AI CLUSTERS

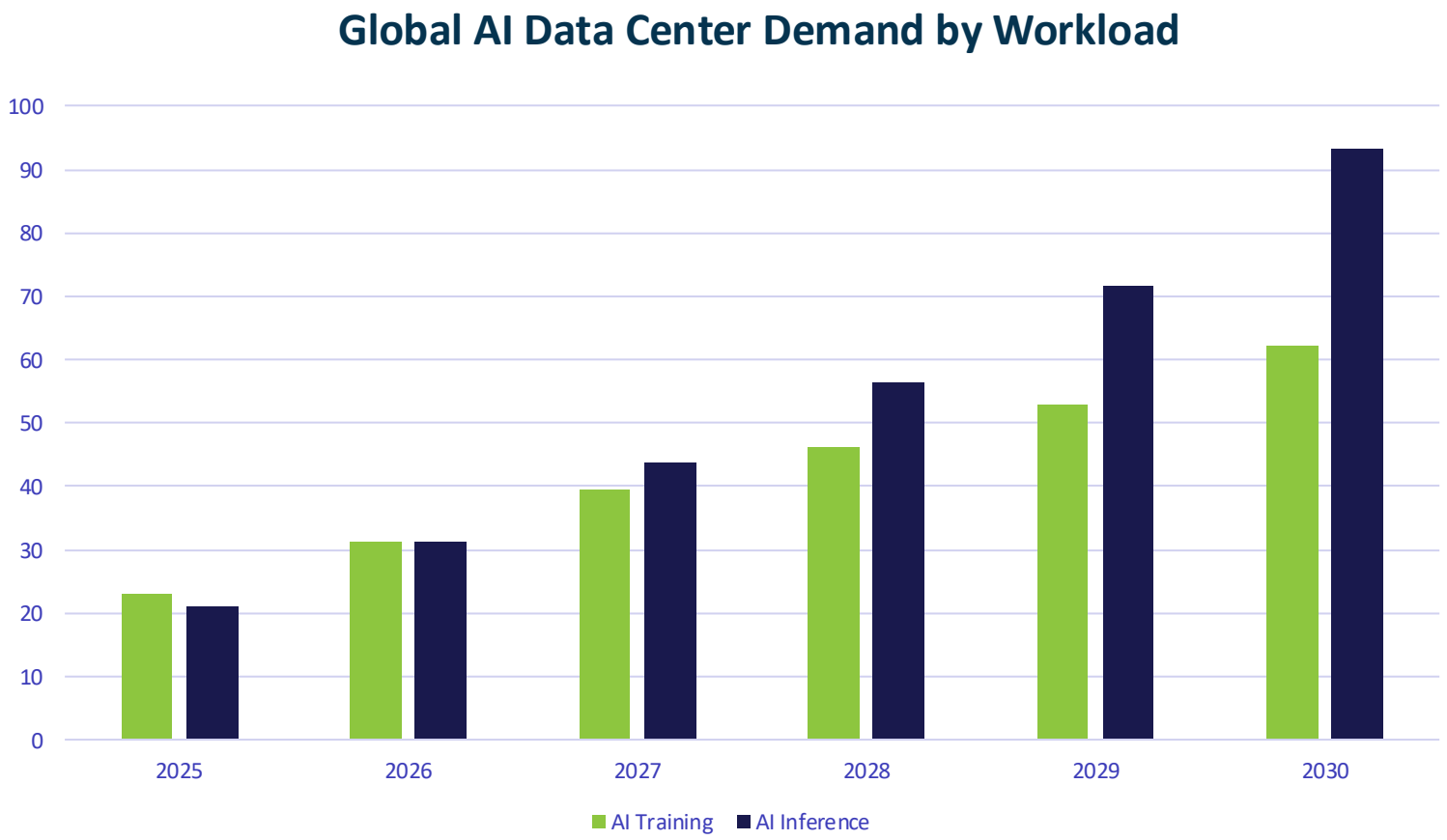
# Choose Your Own XPU: An Open Rack to Maximize Inference Performance and TCO


Chris Petersen

*VP, CTO Office, Astera Labs*


OCP PLATINUM Member

# The Shift to Inference



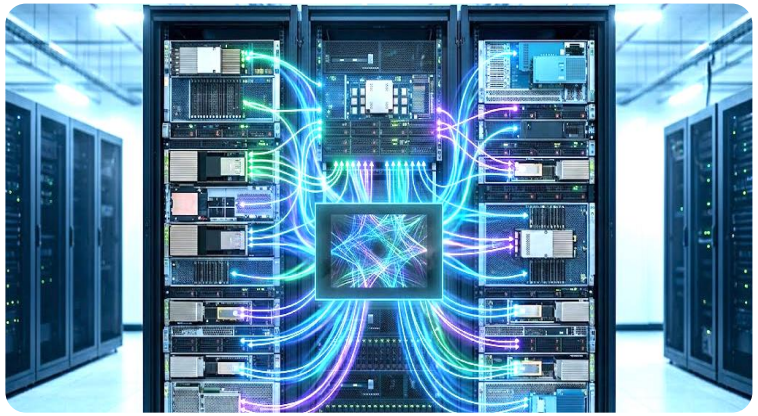

Source: McKinsey Data Center Demand Model: https://www.mckinsey.com/featured-insights/charts/the-future-of-ai-workloads

**Inference spending is projected to surpass training spending for the first time in 2026.**

Inference infrastructure spending will exceed training by > 50% by 2030.

# What Hyperscalers & AI Labs Demand from AI Infrastructure





## Token Economics

**Optimize performance per watt and cost per token**

## Scalability

**Enable diverse accelerators at cluster scale**

## Resilience

**Maximize uptime and utilization**

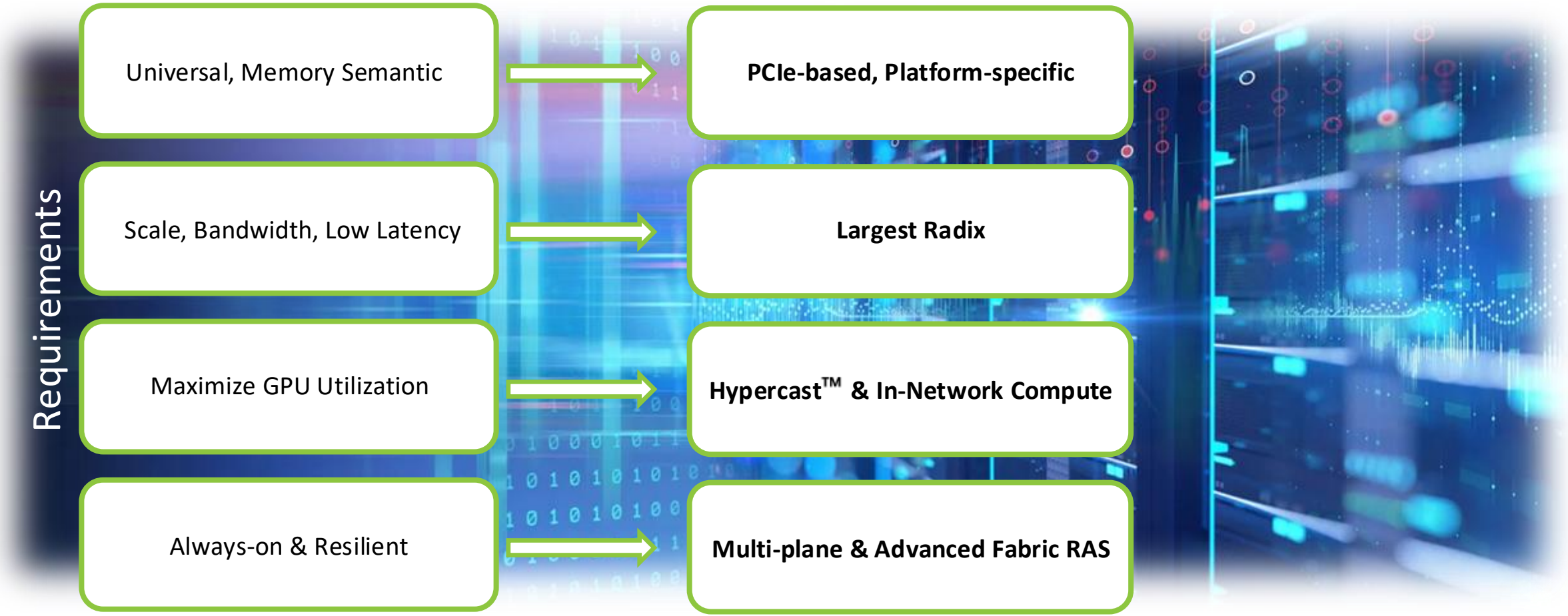
# Universal Scale-up Fabric Solution

Requirements

- Universal, Memory Semantic
- Scale, Bandwidth, Low Latency
- Maximize GPU Utilization
- Always-on & Resilient



# Universal Scale-up Fabric Solution



# Universal Scale-up Fabric Solution

Requirements

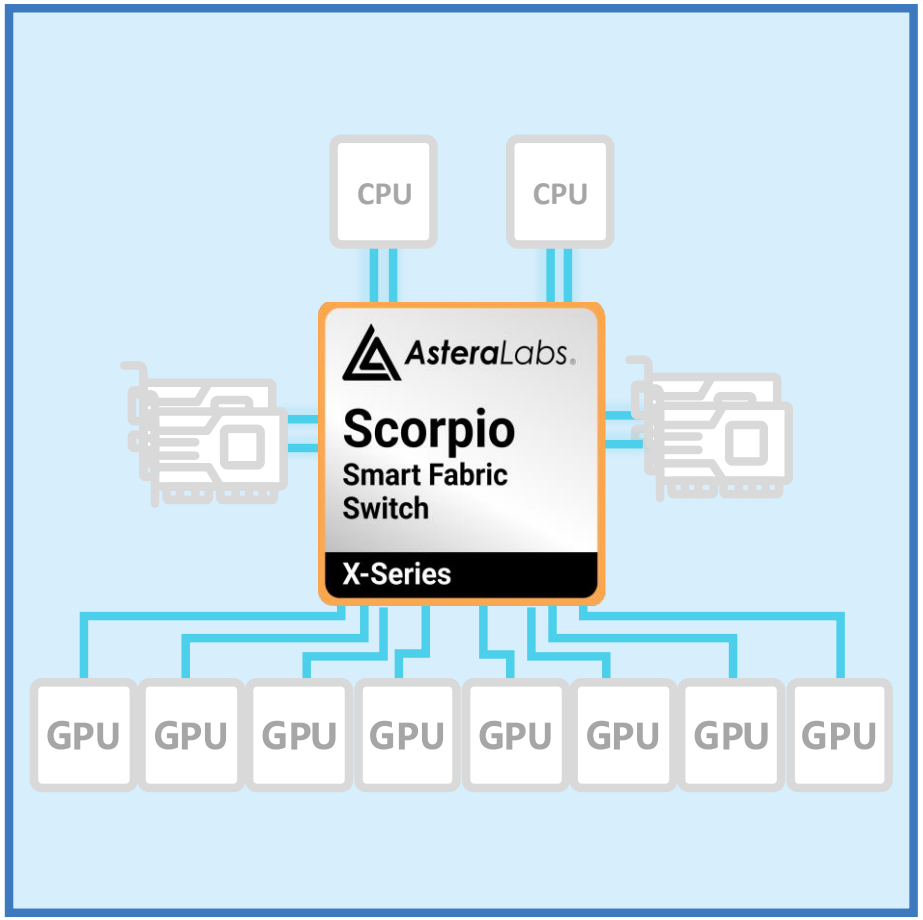
- Universal, Memory Semantic → **PCIe-based, Platform-specific**
- Scale, Bandwidth, Low Latency → **Largest Radix**
- Maximize GPU Utilization → **Hypercast™ & In-Network Compute**
- Always-on & Resilient → **Multi-plane & Advanced Fabric RAS**




**AI Fabric Solution**

# Scorpio X-Series 320 Lane Smart Fabric Switch

## Largest open, memory-semantic switch



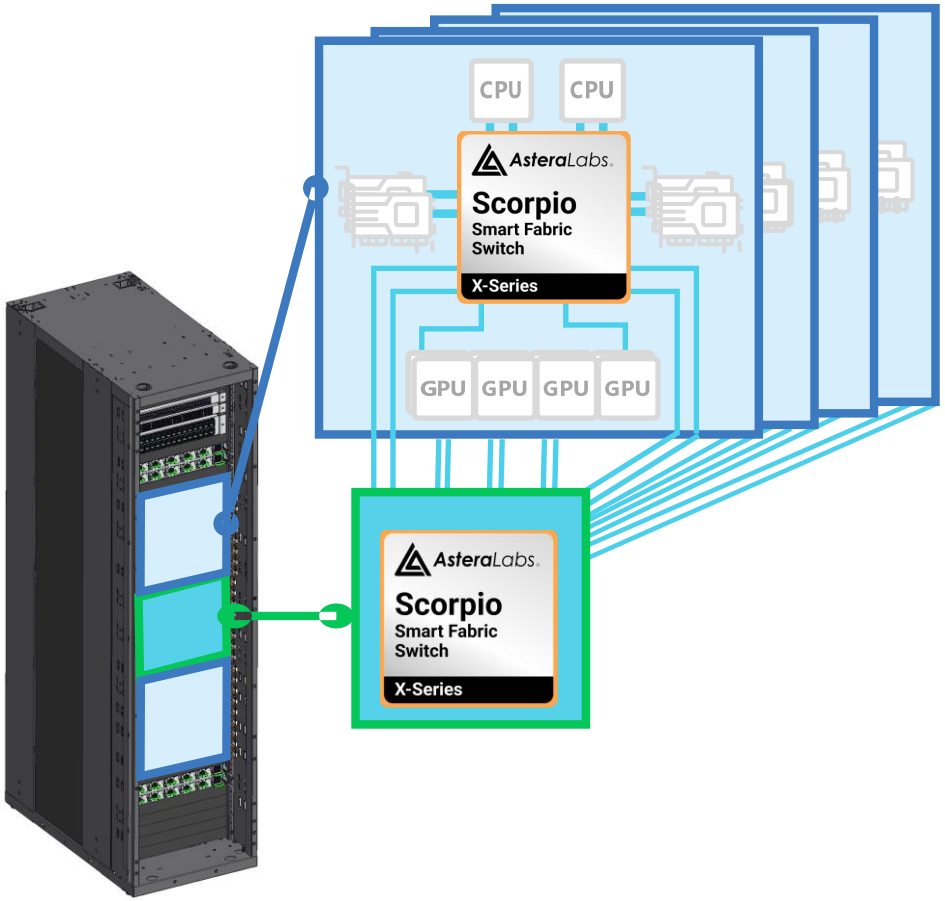

**Orion-CT**

Compute Tray: PCIe GPU Add-in Cards



# Scorpio X-Series 320 Lane Smart Fabric Switch

## Largest open, memory-semantic switch




**Orion-CT**

Compute Tray: PCIe GPU Add-in Cards



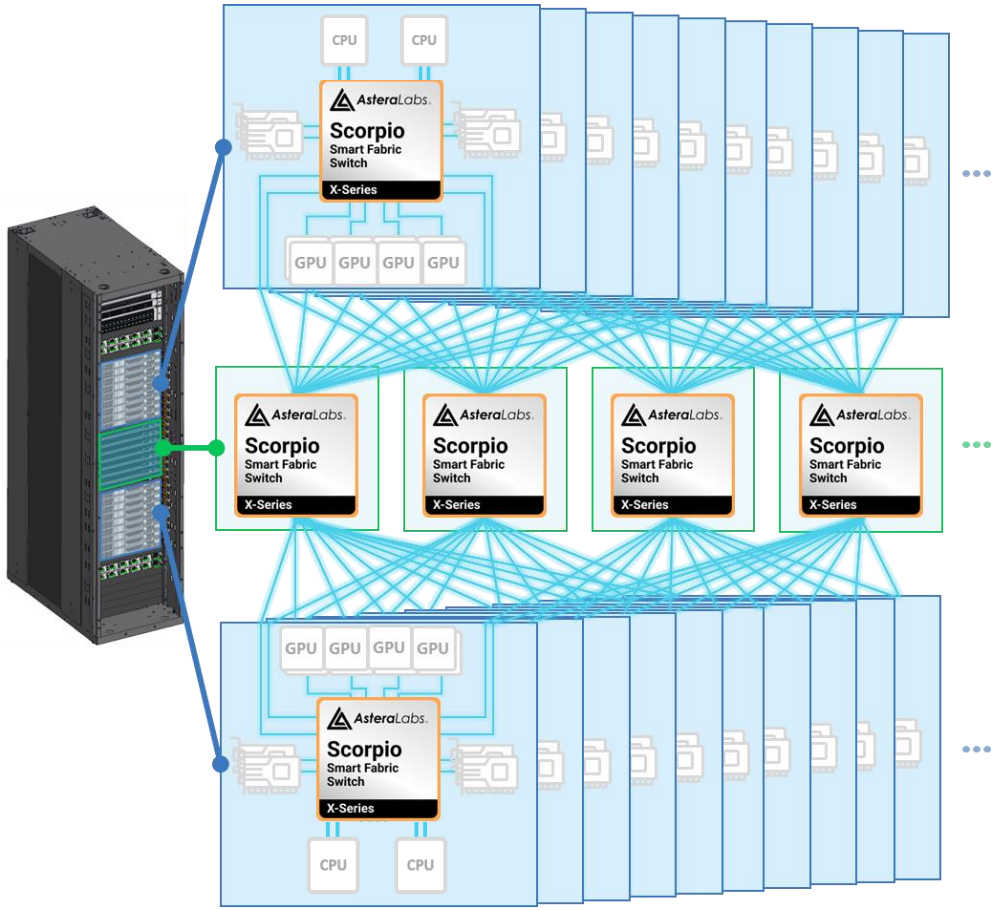
**Orion-FT**

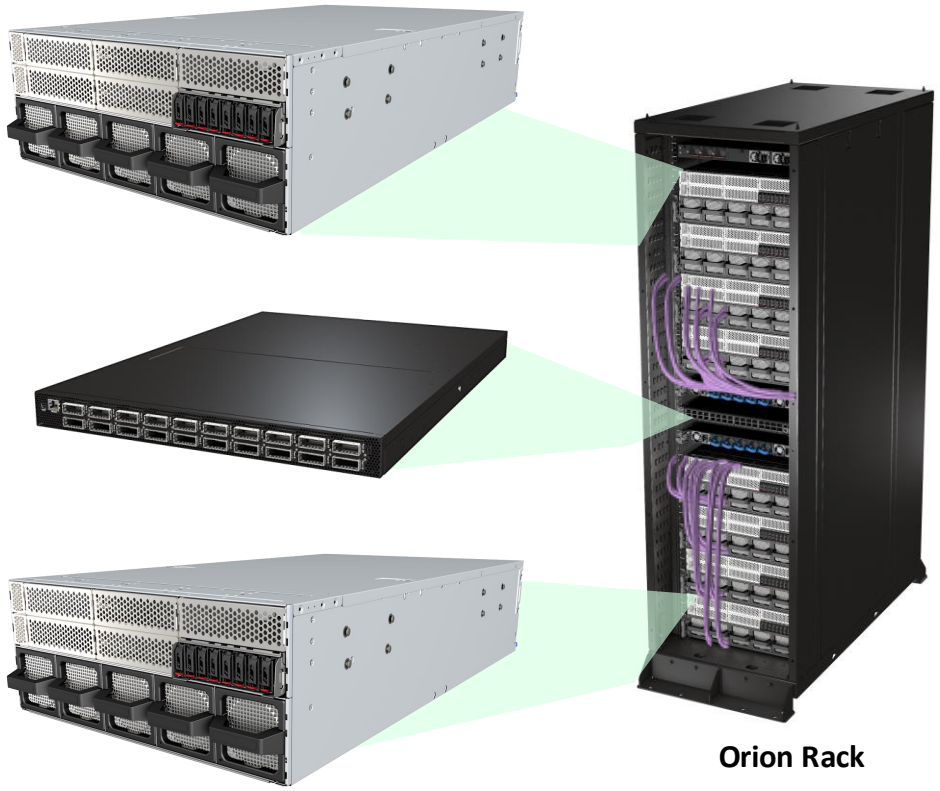
Fabric Tray: High-density scale-up



# Scorpio X-Series 320 Lane Smart Fabric Switch

## Largest open, memory-semantic switch



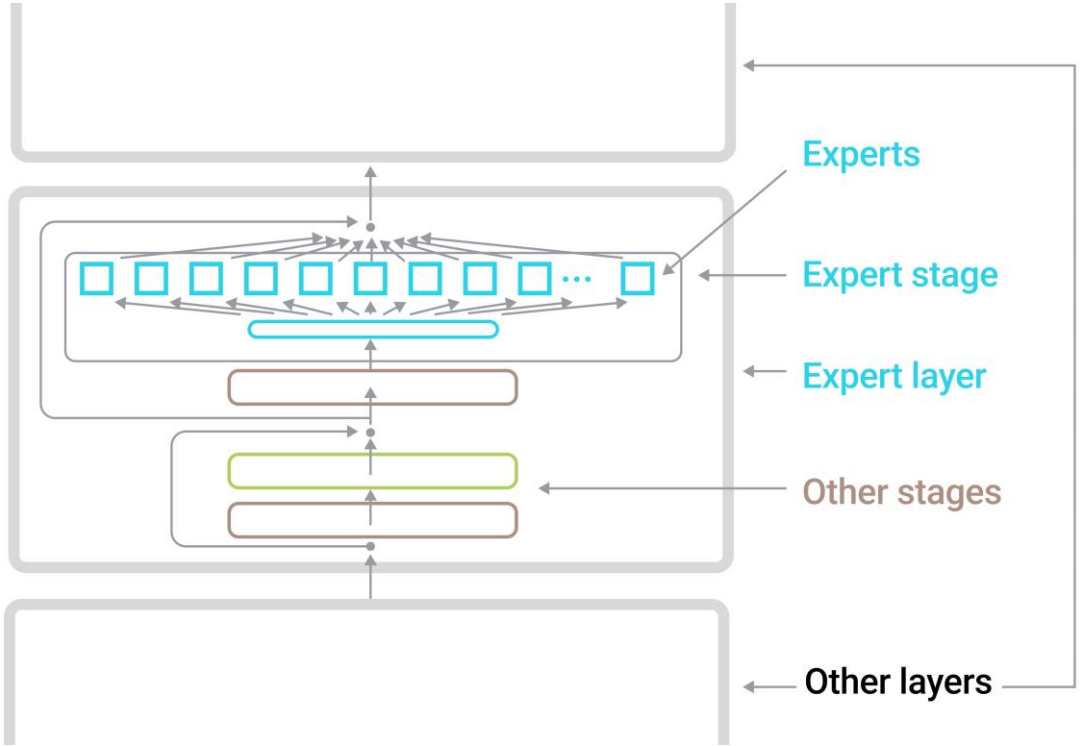
Orion Rack

**Expand GPU scale-up with multi-level switching fabric – up to 160 GPUs**

# Mixture-of-Experts (MoE) Makes Larger Models Practical

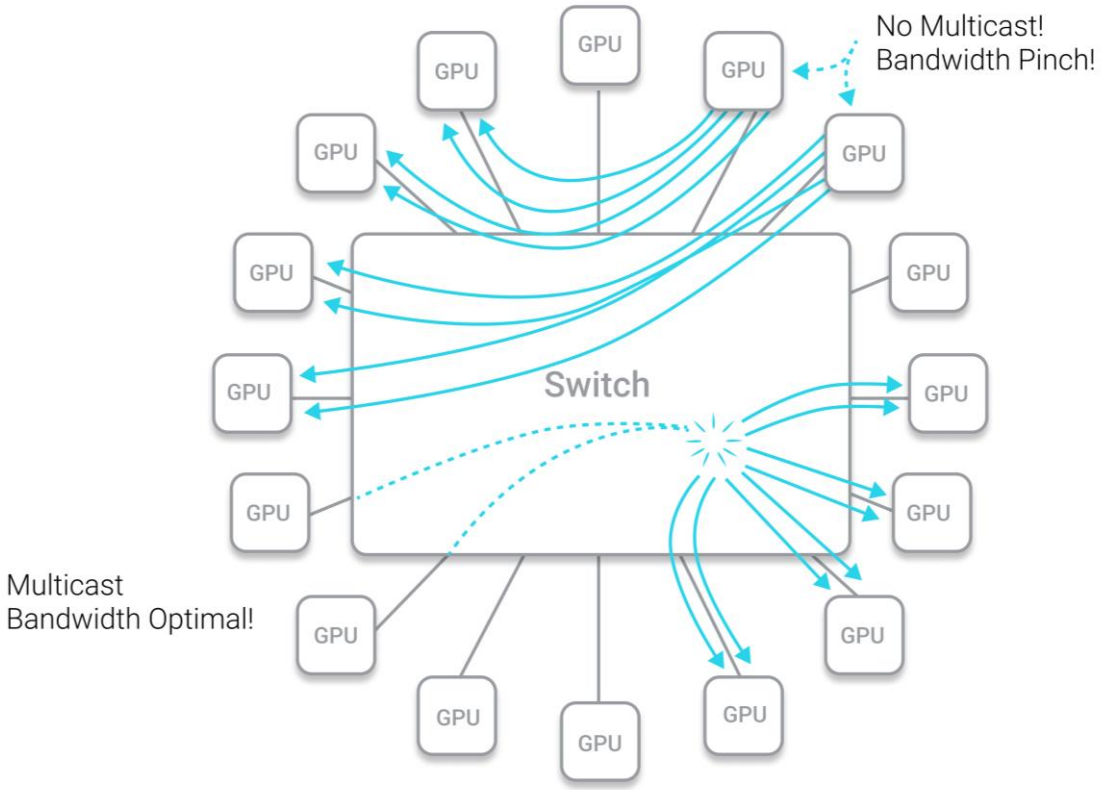
**MoE** improves efficiency by activating only a small subset of experts for each token instead of the full model.

This requires unique multicast groups to target combinations of experts.



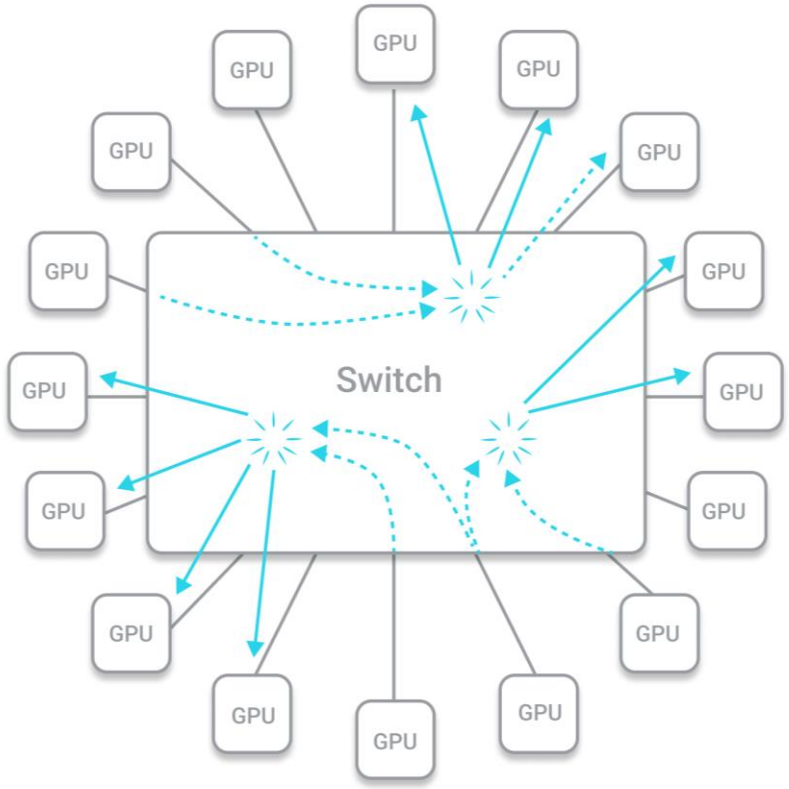
# Legacy Switches Do Not Support Enough Multicast Groups

When only some collectives can be accelerated, the rest fall back to inefficient unicast behavior and become bottlenecks.

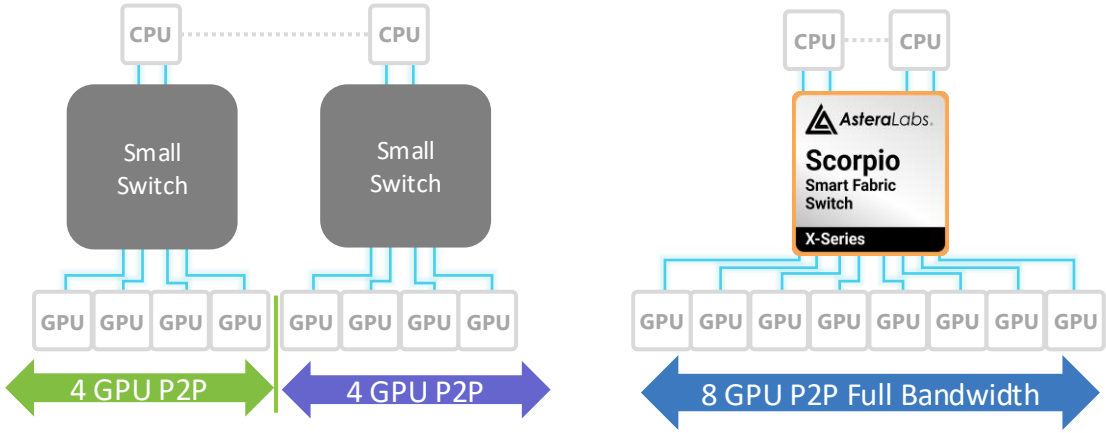


# Astera Labs' Hypercast™ Technology

**Hypercast** is an AI-native multicast technology designed to support both dynamic expert routing and high-frequency collectives.

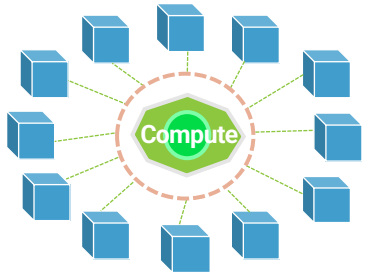


# Size and Intelligence Matter






AsteraLabs. Scorpio Smart Fabric Switch X-Series

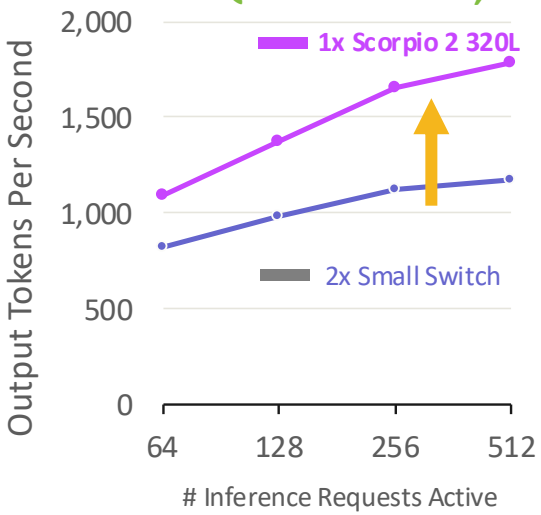



**Hypercast** + **INC**

**Data Replication Engine**
Accelerates one-to-many

**In-Network Compute**
Accelerates many-to-one

**Qwen3.5-122B, Balanced Inference Prefill/Decode**




**+50% Tokens Per Second**
+90% in Prefill heavy workloads

- ↑ GPU utilization
- $ Millions rack-scale value



**Accelerate collective operations by up to 2x**
Faster routing and collectives boost response and Tokens per Second

- ↑ GPU utilization
- $ Millions rack-scale value



**Do more with less: The ultimate differentiator in a constrained environment**

# Resiliency at All Levels



At Scale, Rare Events Aren't Rare – James Hamilton

Bicsi™

99% of your links will work perfectly; the final 1% will take 99% of your deployment time to find and re-seat.

## Fabric

Faster product deployment

COSMOS

Mitigate failures with multi-plane routing and advanced fabric RAS




## Link

Faster product qualification

COSMOS

Minimize error rates and quickly isolate link issues




## Signal

Faster repair & recovery

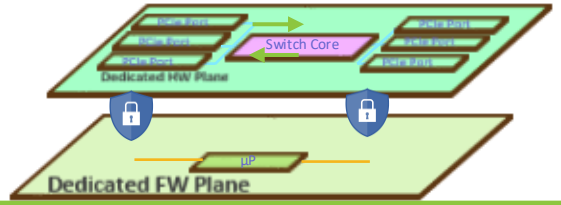
COSMOS

Identify damaged caps, loose cables, and disconnected pins for every signal




## Resilient Architecture

Lower Opex and maximize uptime




Isolated data plane from management/control plane

# Common Scale-up Rack Architecture

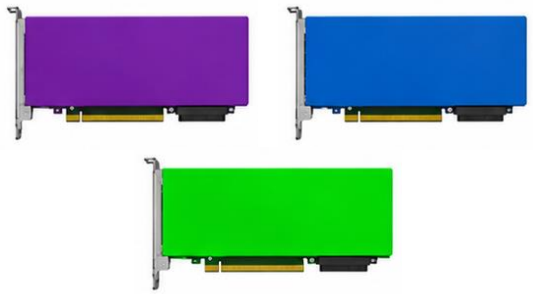
Open Protocol

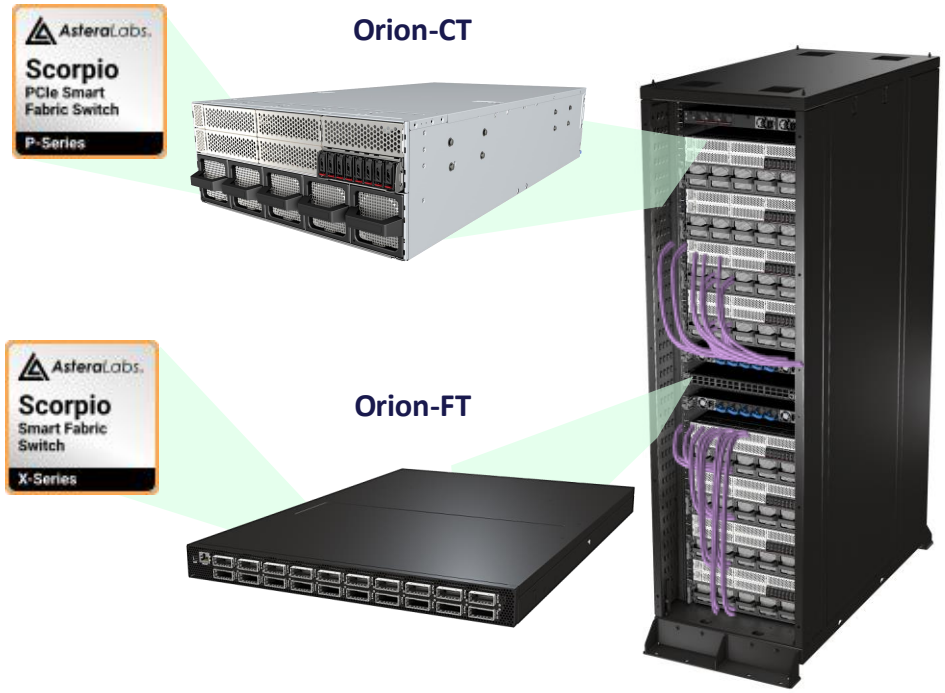
High BW & Low Latency

Hypercast & INC

RAS & Resilience

**XPU Optionality**

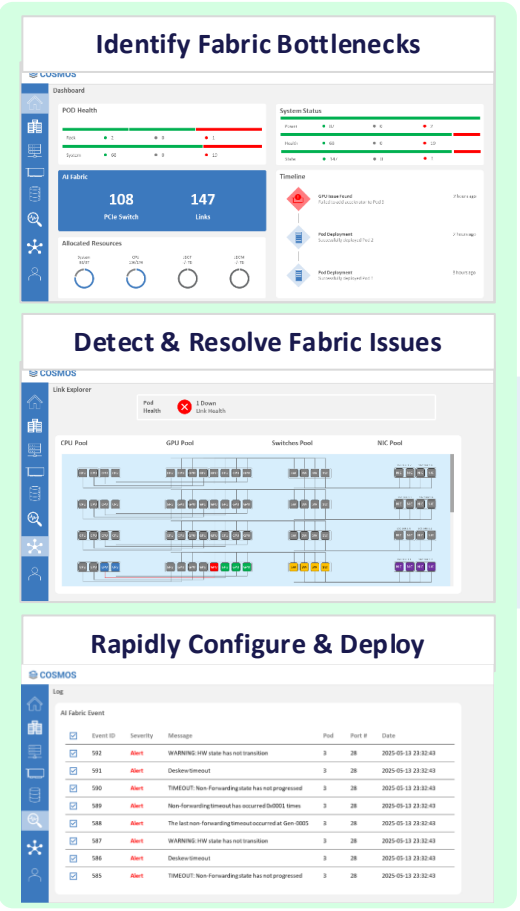





**Orion Rack**
Scale-Up GPUs + Memory

**Leverage common infrastructure across multiple XPUs and generations to reduce TCO**

# Enriching Fabric & Rack-Level Mgmt with COSMOS




**Managing:** System, fabric, and rack-level configuration using open standards (OCP, OpenBMC, DMTF, etc)




**Monitoring:** Performance metrics, error events, fabric notifications

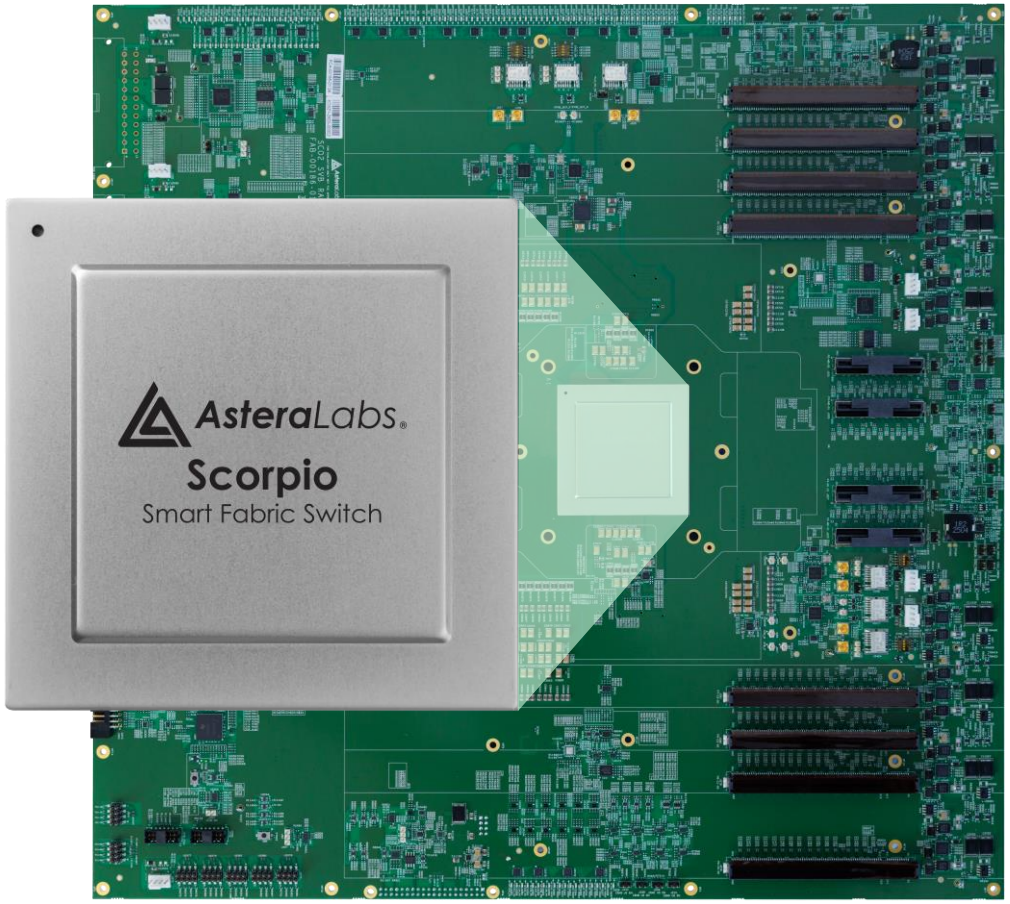



**Observability, diagnostics and lifecycle management across the entire fabric**

# Key Takeaways

Astera Labs is delivering **AI Your Way** with purpose-built rack-scale connectivity.

PCI EXPRESS · CXL Compute Express Link · ULTRA ACCELERATOR LINK™ · Ethernet · NVLink™ Fusion




## Scorpio and Leo Smart Fabric Solution

- Largest open, memory-semantic switch
- Full end-to-end rack-scale solution
- Collective acceleration via Hypercast and INC
- Optimized performance per watt and cost per token
- Interop-tested and validated with leading XPUs

## A full portfolio of copper & optical AI connectivity solutions!

| AsteraLabs Leo | AsteraLabs Aries | AsteraLabs Taurus | AsteraLabs Scorpio |
|---|---|---|---|
| Smart Memory Controller | PCIe/CXL Smart Retimer | Ethernet/UALink Smart Retimer | Smart Fabric Switch |

# Call to Action

- Now is the time for a common scale-up infrastructure that works with any XPU!
- Get involved with OCP Projects
  - Open Adapter Card (OAC), Hardware Management, RAS, DC-SCM, etc.

Learn More:

AI Your Way. Purpose-built Rack-scale Connectivity.

Why Your Mixture-of-Experts Model is Only as Good as Your Fabric

Supercharging Agentic AI: How Leo X-Series Smart Memory Controllers Accelerate Inference with KV Cache Offload

Visit Astera Labs in Booth B34

AsteraLabs.

# Thank You!

