires an interconnect that behaves less like a traditional network and more like a system backplane or bus. This imperative for ultra-high bandwidth, ultra-low latency, and memory-semantic communication is what makes the scale-up domain the epicenter of the interconnect challenge and the primary driver for the adoption of the advanced short-reach optical interfaces detailed in this report.

4.6 Scale-Up Switching

Scale-up networks support workloads characterized by dense communication patterns by providing flexible, high-throughput connectivity to realize tightly coupled communication domains that behave as a single logical system. In such domains, memory-semantic communication may enable direct data exchange across XPUs. When multiple workloads or tenants share a scale-up domain, this tight coupling introduces isolation challenges. Because participating XPUs are interconnected through a shared communication fabric, workloads may implicitly share communication resources and memory. Without sufficient isolation mechanisms, this may lead to performance interference, fault propagation, and security concerns [22].

Conventional scale-up network architectures assume static interconnect configurations and typically employ low-latency switches. Switching allows the interconnects to be modified which can be helpful in several ways.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

1) Reconfiguration in the event of a hardware failure
2) Configuration prior to the start of a job to “right size” the hardware to the job’s needs

There are two basic approaches discussed in the following subsections.

4.6.1 Electrical Switching

The industry has developed electrical switches for scale-up utilizing a low-latency protocol. These switches separate the optical link into two segments: one between the XPU and switch and one between the switch and another XPU as shown in Figure 4-7. The interconnecting links may be optical or electrical. In either approach, there is interface power consumed in the electrical switch to support the two I/O interfaces.

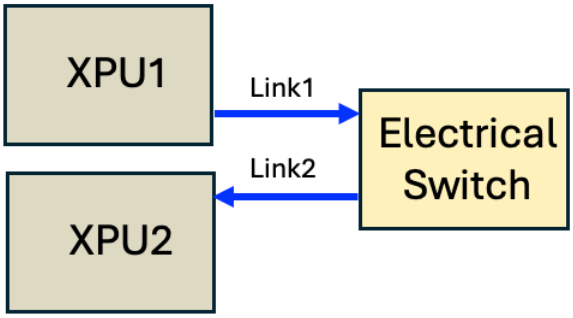


Figure 4-7 - Electrical Switch with Optical Links

4.6.2 Optical Circuit Switching (OCS)

Optical Circuit Switching (OCS) provides the opportunity to decouple logical resource allocation from physical connectivity constraints and provides a mechanism to achieve through dynamic, fiber-level connectivity between endpoints.

OCS switches are inherently low latency as the optical link is not terminated before reaching the destination XPU as shown in Figure 4-8. Note that there is no interface power consumed by the optical switch, and, therefore, optical link loss budget expands to include two optical link segments as well as the optical loss through the optical switch as shown in Figure 4-8.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

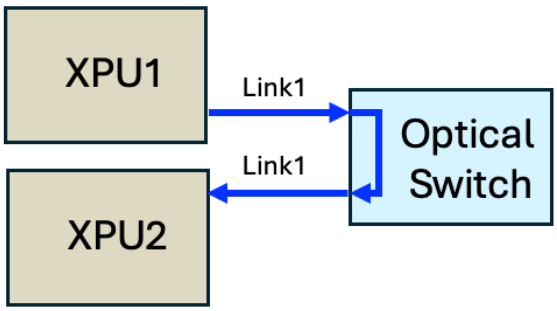


Figure 4-8 - Optical Circuit Switch

In this model, the optical interconnect is no longer limited to a static transport role but instead functions as a programmable connectivity fabric. This capability is enabled by the properties of optical interconnects. Unlike electrical packet-switched networks, OCS can establish dedicated, high-bandwidth paths without intermediate buffering or contention. This allows connectivity to be reconfigured at the fiber-level while preserving the performance characteristics required for scale-up communication. As a result, reconfigurable scale-up fabrics can be realized as an extension of photonic interconnect technologies.

Overall, the introduction of OCS extends the role of short-reach optical interconnects beyond fixed connectivity, enabling a transition toward dynamically configurable optical interconnects that support both efficient resource utilization and workload isolation in multi-tenant scale-up systems.

Several OCS implementation technologies have been explored and are shown in Table 4-8. Each implementation provides a different set of trade-offs in switching speed, scalability, and optical performance [23], [24].

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Table 4-8 -Quantitative Comparison of OCS Implementation Technologies

Category	Robot-arm	Free-space MEMS	Liquid Crystal on Silicon (LCoS)	Piezoelectric	Waveguide
Switching Principle	Physical fiber re-plugging	Micro-mirror beam steering	Liquid crystal on Silicon	Beam steering	On-chip waveguide switching
Switching Speed	Seconds – minutes	ms range	ms range	ms range	~ns – µs
Port Count	< 2000+	100–1000+	~300-512	~300-576	32–256 (typ.), scalable with integration
Insertion Loss	~0.5 dB	~1.0 – 3 dB	~1.0 – 3 dB	~1.0 – 3 dB	~3.0 – 10.0 dB
Crosstalk	< −60 dB (negligible)	~−40 to −55 dB	~−40 to −55dB	~−40 to −55dB	~−20 to −40 dB
Optical Bandwidth (nm)	High	High	Moderate	High	Moderate (design specific)
Fiber type supported	POF (plastic optical fiber), SMF, MMF	SMF	SMF	SMF and MMF	SMF
Power Consumption	Very low (static)	Moderate (~W-level per switch)	Moderate (~W-level per switch)	Moderate (~W-level per switch)	Low (sub-W – W level, depending on technology
Reliability	Mechanical wear risk	Mature (MEMS reliability proven)	High	High	High (no moving parts)
Implementation size	Very large	Large	Large	Large	Very small

The increased optical link insertion loss due to the OCS must be compensated for. One such approach is to reduce the optical data rate per lane combined with multiple λs on a single fiber to achieve the same overall data rate for the fiber. As an example, halving the data rate ideally results in a 3dB improvement in the optical link budget. One such multi-λ implementation is defined by the OCI-MSA [21].

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Different OCS technologies exhibit different insertion loss characteristics, where the insertion loss is defined as the sum of intrinsic switch loss and I/O coupling loss at the switch ports (e.g., connector or fiber coupling loss, typically ~0.5–3 dB). For example:

- Robot-arm switching is dominated by connector loss (~0.2–0.5 dB per connection).
- MEMS-based systems include free-space coupling and mirror losses (~1–3 dB).
- Waveguide-based switching includes propagation, crossing, and device loss (~3–10 dB depending on implementation).
- Liquid crystal switching system is dominated by the optical coupling and alignment loss, 1.5~3dB
- Piezoelectric switching is dominated by angular misalignment of the piezoelectric actuators.

Note that in a multiple-λ link, wavelength dependent losses need to be considered.

Crosstalk is another parameter of importance for which the technologies have differing performance.

- Robot-arm switching provides near-perfect isolation due to physical fiber separation (< −60 dB).
- MEMS-based switching typically achieves ~−40 dB class crosstalk.
- Waveguide-based switching may exhibit higher crosstalk (~−20 dB worst case), depending on integration density.
- Liquid crystal switching's crosstalk is mainly caused by diffraction.
- Piezoelectric switching can achieve ~-50dB crosstalk due to spatial and mechanical phenomenon.

Finally, dynamic reconfiguration speed depends on the OCS technology.

- MEMS/LCoS/Piezoelectric-based systems operate at millisecond timescales and are suitable for coarse-grained reconfiguration.
- Waveguide-based switching enables sub-microsecond reconfiguration, making it a promising candidate for fine-grained, runtime-adaptive photonic fabrics.

Robot-arm, MEMS, LCoS, and Piezoelectric-based OCS represent solutions used in data centers. Waveguide-based switching schemes are being developed; they provide a potential path towards increasing the dynamics of photonic interconnects aligned with future workload requirements.

Date: August 5, 2026

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

OCS architectures are no longer limited to a static transport role but instead functions as a programmable connectivity fabric. This capability is enabled by the properties of optical interconnects. Unlike electrical packet-switched networks, OCS can establish dedicated, high-bandwidth paths without intermediate buffering or contention. This allows connectivity to be reconfigured at the fiber-level while preserving the performance characteristics required for scale-up communication. As a result, reconfigurable scale-up fabrics can be realized as an extension of photonic interconnect technologies.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

5 Copper Implementation

As mentioned before, there are multiple IO technologies that could be deployed to build the Scale-Up Fabric. These technologies include UALink and ESUN for node-to-node connections and PCIe for CPU-to-XPU and CPU-to-NIC connections. Since both UALink and ESUN are based on ethernet specification, the data rate of interest is 212.5 Gb/s PAM4 (IEEE 802.3dj). For PCIe, the current state-of-the-art is Gen7 at 128Gb/s PAM4.

In the following sub-sections, we will have a high-level summary of the electrical parameters and approximate physical channel lengths based on the IEEE 802.3dj and PCIe specifications. Then we will have a deeper look into Scale-Up Network's physical topologies. We will close this section with a brief overview of Scale-Up Network using optical interconnect.

5.1 Data Rate and Insertion Loss Summary

Table 5-1- Electrical Interfaces

Specification	Data Rate (Gb/s) (1)	Nyquist Freq. (GHz)	System insertion loss budget (dB) (2)	Passive copper cable assembly loss budget (dB) (2)	PCB loss assumption (dB/in) (2)
PCIe Gen7	128	32	-24 (3)	-11 (5, 6)	-1 (approx.)
IEEE 802.3dj	212.5	53.125	-40 (4)	-17 (5)	-1.25 to -1.5 (approx.)

Notes for Table 5-1:

1. PAM4 modulation
2. Measured at Nyquist frequency
3. Pin-to-pin loss budget, excluding packages, approximation based on PCIe Gen7 Base Spec Table 8-8, Calibration Channel IL Limits
4. Total loss budget, including packages
5. One meter cable length

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

6. Internal cable

As the data in Table 5-1 suggests, with only passive copper cable, the total channel lengths of both PCIe Gen7 and UALink/ESUN are limited to ~1.5m. The standard channel topology, however, can be modified to allow for longer reach before an optical interconnect must be used. For example, low loss fly-over cables can be used to leave more loss budget to the external cable. Active components such as re-driver or re-timer can enable channels with loss that is larger than spec allowed.

Since the PCI-SIG™ governs the PCIe specifications, readers should consult PCI-SIG literatures for PCIe details. The following discussions are focused on the IEEE 802.3dj interconnect.

5.2 Channel Topology

Backplane (BP) Topology: Connect two different chips, e.g. GPU and switch ASIC, that are in different trays of the same server rack.

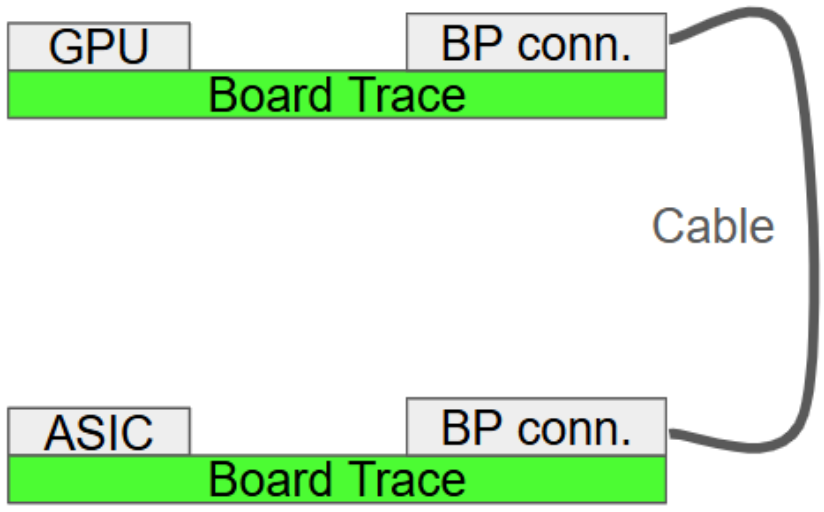


Figure 5-1 - Backplane Cabling

Estimates from the OCP Extended PCIe Connectivity White Paper suggests 3 m cables are needed to connect trays from the top and the bottom of the rack while 2 m cables are sufficient for connecting trays from the top or bottom of the rack to a tray at the middle of

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

the rack. If the cables do not need to go through the rack rails, then the length requirements can be relaxed.

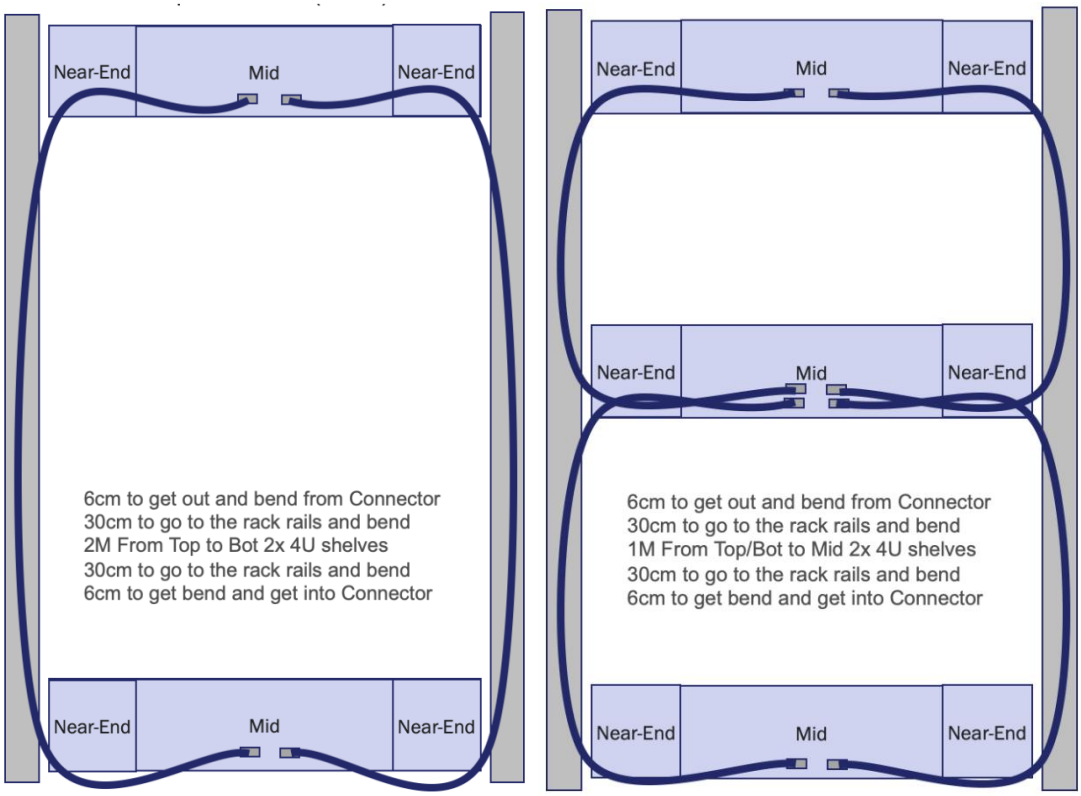


Figure 5-2 - Backplane Reaches

As mentioned already, internal fly-over cables can be used to reduce insertion loss between the chips and the backplane connectors and thus allowing the external cable to be longer.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

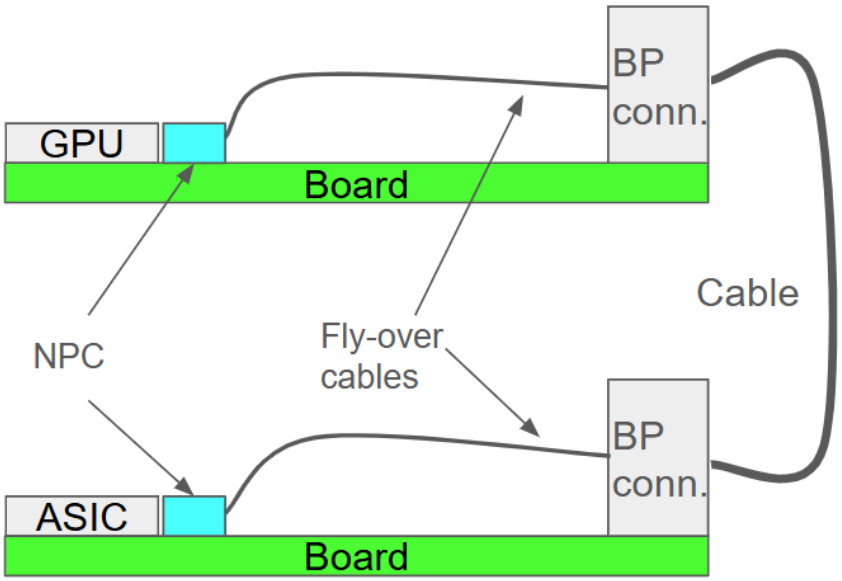


Figure 5-3 - NPC Fly-Over Cables to Backplane Connector

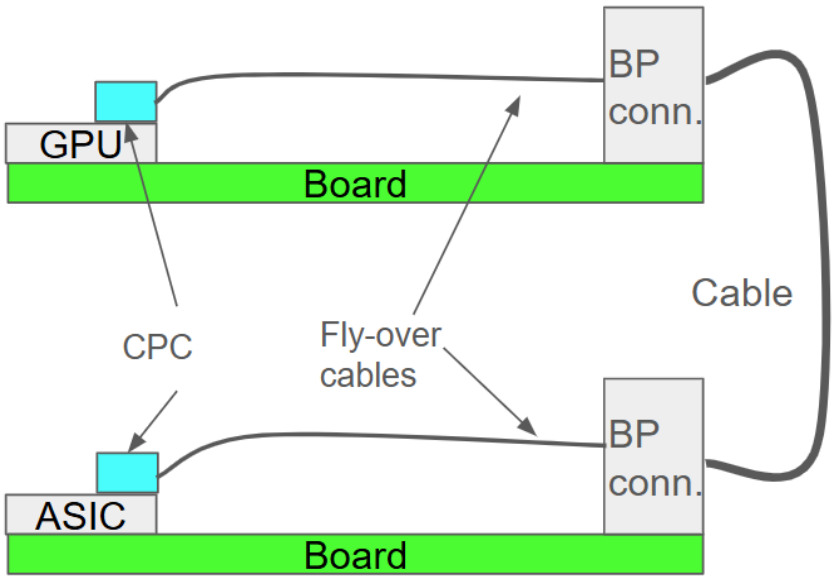
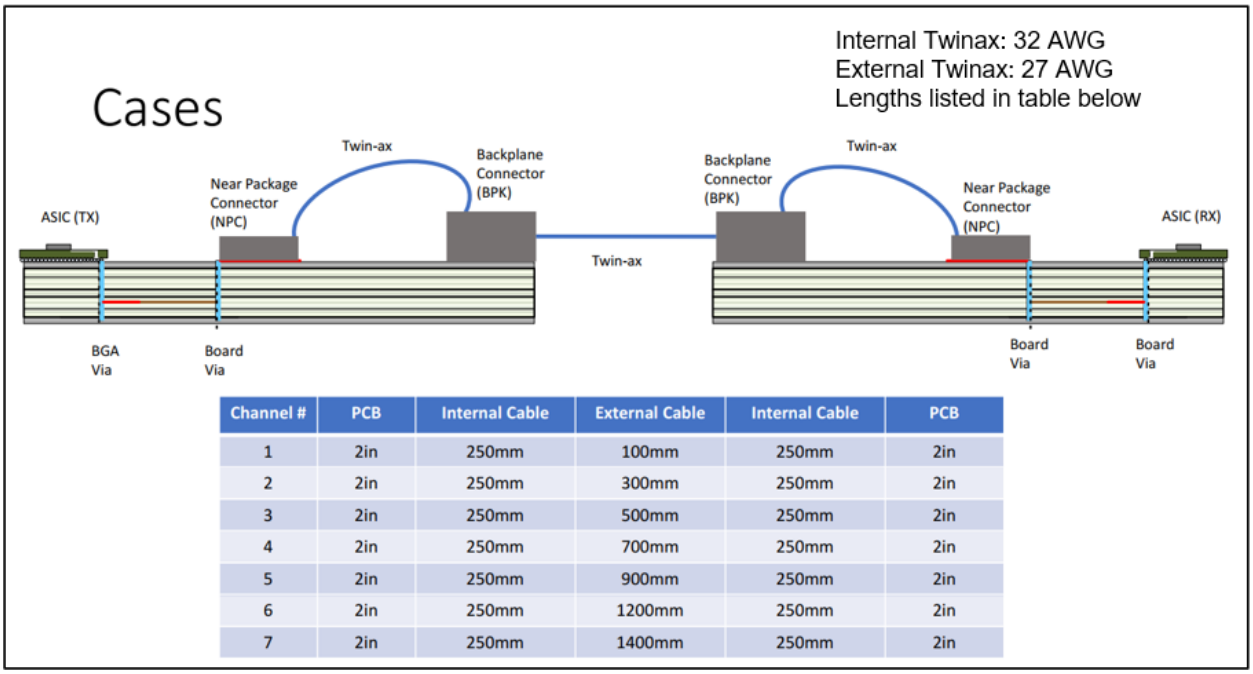


Figure 5-4 - CPC Fly-Over Cables to Backplane Connector

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

The total length of the channel depends on the details of each component; but a study done in the IEEE 802.3dj Task Force shows NPC with fly-over cable can enable 1.4 m external cable at 212.5Gb/s.



Channel #	PCB	Internal Cable	External Cable	Internal Cable	PCB
1	2in	250mm	100mm	250mm	2in
2	2in	250mm	300mm	250mm	2in
3	2in	250mm	500mm	250mm	2in
4	2in	250mm	700mm	250mm	2in
5	2in	250mm	900mm	250mm	2in
6	2in	250mm	1200mm	250mm	2in
7	2in	250mm	1400mm	250mm	2in

(Source: akinwale_3dj_elec_01_2310.pdf)

Figure 5-5 - 200GbE Reach Cases

If CPC connectors are used to replace the 4 inches of PCB routes, it is likely the backplane cable can be 1.5m or longer.

To allow for even longer cable length, repeaters can be used.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

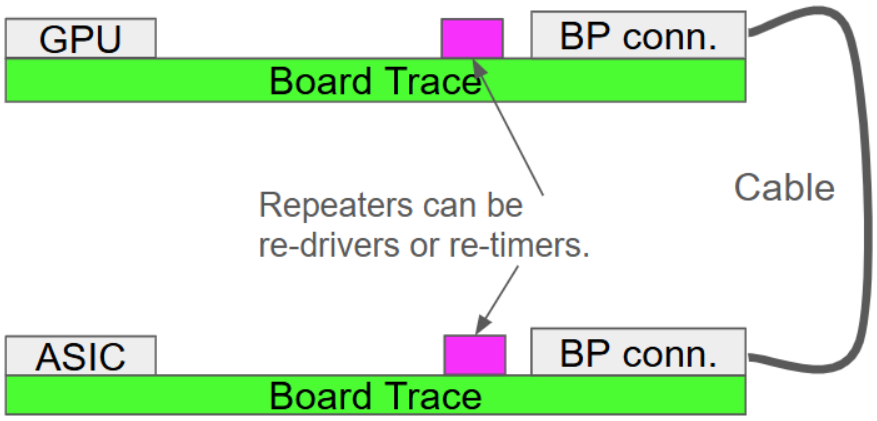


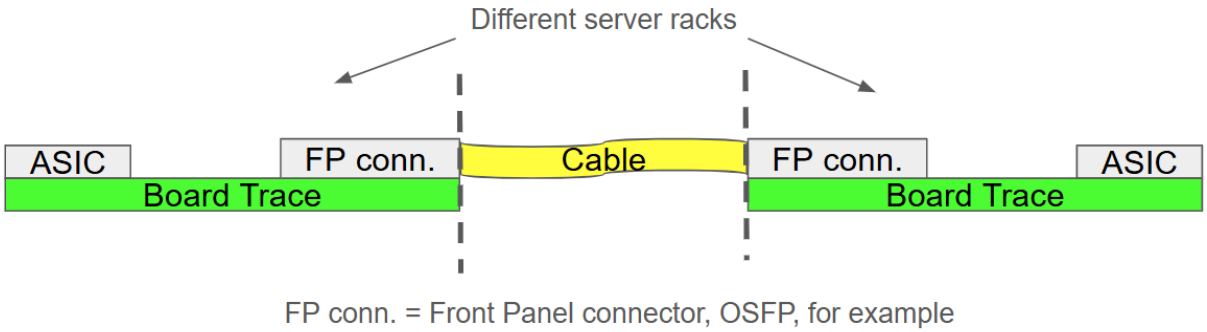
Figure 5-6 - Backplane with CPC

Assuming roughly symmetric insertion loss budget, using re-timers before and after the backplane cable can potentially enable the cable assembly to have the entire spec loss budget, which can translate to roughly 2.5 m to 3.0 m cable length.

Since re-timers are more expensive, consume more power, and may require thermal management, re-drivers are more economical. The amount of insertion loss compensation benefit, however, is smaller than what re-timers can provide.

For the portion of the scale-up network that connects trays on the same server rack using backplane cables, it is actually feasible to use copper interconnect to support 212.5Gb/s.

Pluggable Topology: Connect different trays, switch ASIC trays, for example, in different server racks.



This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Figure 5-7 - Pluggable Topology

Again, referencing the OCP Extended PCIe Connectivity White Paper, length requirements for connecting nearby racks range from 3 to 7 meters.

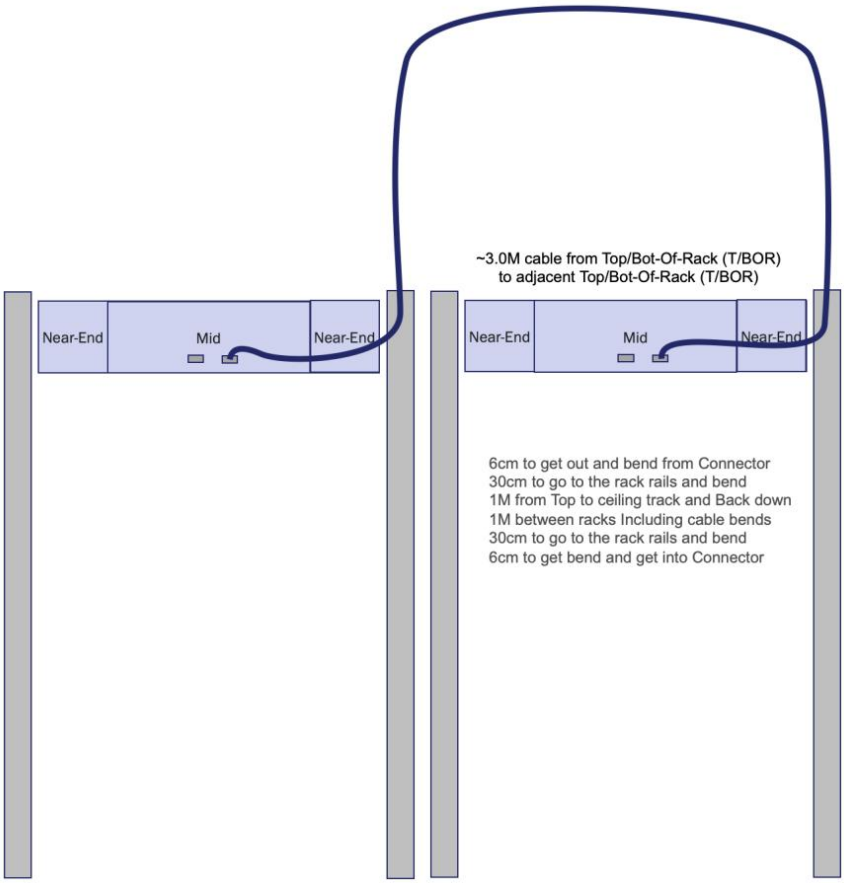


Figure 5-8 - Connectivity Between Top of Racks

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

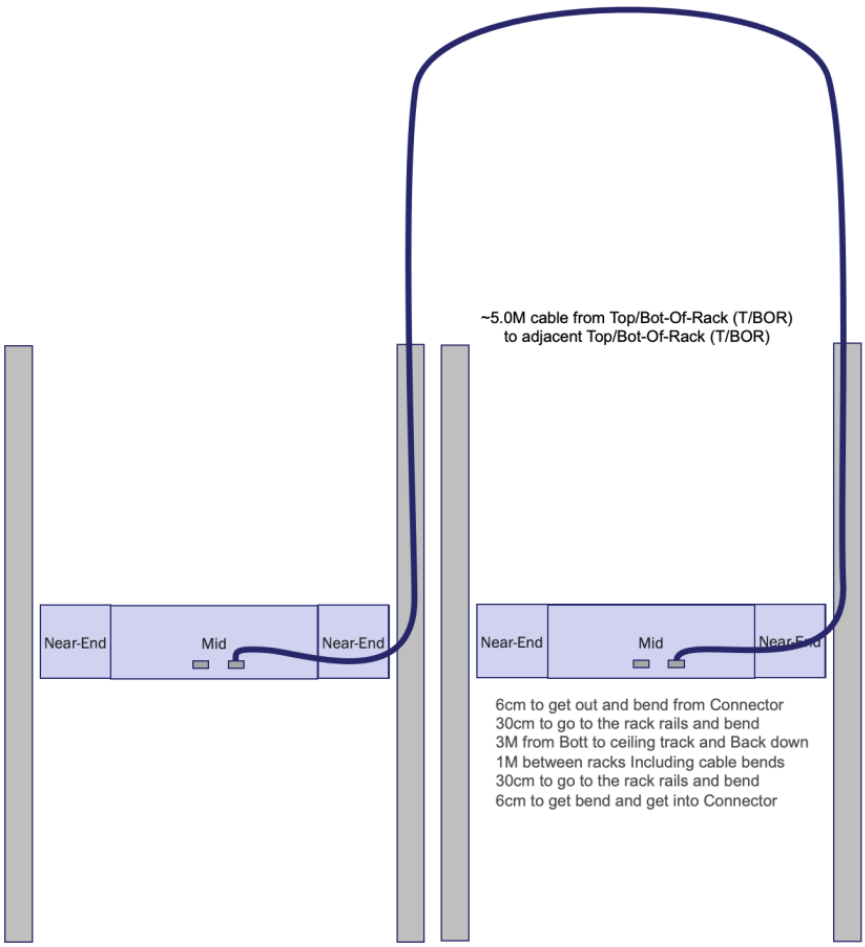


Figure 5-9 - Connectivity Between Middle of Racks

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

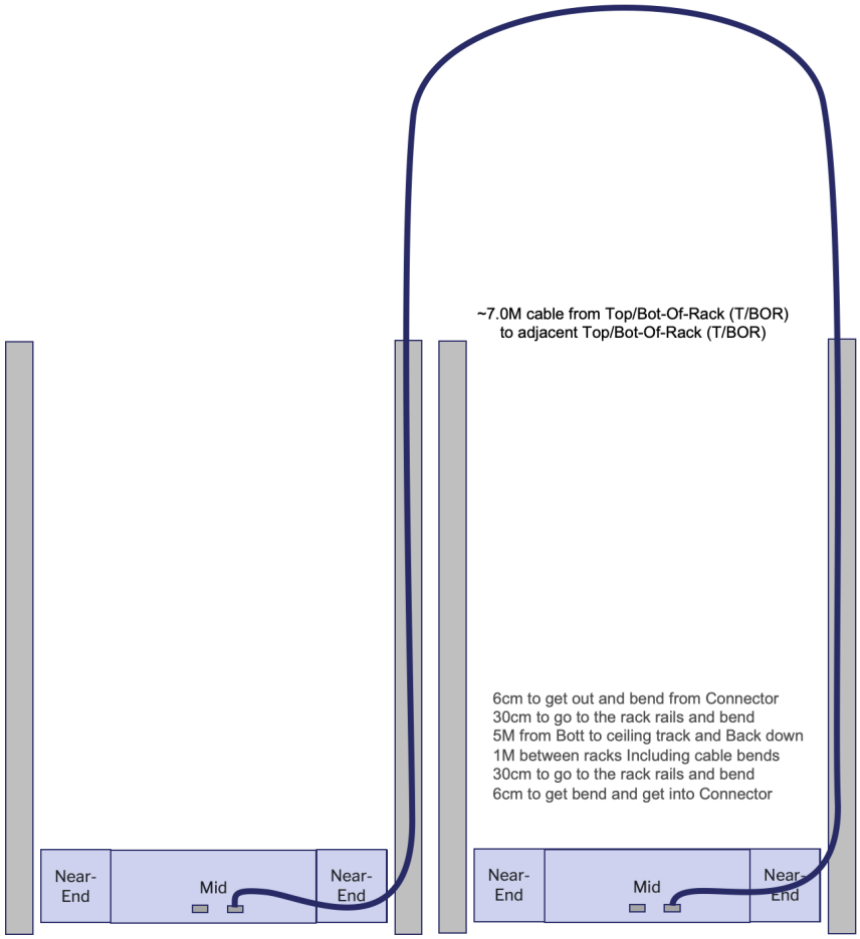
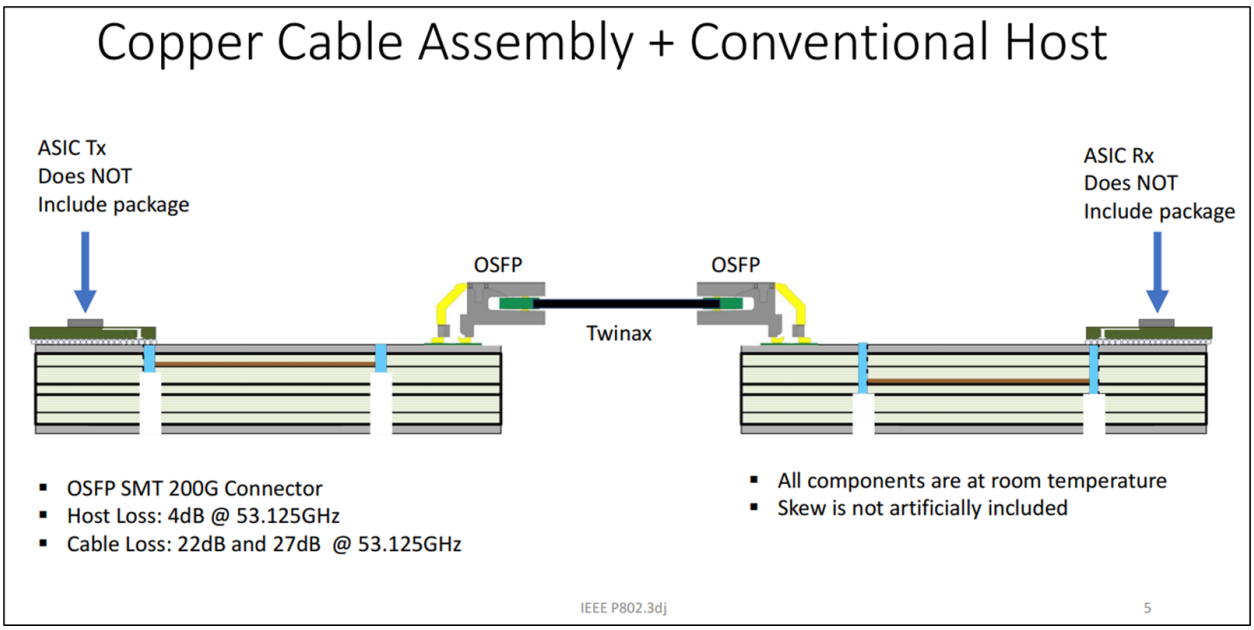


Figure 5-10 - Connectivity Between Bottom of Racks

As mentioned before, if the cables do not need to go through the rack rails, then the length requirements can be relaxed.

With passive copper channels, a study done in the IEEE 802.3dj Task Force shows OSFP™ cable is feasible with “conventional” hosts at 212.5 Gb/s.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.



(Source: akinwale_3dj_01_2311.pdf)

Figure 5-11 - Copper Cable Assembly

With OSFP, or similar, connectors, repeaters can be placed on the paddle cards of the cable and thus turning a passive cable into an active cable. Adding repeaters can potentially increase the cable length to 2.5 m or 3 m. Using internal fly-over cable plus NPC or CPC can extend the total channel length slightly. For a Scale-Up network that has longer than 3 m of connection length requirement, however, optical interconnects are likely required.

5.3 Optical Channel Topology

At 212.5 Gb/s, it is expected the loss of optical interconnect to be very small, and thus the copper routings inside the two trays should have sufficient loss budgets. The placement of the Optical Engine where electrical to optical, and vice versa, transition occurs, is thus determined by tradeoffs between reliability, power, and cost. While these tradeoffs are outside the scope of this section, high-level descriptions of the topologies are still informative.

For the backplane, the optical engine can be placed on the package, near the package, or near the backplane connector.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

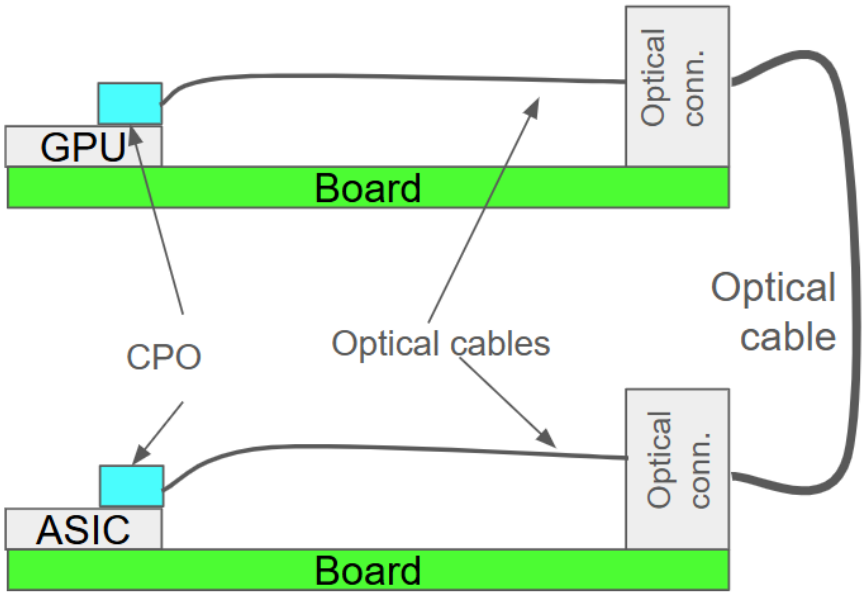


Figure 5-12 - CPO Backplane Topology

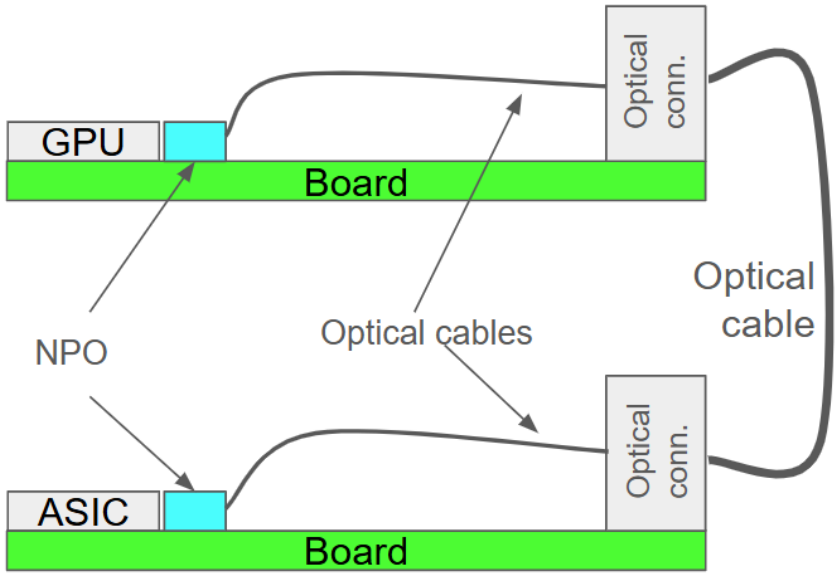


Figure 5-13 - NPO Backplane Topology

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

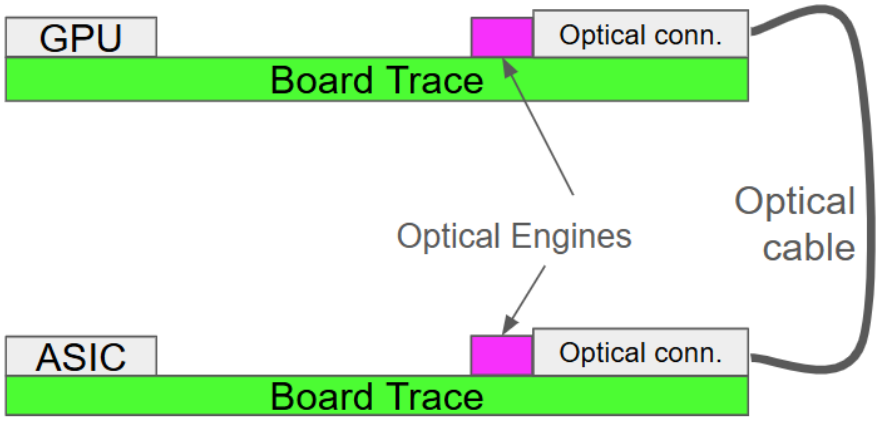


Figure 5-14 - OE placed near the backplane optical connectors

For the pluggable topology, the simplest (from serviceability's perspective) choice is probably having the Optical Engine on the paddle card of the cable.

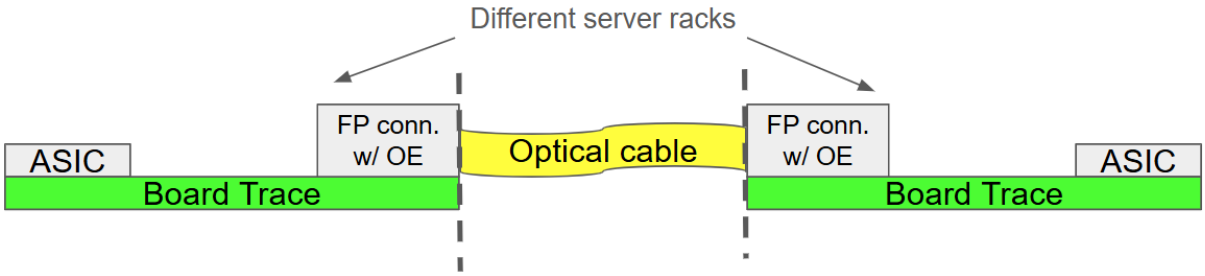


Figure 5-15 - Faceplate Pluggable

Of course, it is completely feasible to use CPO or NPO instead.

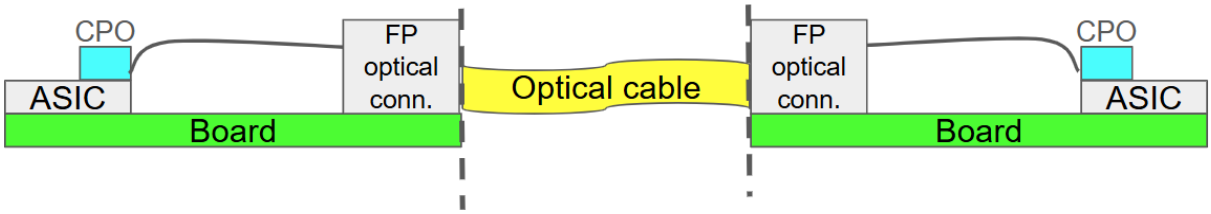


Figure 5-16 - CPO

Date: August 5, 2026

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

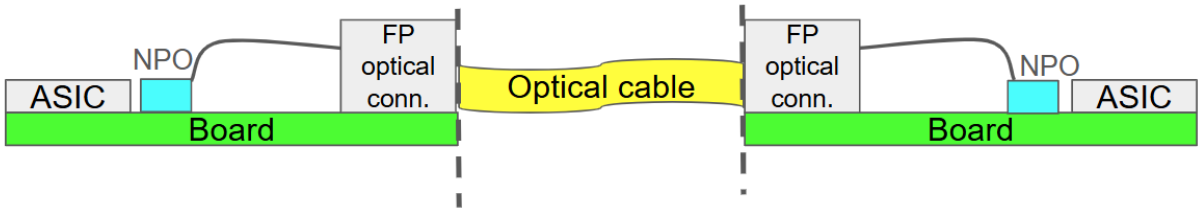


Figure 5-17 - NPO

Again, the design decisions on how to implement an optical interconnect are outside the scope of this section of the white paper.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

6 Optical Key Metrics

In this section, we define the key metrics and discuss their priority of importance. We recognize that the tradeoffs in the face of technology constraints are complex, particularly as multiple metrics are intertwined in technology and design choices.

For general system architecture enablement, the top three priorities for any I/O are cost, reliability and power consumption, the latter having only recently risen to this level of prominence. For scale-up specifically, latency is also a top priority. Given the need for rapid new technology introduction, time-to-market is also part of the tradeoff. Table 6-1 summarizes these top-level metrics and their target values in two years for scale up networks to adopt optics broadly.

Table 6-1 - Key Metrics

Top-level Metric	Threshold Target	Rationale
Cost	Trend towards copper in $/Gbit	Competitive with longest-reach copper (AEC)
Reliability	2 FIT, 0 link-flaps/M-hr	Feasible for optics
Power Consumption (One end, optics only)	Single digits expressed in pJ/b	Matches longest-reach copper (AEC)
Latency	Removing 1 retimer	Equivalent to time of flight for 10 m

Let us distinguish between these top-level system goals and a second set of technology metrics. In this whitepaper, we are primarily concerned identifying technology to develop that will enable the adoption of optics within the scale-up domain once key metrics are met.

6.1 Targets for Top-Level Metrics

Defining the target for metrics that can be allocated to optics in the scale-up domain is a new challenge. Previously interconnect, even for traditional scale-out networks was primarily passive copper with minimal cost and zero power. Clearly it is unreasonable to hold optics or even active copper to the same targets. The real calculus is in terms of optimization of the top-level system goals.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

One threshold for the adoption of optics is when the cost, power and latency of advanced copper rises to meet that of optics.

Note that the full XPU-XPU or XPU-switch link consists of more than the copper or optic cables. The search for improvements in cost, power and latency for the entire link at the system level includes the electrical interfaces, retimers/redrivers, system cooling, latency of FEC, etc. As new forms of optics such as LPO, CPO and in-die optical I/O have started to close the gap between traditional passive copper and traditional front-panel optics, increasing adoption of active copper is starting to close the gap from the other side

Figure 6-1 illustrates how the traditional black-and-white choice between (passive) copper and (front-panel pluggable) optics has expanded into almost a continuum with multiple types.

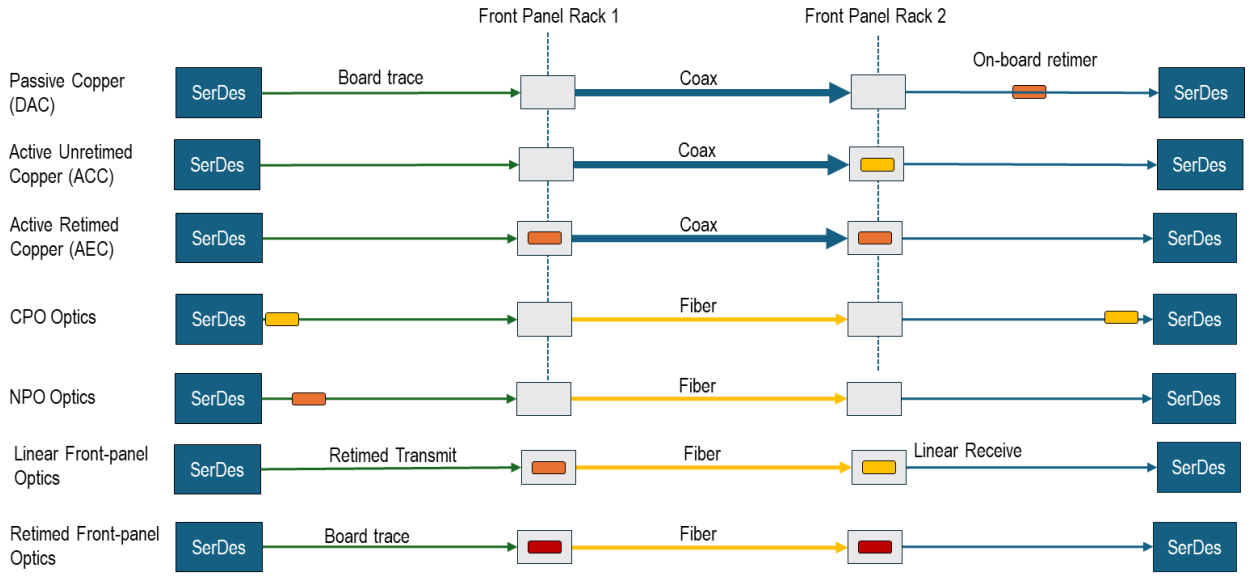


Figure 6-1 -Continuum of Link Options inclusive of Serdes, potential on-board retimers, active copper and optics categories.

6.1.1 Power

In Figure 6-2, Broadcom has postulated 10 pJ/bit for each end of the link (Serdes + Optics/Copper) is the threshold for adoption of optics in the 200G/lane generation with the

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

highest-power version of copper and the lowest-power versions of options converging on that point.

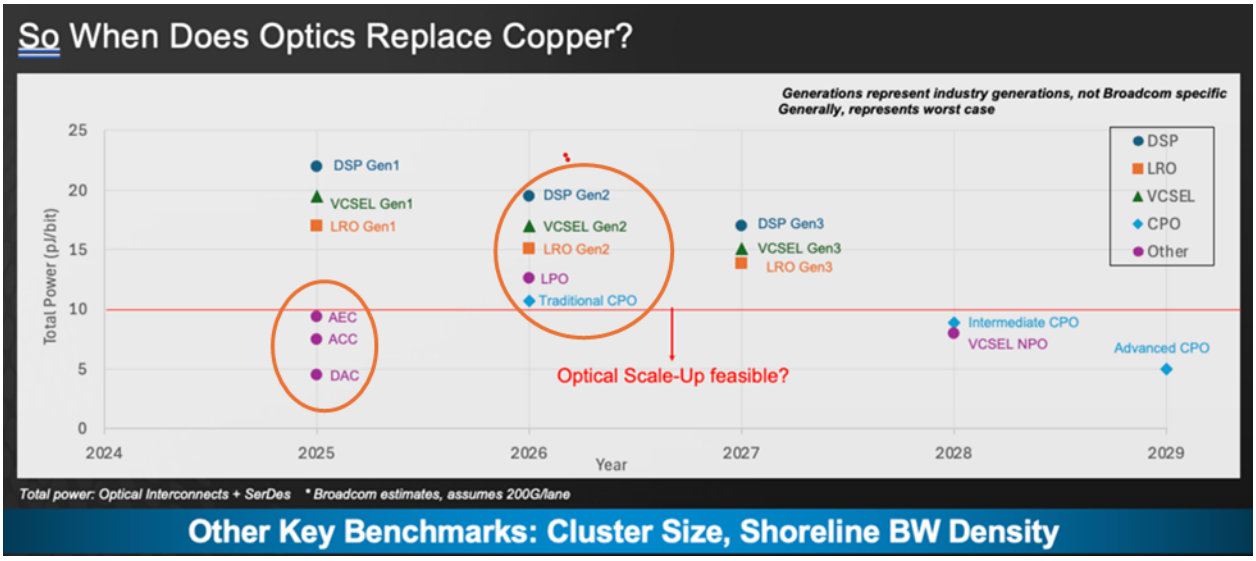


Figure 6-2 -Mark-up of Broadcom chart showing 10 pJ/bit as the boundary between copper and optics implementations at 200G/lane

6.1.2 Cost

Many operators state that cost per bit is the highest priority, and that it is the large gap between the cost of passive copper and the cost of optics that has precluded the use of optics in scale-up networks to date. Roughly, the industry believes there is an order of magnitude gap between passive copper and optics, with passive copper being roughly $0.05/Gbit and optics being order of magnitude $0.50/Gbit in hyperscaler volumes.

But is cost/bit the right metric? The cost of copper scales linearly with length as the cable itself is a significant cost driver. If we consider instead a cost/bit/m metric, the cost of copper at optics (30m) lengths is not so far out of line. On the other hand, the cost of the longer copper cables (along with weight) has also likely been a deterrent to their use. The driver for optics adoption is the near-term is again the decreasing reach of copper and the greater cost of active copper. Analysts believe the premium for active unretimed copper cables and active retimed copper cables are 30 – 50% and 50 – 100% respectively. The total system cost comparison is also impacted by the need to increase cooling and power

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

delivery density if links are constrained to be half as long as in the previous generation such as in Nvidia's ½-height but doubled-capacity Kyber rack. These datapoints suggest that the cross-over cost target cost optics to a low-cost multiple of passive copper.

Cost of ownership includes the fiber infrastructure cost. Optical formats such as bidirectional optics (BiDi) offers a way to reduce the fiber count by ~ ½ as the transmit and receive signals share a single fiber. One such example is the OCI-MSA specification [21] which combines BiDi with WDM transmission.

6.1.3 Hardware Reliability

As described in more detail in section 7, link reliability metrics can be separated into hardware FIT rates and link errors that require retransmission. FIT rates of traditional optics have been a visible barrier to broader adoption but some of the perceptions around optics failures are carryovers from technologies that are no longer used such as DML and more manual assembly practices from lower-volume days. Nonetheless, the huge scale and failure intolerance of scale-up means stringent reliability targets for both hardware FIT and other types of link errors.

FIT rates are traditionally associated with permanent hardware failures. Google has advocated more extensive testing, including more burn-in which weeds out of infant mortality, which in turn could potentially lower the bottom of the bathtub curve. Recently there has been good news that CPO data from Meta showing 5x lower unserviceable failures for CPO than pluggable optics, due to more semiconductor-style manufacturing and integration.

Anecdotally, it appears that laser failures can approach 2 FIT under well-controlled conditions. This may be a reasonable target for raw hardware reliability. However, additional resiliency by means of redundancy and predictive resource allocation is also needed.

One point which has been published but is not widely appreciated is that retimers can increase errors despite their purpose being the opposite. Each retimer makes a hard decision and therefore has some finite probability of locking in a wrong decision.

Concerns about transient link failures have changed the way link performance is measured. Traditionally, pre-FEC bit error rates are measured and an assumption is made as to the corrected error rate. An uncorrected error which requires retransmission is essentially a transient link failure. Newer optical specifications include requirements on "FEC tail" or the number of corrected errors, a metric of how close the link is to having uncorrectable errors.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Broadcom has claimed initial results of no statistics above the correctable number of errors or zero link-flaps in 1M device hours. Again, now that this level has been shown to be feasible, it could be the basis for an initial target across the industry.

Maintaining operating temperature of the optics at something like 40 °C improves reliability compared to conventional datacenter optics allowed to go up to 70 °C.

6.1.4 Latency

Latency requirements are highly application dependent. The other complication is that in addition to intrinsic round-trip latency there is tail latency due to retries necessitated by uncorrected errors and so forth. Again, taking the full link into consideration, latency arises from the Serdes, any retimers in the link whether on the board or in optics as well as time of flight in the channel. Minimizing retimers will always have the biggest impact on minimizing added latency. Each retimer adds 10's of nsec of latency, a typical number on the order of 50 ns. In comparison, each meter of fiber adds 5 ns of latency. Thus, the additional latency due to extending the link by 10 m of fiber can be offset by removing one retimer by moving the optics sufficiently close to the link ends.

6.2 Technology Metrics

Table 6-2 - Technology Metrics for Optical Scale-Up Network

Key Metric	Scale-up Target	Impact and Tradeoffs
Radix	>1024	Size of cluster/switch hop
Linear BW density @ ASIC	Matching ASIC 1-2 Tbps/mm	Impact to power and error rate if density is less than that of the ASIC
Areal BW density		Enable CPO to fit on chip package
Fiber density at faceplate	10k fibers per rack	Rack density

6.2.1 Interconnect Radix

Radix is a key metric which impacts latency. The ideal in a scale-up network is for every XPU or memory element to be directly connected to every other in a mesh. Thus, the radix is ideally the number of XPUs in the scale-up network with each single lane rates ideally the

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

full bandwidth requirement of the XPU. As the size of the scale-up network grows, the total bandwidth for existing XPUs becomes prohibitive.

In practice, there are often one or more switches between XPU's, a trade-off that increases latency but allows the network size to be constrained by the switch radix rather than the XPU radix.

For a bank of N XPUs connected by a set of M switches, the radix of the connectivity from each XPU is reduced to M so that it can connect to all switches. The radix of connectivity from each switch is the full number N of XPU. The scale-up network size is constrained to the total radix of the switch which is usually limited by its package escape. This is the same driver that forces the aggregation of I/O into a smaller number of higher speed ports. Thus, radix of the switch is currently limited to 1024 with each link ideally running at the SerDes lane rate. The importance of radix favors leaving optical links on separate fibers (no WDM). If WDM is required to achieve linear or areal density at the ASIC edge, demultiplexing prior to fiber shuffle for routing rack-to-rack may impact rack density and cost.

6.2.2 Linear and Areal BW Density at XPU

The targets that enable CPO, with the best signal integrity and therefore lowest power, requires the interconnect to be co-located on the chip package or interposer. Next best would enable NPO. For CPO, the constraint is the size of the substrate or interposer. Note that the ASIC may be larger than a single reticle as it could be two or even four chips.

Linear density of the ASIC drives the requirements to increase linear density of the optics. However, linear density has knock-on implications for the areal density as well. For the layout in Figure 6-3, the ASIC I/O density is BW/X1. Ideally, X2 would be zero. The orange block represents optics which supports the full BW but does not quite have matching linear density at BW/X. The worse the mismatch is, the larger X2 becomes. The optics cannot be placed closer than X2 to ASIC, and this dimension X2 also constrains the Figure 6-3 density allowed for the optics to fit on the substrate of size S. Multi-reticle ASICs may be up to 65 mm per edge while substrates are constrained to around 90 mm per edge.

The ASIC bi-directional density BW/X1 is 1.7 Tbps/mm today, resulting in a next-gen target of ~2 Tbps/mm with a continued doubling each generation. Since the density of the substrate or interposer as well as the chips must support this density, CPO becomes necessary. Moving even further, slow/wide D2D interfaces such as UCIe enable I/O densities of the chips are 10 Tbps/mm but the reach becomes only 2 mm which also impacts areal density.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

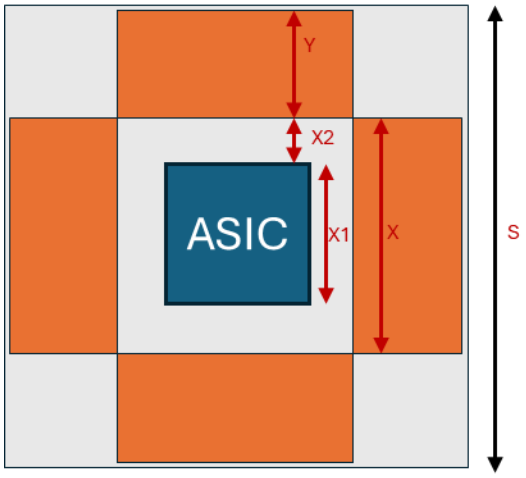


Figure 6-3 - Linear and Areal Density

For reference, a full-reticle die such as Broadcom Tomahawk 6 (TH6) is packaged with I/O chiplets that match its long side of 30 mm with 25.6 Tbps/side. The corresponding BW density, defined as (Tx+Rx) per mm is 1.7 Gbps/mm. Note that definitions of BW differ by technical specialty. Link designers typically count BW bidirectionally whereas I/O designers typically add Tx and Rx. By the latter definition, the industry's current ASIC BW density as exemplified by TH6 is 1.7 Tbps/mm. Taking into account a desire for doubling of density for next generation, the target BW/mm bidirectional for I/O in immediate proximity to the ASIC edge becomes 2 Tbps/mm.

If the optical interface cannot be co-packaged, then the ASIC package edge density becomes the benchmark. Assuming package dimensions are limited to 60 – 90 mm vs 30 mm chip size, the optical I/O density is correspondingly relaxed 2 – 3 times as well. Thus 1 Tbps/mm is an appropriate target for NPO.

6.2.3 Areal Density at Faceplate

An estimate of the fiber cabling required to support a 512 GPU cluster at 64 GPU/rack with the required switches came to over 10k fibers per rack [25]. For an Open Compute 44 OU rack, and with rounding an OU to 50 mm tall and 500 mm wide, this corresponds to only an average density of 1 fiber per cm^2 of front panel area. One might think this density requirement is well under the capabilities of modern fiber connectors. However, this average understates density requirements at the pinch points where the cables exit from the

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

racks as well as the need to accommodate fiber shuffles within fiber management trays. A generation of higher density connectors and cables are now being deployed, and a number of innovations are underway to create even denser connectors, cables and fiber shuffles.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

7 Reliability

Today, local accelerator to accelerator links within a Pod are primarily passive copper with little or no redundancy. Optics brings new challenges as the reliability of optics modules is typically worse than passive copper cables. In addition, energy efficient solutions may not have all desired diagnostic features to limit downtime during repair.

There are broadly two classes of reliability to consider:

- Link reliability in which at data received from across the link is corrupted and requires retransmission
- Hardware reliability in which a link or lane can no longer be used to transfer data until repaired

Each of these will be discussed in the following sections.

7.1 Link reliability

The local accelerator links support memory semantics (load/store operations) which are essential to ensure memory consistency (I/O operation ordering). For example, a write must be completed before a read is allowed in the same location.

When the data link layer receives a data transfer, the FEC corrects any errors present. A CRC is performed on the FEC corrected data. If there are no errors detected, the data link layer sends an acknowledgment back to the sender who is then free to send the next transfer. However, if an uncorrectable error is detected, the data link layer requests a retransmission (replies with a “not acknowledged”) initiating a retry event. During a retry event, the transmitter resends the data and waits for an acknowledgement before continuing.

These bandwidth impacting situations are often classified as uncorrectable random bit errors or burst errors. In addition, link flaps in which the link rapidly goes up and down also impact throughput. System implementors will likely specify link reliability targets such as frame or FLIT loss ratio (FLR) or a maximum acceptable level of bandwidth reduction. These targets will depend on the system architecture and performance targets. For example, a particular PCIe6 link might designed to target less than 1% overall bandwidth reduction due to retries.

These errors introduce latency and reduce the effective link bandwidth as the transaction layer waits for the retry round trip to complete. Due to the synchronous nature of local compute, an effective link bandwidth reduction impacts the computation rate of the entire

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

local Pod. Given the significant investment in the AI infrastructure, it is therefore important to employ links with excellent link reliability.

7.2 Hardware reliability

Hardware errors have many causes and include underperforming links (high error rate, lowered throughput) or dead links. This class of failures will require hardware replacement.

7.2.1 Naively replacing standard pluggable optics for passive copper links

Figure 7-1 shows a representative seventy-two-accelerator system in a single rack. In the current generation, the local links are interconnected with copper cables consisting of over 5,000 differential pairs carrying 200Gb signals.

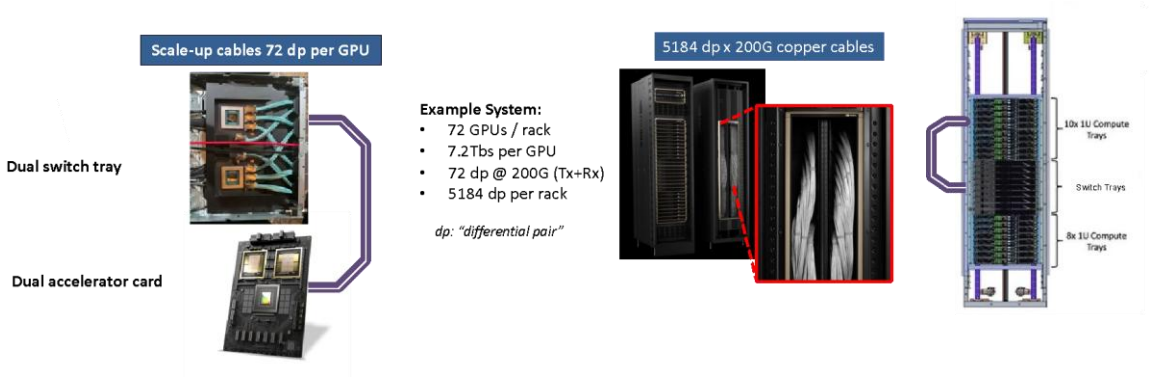


Figure 7-1 - Local AI compute example

It is interesting to see what happens if one naively replaces the current copper links with 800 Gb/s retimed optical modules. The rack would use 1296 optical 800GbE (4x 200G) retimed modules with an additional power consumption approaching 16 kW assuming 12.5 W per module, an astonishing amount of additional power for a rack. In addition, traditional optical modules are less reliable than the passive copper cables, so the system failure rate would increase to ~100,000 FIT assuming ~ 75 FIT per module. Clearly an optimized approach is needed to deploy optics for the local accelerators' interconnects for the next generation of systems.

7.2.2 Hardware reliability regimes

Hardware failures are broken down into three regimes: infant failure, random failure, and wear-out failures as shown in Figure 7-2.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

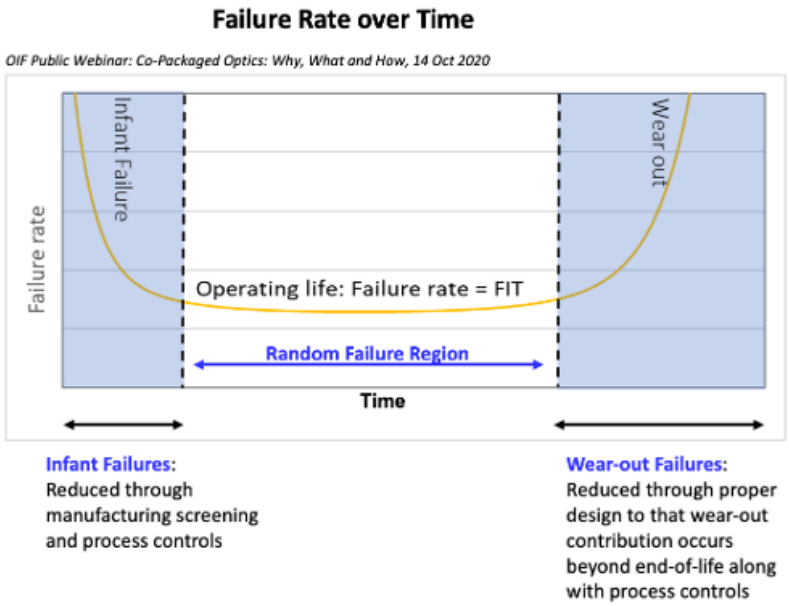


Figure 7-2 – Failure rate over time

Failures are often specified in units of "FIT" which represents the number of failures expected with $1e^9$ device-hours.

It is assumed that most or all the infant failures can be eliminated through a combination of process controls and manufacturing screening so these failures will be largely eliminated before installation into a system.

Wear-out failures likewise can be kept from making a significant contribution though careful design and qualification and will keep the significant wear-out failure region beyond operating lifetime.

Random failures occur unpredictably and without a pattern and are typically modeled as a constant failure rate over time.

Given the complexity of a compute Pod with many elements and given the consequence of a failure, the random failure target should be lower than typical Ethernet applications. In addition, it might make sense for the reliability target for co-packaged and on-board solutions to be higher than pluggable solutions due to increased time and cost to repair of these behind-the-front-panel architectures.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Scale-up links use mostly copper today and copper reliability sets the base failure rate expectation. As previously mentioned, replacing copper links with the current class of pluggable optical transceivers would negatively impact the reliability of the Pod. Pluggable modules have failure rates of perhaps 50 FIT and typically much higher. A review of the failure rate estimates shown Telcordia's SR332 "Reliability Prediction Procedure for Electronic Equipment" suggest that highly integrated solutions will have better reliability. This is in part due to using a well-controlled integrated circuit process with CPk greater than 1.0 (a ±3-sigma process). A highly integrated solution also reduces the number of mechanical interfaces where thermal and mechanical stress can initiate various failure modes. A co-packaged transceiver would need a reliability level commensurate with silicon integrated circuits.

A related issue is the introduction of new optical technologies into scale-up links. It is suggested that new technologies demonstrate a track record in more fault tolerant applications before deploying in scale-up applications, especially in co-packaged or on-board locations.

7.2.3 Repair

Periodically during operation, the Pod saves snapshots of the computation. Once a failure is diagnosed and repaired, the Pod can resume from the last saved snapshot. During the repair cycle, computation is halted stranding the investment in the Pod.

Repair is an important consideration, and the ease of repair is dependent on the architecture used. A failed pluggable transceiver can more easily be replaced given the easy access to the front panel whereas an on-board transceiver or a co-packaged transceiver failure may be more time consuming as it requires the card to be removed and replaced. The card or co-packaged assembly can later be repaired at the manufacturing facility.

7.2.4 Importance of diagnostics

The availability of robust diagnostics is also an important consideration. A key metric is the time to repair as the entire Pod is down during the repair cycle. Being able to quickly diagnose and determine what needs to be repaired is an important part of the repair cycle.

Diagnostics provide important clues in determining the cause of the failure. For example, diagnostics can be used to determine if the problem is in the electrical interfaces or in the optical interfaces and which end of the segment is at fault. The diagnostics can highlight link elements responsible for poor performance as well as pinpointing failed elements.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

7.2.5 Resiliency

Redundancy can be an effective path to reducing the failure rate. For example, introducing redundant lanes and components or redefining performance targets can effectively address these challenges with minimal overhead.

When a component fails, the transceiver can switch to using a spare component. There may be a short down time during the detection and switch over event but at least no components need to be replaced.

Redundancy promises significant improvements in reliability. For example, assume two redundant components each with a failure rate probability of p. The chance of both components failing would p^2, a much smaller values than a probability p without redundancy. However, it is important to consider if the two components were made in the same production process and lot. If so, the failure rate of the backup component may have significant correlation with the failure of the primary component. Such correlation can significantly reduce the advantages of redundancy. The improvement from redundancy depends on the degree of correlation and of the failure distribution of the failure mode involved.

7.2.6 The importance of operating temperature

Optical component FIT rates increase exponentially with junction temperature as shown in equation 7-1. The term EA is the activation energy related to the specific failure mechanism, k is Boltzmann's constant, TREF is the reference temperature in kelvin that the FIT_{REF} was specified at, and T is the operating temperature in kelvin.

Equation 7-1: $$FIT_T = FIT_{REF} \cdot exp\ exp\left(\frac{E_A}{k} * \left(\frac{1}{T_{REF}} - \frac{1}{T}\right)\right)$$

Depending upon the choice of technologies employed, the failure rate may be significantly improved when the optical components are kept cool. Operating at a higher temperature can significantly increase the failure rate.

7.2.7 Lasers

Optical transceivers need a laser source for the light generation. Many best-in-class lasers have a field based FIT rate in the range of a few FIT at 40 °C. The reliability contribution of multiple lasers in a transceiver can contribute a significant portion of the random failure

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

rate. As mentioned previously, higher failures rates for optical transceivers located behind the front panel is not desirable as the replacement is more disruptive.

There are two approaches to mitigate potential laser failures. One option is to locate the laser to a front pluggable external laser source such as the OIF's ELSFP modules thereby allowing easy repair. A second option is to locate the lasers behind the front panel but include sufficient laser redundancy. It is possible even in the presence of correlated failures to reduce the failure rate by 10x or more through redundancy.

7.2.8 Qualification Regimen

Historically, optical transceivers are qualified to the Telcordia standards including the well-known GR-468 as well as JEDEC standards. Applications which require higher reliability such as co-packaged silicon photonics may consider leveraging the Automotive Electronics Council qualification standards. These standards leverage JEDEC and the MIL standards and set higher hurdles that is common for the optical industry. Also, JEDEC is currently developing reliability standards for co-packaged optics.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

8 Scale-up Optical Transceiver Options

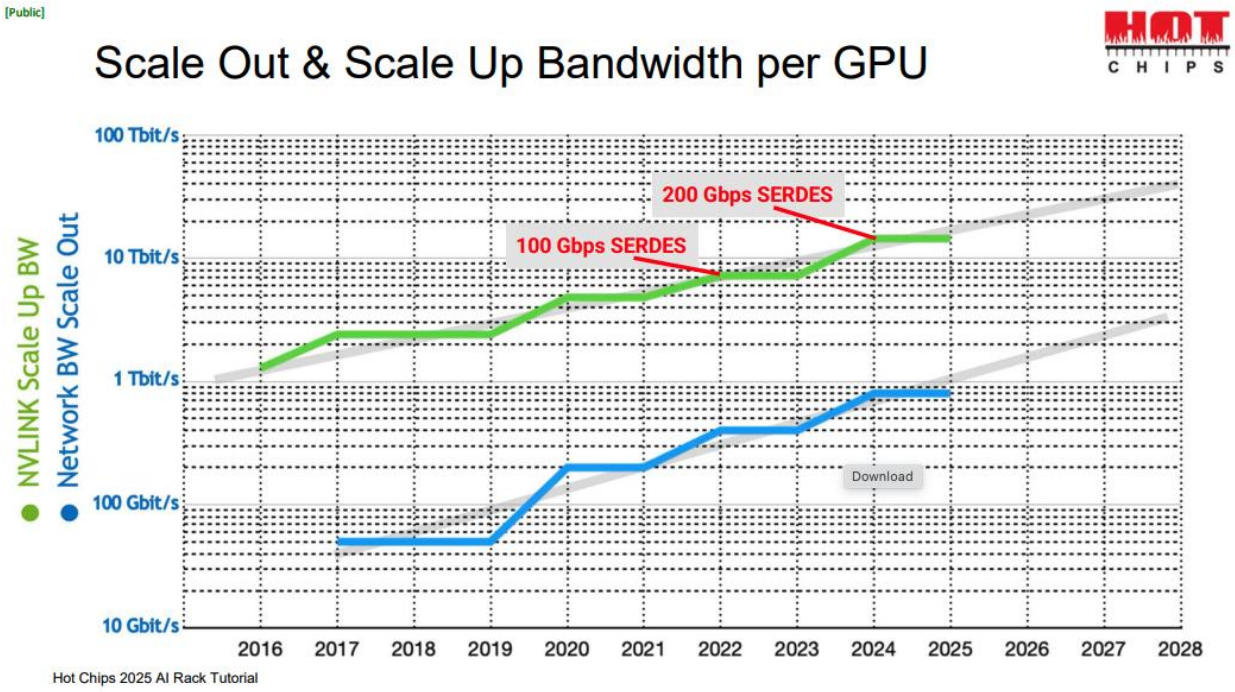


Figure 8-1 - Scale-out and Scale-up Bandwidth

Scale up connectivity bandwidth is about ~10x of scale out bandwidth, leading-edge scale up connectivity per GPU at 7.2 Tbps (e.g., NV GB200/300), and transitioning to 14.4 Tbps (e.g., NV VR200). This doubling of bandwidth every 2 years will continue in the foreseeable future. Current physical I/O is built with 200G-PAM4 per lane and combines two lanes to construct a 400G pipe. With the increasing bandwidth growth, 800G pipes will be necessary around 2028. These pipes can be constructed with 2x 400G, 4x 200G, or perhaps slower rates such as 16x 50G. The choice will be driven by the system architectural needs.

Majority of today's scale up is with copper cables which provided connectivity within a single rack. While system designers will continue to push the limit with copper cable, AI compute scalability demand will drive optical connectivity inevitable: larger AI model and inferencing service will demand ever larger scale up domain size from 64-72 XPUs to larger size such as ~ 500-1000, so multi-rack systems with optical connectivity solutions will be required. One

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

example is NVL576 that will use copper connectivity within rack, and optical connectivity between racks.

Scale up optics requires high bandwidth scalability, high reliability and power efficiency. There are multiple approaches to deploying optics as shown in Figure 8-2.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

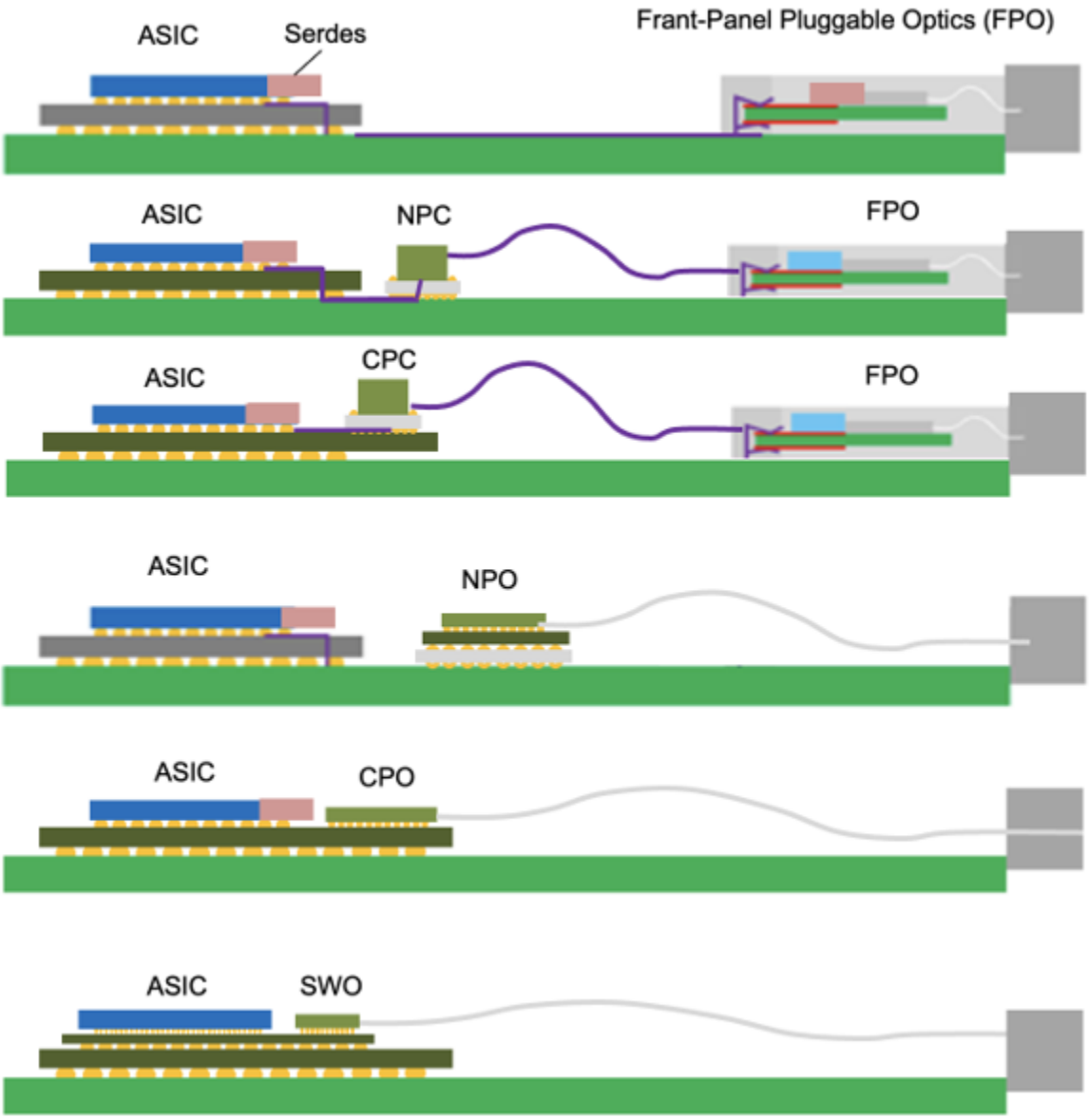


Figure 8-2 - O/E Placement Options

Trade-offs between these options are summarized in Table 8-1.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Table 8-1 -O/E Placement Tradeoffs

Solution Type	Bandwidth Scalability	Reliability	Power Efficiency	Deployment Considerations
Front Panel Pluggable Optics (FPLO)	1.6Tbps with 200G-PAM4 per lane per module	Traditional pluggable optics with discrete components with nominal FIT ~ 100, silicon photonics promising to significantly improve FIT	Pluggable optics with fully retimed DSP power dissipation ~ 15 pJ/b Half-retimed (RTLR) ~ 10pj/b Linear optics (LPO) ~ 6pj/b	OSFP MSA supports Open ecosystem Good serviceability and easy repair/replacement
	Scaling to 3.2T+ with 400G-PAM4	New pluggable Trx built with SiPho PIC and advanced packaging can improve FIT to <1-10 with high volume field deployment data support	Fully retimed power dissipation ~ 13pj/b Half-retimed (RTLR) pluggable TRx power ~ 8pj/b; Fully linear pluggables (LPO) power dissipation ~ 5pj/b optics	New MSA upgrade needed. Anticipate Good serviceability and easy repair/replacement
	12.8T pluggable optics with 64x200G, scaling to 25.6T per module with 400G-PAM4	With embedded liquid cooling, and more aggressive silicon photonics integration, reliability expected to improve vs. OSFP	Emerging NPC/CPC combined with PLO reduce loss between ASIC and PLO, making LPO viable at high baud rate Power dissipation similar to 1.6T OSFP, LPO type approaching ~ 5pj/b	XPO MSA supports open ecosystem Can support wide range of technologies such as Silicon Photonics, VCSELs, uLED, etc. Can achieve 4x higher faceplate density vs. OSFP Maintain Good serviceability and easy repair/replacement

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Solution Type	Bandwidth Scalability	Reliability	Power Efficiency	Deployment Considerations
Near Package Optics (NPO)	200G-PAM4 per lane today, and potentially scale to 400G per lane 6.4T NPO (32x200G) are in development today, 12.8T (64x200G, or 32x400G) coming soon	Mainly dependent on implementation technologies (e.g., VCSEL array, or Silicon photonics). Can effectively leverage on well developed silicon photonics technologies and advanced packaging for high reliability	Flexibility of half retimed or full linear (lowest power) with reduced PCB trace length and loss between ASIC and optics optics + 5pj/b	Open CPX MSA supports open eco system Require open Box repair/replacement at modest cost
Co-packaged Optics (CPO)	200G-PAM4 per lane today. Mostly are solder-down on ASIC substrate.	Silicon photonics with advanced packaging, with external CW laser source. SiPho Engine can be developed into high Reliability chiplets, over all determined by end-to-end system integration.	With close proximity of ASIC to CPO, over all power dissipation can be reduced with modest equalization expense. Power dissipation ~ 4-5 pJ/b	Mostly proprietary vertical integrated solution, there is potential for multi-source agreement with technology maturing Open CPX MSA addressing open socket CPO type, improve multi-sourcing supplies Still require openBox repairment, but with improved repair/replacement at high cost
Dense wavelength division multiplexing (DWDM)	SiPho MRM/MRR 25-50Gbps NRZ x N lanes	Reliability with CMOS like semiconductors and packaging	Going from serial PAM4 to multi-lane NRZ power, ~ 3pj/b optics + 2-3 pJ/b	OCI MSA defines a 4x50G-NRZ WDM BiDi PMD May deploy first as building block for pluggable optics, NPO, etc. Eventually may get deployed directly as a

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Solution Type	Bandwidth Scalability	Reliability	Power Efficiency	Deployment Considerations
				Die-to-Die interface to achieve density and lowest power
Slow and Wide Optics (SWO)	VCSEL: 25-50 Gbps-NRZ x N Lanes	Hybrid CMOS EIC + VCSEL reliability	Reverse gear box Serdes ~ 3pj/b optics + 2 pJ/b	Higher density parallel MMF and connector systems Can be implemented as one form of Open CPX MSA PMD type
	GaN uLED: 2-4Gbps-NRZ x N Lanes	Leverage on consumer GaN LED technology process and reliability, require "reverse gearbox" EIC	Reverse gear box Serdes ~ 3pj/b Optics + 1-2 pJ/b	Need to develop massive high count MCF, reach limit to 20-30meters Can be implemented as one of the Open CPX MSA PMD type

Surveying these options, two trajectories emerge as most viable for near-term scale-up deployment. Pluggable optics in the LPO and RTLR configurations offer the most accessible entry point. NPO represents the near-term option offering the best balance of power efficiency, density, and reliability improvement, particularly as 6.4T configurations enter development and silicon photonics packaging matures. CPO offers the lowest achievable power (4–5 pJ/b) and best signal integrity, but its predominantly proprietary integration model constrains near-term adoption to system designs where these tradeoffs are acceptable. SWO and VCSEL-based approaches are technically promising for the longer term, particularly where lane density and power dominate over reach, but manufacturing ecosystems at scale are not yet established.

Across all options, the decision is dominated by three coupled trade-offs:

1) Power efficiency versus serviceability: CPO and NPO gain efficiency at the cost of repair accessibility.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

2) Integration depth versus supply chain maturity: Tighter integration improves performance but concentrates sourcing risk.
3) Bandwidth scalability versus near-term reliability data: Higher-density solutions lack the field track record that scale-up's stringent FIT requirements demand.

System architects should treat the near-term landscape as a two-tier deployment model: LPO and RTLR for initial multi-rack Pod deployments where serviceability is prioritized, with a planned transition to NPO or CPO as reliability qualification data accumulates and optical interface standardization through efforts such as the CPX MSA matures.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

9 Looking Forward: IMDD Optical Transmission at 400G

From a systems perspective, it is desirable for the optical modulation format to mirror the electrical modulation format, the most significant benefit being that a gearbox to shift data formats and rates at the optical transceiver would not be required. It is unfortunate that as data rates shift from 200G to 400G, electrical signals over copper differential pairs, and IMDD optical signals over fiber have quantitatively different physical performance impairments. This leads to a curious situation in which the electrical and optical signaling at 400G have distinctly different tradeoffs for choosing between PAM4, PAM6 or PAM8. This section briefly reviews the different physical effects which lead 400G PAM4 to be most straightforward optical modulation format to implement by a reasonably large margin compared to PAM6 or PAM8.

9.1 Optical impairments

The first optical characteristic that limits optics at 400G is the optical modulation bandwidth of the optical transmitter. As detailed in section 9.2, this is highly dependent on the material system used to modulate the optical signals.

Even in a world where the optical modulation bandwidth does not limit the performance of high-speed optical systems, other specific impairments may impact optical performance in a way that is quantitatively different than those limiting the electrical performance.

First, as operating bandwidth increases so does the received optical power required to maintain a constant BER (or SNR). Similarly, as modulation order increases, the received optical power required to maintain a constant BER (or SNR) increases. Increased power requirements at the optical receiver require higher-power (more costly) optical transmitters.

Secondly, optical systems are subject to intensity-dependent noise. At lower received optical powers, thermal noise dominates, but, as received optical power increases shot noise and relative intensity noise (RIN), of which both depend on optical intensity begin to dominate. While RIN can be mitigated by using a higher-quality, lower-noise (more expensive) laser source, it ultimately asymptotically limits the how much BER benefit that can be achieved by increasing the received optical power. These limitations will be more serious for the higher modulation orders.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Finally other optical impairments such as interferometric intensity noise, affect higher order modulation compared to lower order modulation. As a result, PAM8 modulation would be subjected to much stricter optical return loss requirements compared to PAM4 modulation.

These optics-specific impairments mean that 400G PAM4 modulation will be the optimal optical modulation format compared to 400G PAM8 or PAM6 (if modulation bandwidth is ignored.)

9.2 Materials issues

Earlier standards for NRZ and PAM4-based have benefited from a wide variety of optical transmitter options. With the exception of multimode vs. single mode standards, optical transceiver architectures based on different materials could utilize different materials and physical modulation approaches while remaining interoperable. The most widely deployed approaches (at 50G / lane and above) are externally-modulated DFB lasers for single-mode fiber and directly modulated VCSELs for multi-mode fiber. Manufacturable externally-modulated transceivers based on silicon photonic ring or Mach-Zehnder interferometers or indium phosphide- based electro-absorption (and occasionally Mach-Zehnder) modulators have worked well up to today's 200G data rate. Historically, the doubling of data rate and/or symbol rate have required at least some improvement / engineering to the optical materials and electronic structures meet performance requirements

As of this writing, the optical materials issues for 200G to 400G seem more profound than in previous speed transitions. Firstly, the modulation bandwidth for the plasma dispersion effect in silicon seems too low to allow 400G PAM4 modulation. Indium phosphide and new class of thin film crystals such as lithium niobate and barium tantalate have all demonstrated their ability to support 400G PAM4. The only approaches to use silicon photonic-based modulators for 400G PAM4 require development of new hybrid integration of silicon photonic waveguides with other high-speed materials. It is unclear how friendly these hybrid approaches will be to either existing silicon photonics manufacturing processes or if the small per lane device dimensions needed for co-packaged silicon photonics can be preserved.

Efforts to standardize VCSELs for 200G PAM4 have commenced within IEEE 802.3 (Ethernet) just as the Ethernet standards for 200G PAM4 single-mode optics are approaching final approval. VCSELs have always seemed to require significant changes in the design of the semiconductor laser structure at each speed increase which resulted in a lag in time-to-market compared to corresponding single-mode solutions. At present a viable 200G PAM4

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

VCSEL manufacturing ecosystem seems ready to emerge. While it is clear that it will require a substantial effort to expand VCSEL bandwidth to support any 400G modulation format, it is too early to assess the manufacturing viability.

To date commercial prototypes of 400G PAM4 optical transceivers were demonstrated at both the Optical Fiber Communications Conference, and at the European Conference on Optical Communications in 2025. Barriers to use of PAM6 or PAM8 optical modulation formats could be lowered only by significant loosening of pre-FEC BER requirements in the systems.

There are other IMDD concepts for 400G optical modulation, but most are commercially speculative at this moment.

9.3 Emerging Alternatives

At the same time, the broader roadmap for 400G/lane optics is no longer focused exclusively on IMDD. For longer-reach links, or for links traversing optical circuit switch (OCS) fabrics, coherent-lite is emerging as a complementary approach that trades higher DSP complexity for greater reach, improved link margin, and better tolerance to channel impairments [28]. Recent industry developments around 1.6T coherent-lite pluggables suggest that this approach may become increasingly relevant for campus interconnect and intra-data-center optical applications.

In parallel, the Optical Compute Interconnect (OCI) MSA [21], announced in March 2026, points to a broader architectural shift toward an open, silicon-centric optical PHY for AI scale-up. OCI defines an NRZ-plus-WDM roadmap and supports pluggable, on-board, and co-packaged implementations, reflecting a wider industry transition away from copper-reach-limited, module-centric interconnects and toward more tightly integrated optical solutions. OCI also represents a shift from narrow-and-fast signaling to slow-and-wide optical parallelism, with bandwidth scaled through multiple wavelengths per fiber rather than ever-faster lanes.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

10 Looking Forward: Copper Transmission at 400G

Similar to section 9, this section presents a forward-looking discussion on transmitting data at 400G. Instead of optical fibers, however, this chapter focuses on copper interconnects.

Again, somewhat like the discussion in section 9, the discussion of 400G copper interconnects centers on modulation scheme; specifically, PAM4 vs. PAM6 modulation (as PAM8 is too speculative for copper at this moment). At a fixed data rate, the modulation scheme determines the Nyquist frequency of the IO. For example, at 448 Gb/s data rate, PAM4 modulation leads to 112 GHz Nyquist while PAM6 modulation leads to 89.6 GHz Nyquist (baud rate is roughly 179.2 Gb/s). For copper interconnects, insertion loss worsens at higher frequency, a lower Nyquist frequency means a copper interconnect can go longer at a fixed insertion loss budget. So, from copper interconnect's perspective, PAM6 is more desirable. But, as discussed in section 9, optical interconnects prefer PAM4 due to various reasons. Since the discussion on PAM4 vs. PAM6 is on-going, and we cannot predict the outcome, for the rest of the section, data and concepts will be presented with a PAM4 or PAM6 distinction, wherever appropriate. This approach should provide the readers with the most data for informed decisions down the road.

10.1 Co-Packaged Copper Topology Length and Insertion Loss at 400G

As mentioned in our 200G discussion in §5, a typical copper interconnect design target is the 1 m cabled topology. At 400G, simply based on experience, the easiest topology for maximum chance of success is the co-packaged copper (CPC) topology. Indeed, reference [18] contains a 400G PAM4 end-to-end channel simulation study on this exact topology:

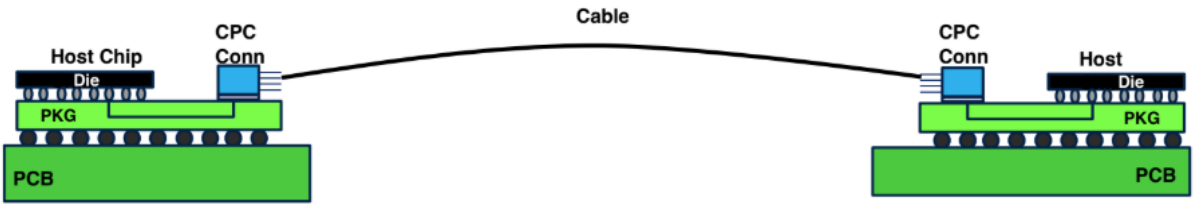


Figure 10-1- Copper Interconnect

As the authors mentioned, this study is the continuation of similar studies done and published in reference [19]. As explain in reference [19], the die plus package loss is scaled from 224G reference die and package to maintain a constant loss of -6.23dB at Nyquist.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Furthermore, the test channels, which include 1 m cable assembly, are also scaled to be within the range of -16 to -28dB of loss.

Table 10-1-Copper Losses

Link	Insertion Loss (dB)	
	PAM4 (112 GHz Nyquist)	PAM6 (89.6 GHz Nyquist)
224G Ref. Die+Pkg (33 mm)	-41.78	-21.29
448G Die+Pkg (scaled)	-6.23	-6.23
448G Test Channel (scaled)	-16 to -28	-16 to -28
448G Max End-to-End Loss (scaled)	Roughly -40	Roughly -40

Embedded in these numbers are assumptions that technology will improve to enable die plus package and 1 m cable assembly loss to become roughly -6 dB and -28 dB worst case, respectively, at Nyquist. While currently there are no data on how die and package technologies may scale to 400G; there are some publicly available data on CPC. From reference [20], measured 400 mm CPC cable assembly shows roughly -31 dB and -37 dB of loss at 90 GHz and 110 GHz, respectively. It is, however, unclear whether loss from measurement fixture has been de-embedded or not.

The conclusion from these three references is 448 Gb/s PAM4 copper interconnects can support endpoints connected by about one meter of CPC-to-CPC cable [18]. Since important assumptions such as improvements in packaging technology and SerDes design are currently unclear, this conclusion may be optimistic. The information, however, is still valuable since it shows copper can still play an important role in 400G network, and CPC is probably the only viable topology.

10.2 Summary

References [18], [19], and [20] focus on 448G PAM4 modulation. As a first order approximation, PAM6 data can be obtained by scaling PAM4 data with the ratio of their Nyquist frequencies, which is 1.25. As mentioned previously, the conclusion of 1 m CPC cable topology being viable at 448G PAM4 could be optimistic. In a speculative manner, 0.5

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

m CPC cable may be more realistic, and 0.5 meter of internal cable is still very meaningful in enabling a 400G network.

Some final comments:

- Before the end-to-end loss budget becomes clear, discussion in using re-timers to enable completely copper 400G network is too speculative.
- Near-package copper cable may be viable at 400G with PAM6, but currently it is also too speculative due to uncertainty in package technology improvement.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

11 Conclusion

The transition from passive copper to optical interconnects in the AI scale-up domain is a reach-driven architectural evolution. The relevant variables are timing, architectural alignment, and disciplined implementation. As Pod configurations expand from single-rack systems of 64 to 72 accelerators to multi-rack deployments of 500 to 1,000 or more, copper reach at high lane rates such as 200 Gb/s, typically limited to approximately 1 to 1.5 meters even with fly-over cables and active conditioning, becomes a structural constraint. Optical interconnect therefore becomes an architectural requirement rather than an incremental enhancement.

This analysis establishes several key conclusions.

1. **SROI represents a distinct optical reach regime**.

Fiber attenuation and chromatic dispersion assumed to be negligible relative to system margins for the short reach regime (up to 30 m). The dominant design objectives shift to endpoint integration parameters: optical lane density, energy per bit, cost per lane, and reliability across rack-scale link counts. These objectives differ materially from those governing medium- and long-reach optical specifications and require correspondingly different architectural priorities.

2. **The performance and economic targets are defined and stringent.**

Energy per link end, inclusive of SerDes and optical elements, must converge toward single digits when expressed in pJ/bit to remain within system power envelopes at projected lane counts. Cost must trend towards passive copper on a per-bit basis when evaluated at total system level, including cooling and power delivery impacts. Reliability must approach semiconductor-grade FIT rates and must minimize transient error events that degrade synchronous, memory-semantic traffic.

3. **High levels of integration are likely necessary.**

Co-packaged and on-board optical architectures offer advantages in energy efficiency, density, and reliability relative to pluggable implementations in scale-up environments, with tradeoffs in serviceability. Availability must therefore be achieved through architectural techniques including lane redundancy, component redundancy, diagnostics, predictive management, and controlled operating temperatures. Operating optical components near

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

approximately 40°C rather than traditional datacenter optical temperatures near 70°C provides measurable reliability improvement under established acceleration models.

4. **The benign channel characteristics of the SROI regime expands architectural flexibility.**

In medium- and long-reach optics, channel impairments and optical budgets favor fewer, higher-rate lanes because each lane carries fixed overhead in optics, fiber, and amplification. In the short-reach regime, where channel impairments are negligible, this constraint relaxes. Aggregate bandwidth can be partitioned across more lanes at lower per-lane rates if doing so improves energy efficiency, cost, reliability, manufacturability, or supplier diversity.

For example, an 800G port may be implemented as 4 × 200G or 2 × 400G or 8 × 100G and so on. A wider, lower-rate configuration reduces modulation bandwidth requirements on the photonic platform and may broaden the set of qualified suppliers while simplifying SerDes and driver design. A narrower, higher-rate configuration reduces fiber and connector count, which may be advantageous at high port densities. The appropriate balance depends on reach tier, topology, thermal envelope, and supply chain maturity at deployment.

Accordingly, system-level specifications should define aggregate bandwidth per port and an acceptable lane-rate envelope rather than mandating a single lane rate, enabling ecosystem optimization within defined performance, energy, and reliability objectives.

Looking forward, SROI should be understood as more than a copper substitution. When implemented within defined energy, cost, and reliability envelopes, short-reach photonics enable architectural configurations that are impractical under copper constraints. These include rack-scale disaggregated memory, composable I/O fabrics, and tighter chiplet integration supported by optical interconnect across multiple reach tiers. The resulting shift is architectural rather than incremental.

The Open Compute Project community can accelerate this transition through coordinated system-level action in four areas. First, clearly define optical reach categories for intra-rack and inter-rack scale-up aligned with established electrical reach taxonomies, providing structured input to standards development organizations. Second, establish tiered reliability qualification frameworks appropriate to in-package, intra-rack, and inter-rack short-reach regimes. Third, define interoperable optical interface objectives for scale-up fabrics that preserve multi-vendor ecosystems without mandating a specific photonic implementation. Fourth, align optical interface roadmaps with emerging scale-up protocols

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

such as UALink and ESUN and with adjacent transitions including optical PCIe and CXL-based memory and storage fabrics, ensuring that optical readiness does not become the gating constraint for next-generation Pod architectures.

Where feasible, common SROI objectives for lane rate, fiber class, connector interface, and reliability qualification should be harmonized across protocol families to enable unified optical platforms rather than fragmented, application-specific solutions.

AI infrastructure design cycles are compressing. Accelerator bandwidth continues to scale on a one- to two-year cadence. Infrastructure investments made today will define deployed architectures several years forward. The optical reach framework established in the near term will determine whether the scale-up fabric remains the principal system constraint or evolves into a scalable foundation for distributed, memory-semantic computing.

System architects and operators must articulate these objectives clearly and early so that standards bodies and the supply chain can respond in alignment with deployment timelines.

A consolidated summary of the system-level requirements and technology metrics established in this paper is provided in Appendix A for use by optics providers and standards contributors.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

12 Glossary

Accelerator: Specialized hardware designed to speed up AI workloads

Activation Energy (EA): The minimum energy for a process to proceed by having sufficient energy to move across an energy barrier

ACC: Active Copper Cable which employs electronics in the cable to improve signal integrity

AEC: Active Electrical Cable

AI: Artificial Intelligence

AOC: Active Optical Cable which employs electrical to optical conversion to communicate over a fiber optic link

Areal density: A density reflecting the quantity per unit area

ASIC: Application Specific Integrated Circuit

BiDi: Bidirectional traffic on a single media which could be electrical or optical

BW: Bandwidth, usually expressed in Gb/s or Tb/s

CPC: Co-Packaged Copper interconnects electrical copper with the ASIC in the same package

CPk: Process Capability Index is a statistical measure of how well a process can produce output within specification limits, while accounting for how centered the process is

CPO: Co-Packaged Optics interconnects fiber to the ASIC in the same package

CPX-MSA: The Open CPX MSA is developing specifications for the attachment for co-packaged optics (CPO) and near-packaged optics (NPO)

CRC: Cyclic Redundancy Check indicates if a received block contains an error

CWDM: Coarse Wavelength Division Multiplexing indicates that multiple wavelengths of light carry traffic on the same fiber core with wavelength separations in 10's of nm

DAC: Direct Attach Copper is a passive copper cable assembly used for short reach links

DC: Data Center

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

DR: Data Center Reach is an IEEE designation for a single wavelength per fiber core with a typical reach of 500 m

DWDM: Dense Wavelength Division Multiplexing indicates that multiple wavelengths of light carry traffic on the same fiber core with wavelength separations in 100s of GHz

ESUN: Ethernet Scale-up Networking is streamlined Ethernet implementation providing lower latency than standard Ethernet

FEC: Forward Error Correction

FIT: Failures in Time is a measure of a device's failure rate expressed in failure per billion hours

FLIT: Flow Control Unit is a fixed size block for data transfer and flow control for a high-speed serial interconnect

FLOP: Floating Point Operation; FLOPS: Floating Point Operations per Second

FR: Fiber Reach is an IEEE designation for a four wavelengths per fiber core with a typical reach of 2 km

JEDEC: Joint Electron Device Engineering Council

GAN: Gallium Nitride

GPU: See accelerator

HBM: High-bandwidth Memory

IEEE: Institute of Electrical and Electronics Engineers

IL: Insertion loss

InfiniBand: A high-performance networking technology designed for very low latency, high bandwidth, and scalable interconnects, especially in HPC and AI systems

LED: Light Emitting Diode

LTLR: Linear Transmitter Linear Receiver used for NPO and pluggable optics (LPO)

LPO: Linear Pluggable Optics is a pluggable module with a linear transmitter and a linear receiver

MIL: Military based standards

MMF: Multi-mode Fiber

Date: August 5, 2026

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

NIC: Network Interface Card

Node: A node is a compute element containing one or more accelerator and one or more CPUs

NPC: Near-Packaged Copper is connection to a copper cable placed in close proximity to the ASIC but not co-packaged with the ASIC

NPO: Near-Packaged Optics are modules placed in close proximity to the ASIC but not co-packaged with the ASIC

NVLink: NVLink is a high-bandwidth, low-latency interconnect developed by NVIDIA for direct communication between GPUs and between GPUs and CPUs

OCI-MSA: An industry Multi-Source Agreement (MSA) that defines standard electrical and mechanical specifications for high-speed Optical Compute Interconnect (OCI) modules, primarily used in AI systems

OIF: Optical Internetworking Forum

OIO: Optical I/O

OSFP: Octal Small Form factor Pluggable

OU: Open Unit is a standardized vertical height unit use by OCP for racks

PAM4: Pulse Amplitude Modulation with 4 levels

PAM6: Pulse Amplitude Modulation with 6 levels

PAM8: Pulse Amplitude Modulation with 8 levels

PCB: Printed Circuit Board

PCIe: Peripheral Component Interconnect Express

Pipe: Represents a capacity for an interconnecting link independent of the implementation. For example, a 400G pipe could consist of 1x 400G signal, or an 8x50G signals across one or more fibers

PLO: Pluggable Optics

POD: A self-contained, repeatable building block of AI infrastructure that groups compute, networking, and power/cooling to operate as a single scalable unit

RDMA: Remote Direct Memory Access

Date: August 5, 2026

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

RTLR: Retimed Transmitter, Linear Receiver sometimes referred to as half retimed. When used in pluggable modules, it is often referred to as LRO

RTRR: Retimed Transmitter, Retimed Receiver sometimes referred to as fully retimed.

RTT: Round Trip Time

SerDes: Serializer Deserializer

SiPho, SiPh: Silicon Photonics

SROI: Short Reach Optical Interconnect

SuperLink: A very high-bandwidth, low-latency link used to tightly connect GPUs or other accelerators, so they behave like a single scale-up system

UALink: Ultra Accelerator Link

UEC: Ultra Ethernet Consortium

VCSEL: Vertical Cavity Surface Emitting Laser

XPO-MSA: The XPO MSA is developing specification for liquid-cooled pluggable optical module designed for hyperscale AI data centers with a capacity of 12.8T.

XPU: Generic processing unit designed for AI workloads, could be any of CPU, GPU, NPU or TPU.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

13 References

[1] E. R. Jaime Sevilla, "Training Compute of Frontier AI Models Grows by 4-5x per Year," EPOCH AI, 2024. https://epoch.ai/blog/training-compute-of-frontier-ai-models-grows-by-4-5x-per-year

[2] Llama Team, AI@Meta, "The Llama 3 Herd of Models," Meta, 2024. https://arxiv.org/abs/2407.21783

[3] W. Wang, M. Ghobadi, K. Shakeri, Y. Zhang and N. Hasani, "Rail-only: A Low-Cost High-Performance Network for Training LLMs with Trillion Parameters," IEEE Symposium on High-Performance Interconnects (HOTI), 2024. https://arxiv.org/abs/2407.12644

[4] P. Zhang and S. Lei, "Memory Analysis on the Training Course of DeepSeek Models," arXiv:2502.07846v1, 2025. https://arxiv.org/abs/2502.07846

[5] Pangu Team, Huawei, "PANGU ULTRA MOE: How to train your big MOE on Ascend NPUs," Huawei, 2025. https://arxiv.org/abs/2501.12574

[6] T. Brown et al., "Language Models are Few-Shot Learners," Advances in Neural Information Processing Systems, 33, 1877-1901, 2020. https://arxiv.org/abs/2005.14165

[7] OpenAI, "GPT-4 Technical Report," arXiv:2303.08774, 2023. https://arxiv.org/abs/2303.08774

[8] A. Chowdhery et al., "PaLM: Scaling Language Modeling with Pathways," Journal of Machine Learning Research, 24(240):1-113, 2023. https://arxiv.org/abs/2204.02311

[9] NVIDIA Corporation, "NVIDIA H100 Tensor Core GPU Architecture Whitepaper," 2022. https://resources.nvidia.com/en-us-tensor-core

[10] NVIDIA Corporation, "NVIDIA Blackwell Architecture Technical Brief," 2024. https://resources.nvidia.com/en-us-blackwell-architecture

[11] Google Cloud, "Cloud TPU v5p architecture and supported configurations," Google Cloud Documentation, 2024. https://cloud.google.com/tpu/docs/v5p

[12] Advanced Micro Devices, Inc., "AMD Instinct™ MI300X Accelerator Data Sheet," 2023. https://www.amd.com/en/products/accelerators/instinct/mi300/mi300x.html

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

[13] D. Narayanan et al., "Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM," Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC21), 2021. https://arxiv.org/abs/2104.04473

[14] S. Rajbhandari et al., "ZeRO: Memory Optimizations Toward Training Trillion Parameter Models," SC20: International Conference for High Performance Computing, Networking, Storage and Analysis, 2020. https://arxiv.org/abs/1910.02054

[15] NVIDIA Corporation, "NVIDIA NVLink and NVSwitch: The building blocks of advanced multi-GPU communication," Architecture Whitepaper, 2024. https://www.nvidia.com/en-us/data-center/nvlink/

[16] Ultra Ethernet Consortium, "Ultra Ethernet: High-Performance Networking for AI and HPC," UEC Whitepaper, 2024. https://ultraethernet.org/

[17] Ultra Accelerator Link Consortium, "UALink 1.0 Specification Overview," 2024. https://ualinkconsortium.org/

[18] “Next Generation Optical Interconnects” - https://drive.google.com/file/d/1sofkb7CpYGeWYH67lANT10B8HCBOKOlg/view?usp=drive_link

[19] “TX/RX, Channel, and Signaling Investigations at 448 Gbps” - https://www.oiforum.com/wp-content/uploads/448G_signaling_channel_MikeL_Intel_final_040425.pdf

[20] “Extending Copper Interconnects to 448G: Innovations Overcoming Conventional Limitations” - https://drive.google.com/file/d/16lOHUHiWdMJe8ei74L-V-VyCOIgEtc-5/view?usp=drivesdk

[21] OCI-MSA: https://oci-msa.org/

[22] IOWN Global Forum, “Functional Architecture for Optically-Accelerated AI Interconnect,” Tech. Rep., Mar. 2026. [Online]. Available: https://iowngf.org/wp-content/uploads/2026/04/IOWN-GF-RD-Functional-Architecture-for-Optically-Accelerated-AI-Interconnect-1.0.pdf

[23] M. Stepanovsky, “A comparative review of MEMS-based optical cross-connects for all-optical networks from the past to the present day,” IEEE Commun. Surveys Tuts., vol. 21, no. 3, pp. 2928–2946, 2019.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

[24] M. A. Mahdian, et al., “A Comprehensive Survey on Integrated Photonic Switch Fabrics: From Device to Network Topology and Controller Design,” IEEE Access, 2026.

[25] D. Piehler, “Optical Opportunities for the AI Scale-up Fabric”, 2025, https://drive.google.com/file/d/1TR5hPe34dyof2xGjQ492jYlyb-1OW5zX/view?usp=drive_link, https://www.opencompute.org/events/past-events/2025-ocp-global-summit#fti-short-reach-optical-interconnects

[26] OCP Foundation, “Optical Circuit Switching for AI and Hyperscale Data Centers,” White Paper, Apr. 2026.

[27] R. Urata, et al., “Mission Apollo: Landing Optical Circuit Switching at Datacenter Scale,” arXiv, 2022.

[28] X. Zhou, “IM-DD vs. Coherent in Datacenters: A Revisit in 2025,” OFC 2025, M4C.1.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

14 License

Creative Commons

OCP encourages participants to share their proposals, specifications and designs with the community. This is to promote openness and encourage continuous and open feedback. It is important to remember that by providing feedback for any such documents, whether in written or verbal form, that the contributor or the contributor's organization grants OCP and its members irrevocable right to use this feedback for any purpose without any further obligation.

It is acknowledged that any such documentation and any ancillary materials that are provided to OCP in connection with this document, including without limitation any white papers, articles, photographs, studies, diagrams, contact information (together, “Materials”) are made available under the Creative Commons Attribution-ShareAlike 4.0 International License found here: https://creativecommons.org/licenses/by-sa/4.0/, or any later version, and without limiting the foregoing, OCP may make the Materials available under such terms.

As a contributor to this document, all members represent that they have the authority to grant the rights and licenses herein. They further represent and warrant that the Materials do not and will not violate the copyrights or misappropriate the trade secret rights of any third party, including without limitation rights in intellectual property. The contributor(s) also represent that, to the extent the Materials include materials protected by copyright or trade secret rights that are owned or created by any third-party, they have obtained permission for its use consistent with the foregoing. They will provide OCP evidence of such permission upon OCP’s request. This document and any "Materials" are published on the respective project's wiki page and are open to the public in accordance with OCP's Bylaws and IP Policy. This can be found at http://www.opencompute.org/participate/legal-documents/. If you have any questions, please contact OCP.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

15. About Open Compute Foundation

The Open Compute Project (OCP) brings at-scale innovations and hyperscaler best practices to all, spanning technology domains from the data center to the edge, and the technology stack from silicon, to systems, to site facilities and services. The international OCP Community is made up of organizations and people from hyperscale and tier-2 cloud data center operators, communications providers, colocation providers, diverse enterprises, and technology vendors. The OCP Foundation fosters collaboration within the Community to tackle market challenges in an array of OCP Projects that create open contributions such as specifications, designs, and more. The OCP Solution Provider Program then recognizes the application of those contributions in compliant products, solutions, and data center facilities and services with designations like OCP Inspired, OCP Accepted and OCP Ready. With the tenets of openness, impact, efficiency, scale and sustainability, the OCP engages and educates thousands of engineers every year through many events and webinars. Across many initiatives the OCP Foundation and Community are meeting the market today and shaping the future.

Learn more at: www.opencompute.org.

Date: August 5, 2026

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Appendix A. SROI Requirements Summary

This appendix consolidates the system-level requirements and technology metrics established throughout this white paper into a single reference for optical transceiver providers, system architects, and standards contributors. Values are drawn directly from the sections indicated. "Threshold" denotes the minimum bar for broad adoption; "Goal" denotes the aspirational target aligned with highly integrated (CPO-class) solutions. All parameters apply to scale-up optical interconnects within the SROI reach regime (≤ 30 m) unless otherwise noted.

Table A-1 — Top-Level System Metrics

Parameter	Threshold Target	Goal	Notes	§
Power efficiency (optics only, one link end)	Single digits pJ/b	≤ 5 pJ/b	Measured at optical interface, exclusive of SerDes. Threshold matches AEC long-reach copper.	6.1.1
Power efficiency (SerDes + optics, one link end)	≤ 10 pJ/b	≤ 7 pJ/b	Full link-end budget. Broadcom 10 pJ/b identified as the copper-to-optics adoption crossover.	6.1.1
Cost per bit	Trending toward passive copper ($/Gbit)	Competitive with AEC on $/Gbit/m basis	Passive copper baseline ~$0.05/Gbit. AEC is the near-term competitive reference.	6.1.2
Hardware reliability (FIT rate)	≤ 2 FIT per optical link end	≤ 2 FIT per optical link end	Traditional pluggable modules: best in class FIT rate of 50–100 . Target reflects laser-limited floor under controlled temperature.	6.1.3 7.2
Link error rate (transient)	0 link-flaps / 1M device hours	Zero FEC tail exceedance; zero uncorrectable errors	Broadcom CPO data cited as feasibility demonstration. FEC tail monitoring required.	6.1.3 7.1
Latency contribution (optics)	Equivalent to removing ≥ 1 retimer	No retimer in optical path	Each retimer adds ~50 ns. Each meter of fiber adds ~5 ns. Retimer elimination is the primary lever.	6.1.4
System round-trip time (scale-up fabric)	< 1 µs	Minimize	Full XPU-to-XPU RTT target including fabric switching. Tail latency from retries is additive.	4.4

Date: August 5, 2026

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Table A-2 — Optical Physical Layer Parameters

Parameter	Threshold Target	Goal	Notes	§
Maximum reach (SROI regime)	≤ 30 m	—	Defines the SROI regime boundary. Fiber attenuation and chromatic dispersion are negligible throughout this range for SMF and MMF.	2
Minimum reach requiring optics	> 3 m (copper limit with active conditioning)	—	Passive copper: ~1–1.5 m at 200G/lane. With re-timers: ~2.5–3 m maximum. Beyond 3 m, optical is architecturally required.	5.2
Fiber type	SMF or MMF	—	Both acceptable within the SROI regime. No modal or dispersion penalty at ≤ 30 m at current and projected lane rates.	2
Maximum operating temperature (optical components)	≤ 70 °C	≤ 40 °C	FIT rates increase exponentially with junction temperature (EQ 7-1). 40 °C goal provides measurable reliability improvement vs. traditional datacenter optics.	7.2.6
Optical modulation format (current generation, 200G/lane)	PAM4	PAM4	PAM4 preferred due to RIN, interferometric noise, and return-loss constraints. PAM6/PAM8 require significant relaxation of pre-FEC BER requirements.	9.1
Optical modulation format (400G/lane, forward-looking)	PAM4 at 400G/lane	PAM4 or NRZ+WDM (OCI-MSA approach)	Silicon photonics bandwidth limitations make 400G PAM4 most tractable. WDM (slow-and-wide) is an emerging alternative with lower per-lane rate.	9.1 9.3
Lane rate (current generation)	200G PAM4 (212.5 Gbps)	—	Based on IEEE 802.3dj. UALink and ESUN both build on this electrical specification.	3.1 5
Lane rate (next generation, ~2028)	400G PAM4	—	Required to support 800G pipes. May be implemented as 2×400G, 4×200G, or 8×100G within the SROI regime depending on system architecture.	8

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Retiming architecture	Non-retimed (LTLR) or half-retimed (RTLR)	Non-retimed (LTLR / LPO)	Fully-retimed (RTRR) increases power by ~3–4× vs. non-retimed. Each retiming stage adds latency and introduces additional hard-decision error probability.	3.3 6.1.4

Table A-3 — Density and Radix Requirements

Parameter	Threshold Target	Goal	Notes	§
Interconnect radix (scale-up switch)	> 1024 links per switch	—	Switch radix currently the binding constraint, limited by package escape bandwidth. Higher radix reduces hop count and latency.	6.2.1
Linear BW density at ASIC edge (CPO placement)	1 Tbps/mm (bidirectional)	2 Tbps/mm (bidirectional)	Current ASIC density (e.g., Broadcom TH6): ~1.7 Tbps/mm (Tx+Rx combined). Density doubles each generation. CPO optics must match or exceed this density.	6.2.2
Linear BW density at package edge (NPO placement)	0.5 Tbps/mm (bidirectional)	1 Tbps/mm (bidirectional)	Package dimensions (60–90 mm) relax the density requirement 2–3× relative to the chip edge (30 mm). 1 Tbps/mm is the appropriate NPO target.	6.2.2
Areal BW density	Enable CPO footprint on ASIC substrate	—	Constrained by substrate size (~90 mm edge) relative to multi-reticle ASIC edge (~65 mm). Specific numerical target under active development.	6.2.2
Fiber count at faceplate (rack level)	≥ 10,000 fibers per rack	—	Estimated for a 512-XPU cluster at 64 XPU/rack with required switches. Drives adoption of high-density connector and cable systems.	6.2.3
Aggregate scale-up BW per XPU (current generation)	7.2 Tbps	—	Reference: NV GB200/B300. Drives the total optical lane count per accelerator. Scale-up BW is ~10× scale-out BW.	8

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Aggregate scale-up BW per XPU (next generation)	14.4 Tbps	—	Reference: NV VR200. Approximately doubles every 2 years on current trajectory.	8
WDM per fiber	Meet the specified link BW, single λ preferred	Single λ per fiber	Radix preservation favors separate fibers per lane. WDM demultiplexing prior to fiber shuffle adds cost and may impact rack density.	6.2.1

Table A-4 — Reliability and Qualification Requirements

Requirement	Threshold	Notes	§
Infant failure elimination	Manufacturing screening required before system installation	Process controls and burn-in must eliminate infant failures. Google has advocated extended burn-in to lower the effective FIT floor.	7.2.2
Wear-out failure regime	Must fall beyond operating lifetime	Design and qualification must keep wear-out failures outside the service window through careful design and process control.	7.2.2
New technology validation	Track record in fault-tolerant applications required prior to scale-up deployment	Especially critical for CPO and on-board placements where repair requires card removal and halts Pod computation.	7.2.2
Qualification standard (pluggable optics)	Telcordia GR-468 and JEDEC standards	Baseline qualification for pluggable modules. Sufficient for front-panel replaceable implementations.	7.2.8
Qualification standard (CPO and on-board)	Automotive Electronics Council (AEC) or equivalent; JEDEC co-packaged optics standard (in development)	Higher bar required due to increased repair cost and computation downtime of behind-panel architectures.	7.2.8
Diagnostic capability	Required — fault isolation to electrical vs. optical interface and near-end vs. far-end	Pod is fully halted during repair. Diagnostic speed and precision directly determine downtime cost. Diagnostics are a key differentiator.	7.2.4

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Lane or component redundancy	Recommended for pluggable; required for CPO and on-board architectures	Correlated failures within same production lot reduce redundancy benefit; design must account for correlation when specifying redundancy gain.	7.2.5
Operating temperature (FIT impact)	≤ 70 °C (minimum); ≤ 40 °C (goal)	FIT rates increase exponentially with junction temperature per Arrhenius model (EQ 7-1). CPO and NPO architectures enable lower operating temperatures than front-panel pluggables.	7.2.6
Laser source reliability	≤ 2 FIT per laser source at 40 °C	Multiple lasers per transceiver compound the system FIT contribution. External laser source (e.g., OIF ELSFP) or on-board redundancy are the two mitigation strategies.	7.2.7

Table A-5 — Module Placement Option Comparison

PLO = Pluggable Optics; LPO = Linear Pluggable Optics; RTLR = half-retimed pluggable; NPO = Near-Package Optics; CPO = Co-Packaged Optics; SWO = Slow-and-Wide Optics.

Parameter	Pluggable (PLO/LPO/RTLR)	Near-Package (NPO)	Co-Packaged (CPO)	Slow & Wide (SWO)	§
Power efficiency (optics only, one end)	LPO: ~5–6 pJ/b RTLR: ~8–10 pJ/b Full retimed: 13–15 pJ/b	~5 pJ/b (linear / LPO-class)	4–5 pJ/b	~3 pJ/b optics +2–3 pJ/b SerDes gearbox	8
Hardware FIT (current)	50–100 FIT (discrete components); improving with SiPho PIC	Dependent on implementation; SiPho enables significant improvement	Approaching silicon-grade; Meta data: 5× fewer unserviceable failures vs. pluggable	CMOS-like reliability; field data limited	7.2 8
Bandwidth scalability	1.6T today (OSFP) 3.2T–12.8T with 400G PAM4 (XPO MSA)	6.4T today (32×200G) 12.8T coming (64×200G or 32×400G)	200G PAM4 today; roadmap tied to ASIC generation	25–50G NRZ × N lanes; scales via lane count and WDM	8

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

Repair accessibility	Front panel; field replaceable	Open-box; board removal required	Open-box; card or substrate removal; factory repair	Dependent on host integration	7.2.3 8
Supply chain / MSA	Established (OSFP, XPO MSAs)	Emerging (CPX MSA)	Primarily proprietary; CPX MSA addressing open socket CPO	OCI-MSA defines 4×50G-NRZ WDM PMD	8 9.3
Near-term deployment readiness	Highest (LPO and RTLR configurations)	High (6.4T in development)	Moderate (proprietary integration; limited multi-source)	Emerging (manufacturing ecosystem not yet at scale)	8
Key trade-off	Serviceability vs. power efficiency gap vs. CPO/NPO	Balance of efficiency, density, and reliability improvement	Lowest power and best signal integrity; highest repair cost and integration risk	Lowest power potential; requires reverse-gearbox SerDes and high lane count fiber	8

Note: The values in this appendix represent system-level targets as of the date of this white paper. They are intended to inform standards development and supplier roadmap planning, not to constitute a formal specification. Specific electrical and optical interface parameters at the transceiver level are deferred to the relevant IEEE 802.3, OIF, JEDEC, and MSA bodies. Targets marked "Goal" reflect integration-class solutions (CPO/NPO) and may not be achievable with pluggable architectures in the near term.

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.