Breaking up the Monolith: An Open Architecture for Multi-rack Scale-up Domains of 5000+ GPUs

Dan Daly *CTO & Co-Founder, Delos Data*

Promode Nedungadi *CTO, Eridu*

OCP STARTUP™ Member

Why this matters now

AI workloads demand multi-rack scale-up infrastructure

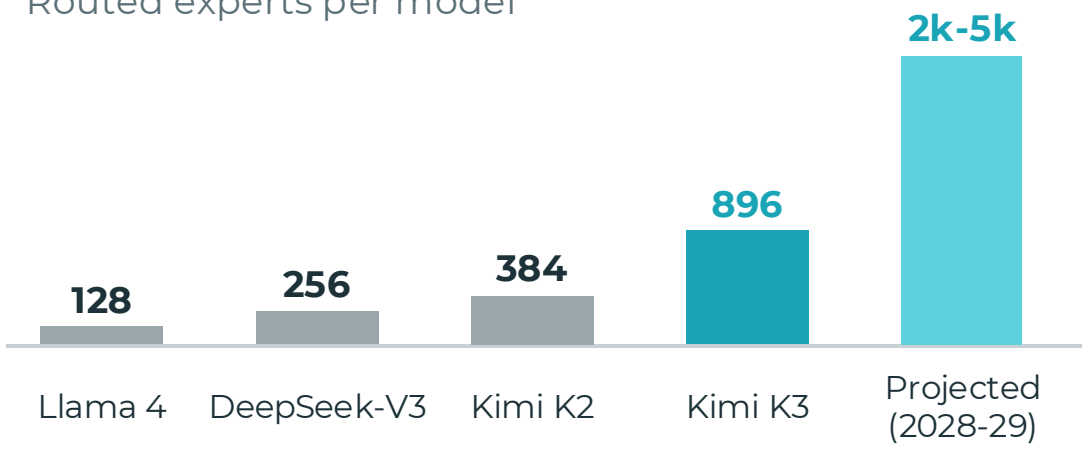
Current approaches leave room for improvement

We propose new approach: open, disaggregated, multi-rack scale-up architectures incorporating high-radix switching for large single-hop domains

AI Scaling Pushes Networks Relentlessly

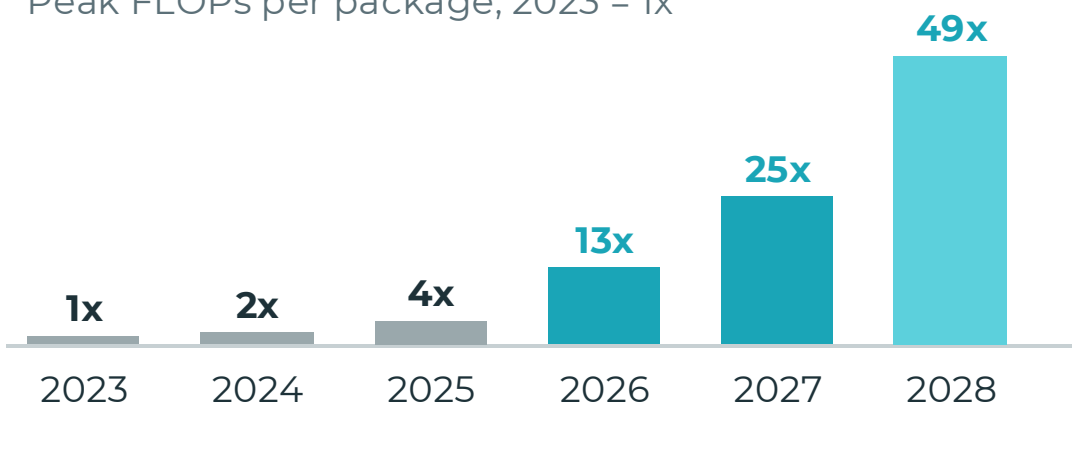
Frontier MoE expert counts are climbing

Routed experts per model



GPU performance scaling faster than networks

Peak FLOPs per package, 2023 = 1x



New Algorithms Increase Data Movement

Larger high-bandwidth domains, clusters
Scale up beyond 1k, scale out beyond 1M GPUs

Evolving models, MoE
Trillions of parameters, thousands of experts

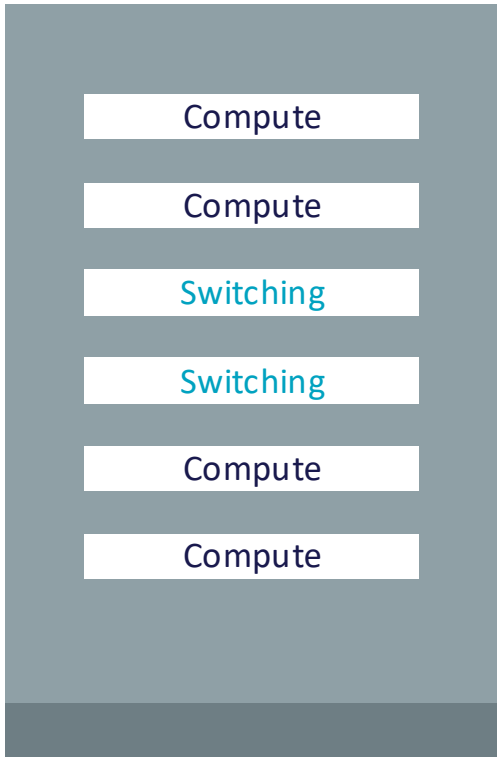
Larger context windows
Memory, personalization, multimodal

Disaggregated inference
Prefill & decode in separate racks, KV cache replication

The Monolithic Rack

PRO

- Extends high-bandwidth scale-up domain beyond one server
- Single vendor control of system-level performance
- Single switch hop for low latency, minimum congestion



CON

- Extreme power density (~MW/rack)
- Heavy & hard to service
- Practical limit of 64 GPUs w/ 8 “spares”
- Poor workload placement into 64-GPU bins
- Closed, proprietary architectures

Growing Scale-up Domains

Strong demand for growing beyond one rack
… but how do you build it?

You Are Here

8 GPUs

72 GPUs

Hundreds of GPUs

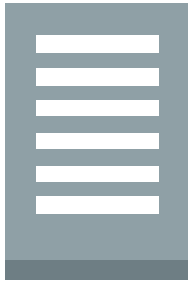
Thousands of GPUs

How to scale single hop, low-latency connectivity across racks & rows?

Single hop, low-latency between GPUs



Server



Rack



Multi-rack



Multi-row

Current Approaches for Larger Domains

Announced scale-up domain size		Single-hop domain
Nvidia NVL72 · AMD Helios	72	72
Nvidia NVL576 (2027)	576	72
Nvidia NVL1152 (2028)	1,152	144
OpenAI Jalapeño	2,048	128
Microsoft Maia 200	6,144	48

Issues:

- Closed architectures
- Nominally large domains are composed of smaller single-hop domains
- Rack-to-rack performance is lower than in-rack (sometimes significantly)
- Topology-aware job placement is required within the scale-up domain

Proposed new approach:
Open, multi-rack scale-up architectures with high-radix switching

Sources: Hot Chips 2026, GTC 2026, AMD Advancing AI 2026, Google ISCA 2023

Building open, multi-rack scale-up domains

Using disaggregated systems and components

Leveraging the OCP open ecosystem

Building Blocks for Open Multi-rack Scale-up Architectures



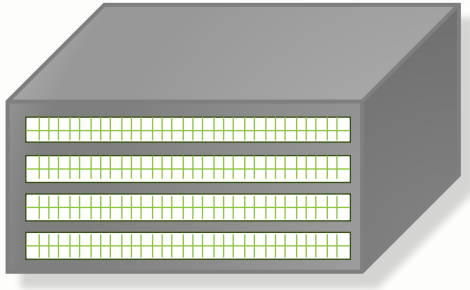
Server w/ Optical Scale Up

- Multi-rack scale-up
- Leverage existing ecosystems
- Avoid resource stranding



Optical MSAs (e.g. OCI)

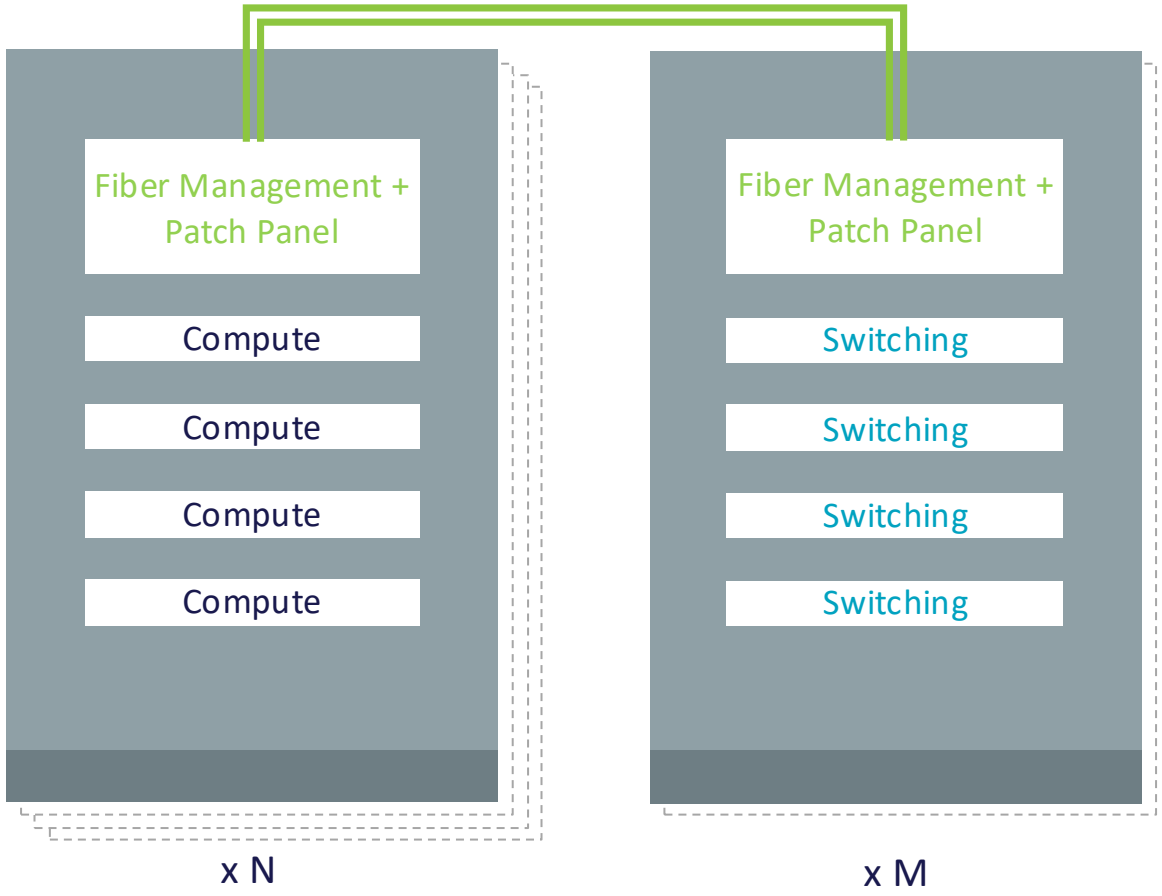
- Growing ecosystem adoption of optical interconnects due to increasing workload demands
- Enabling higher bandwidth, lower power, lower latency



High Radix, Low Latency Switches

- Higher radix → larger single-hop domain
- Open Ethernet standards-compliant (UEC, ESUN)

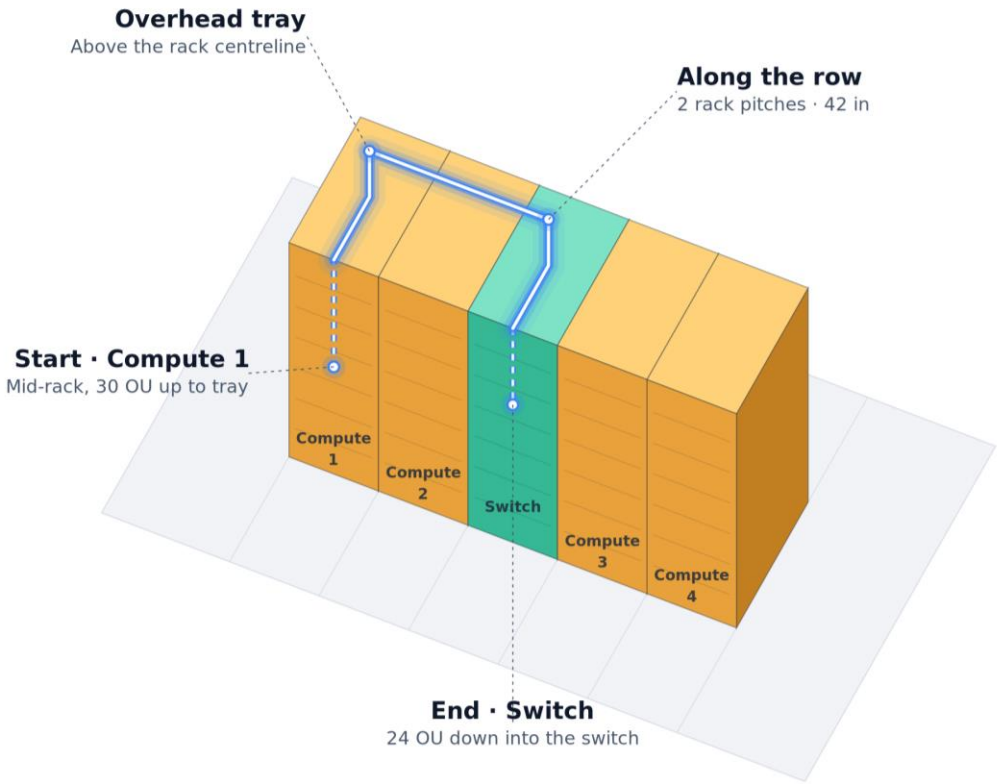
Disaggregation Enables Flexible, Scalable Architectures



Endpoints (xPUs)	Rows	Max Cable Run
128	1	~4m
512	3	~10m
576	3	~10m
2048	5	~15m
5120	12	~30m

128 Endpoint Single Row Design

1 row × (2 compute + 1 switch + 2 compute) · three-quarter overhead view



LONGEST RUN

≈ 12 ft

≈ 3.7 m end-to-end

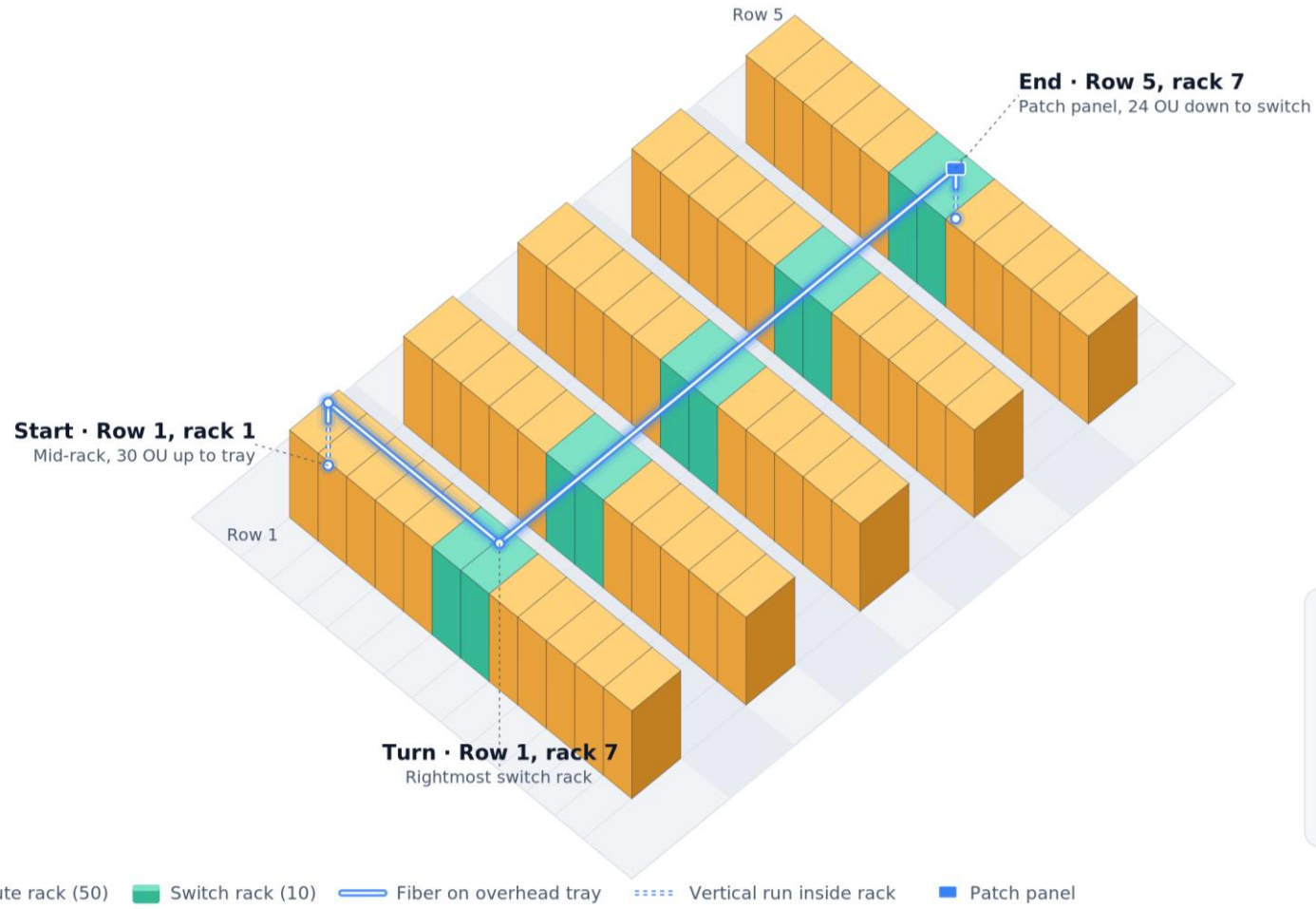
30 OU up (57 in) · 2 racks along (42 in)
24 OU down (45 in)

Racks 21″ × 36″ × 42 OU · 1 OU = 48 mm

Compute rack (4) · Switch rack (1) · Fiber on overhead tray · Vertical run inside rack

2048 Endpoint Five Row Design

5 rows × (5 compute + 2 switch + 5 compute) · three-quarter overhead view



LONGEST RUN

≈ 49 ft

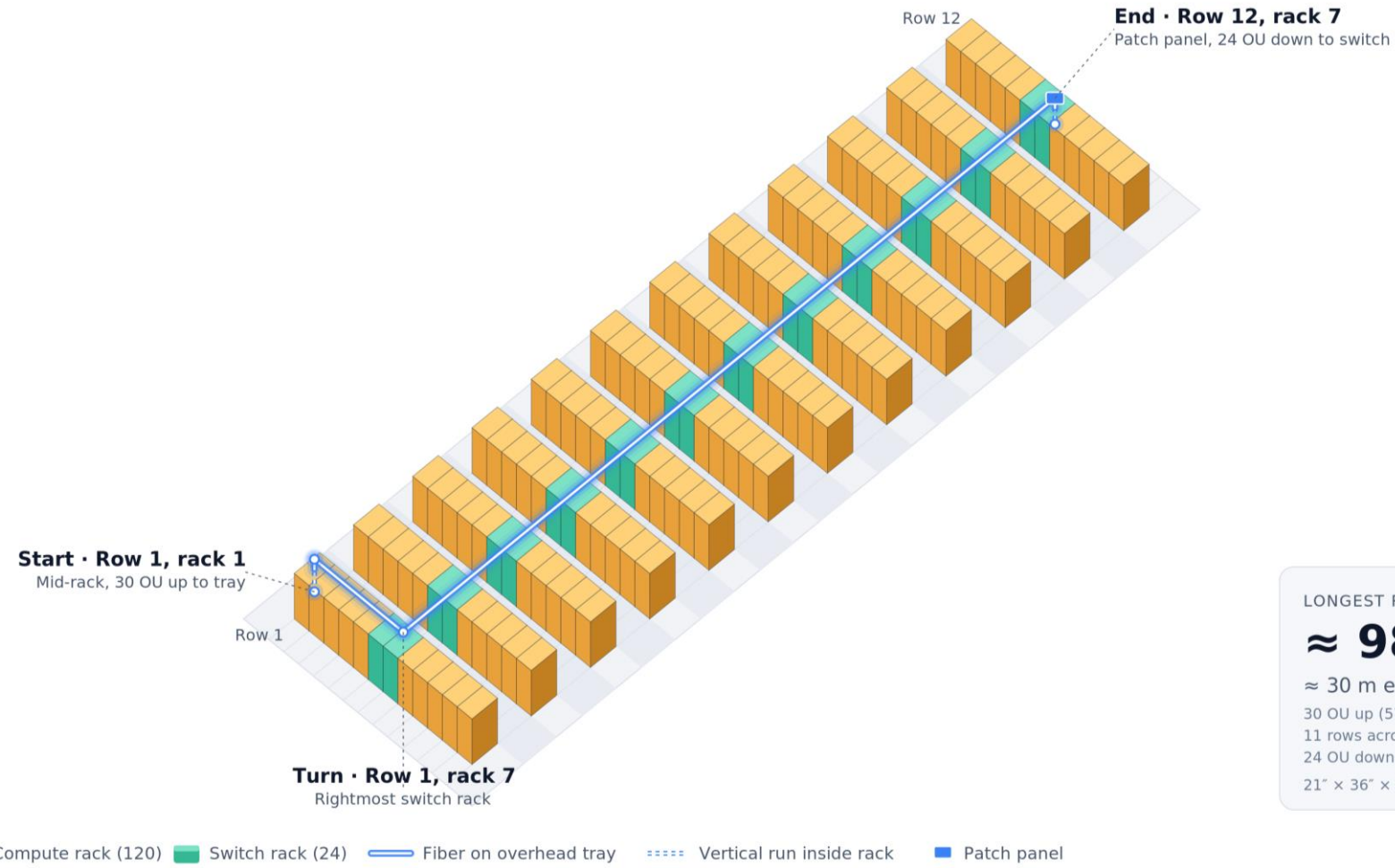
≈ 15 m end-to-end

30 OU up (57 in) · 7 racks along row 1 (147 in)
4 rows across (336 in) · patch panel
24 OU down (45 in)

21″ × 36″ × 42 OU racks · 7 ft row pitch

5120 Endpoint Twelve Row Design

12 rows × (5 compute + 2 switch + 5 compute) · three-quarter overhead view



LONGEST RUN

≈ 98 ft

≈ 30 m end-to-end

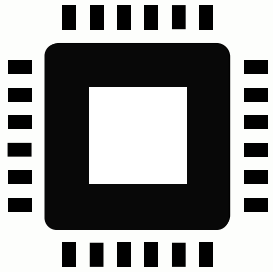
30 OU up (57 in) · 7 racks along row 1 (147 in)
11 rows across (924 in) · patch panel
24 OU down (45 in)

21″ × 36″ × 42 OU racks · 7 ft row pitch

OCP Global Summit · Oct 12, 2026

Building a petabit-class switch

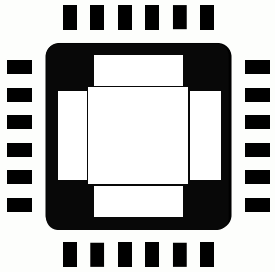
Building a Petabit-class Switch Requires a Shift from Evolution to Revolution



51.2 Tbps
2022*

- Monolithic ASIC
- Embedded I/O
- 5 nm

2x



102.4 Tbps
2026*

- Monolithic core
- Separate I/O tiles
- 3 nm

10x

?

How to build a petabit-class switch?

* Year when switching systems based on this generation of ASICs began volume shipment (Source: 650 Group)

Building a Petabit-class Switch Requires a New Approach

Silicon and System Co-design

Silicon



Architecture

I/O density

Packaging

+

System

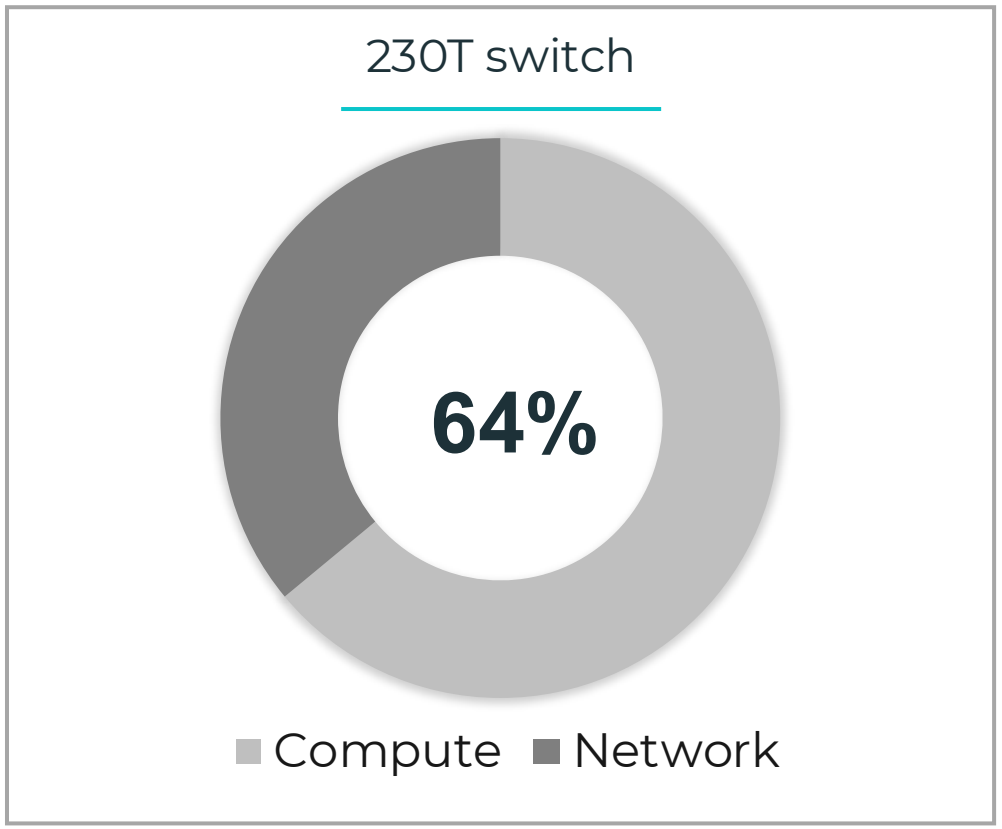


Purpose built

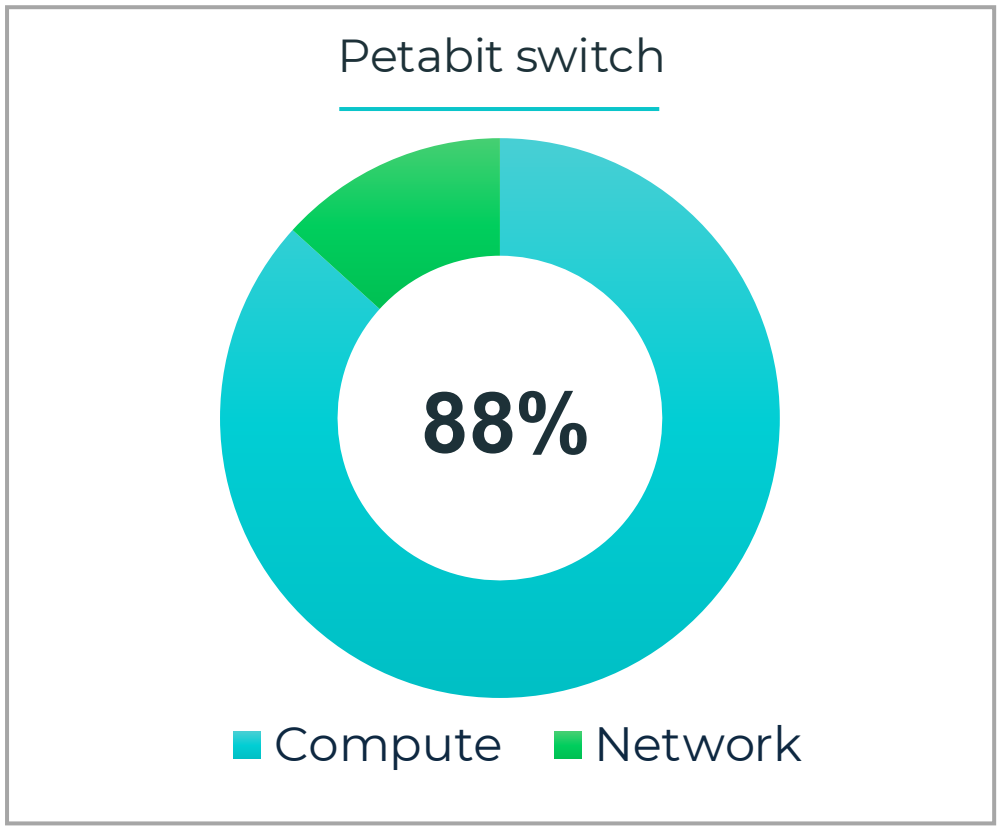
Cooling

Power

Large Flat Scale-up Domains Increase GPU Utilization



38% Higher GPU Utilization



*Training simulation by Turbalance for HotInterconnects 2026
https://hoti.org/2026/assets/slides/09-breaking-through-network-wall.pdf

Key Benefits of Petabit-class Switching in Scale Up

Larger high-bandwidth scale-up domains
with flatter, simpler networks

Economics
Higher GPU utilization
Lower cost & power

Performance
Faster job completion
Lower latency & congestion

Operations
Higher reliability, simpler OA&M, faster deployment

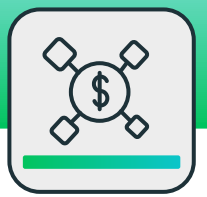
BETTER AI OUTCOMES



Better models & more flexibility



Faster training



Lower cost per token

Key Takeaways

AI workloads demand multi-rack scale-up infrastructure

Hundreds to thousands of XPU endpoints

Multi-rack scale-up needs an open, disaggregated architecture

- ✓ Flexibility
- ✓ Ecosystem choice
- ✓ Heterogeneous compute

Maximize single-hop domain size using high-radix switching

Large flat networks



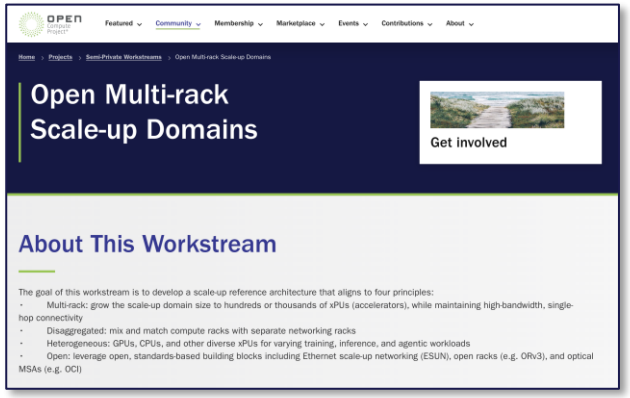
Better AI outcomes

New OCP Workstream

Open Multi-rack Scale-up Domains

Scope

- Develop a **reference architecture** for multi-rack scale-up domains that provides flexibility to scale from hundreds to thousands of accelerator endpoints
- Propose requirements and recommendations to facilitate open and disaggregated scale-up systems that enable heterogeneous compute and networking racks to be combined



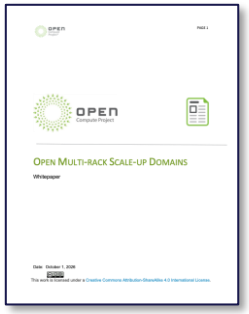
https://www.opencompute.org/projects/semi-private-workstreams/open-multi-rack-scale-up-domains/

Principles

1. **Multi-rack**: single-hop, thousands of xPUs
2. **Disaggregated**: mix/match compute & networking
3. **Heterogeneous**: GPUs, CPUs, xPUs
4. **Open**: ESUN, ORv3, OCI, etc.

Call to Action

- Request to get involved
- Request the whitepaper
- Kickoff meeting in early November *(date/time TBA)*



Thank You!

