Scaling the Rack: Electrical Interconnect for Hyperscaler Compute Demands

Halil Cirit
META

Outline

- Designing data centers around AI
- Targets for AI infrastructure
- 72 GPU scale-up rack system and connection lengths inside
- CPC channel analysis
- Modulation and FEC conversations
- ACC as an alternative solution for reach problem
- Disaggregation an OCI for future scale-up systems
- What hyperscalers need

Re-designing the Data Center for AI

Massive model traffic is forcing a new stack

01 | RACK-SCALE FABRIC

>130 TB/s

Low-latency GPU communications inside one 72-GPU rack [1]

≈ 1 PB every 8 sec

02 | AI NETWORK

200G → 400G

Per-lane network evolution for AI fabrics [2]

UP TO 400 Tb/s

announced switch-system throughput [2]

03 | FACILITY SCALE

>4×

AI-optimized data-center electricity demand by 2030 [3]

945 TWh

projected global data-center electricity demand in 2030 [3]

The next data center is an AI factory: compute, network, power and latency engineered as one system

[1] NVIDIA GB200 NVL72 [2] IEEE P802.3dj [3] IEA Energy and AI, Apr 2025

Targets for AI Infrastructure & Multiplier Effect

Hyperion scales the campus; dense scale-up racks multiply short-reach interconnect

HYPERION

5 GW

Planned compute capacity

$50B+ regional investment

Meta, Jul 2026 [1]

REPEATABLE CAMPUS

1 GW

Prometheus

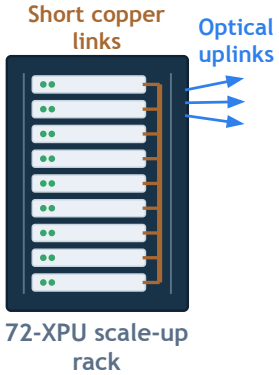
Meta, Feb 2026 [2]

FIBER SUPPLY

UP TO $6B

Meta, Jan 2026 [3]

The rack-level multiplier



≈ 1000 short copper connections per 72-XPU rack

≈ 10.000 connections across a 10-rack row

Cost implication: cheaper per link × far more links = copper can dominate rack-level interconnect spend

Reliability: cheaper passive components × far more reliability (no laser failure) = less maintenance

Design consequence: Everything scales to huge numbers, compute as well as the cost

[1] Meta, Jul 2026 — Hyperion 5 GW / $50B+ [2] Meta, Feb 2026 — Lebanon 1 GW campus
[3] Meta, Jan 2026 — Corning agreement up to $6B

Detailed Look Into a 72 GPU Rack System

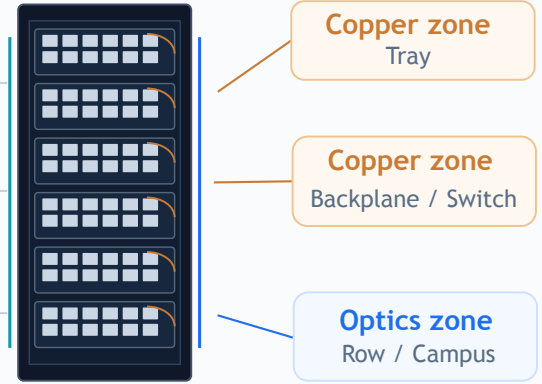
System size
72 GPUs
High-density rack domain

In-rack fabric
1,000+ short copper links
Short reach, high link count

Design rule
Copper first
Best fit for tray / in-rack reach

Rack anatomy [1]

- **9+9 = 18 Compute trays**
  Top and bottom
  4 GPUs across tray modules
- **9 Switch Trays**
  1296 Scale Up Ports
- **1.5m Backplane Cables**
  Connecting accelerators with switches



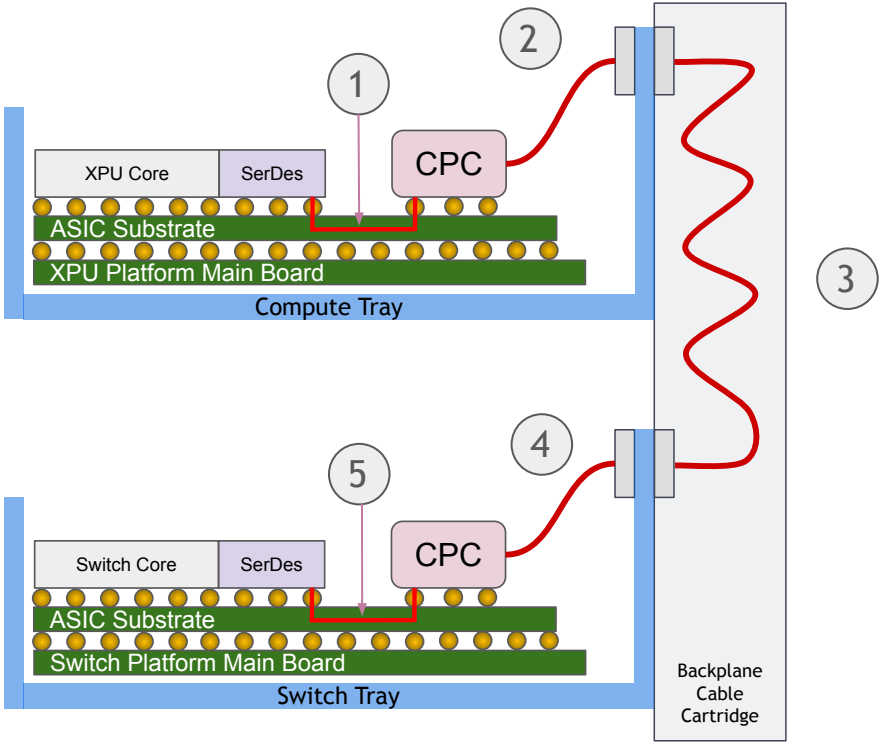
Why copper is efficient inside the rack

Power, cost, and reliability at short reach

- **Power**
  Retimer-free short links can reduce rack power ~15-20% [1]
- **Latency**
  Low latency connections can be achieved with retimer-free links
- **Reliability**
  Simple local links improve serviceability and failure isolation [2]

[1] Marvell CPC (Jan 2026) [2] Marvell Copper & Optical Interconnects (Jun 2024)

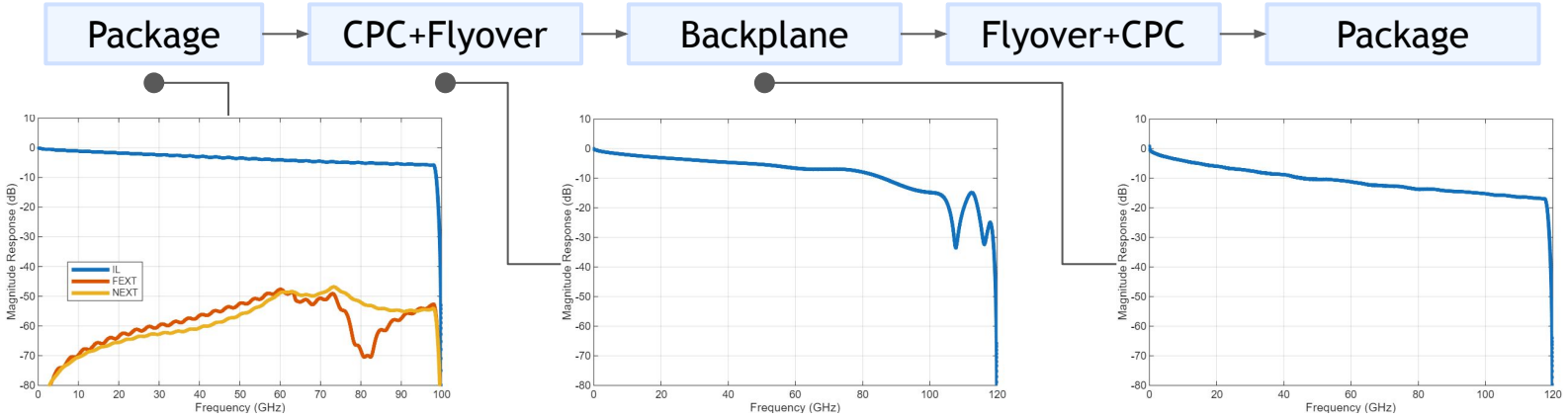
Lengths Inside the Rack



Connection	Length
1) XPU Core to CPC (Package)	≈3.5 cm
2) CPC to Module (Flyover)	40 cm
3) Backplane	170 cm
4) Module to CPC (Flyover)	40 cm
5) CPC to Switch Core (Package)	≈3.5 cm

- First target: Closing the link
- Crucial point: Closing it *without a retimer*

CPC Channel for Next Generation 448G



META Package Model (35mm) | Flyover Cable (AWG31) with CPC [1] (400mm) | Backplane Cable (AWG27) [2]

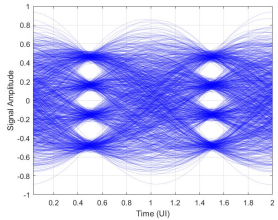
PAM6 is the strong candidate for 448G in CPC channel

	Package	Flyover + CPC	Backplane
IL @ 90GHz	5.5dB	11.75dB	12dB/meter

EQ Type	DFE	MLSE	Power
BP Reach	≈0.5-0.6m	≈1m	1x (Reference)

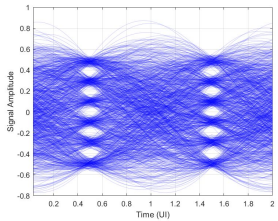
[1] Matched with Huawei's CPC Model [2] Plotted for 1.2m length

Modulation Types and FEC



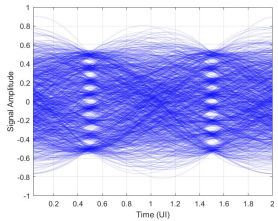
PAM4

- Needs more improvement in CPC and package
- Transition can be made after some time



PAM6

- Strongest candidate with current CPC
- Still some deteriorations exist around 90GHz

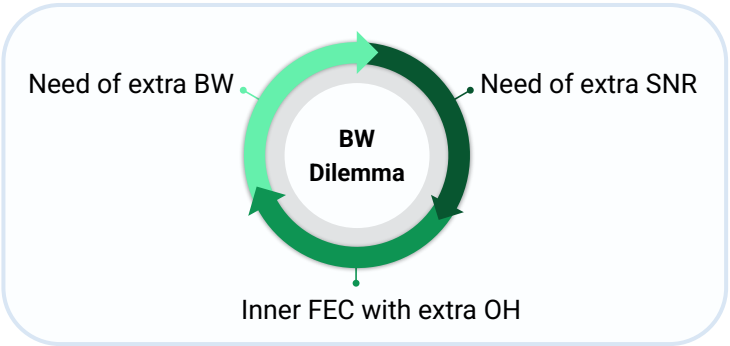


PAM8

- Not being strongly considered at the moment for 400G

FEC Considerations

Need for extra SNR in PAM6 → Inner FEC



5b2s PAM6: Wasting ≈%3 of BW which is critical

- Higher dimensional PAM6 mapping
- Complexity and latency should be carefully considered

Inner FEC Candidates

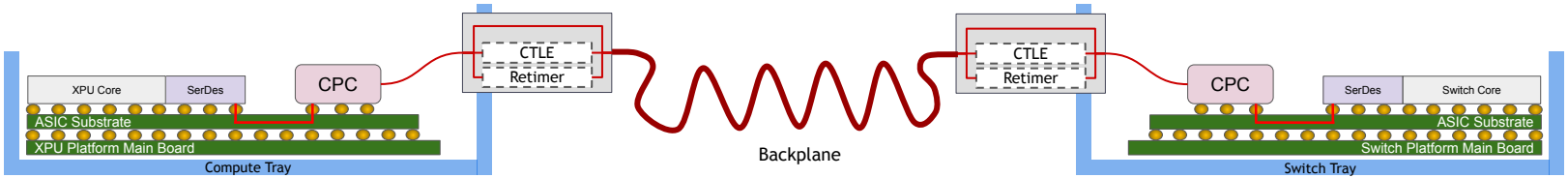
- Hamming Code
- BCH Code (t=1,2)
- Constrained MLSE [1] for joint inner FEC decoding + EQ

[1] 448Gbps: Challenges for Scale-Up & Scale-Out Applications, DesignCon 2026

Another Solution for Reach: Active Copper Cable (ACC)

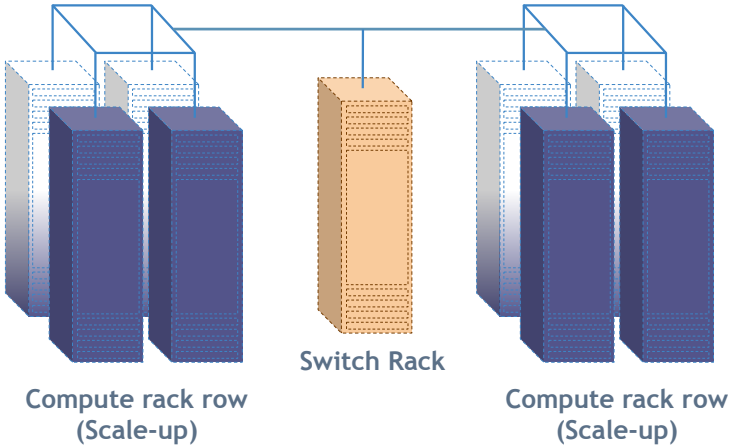
Method: Including extra linear CTLEs for increasing reach in PAM6
Active Electrical Cable (AEC): Instead of CTLE, retimers can be utilized

- Is there a room for retimer with CPC?
- Don't want to double the power at the module



Type	TX Module Active Element	RX Module Active Element	≈DFE Backplane Reach (m)	≈MLSE Backplane Reach (m)	Est. Power Increment
Direct Attach Cable	NA	NA	0.6	1.0	+%0 (Reference)
ACC	NA	CTLE	1.4	1.8	+%25
ACC	CTLE	CTLE	1.8	2.5	+%50
AEC	NA	Retimer	2.1	2.7	+%75
AEC	Retimer	Retimer	3.2	3.7	+%137

Future of Scale-Up: Disaggregation and Optics



- Disaggregation: Achieving large scale-up domain with small, modular units with special optical interconnects
- Flexibility, improved SI, easier thermal management and transportation
- Could be the key to unlocking the next era of high-performance, scalable AI systems

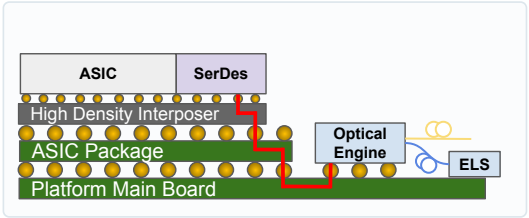
200G Optical Compute Interconnect (OCI) [1]

- Purpose-built for massive AI backend networks
- Delivers low-power, high-density optical interconnects directly to the compute layer
- Solves density challenges with an innovative DWDM architecture

[1] 200G Optical Compute Interconnect Line Interface Specification, v1.0, March 2026

Possible OCI Implementations

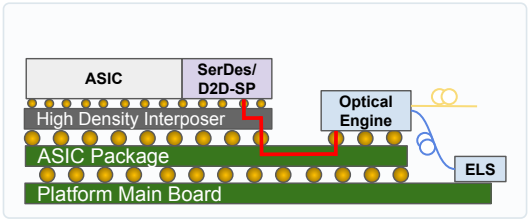
LOWER INTEGRATION



On-board optics (OBO)

Optical engine and ELS sit on the platform board; the ASIC reaches the engine through a board-level electrical path.

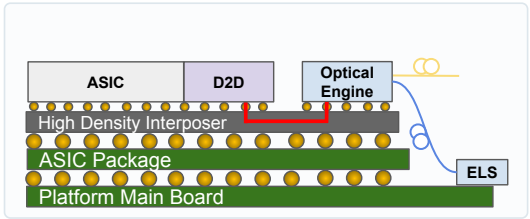
MID INTEGRATION



Package-integrated optics

The optical engine moves onto the ASIC package, reducing the electrical distance between SerDes/D2D and optics.

HIGHEST INTEGRATION



Interposer-integrated optics

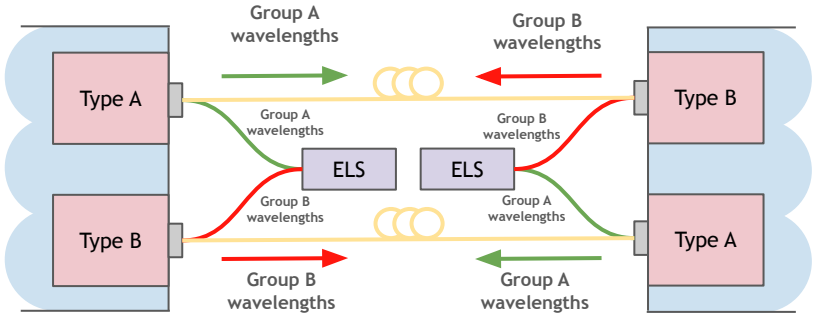
The optical engine shares the high-density interposer with the ASIC for the most tightly integrated implementation.

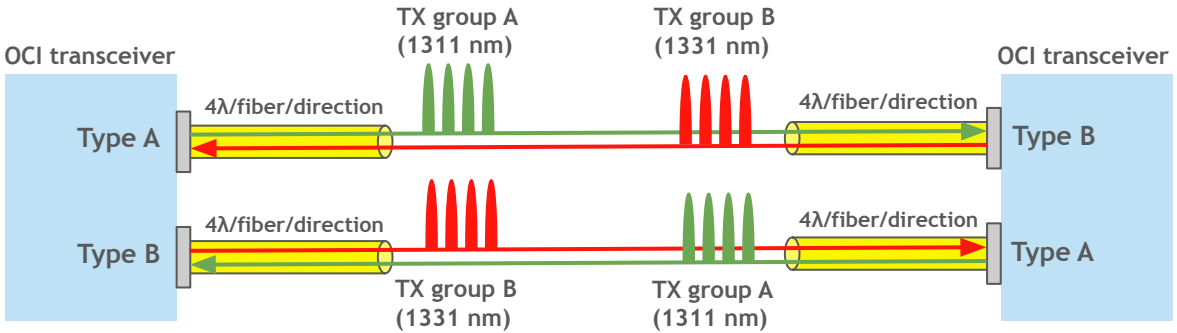
[1] 200G Optical Compute Interconnect Line Interface Specification, v1.0, March 2026

Bidirectional Transmission on a Single Fiber

Solution to the number of fiber connectivity and density challenges of the AI scale-up domain

- **Halving Fiber Count:** Transmit and receive on one cable
- **Counter-Propagating:** Signals travel in both directions simultaneously
- **Wavelength Groups:** Type A & B transceivers use alternating bands





CWDM Grid:

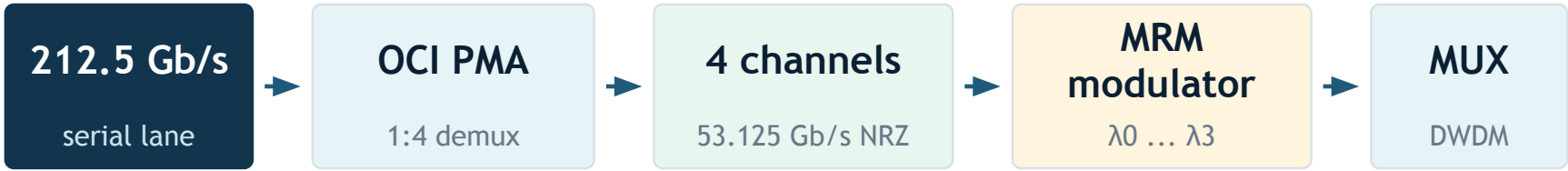
Group A (1311nm)

Group B (1331nm)

[1] 200G Optical Compute Interconnect Line Interface Specification, v1.0, March 2026

Relation of OCI PMD-PMA with Standards PMA

The OCI PMA demultiplexes each 212.5 Gb/s serial stream into four 53.125 Gb/s NRZ optical channels, then remaps them back at the receiver after deskew.



PMA scaling

Same 53.125 Gb/s optical lane

Input lanes → optical lanes

200G	1 : 4
400G	2 : 8
800G	4 : 16
1.6T	8 : 32

[1] 200G Optical Compute Interconnect Line Interface Specification, v1.0, March 2026

What Hyperscalers Need

In short: Settle the 400G in-rack roadmap early enough to keep 2028 systems open and multi-vendor

1) Prototype
400G rack prototypes
Example class: Vera Rubin NVL72

2) First checkpoint
2027 PHY prototypes
Signal integrity + CPC validation

3) System target
2028 full system
If timelines are settled early

CPC risks to de-risk early

Want to know the risks earlier. CPC must be treated as a supply chain and reliability program

CPC shortage
Manufacturing base with high yield
Capacity and supply must be reserved

Component sourcing
Connectors, flyovers, etc.

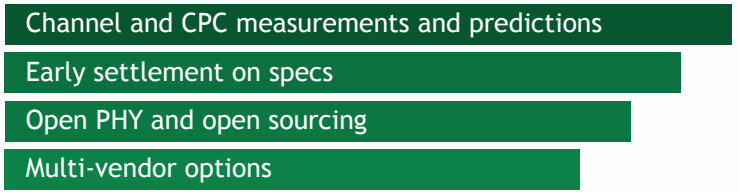
CPC reliability
Field serviceability
Failure rate

Active Copper Cable

Interested in

- will it be required, and
- finding out how and where the active CTLE fits into the link

Progress Path



Delayed settlement results in closed and risky systems

Conclusion

- >1.7m backplane link is needed in scale-up systems for now
- PAM6 with MLSE and additional inner FEC is a candidate for closing the link
- %3 bandwidth waste in PAM6 of 5b2s should be carefully considered
- There are different inner FEC options to be evaluated
- ACC could be a solution for closing the link for desired reach with acceptable power
- Disaggregation and optical connection in scale up could be the future plan with OCI like connections
- Need to know the timelines, CPC risks.
- Need to settle early on spec to achieve multi-vendor sourcing.

Thank You!

Back-up’s

ACC-AEC Simulation Parameters

Parameter	Value
Modulation	PAM6
Bitrate	448G
Nyquist Rate	~90GHz
DAC SNR	41dB
TX SNR	30.2dB
ADC SNR	43dB
Jitter	50fs
CTLEs Peaking Frequency	72GHz (90GHz*0.8)
CTLEs Noise Figure	6dB
FFE Main Length	48
FFE Number of Floating Taps	4
FFE Floating Length	4
FFE Floating Taps Max Distance	256
DFE Taps	1
Target BER	1e-3