Choose Your Own XPU

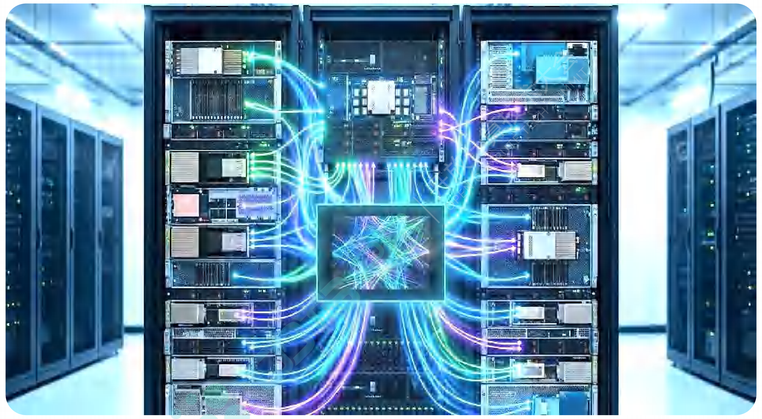
An Open Rack to Maximize Inference Performance and TCO Tokens-Per-Yuan

Chris Petersen, VP, CTO Office, Astera Labs
Board Director, CXL Consortium and UALink Consortium

What Hyperscalers & AI Labs Demand from AI Infrastructure

Token Economics

Optimize performance per watt and cost per token



Scalability

Enable heterogeneous accelerators at cluster scale

Resilience

Maximize uptime and utilization

Universal Scale-up Fabric Solution

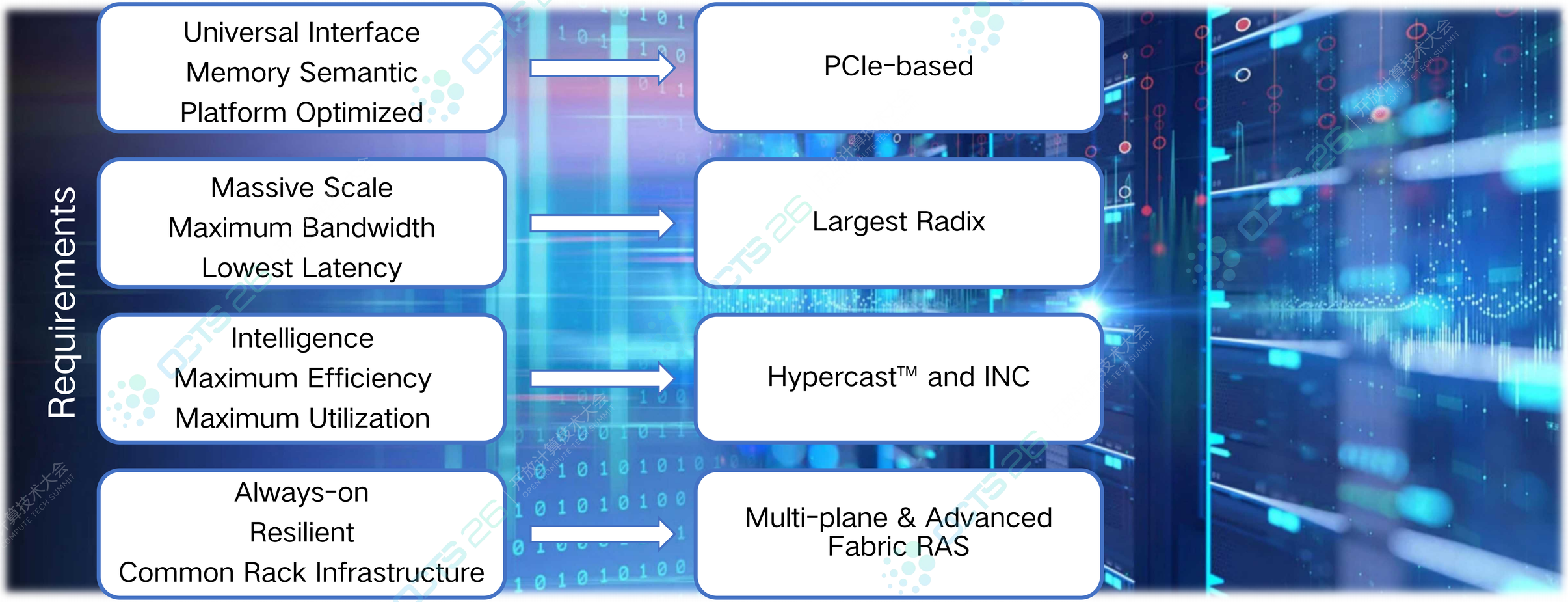
Requirements

- Universal Interface
  Memory Semantic
  Platform Optimized
- Massive Scale
  Maximum Bandwidth
  Lowest Latency
- Intelligence
  Maximum Efficiency
  Maximum Utilization
- Always-on
  Resilient
  Common Rack Infrastructure

Universal Scale-up Fabric Solution

Requirements

Universal Interface Memory Semantic Platform Optimized	PCIe-based
Massive Scale Maximum Bandwidth Lowest Latency	Largest Radix
Intelligence Maximum Efficiency Maximum Utilization	Hypercast™ and INC
Always-on Resilient Common Rack Infrastructure	Multi-plane & Advanced Fabric RAS



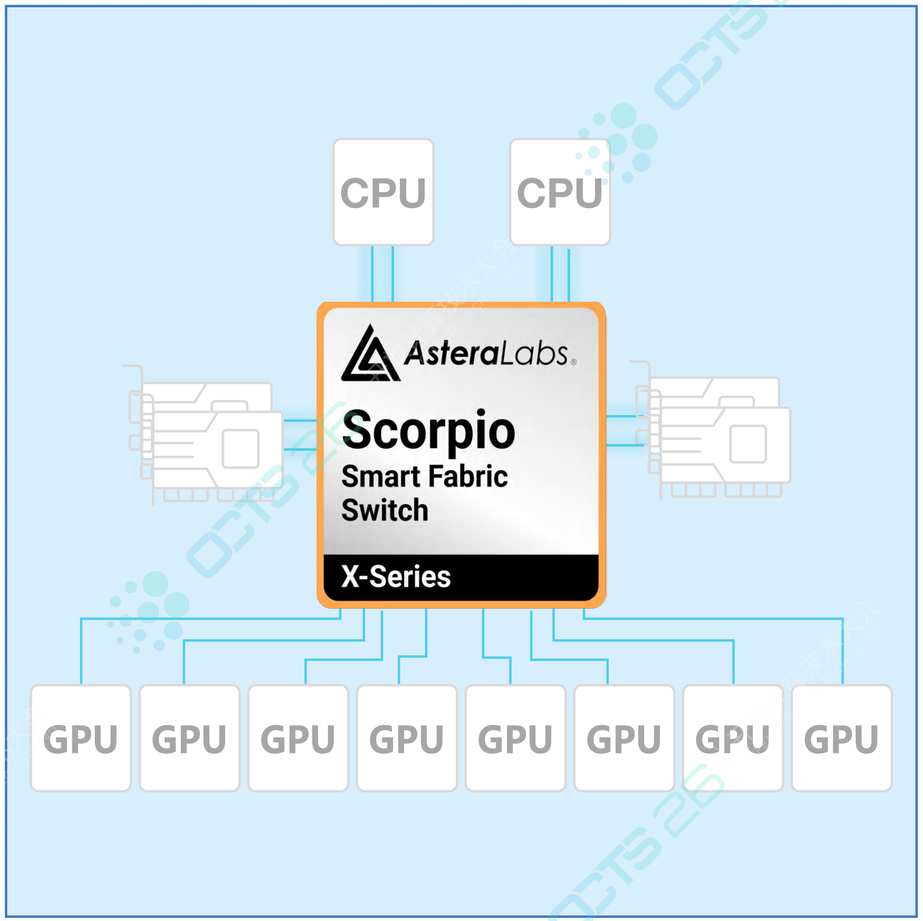
Universal Scale-up Fabric Solution

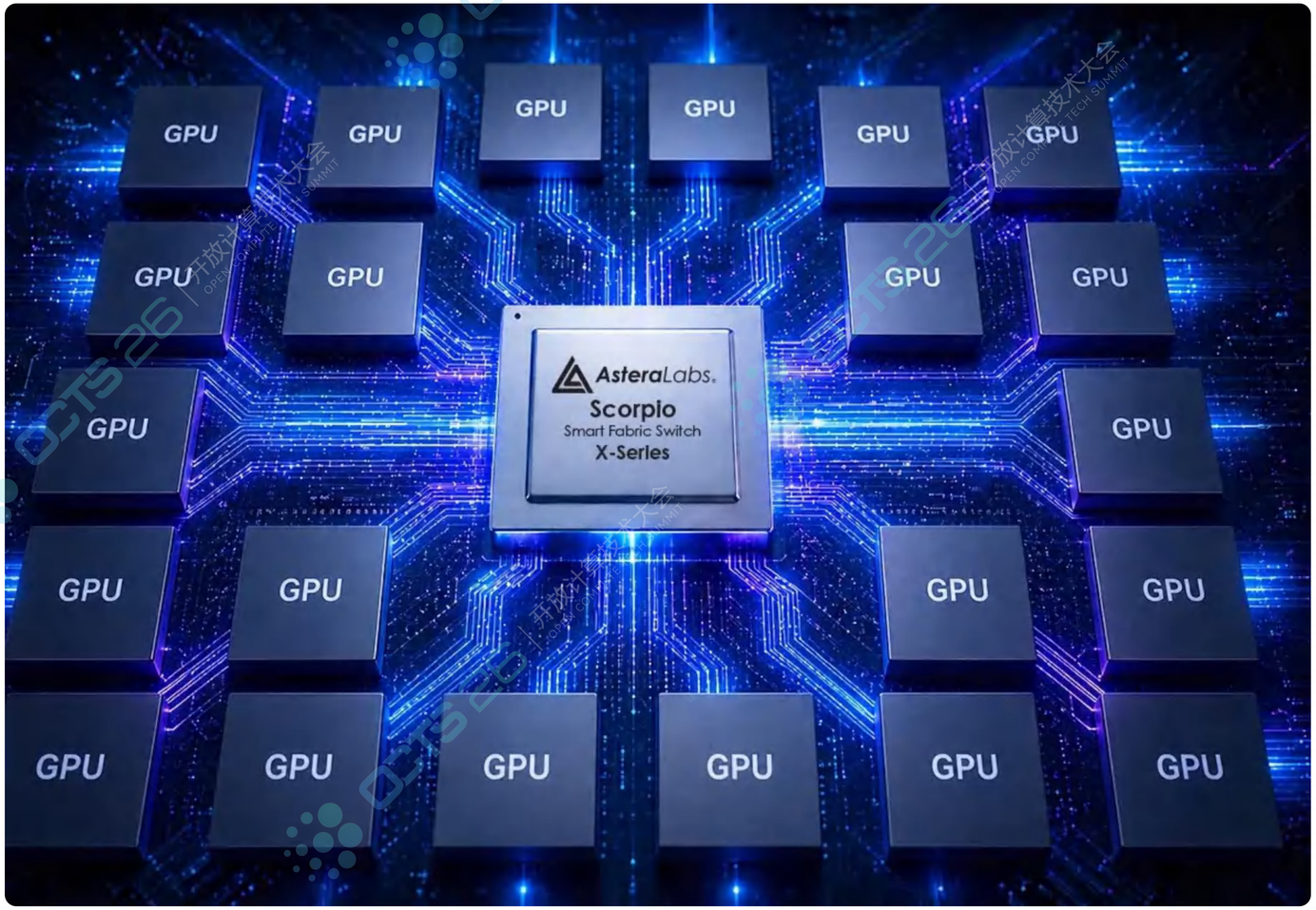
Requirements

Requirements	
Universal Interface Memory Semantic Platform Optimized	PCIe-based
Massive Scale Maximum Bandwidth Lowest Latency	Largest Radix
Intelligence Maximum Efficiency Maximum Utilization	Hypercast™ and INC
Always-on Resilient Common Rack Infrastructure	Multi-plane & Advanced Fabric RAS

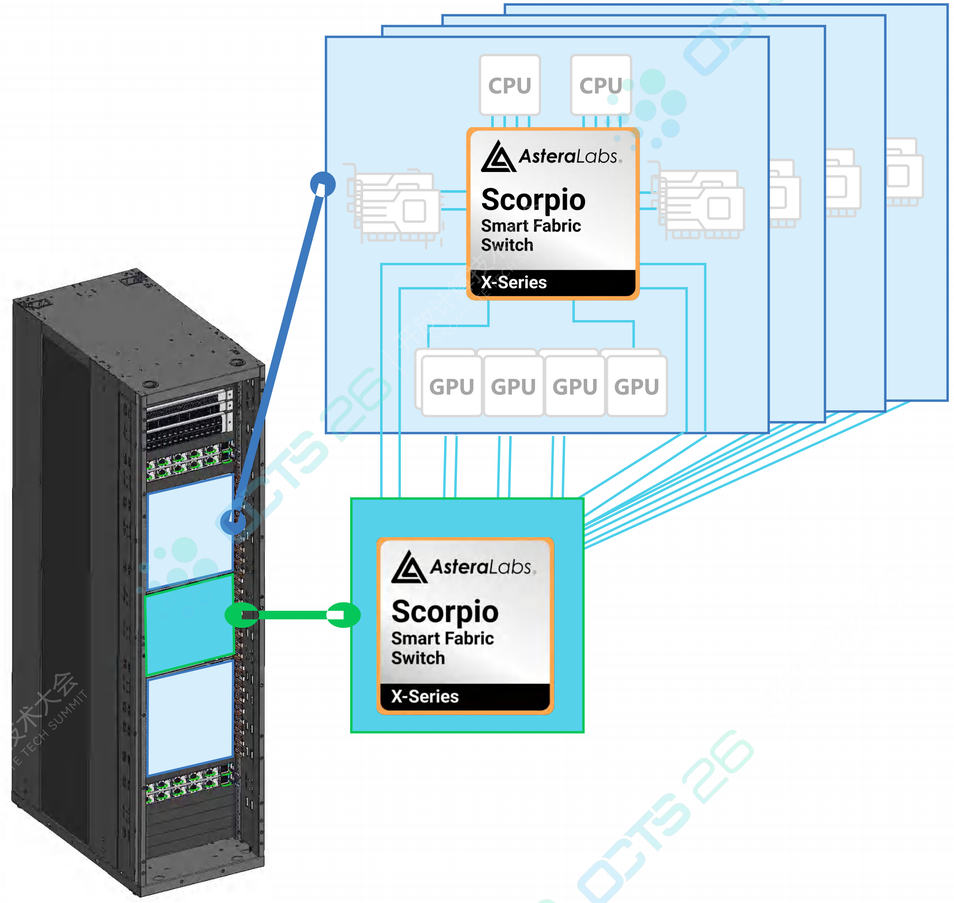
AI Fabric Solution

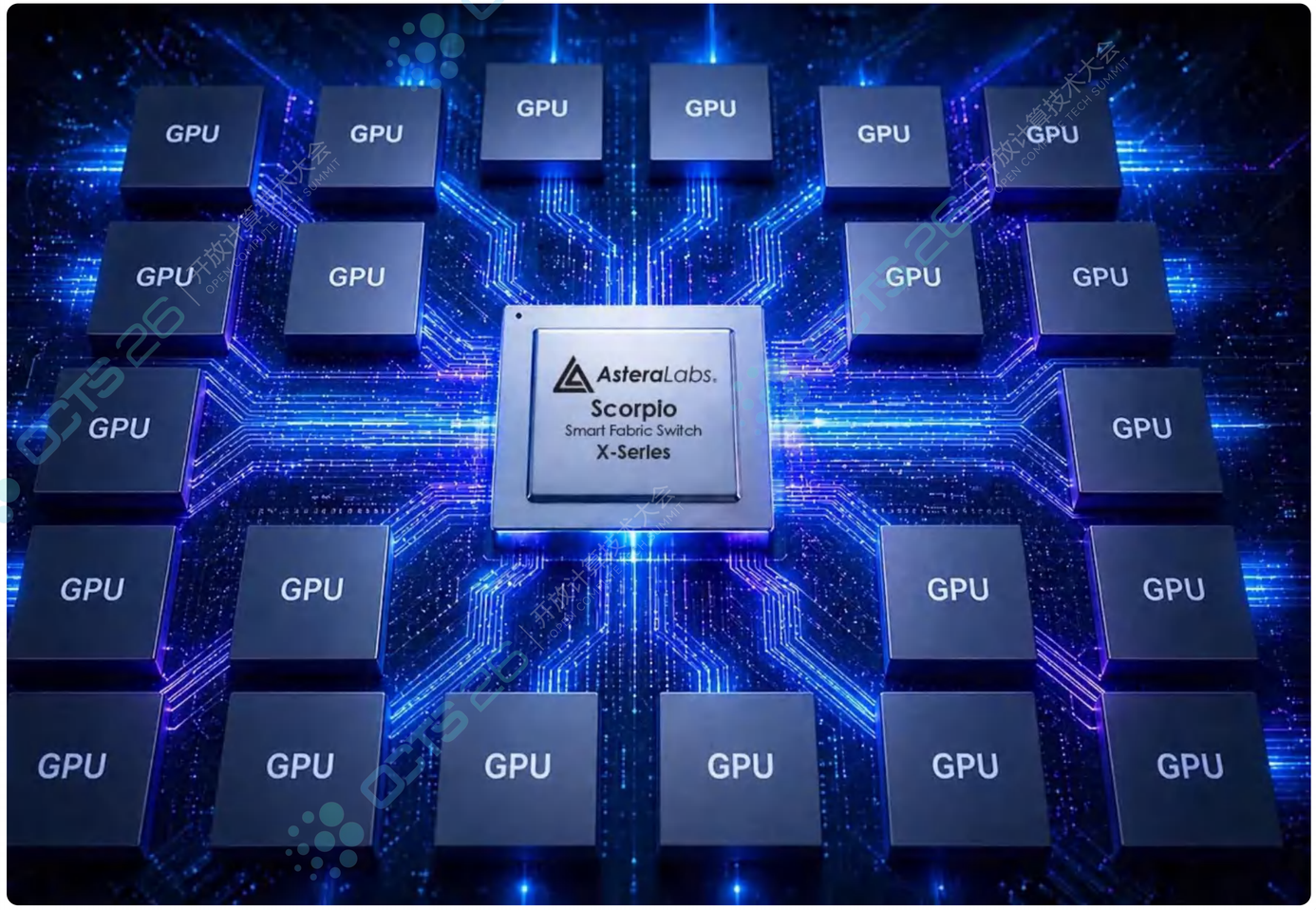
Scorpio™ X-Series 320 Lane Smart Fabric Switch
Largest open, memory-semantic switch



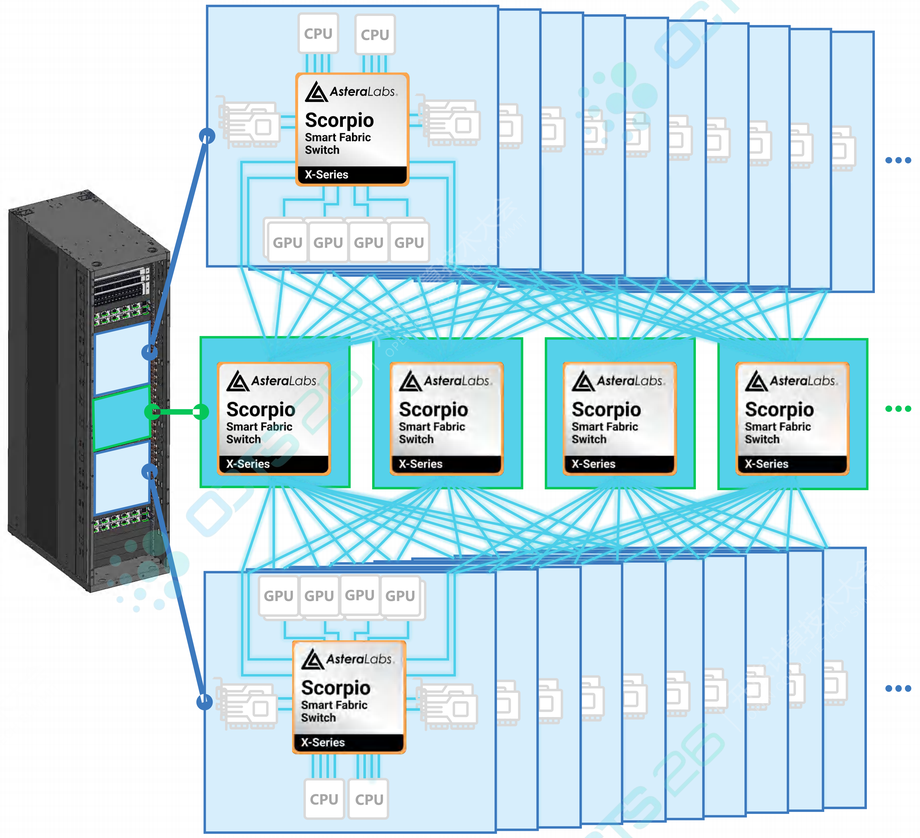


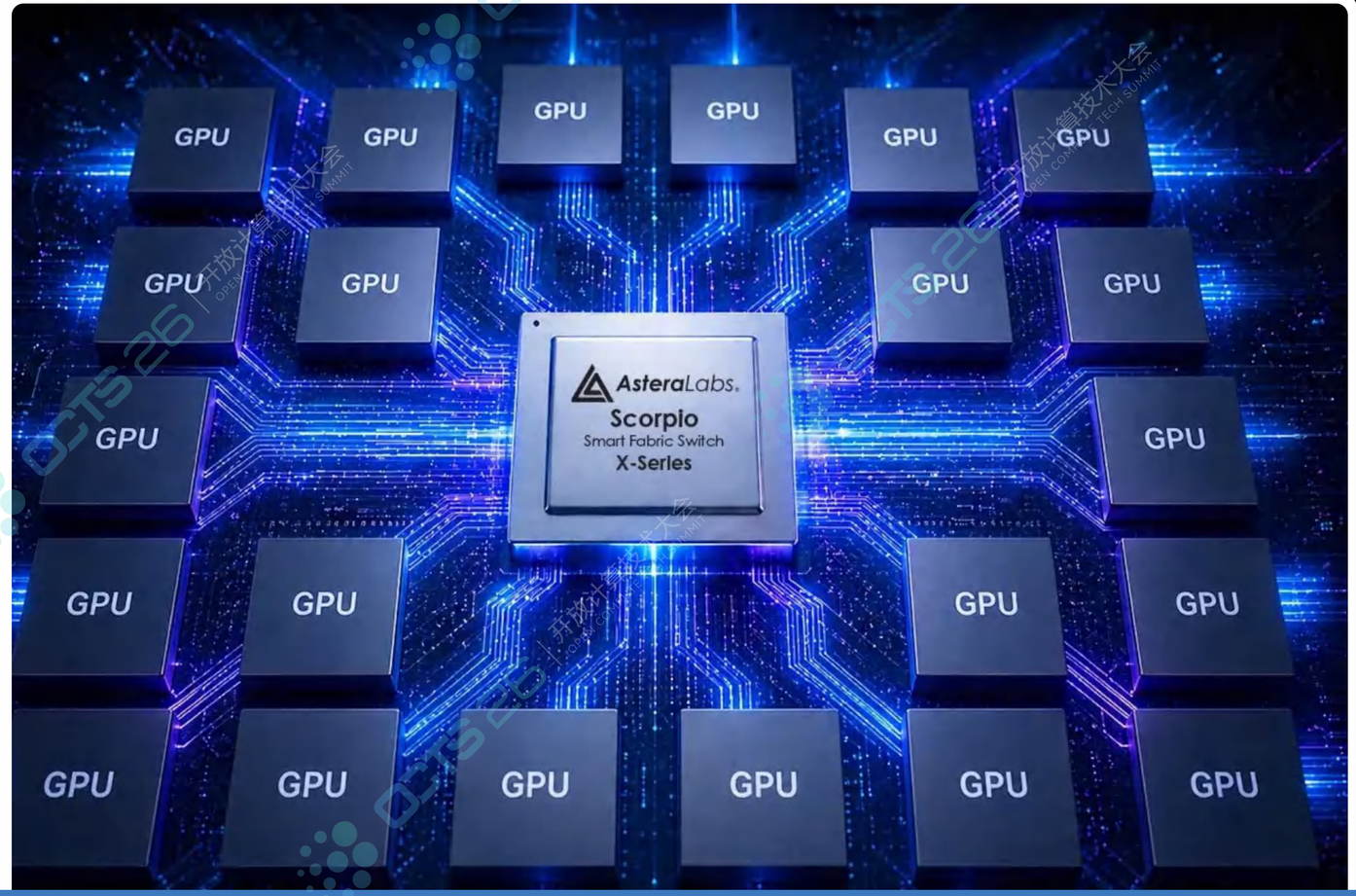
Scorpio™ X-Series 320 Lane Smart Fabric Switch
Largest open, memory-semantic switch





Scorpio™ X-Series 320 Lane Smart Fabric Switch: Largest open, memory-semantic switch





Expand GPU scale-up with multi-level switching fabric – up to 512 GPUs

Introducing Hypercast™ and In-Network Compute (INC) Engines

Reducing GPU I/O Translates to Real Data Center Savings

Hypercast™ Data Replication Engine

Accelerates all-gather, all-scatter and all-to-all

"Share all my weather logs with all stations"

COSMOS

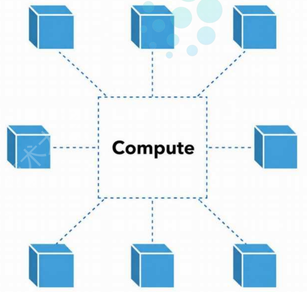
Accelerate collective operations by up to

2X

In-Network Compute Engines

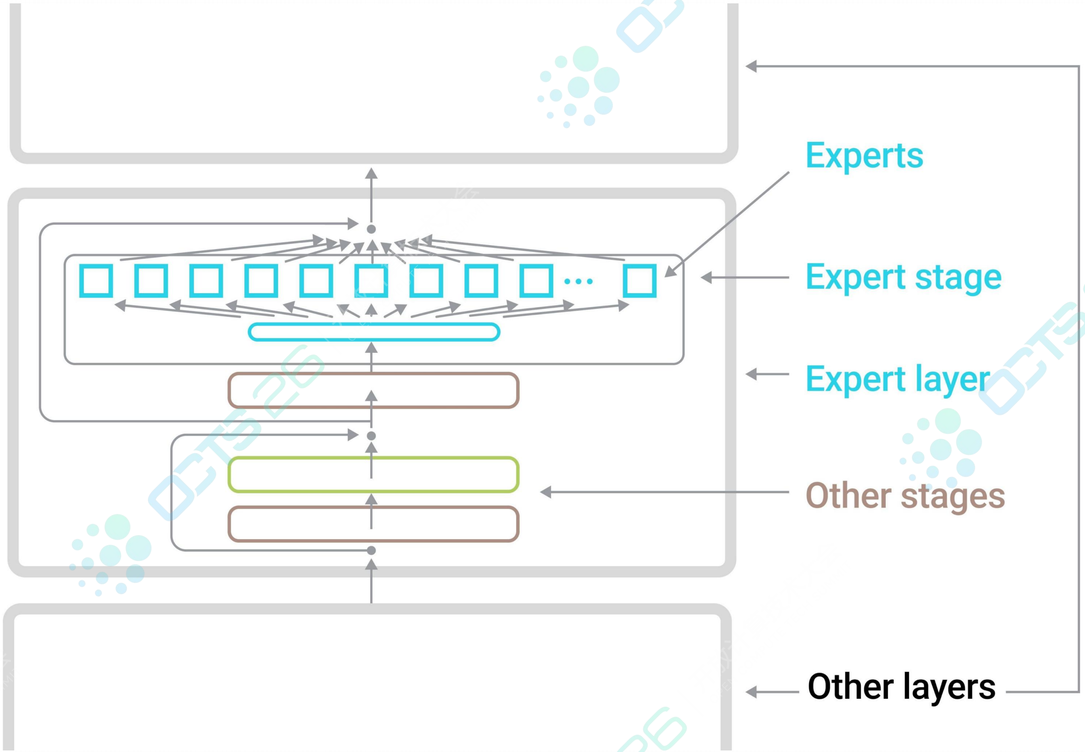
Accelerates all-reduce and reduce-scatter

"Calculate average temp across all weather stations"



COSMOS

What is MoE? MoE Makes Larger Models Practical!



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

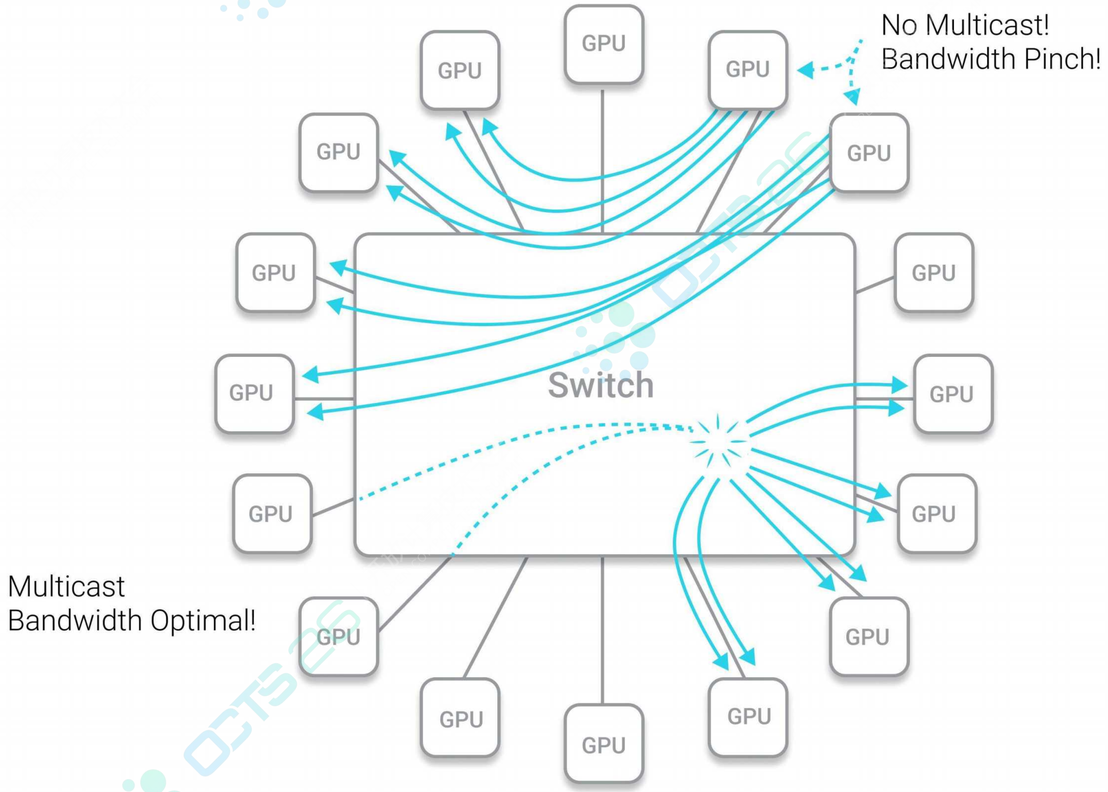
This requires unique multicast groups to target combinations of experts.

MoE 为每个Token只会启用少数专家，因此能以更高效率执行更大型的 LLM 模型。

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.

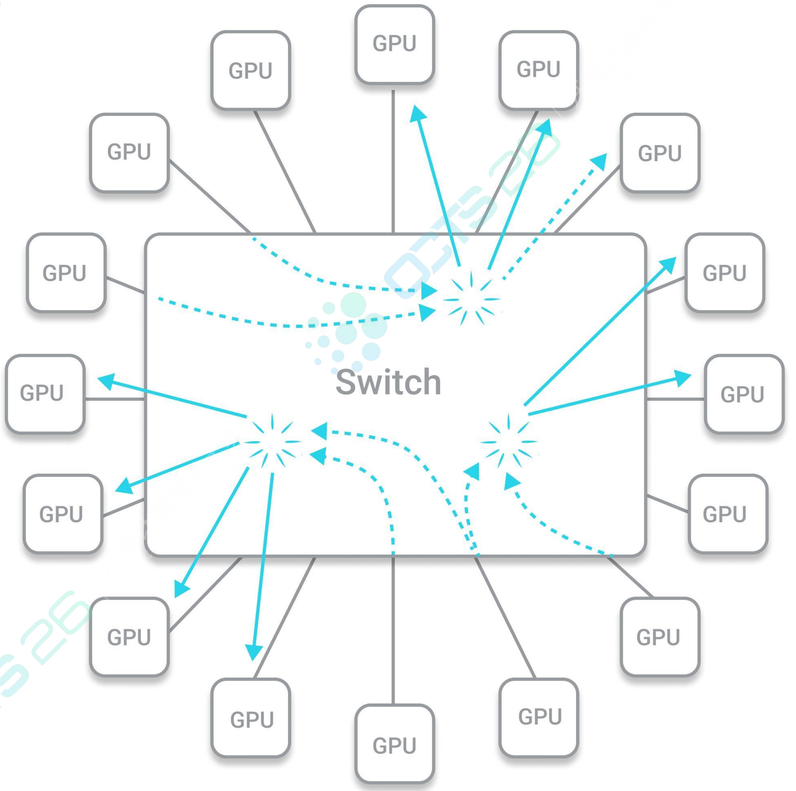
当只有部分集合操作能被加速时，其余操作就会退回到低效率的 Unicast 路径，进而造成瓶颈。



Astera Labs Hypercast™ Technology

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

Hypercast 是一种为 AI 原生打造的 Multicast 技术，其特定设计为同時支持动态专家路由与高频集合操作。



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

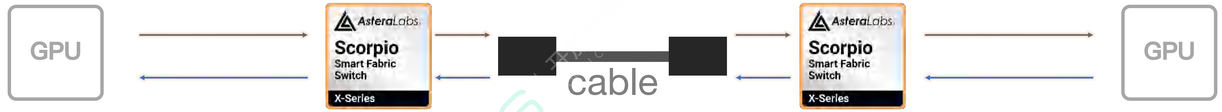
Minimize error rates and quickly isolate link issues

Signal

Faster repair & recovery

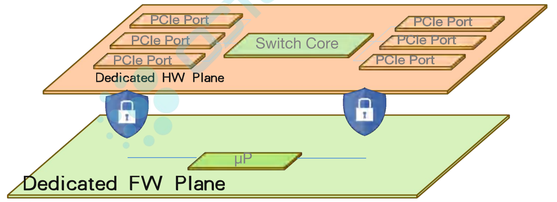
COSMOS

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

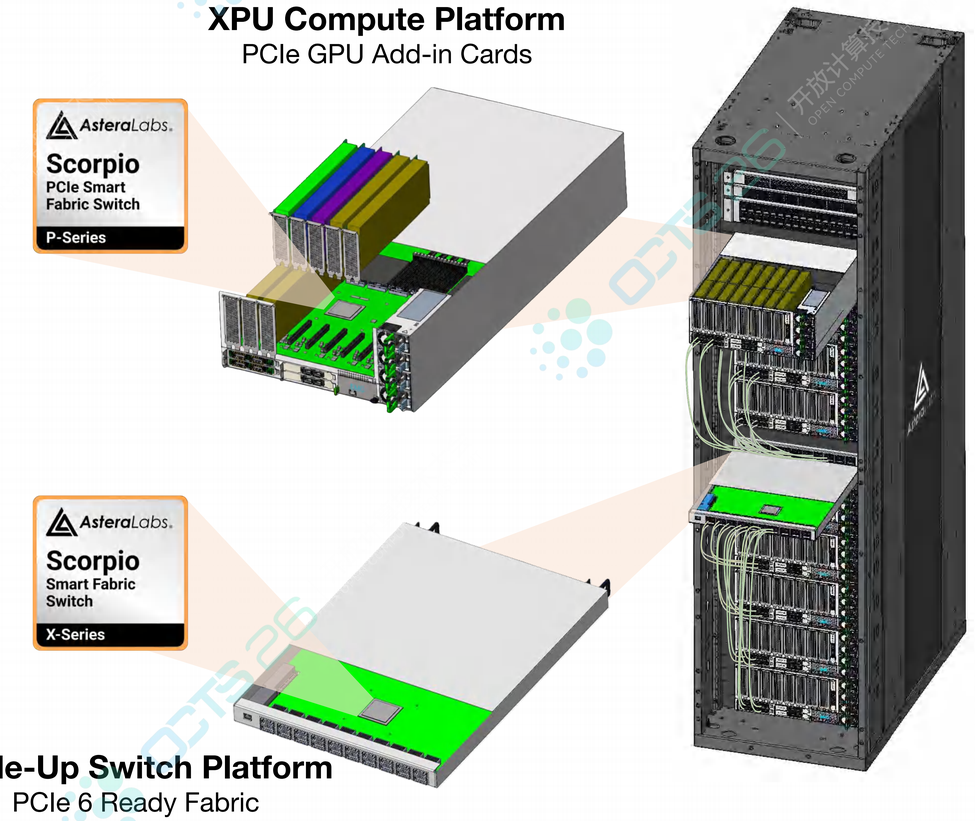
Common Scale-up Rack Architecture

XPU Platform Specific

Open Protocol
Hypercast
INC
Resilience

XPU Optionality

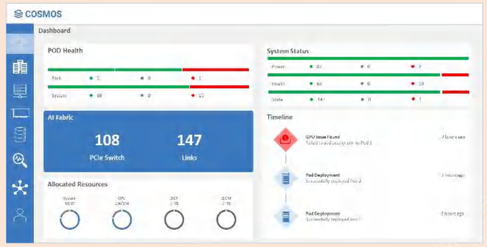
XPU | XPU | XPU



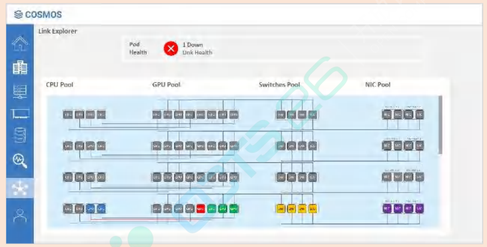
Leverage common infrastructure across multiple XPUs and generations to reduce TCO

Enriching Fabric & Rack-Level Management with COSMOS

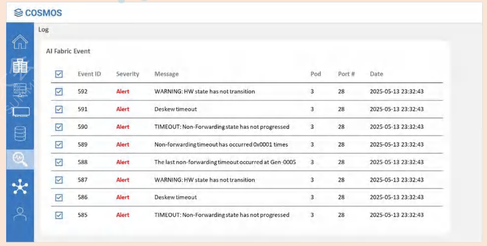
Identify Fabric Bottlenecks



Detect & Resolve Fabric Issues



Rapidly Configure & Deploy

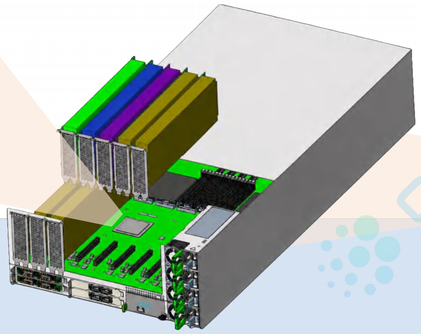


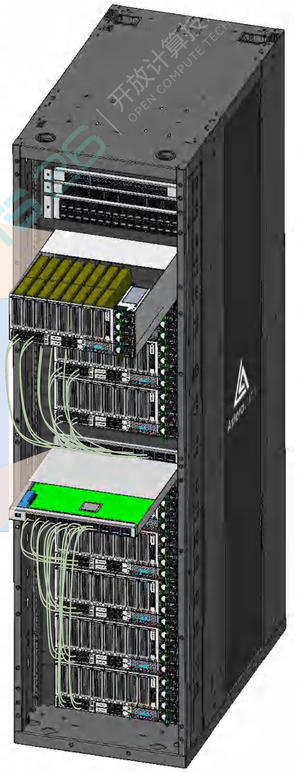
Managing: System, fabric, and rack-level configuration using open standards (OpenBMC, DMTF, SAI, etc)

Monitoring: Performance metrics, error events, fabric notifications

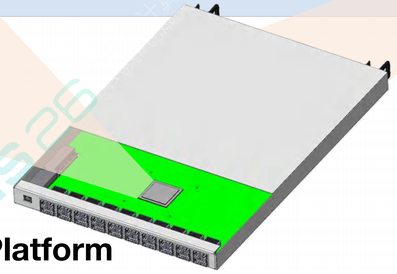
XPU Compute Platform
PCIe GPU Add-in Cards

AstraLabs.
Scorpio
PCIe Smart Fabric Switch
P-Series





AstraLabs.
Scorpio
Smart Fabric Switch
X-Series



Scale-Up Switch Platform
PCIe 6 Fabric

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

Scorpio Smart Fabric Solution

- Largest open, memory-semantic switch
- Intelligent rack-scale infrastructure
- 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
CXL Smart Memory Controller	PCIe/CXL Smart Retimer	Ethernet Smart Retimer	Smart Fabric Switch

Now is the time for a common scale-up infrastructure that works with any XPU

THANK YOU!