



**Hewlett Packard
Enterprise**

講題 – AI Data Center 之介紹

Networking for AI

**CT Hu
Technical Manager, Taiwan**

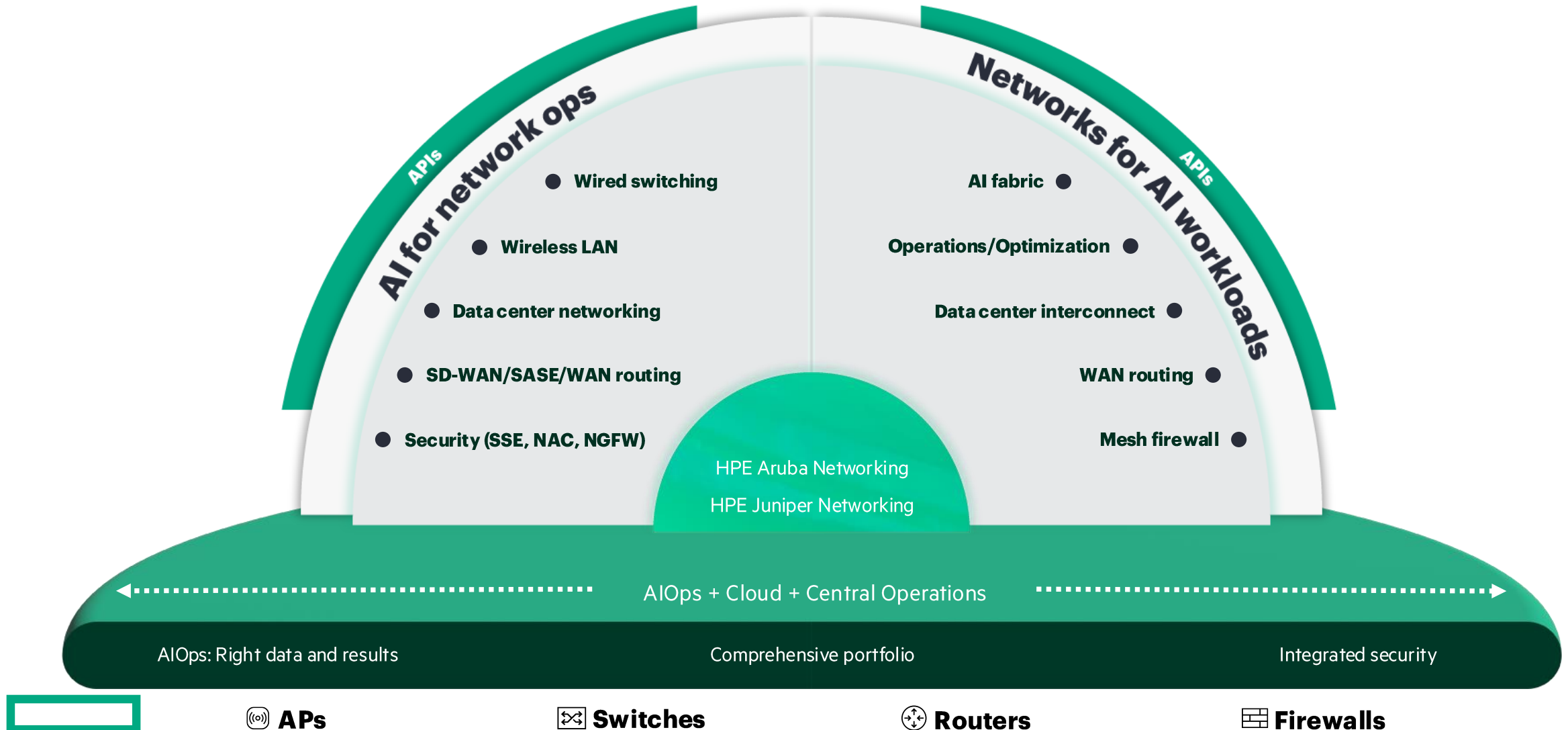


Agenda (大綱)

- Agentic AI 趨勢
- AI Data Center 解密
- 開放的HPE AI DC網路方案
- AI cluster的設計建議
- 總結



The secure AI-native network



Recognized as a leader

A LEADER
Wired/Wireless:
Furthest in Vision, Highest in Execution
#1 in all critical capability use cases

Figure 1: Magic Quadrant for Enterprise Wired and Wireless LAN Infrastructure



Gartner

A LEADER
Data Center Networking
#1 in Enterprise Build Out use case

Figure 1: Magic Quadrant for Data Center Switching



Gartner



The Thinking Game | Full documentary | Tribeca Film Festival official selection



Google DeepMind 和3

訂閱

26萬



分享

提問

儲存



觀看次數：4.4億次 8 個月前

The inside story of the AI breakthrough that won a Nobel Prize.

AI's impact spans every industry

Know how AI can enhance your customer's business

Finance



Fraud detection
Personalized banking
Investment insights

Healthcare



Molecule simulation
Drug discovery
Clinical trial data analysis

Retail



Personalized shopping
Automated catalog descriptions
Automatic price optimization

Telecommunications



AI virtual assistants
Network performance tuning
Remote support capabilities

Media and entertainment



Character development
Video editing and image creation
Style augmentation

Manufacturing



Factory simulation
Product design
Predictive maintenance

Federal



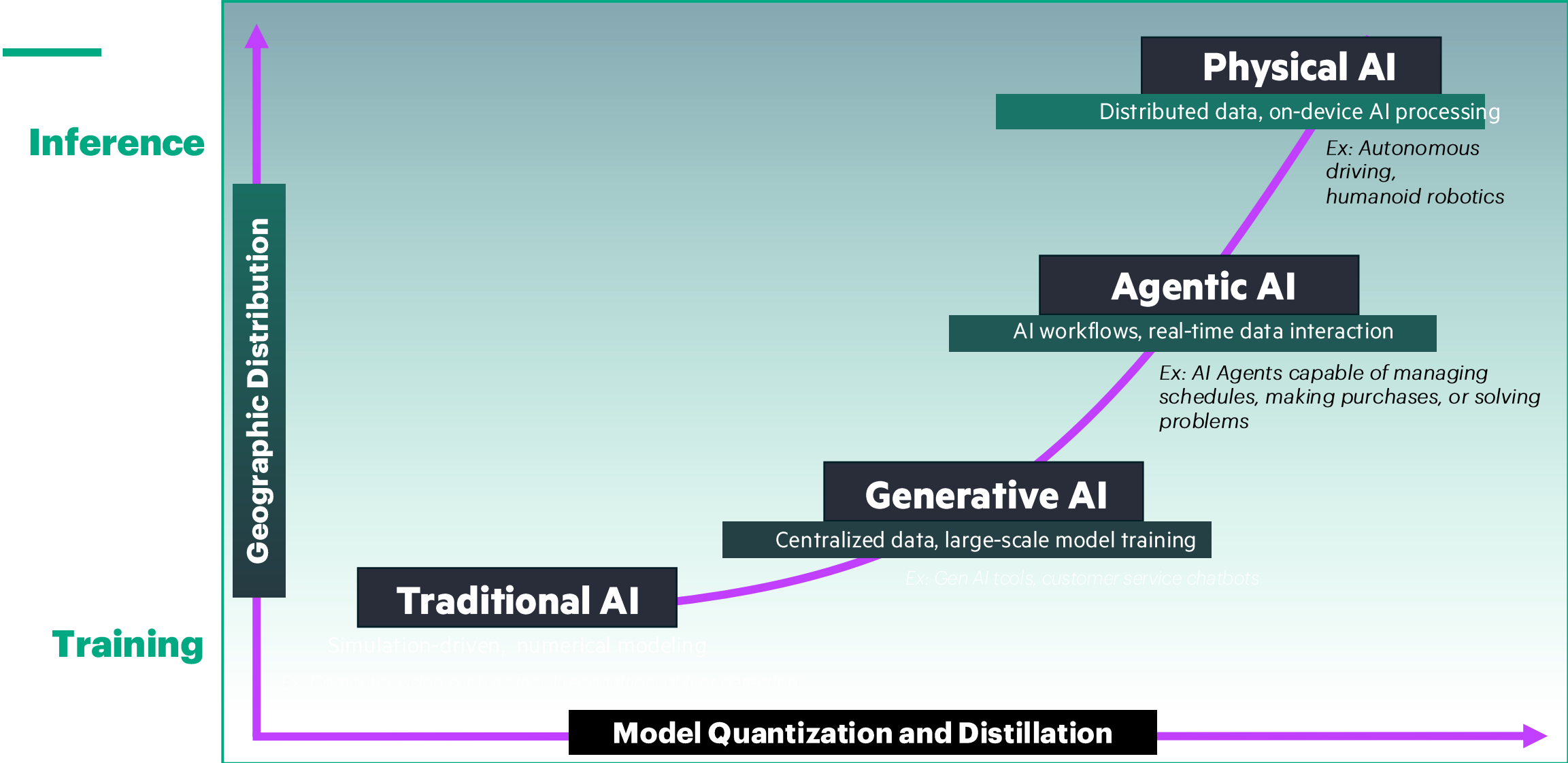
Document summarization
Audit compliance
AI virtual assistants

Energy

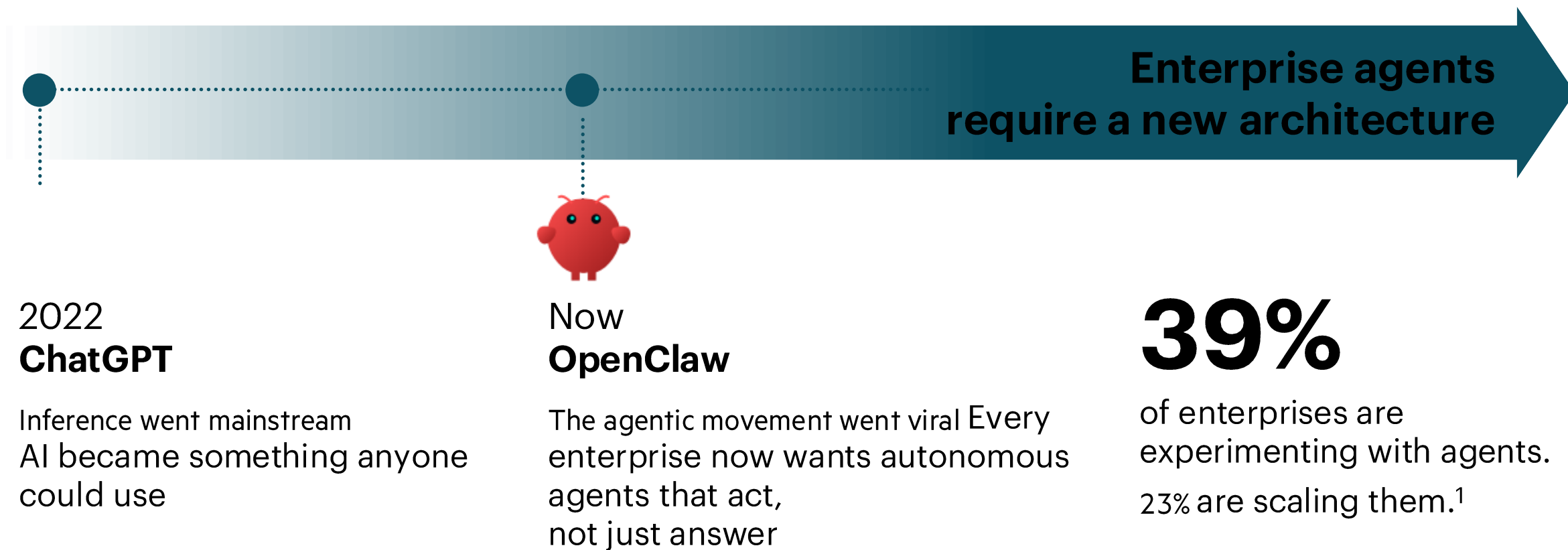


Knowledge base Q&A
Predictive maintenance
Customer service

AI 的演進



The agentic movement has made it to the enterprise

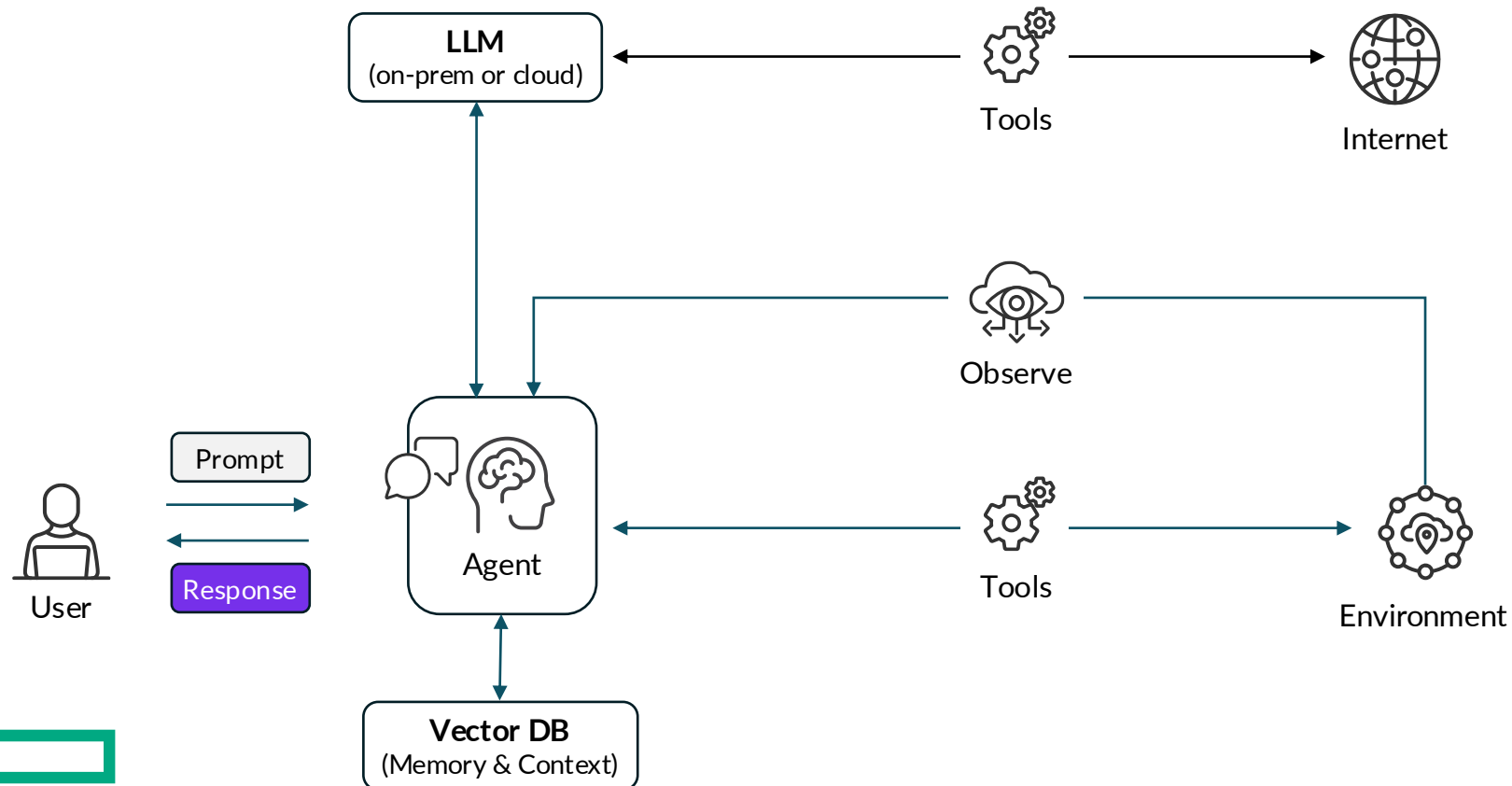


 1. McKinsey State of AI Global Survey, 2025.

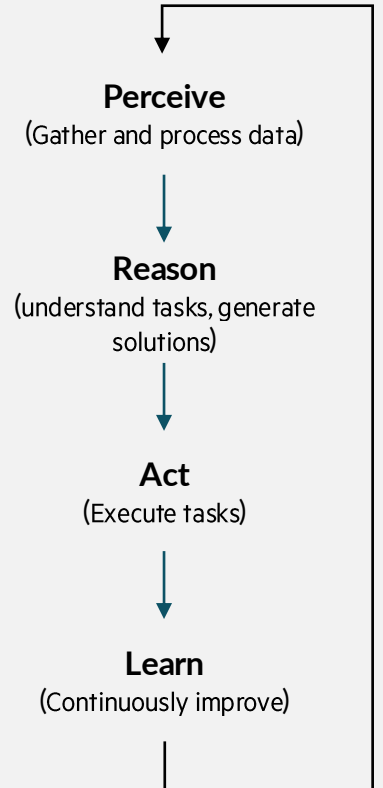
Agentic AI

Beyond traditional AI:

- Agentic AI doesn't just respond—it **plans, executes, and adapts** to achieve complex objectives.
- **Decompose complex goals** into actionable steps,
- **Maintain context** across multiple interactions, and iteratively refine its approach based on results.



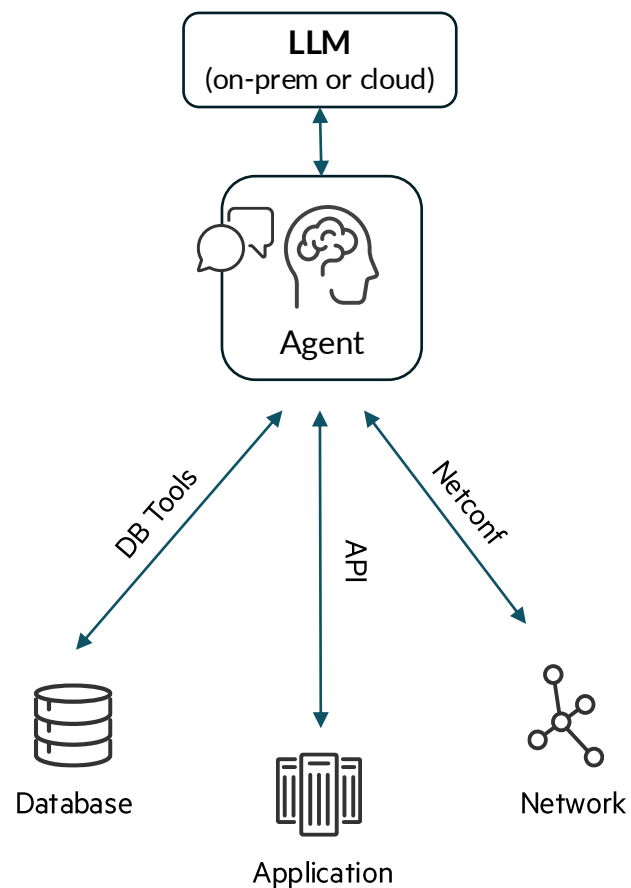
Agentic AI process



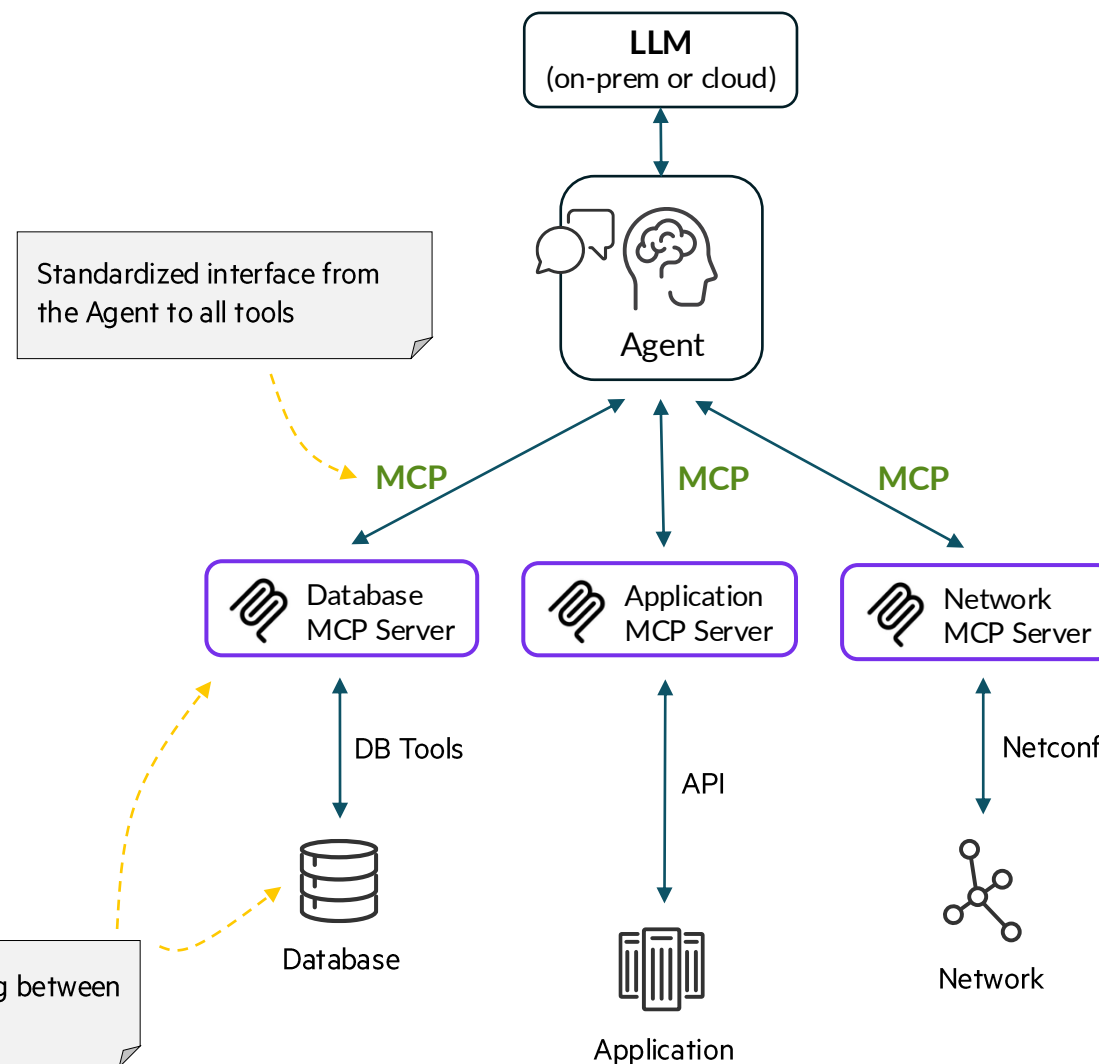
Model Context Protocol (MCP)



Without MCP



With MCP



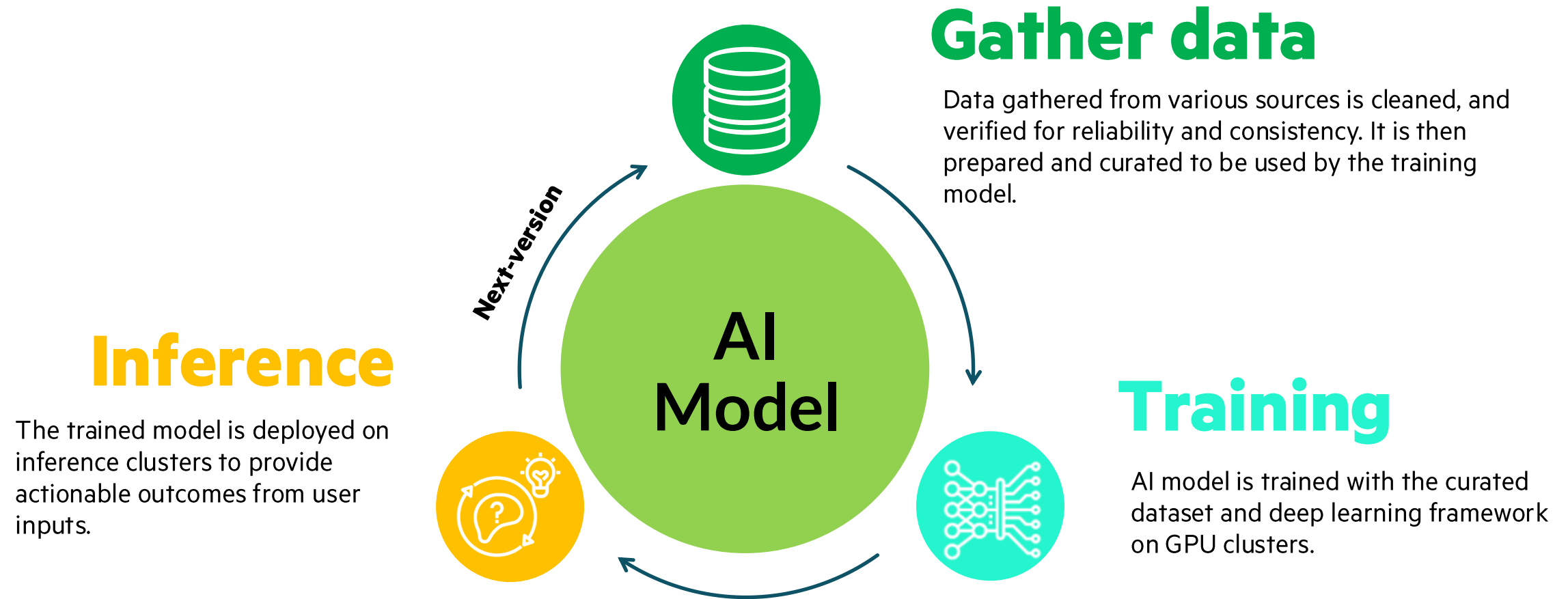


AI Data Center 解密

- AI DC 需求與挑戰

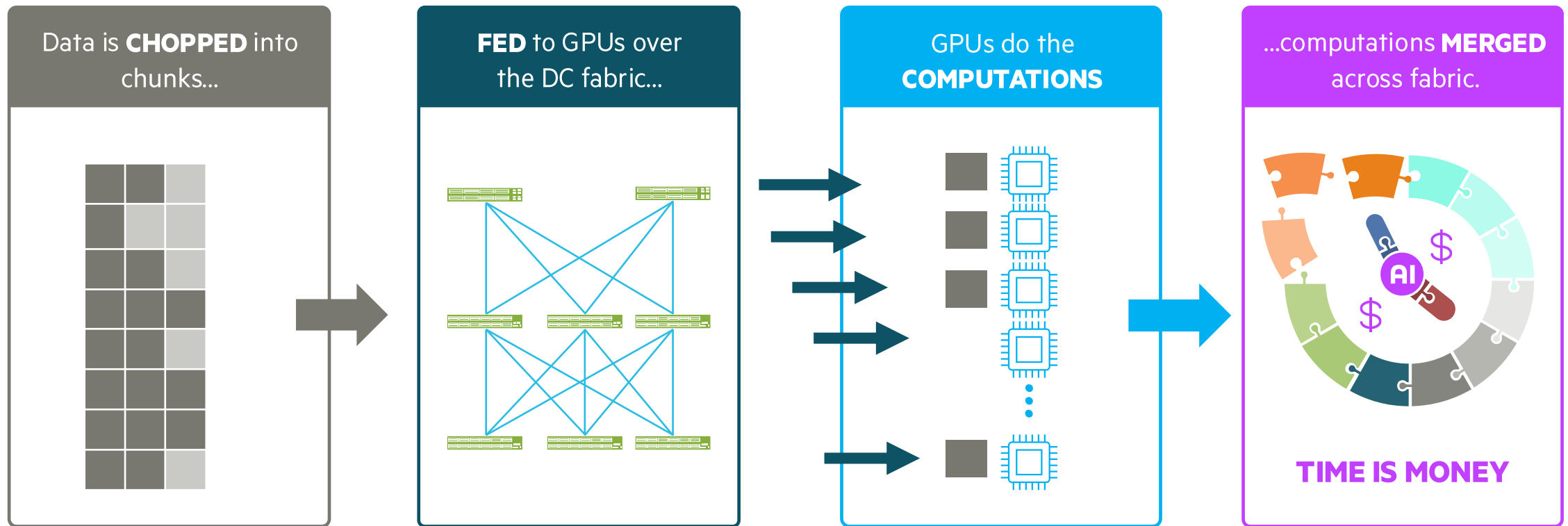


Lifecycle of an AI Model



The AI/ML app lifecycle can be a continuous, iterative process of designing and developing models, training and validating them with curated data, and deploying them into production while monitoring their performance for constant refinement and improvement.

AI model performance and economics based on minimizing Job Completion Time (JCT)



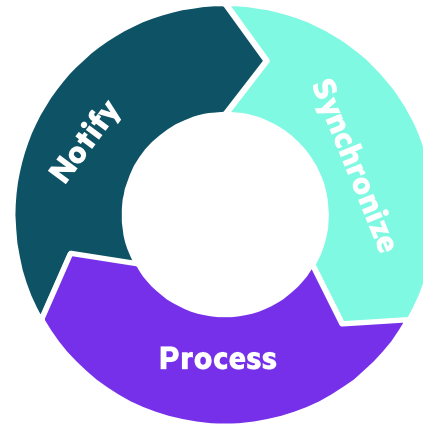
The progression of all nodes can be held back by any delayed flows (tail latency)

AI/ML Training Workload Challenge

How AI/ML traffic differs from traditional DC traffic?

2. Send results of computation

All-to-All Collective (Several Methods)
(Everyone sends to everyone)



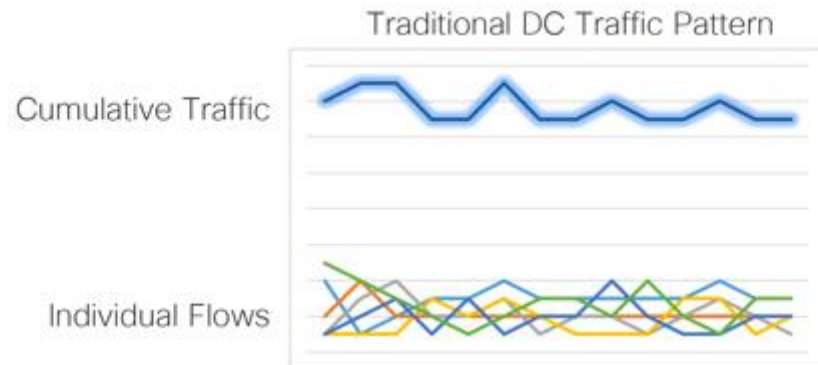
1. Execute Instructions on GPUs

High bandwidth can saturate network links

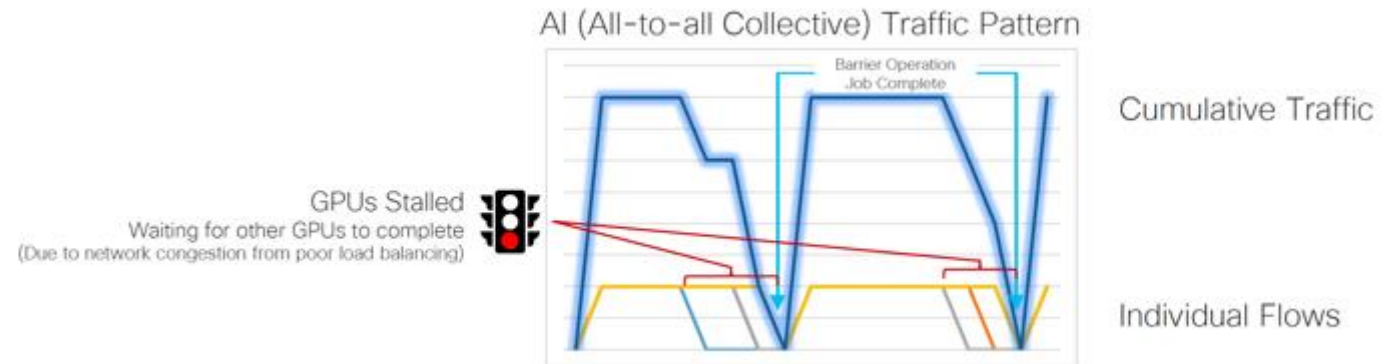
The Network is fundamental to AI/ML applications

3. Wait for everyone to complete

Creates synchronization between GPUs
Computation stalls waiting for the slowest path
Job Completion Time is based on the worst-case tail latency

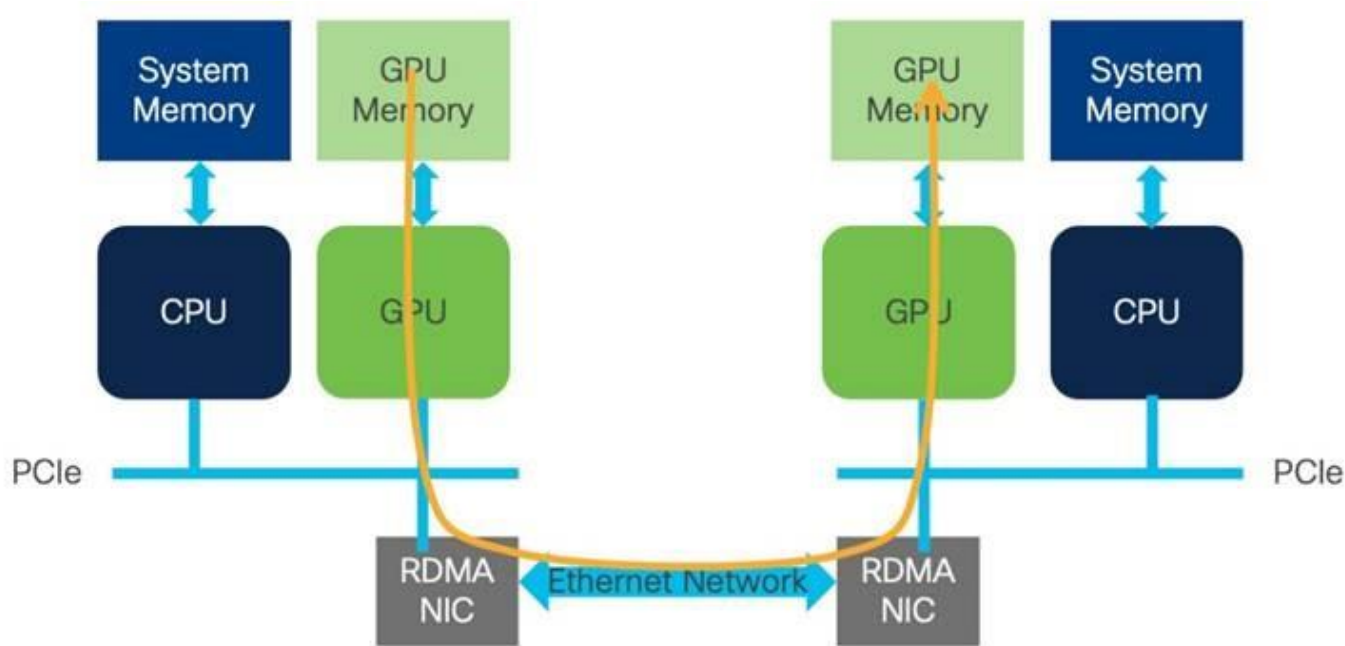


Many asynchronous small BW flows (TCP, UDP)
Chaotic pattern averages out to a consistent load



Few synchronous high BW flows (**RDMA, Remote Direct Memory Access**)
Synchronization magnifies long-tail latency & bad load-balancing decisions

Remote Direct Memory Access (RDMA) Overview



- A technology that makes it possible to read data directly from the main memory of one host and write that data directly to the main memory of another host.
- RDMA data transfers bypass the kernel networking stack in both computers, and as a result improves throughput, latency, and performance
- RoCE and InfiniBand are network protocols that support RDMA to improve the efficiency and speed of data transfers
- RDMA requires a lossless network



AI is driving explosive growth in networking

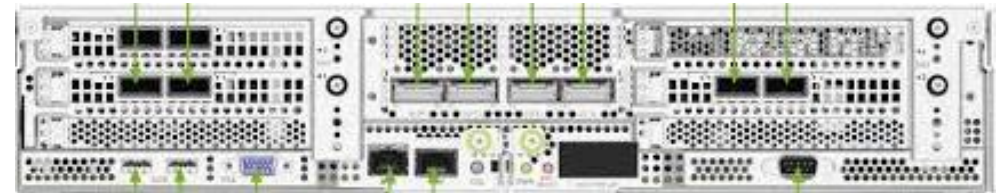
General Purpose Server



2x25G still the most
common configuration
we sell

50G per server

NVIDIA DGX H100 Server



8x400G Backend
GPU Network
CX7, BF3 NIC
support both
Infiniband and
Ethernet

2x100G
Frontend
Network

4x200G Storage
Network

~4Tbps per server



Portfolio of purpose-built servers for AI model training, tuning and inferencing

8-way GPU servers for accelerated AI

HPE Cray XD670



- 8X NVIDIA H200 GPUs
- 2x 5th Gen Intel Xeon Scalable processors
- Air-cooled and liquid-cooled
- 5U
- #1 in 6 different scenarios in MLPerf Inference: Datacenter v5.1 benchmark^[1]
- HPE Performance Cluster Manager (HPCM)

HPE ProLiant Compute XD685



- Choice* of
- 8x NVIDIA H200 GPUs or NVIDIA B200 GPUs or 8x AMD Instinct™ MI355X
 - 2x 5th Generation AMD EPYC™ Processors
 - 5U Direct liquid-cooled or 6U air-cooled
 - HPE ILO, HPE Performance Cluster Manager
- * Rolling GPU Releases

HPE Compute XD690

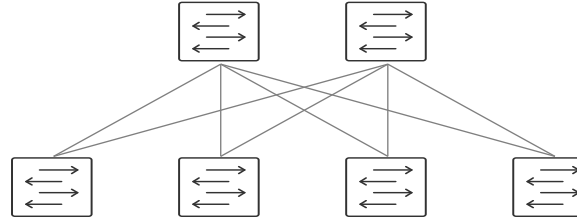


- 8x NVIDIA Blackwell Ultra GPUs (NVIDIA HGX™ B300)
- 2x Intel® Xeon® 6 processors with P-cores
- 10U, air-cooled
- HPE Services & Support
- 8 x 800G NIC (CX-8) for GPU cluster

^[1] HPE delivers several world records in latest MLPerf Inference benchmarks, September 2025

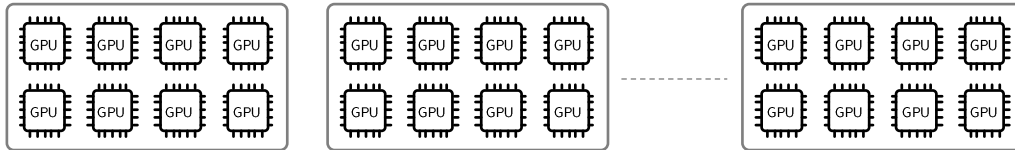
AI Data Center fabrics

Front-end Fabric / Inference Fabric

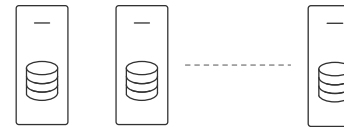


- User to workload connectivity
- Typically, 10/25/100G
- Ethernet

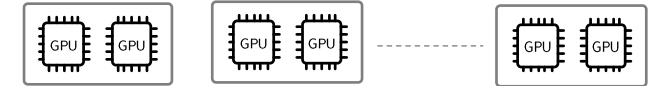
Training nodes



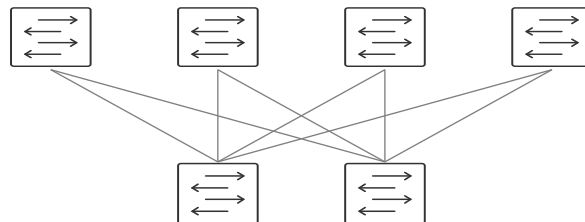
Storage nodes



Inference nodes

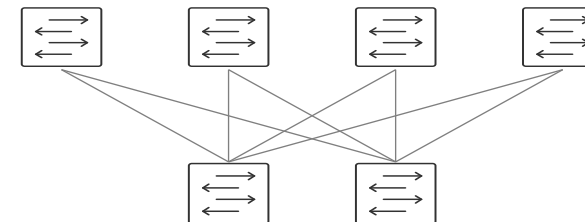


Training Fabric (Backend)



- Inter-node GPU communication
- Typically, 400G
- InfiniBand or Ethernet

Storage Fabric (Backend)



- GPU-to-Storage communication (Training & Inference)
- Typically, 100/200G
- InfiniBand or Ethernet

開放的HPE AI DC網路方案



HPE Juniper Networking: A Leader in Data Center Switching

Ranked #1 in Enterprise Data Center Network Buildout and #2 in AI Ethernet Fabric Buildout

2025 Gartner® Critical Capabilities™

Gartner DC MQC

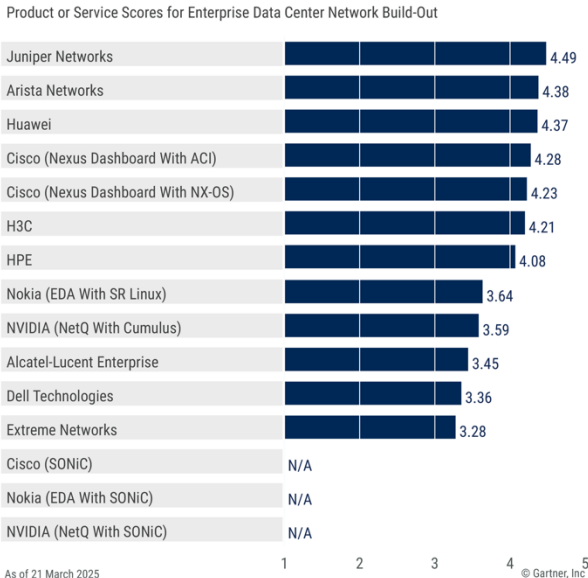
Figure 1: Magic Quadrant for Data Center Switching



Gartner

Enterprise Data Center Network Buildout

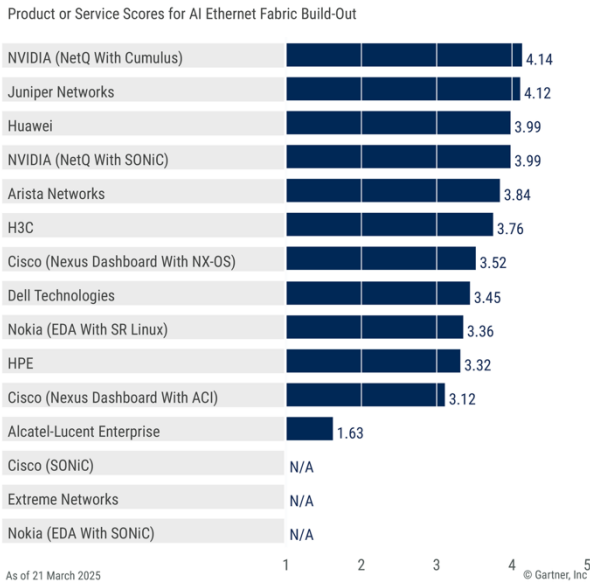
Figure 1: Vendors' Product Scores for Enterprise Data Center Network Build-Out Use Case



Gartner

AI Ethernet Fabric Buildout

Figure 2: Vendors' Product Scores for AI Ethernet Fabric Build-Out Use Case



Gartner

HPE Discover Las Vegas June 2026

Announcements for complete “Networks for AI” portfolio

Only vendor with ethernet networking for all use cases



Q4 CY2026

HPE Networking QFX5252: Industry-first ethernet scale-up solution for AMD Helios





Shipping Now!

HPE Networking QFX5250: Industry's highest performance 100% liquid cooled switch





New! Shipping Now!

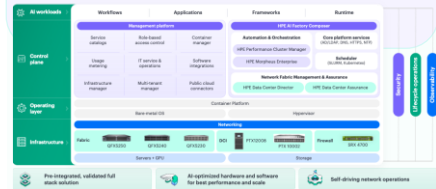
HPE Networking QFX5140: Purpose-built switch for Inference clusters





New! Available Now!

HPE AI Factory Solution available with HPE Networking

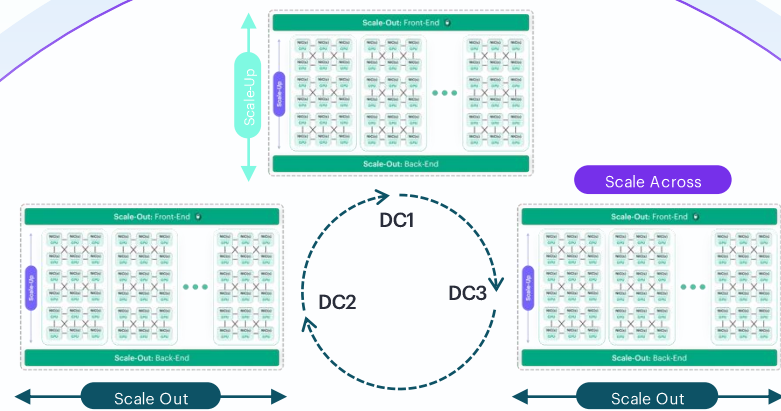




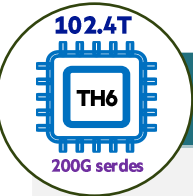
HPE Networking MX Routers
HPE Networking PTX Routers



AI Optimizations in Software



QFX5250-64OE-L : Liquid Cooled 20U



QFX5250-64OE-L

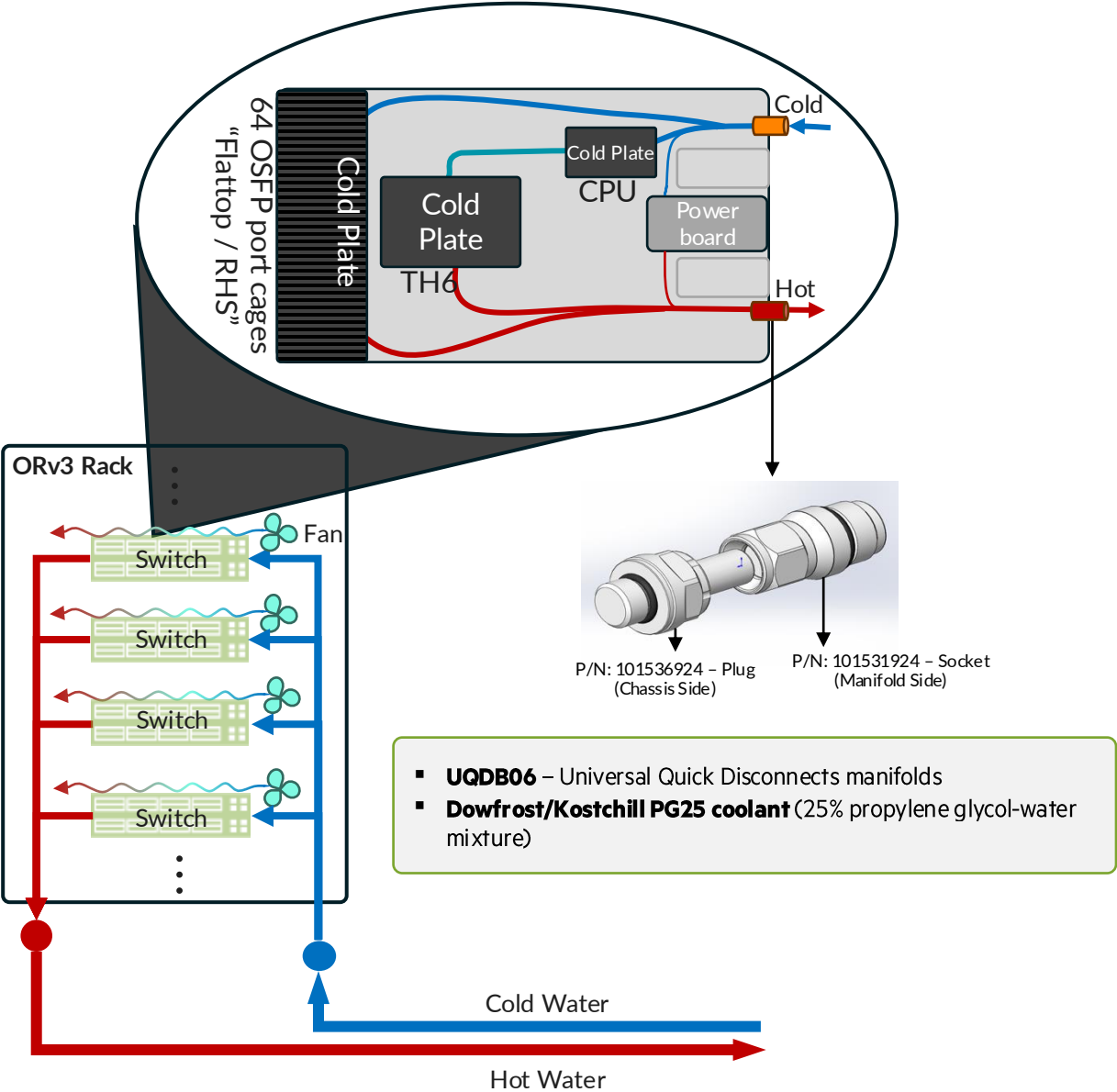
Target
Q1'26



Direct-to-chip liquid-cooled SKU

64x1.6T OSFP-RHS | 20U (ORv3)

- OSFP1600 (8x200G electric lanes)
- 64x1.6T, 128x800, 256x400, 512x200G, 512x100, 512x50G
- OSFP-RHS : VR, DR, FR, LR; DAC, AEC, AOC cables; LPO/LRO
- Direct-to-Chip Liquid-Cooled, Integrated DC PSU circuitry
- Hybrid cooling with majority of cooling taken care with liquid
- Designed for 21" ORv3 form factor

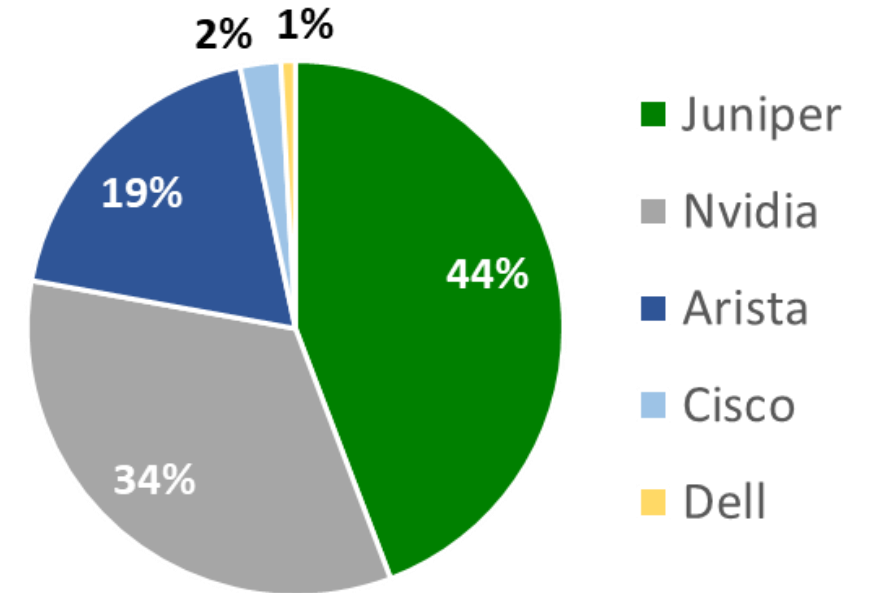


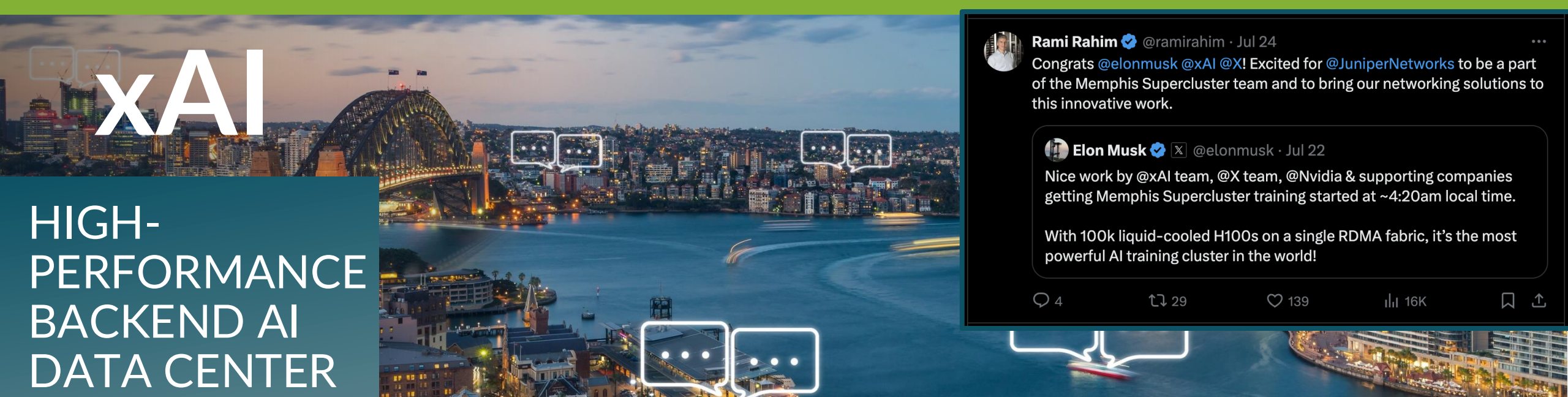
Junos hardware / software + ecosystem

Any self-driving vehicle needs a solid road – critical infrastructure

- Juniper - first to deliver 800G for AI use cases
 - Maximizing GPU performance by (Lossless Ethernet)
 - Continuously monitoring RoCE congestion metrics
 - Auto tuning DCQCN real-time (PFC / ECN)
- Inheriting 25 years of Protocol development and hardening from Juniper's Service Provider heritage

800GbE OEM Market Share, 2024





xAI

HIGH-PERFORMANCE BACKEND AI DATA CENTER NETWORKING

THE CHALLENGE

- Accommodate large-scale and efficient GPU cluster and server connectivity.
- AI training is compute-intensive, with traffic flows that break traditional data center networking
- Efficient use of GPU cycles. (Job completion time)

THE SOLUTION

- Juniper QFX5240 for 800G leaf spine connectivity
- Enabling the same form factor for both Front and Back-end AI clusters
- Efficient Load balancing and congestion control for RDMA traffic

WHY JUNIPER?

- **Junos feature richness** for congestion management with ROCEv2
- **AI Lab for POC** for various load balancing scenarios
- **Lightening responsiveness to needs**
- **Better supply chain** for switch and optics with Juniper



Rami Rahim @ramirahim · Jul 24

Congrats @elonmusk @xAI @X! Excited for @JuniperNetworks to be a part of the Memphis Supercluster team and to bring our networking solutions to this innovative work.



Elon Musk @elonmusk · Jul 22

Nice work by @xAI team, @X team, @Nvidia & supporting companies getting Memphis Supercluster training started at ~4:20am local time.

With 100k liquid-cooled H100s on a single RDMA fabric, it's the most powerful AI training cluster in the world!

4

29

139

16K

AI Technology on Juniper

CORPORATE PROFILE

- Founded 2017; Series D (backed by SoftBank Google Ventures, BlackRock, et al.)
- Builds AI hardware and integrated systems to run AI applications from the data center to the cloud
- Purpose-built enterprise-scale AI platform is the technology backbone for the next generation of AI computing

THE SOLUTION

- Juniper QFX and PTX series fabric for high density, performance & scalability to move massive volumes of data
- Automation across design, deployment & operations of the network lifecycle with Juniper Apstra

THE RESULTS

- **Accelerate high-performance ML model building across industries** enabling customers to deploy AI in days, not months
- **Eases the construction of ML infrastructure & delivers enhanced capabilities and efficiency** for ML model training, inference, and high-performance computing
- **Five times better performance** than traditional GPU architecture

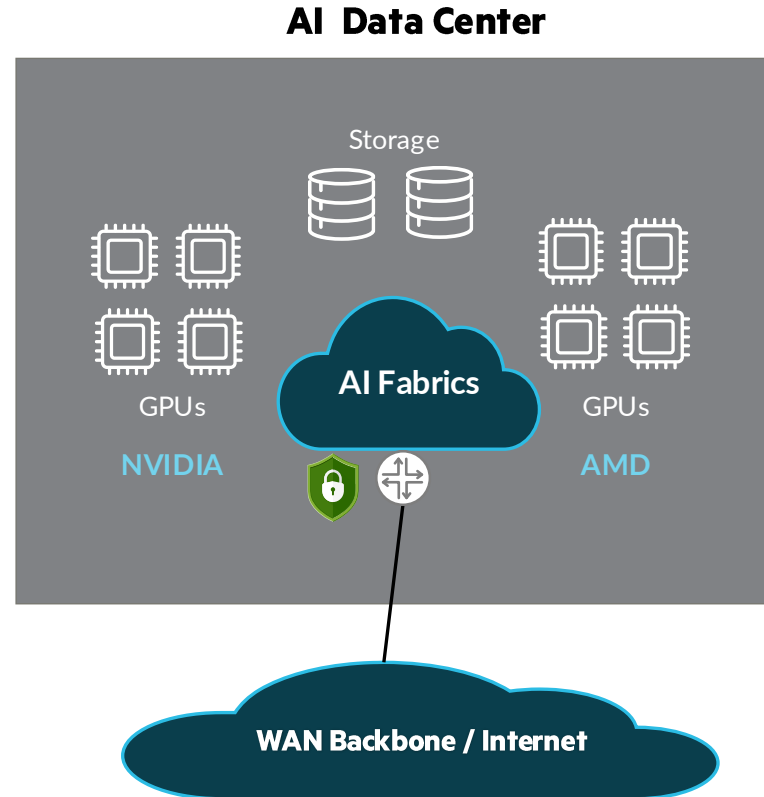
"In AI, data flow is king. We need the lowest network latency and the highest bandwidth possible, and the performance of the Juniper QFX5200 switches has been phenomenal,"

Vijay Tatkar, Director of Product Management

Proven success and measurable outcomes for our customers



Digital Ocean is 3rd largest Cloud infrastructure provider based on number of customers - has multiple data centers worldwide

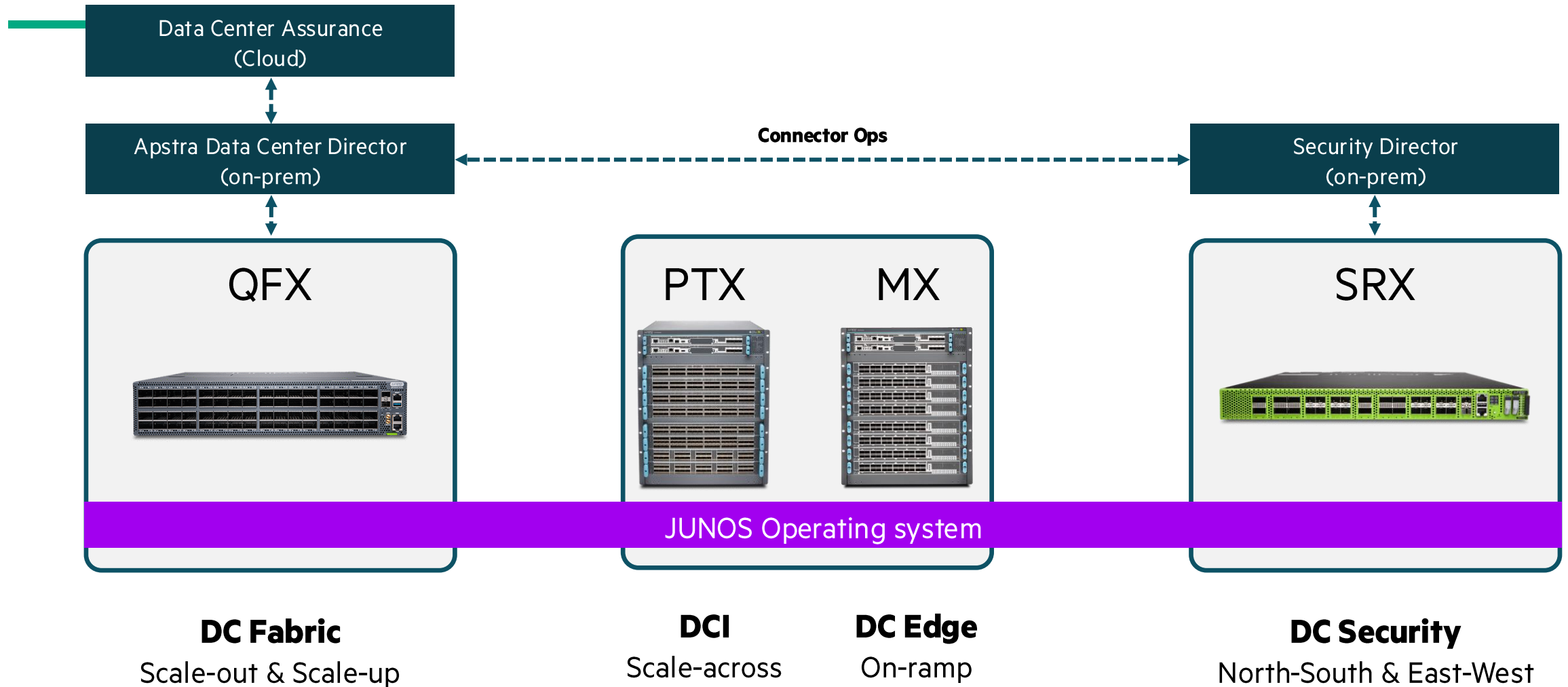


Outcome

- Juniper in front-end & back-end AI Fabrics
 - 400G/800G QFX switches deployed
- Networking for both Nvidia and AMD GPUs
- Also includes PTX for DCI & SRX for security

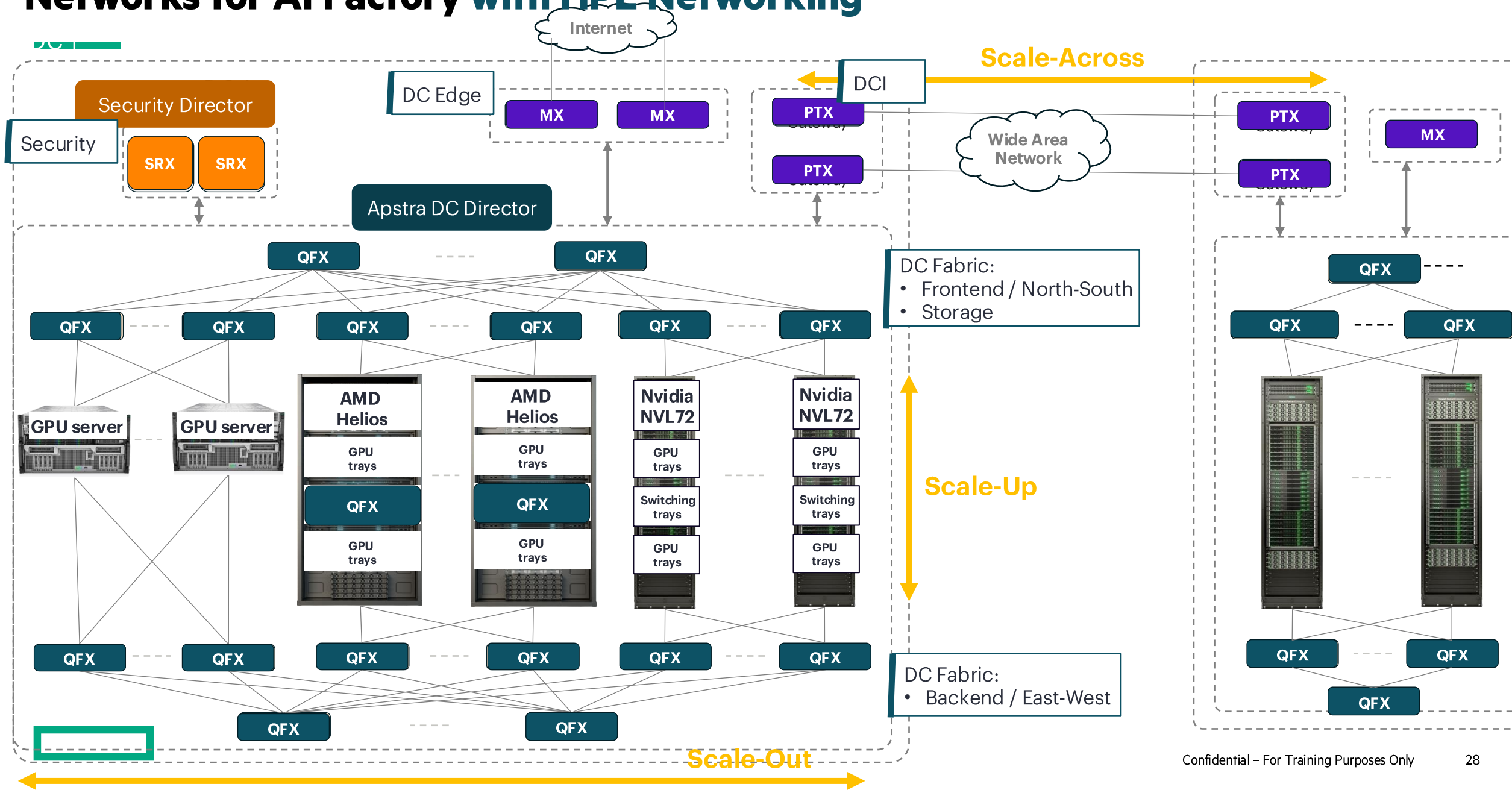
Juniper delivers comprehensive, scalable and easy-to-manage network solutions

HPE Networking AI Data Center Networking portfolio



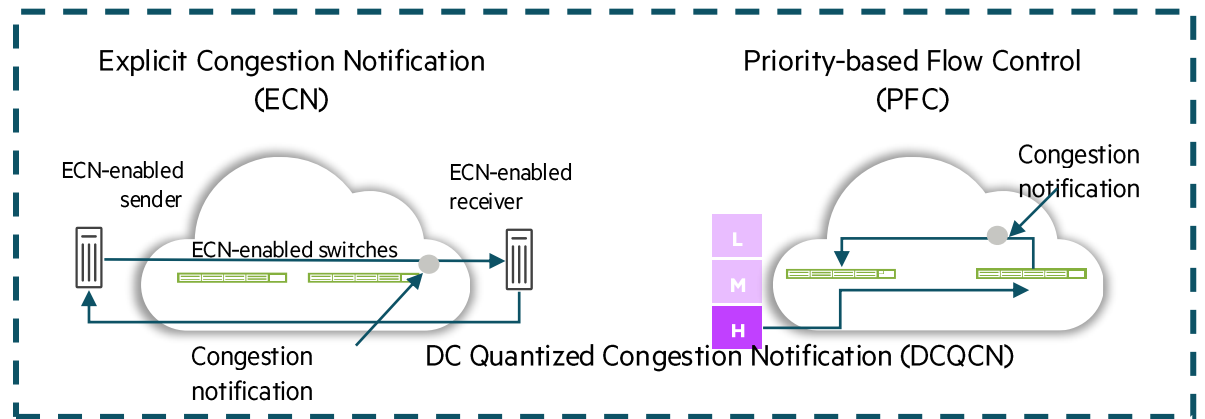
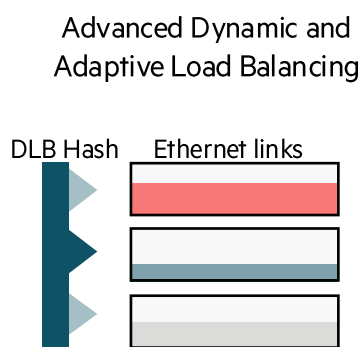
Comprehensive networking + security portfolio for AI data centers

Networks for AI Factory with HPE Networking



HPE Networking AI DC Ethernet: InfiniBand performance w/o the high cost

Ethernet ecosystem will continue to push down costs and drive further innovation



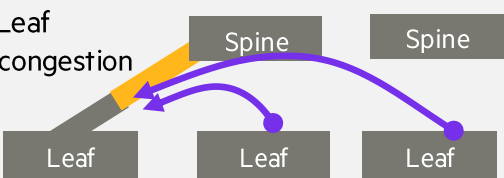
Juniper Lab Benchmark tests show 'AI Optimized Ethernet' performance comparable to IB.

HPE Networking RoCEv2: 33% lower TCO vs. NVIDIA InfiniBand.

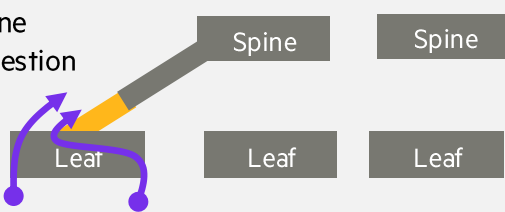
Benefits of Ethernet AI load balancing

The problem

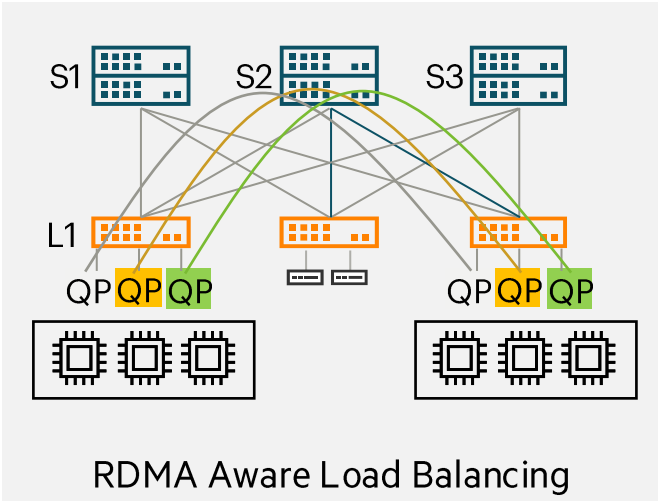
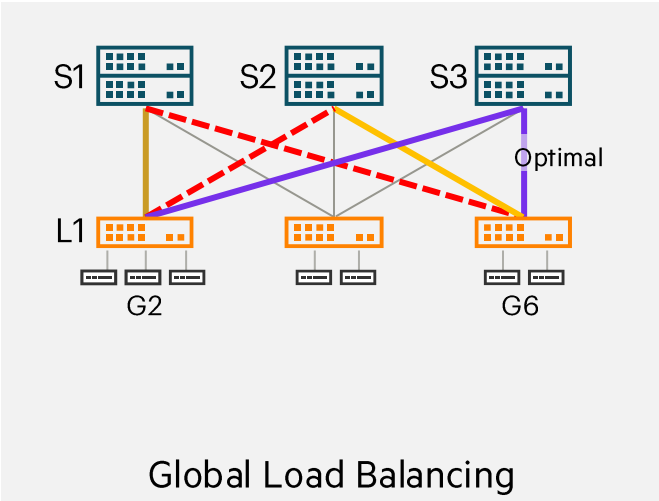
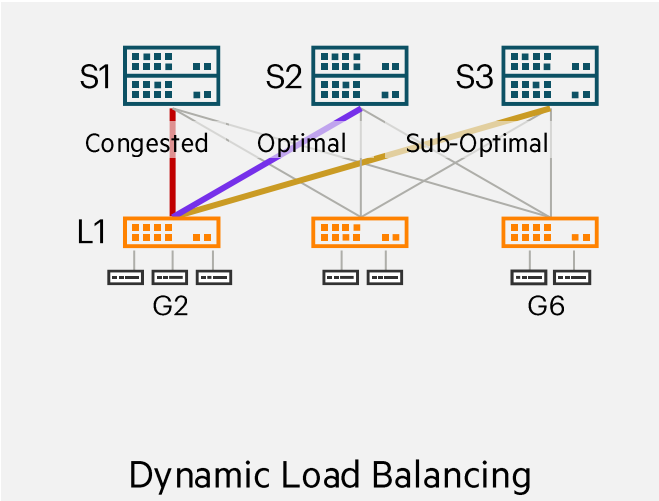
Spine-to-Leaf
downlink congestion



Leaf-to-Spine
uplink congestion



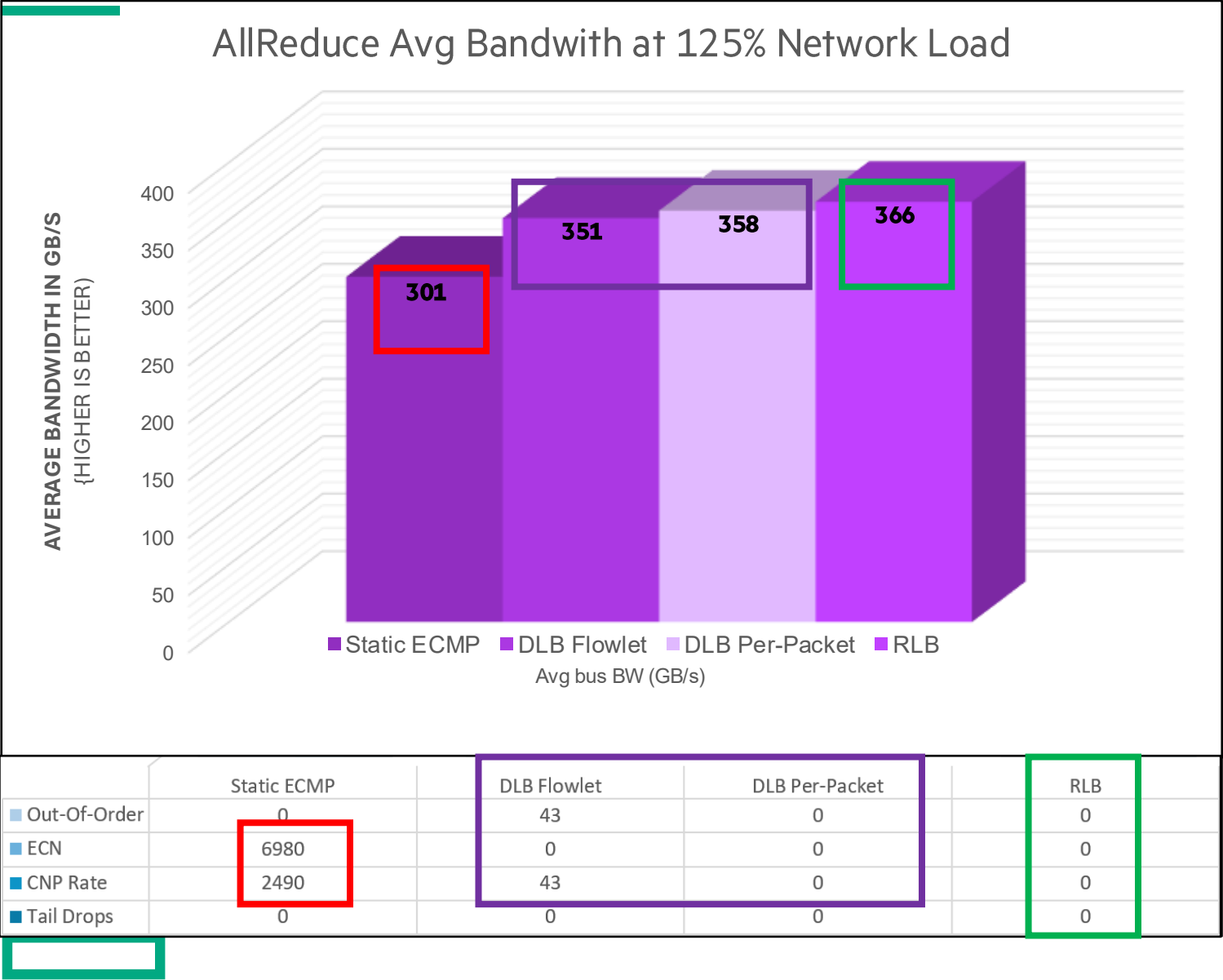
There are fewer large-sized elephants flows in AI training networks making the problem worse



50% improvement over standard ethernet
Performance comparable to any proprietary technology at lower TCO



Load Balancing performance results



ECMP sees the highest network congestion (expected)

ECN(Explicit Congestion Notification), switch to server. Congestion detection.

CNP (Congestion Notification Packet), server to server. Smart NIC Slow down data flow.

DLB-Per-Packet shows no congestion in network. Will require proprietary NIC support (till UEC standardization)

DLB-Flowlet NOT markedly lower than DLB-Per-Packet. Plus, job performance (JCT) may be identical if performance is NOT network-bound

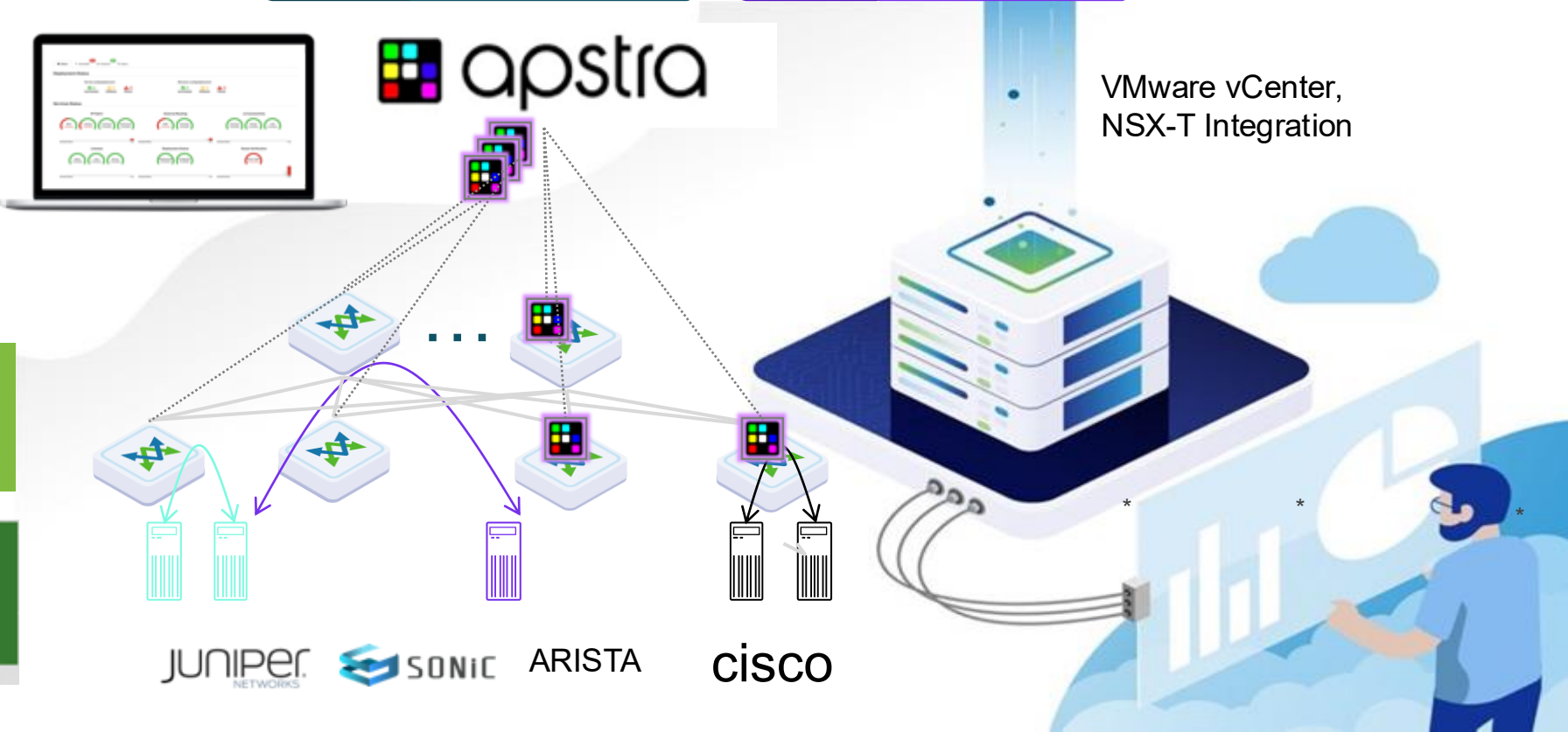
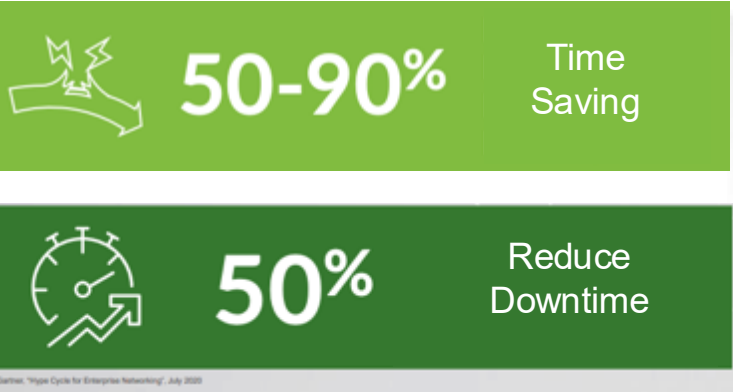
RLB delivers zero congestion, the best performance without the use of proprietary technologies

Juniper Apstra - Intent-based Networking 意圖式網路管理

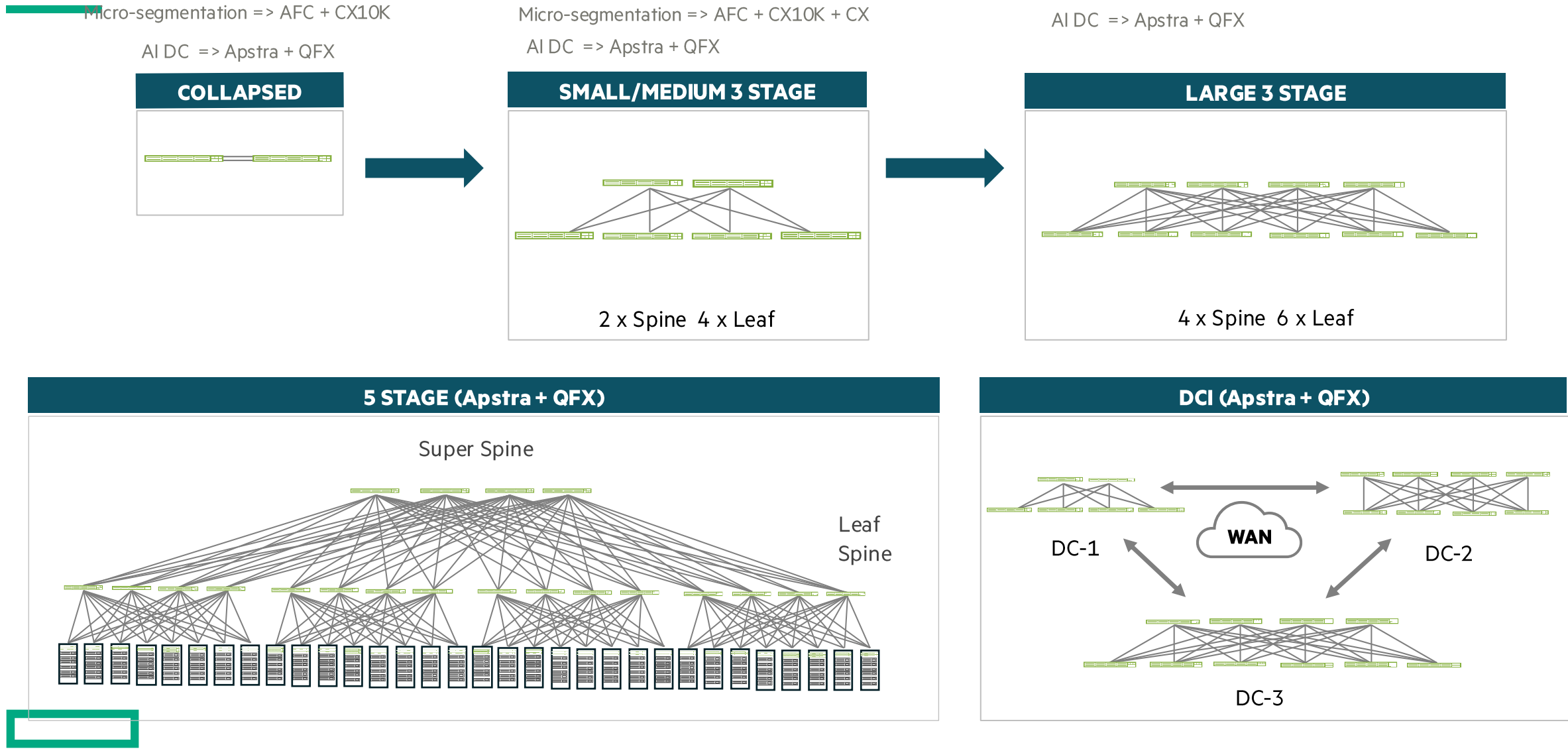
From the required network design to automated construction, monitoring, and diagnostics



Reference Architectures
IP Fabric
EVPN/VXLAN
AI DC

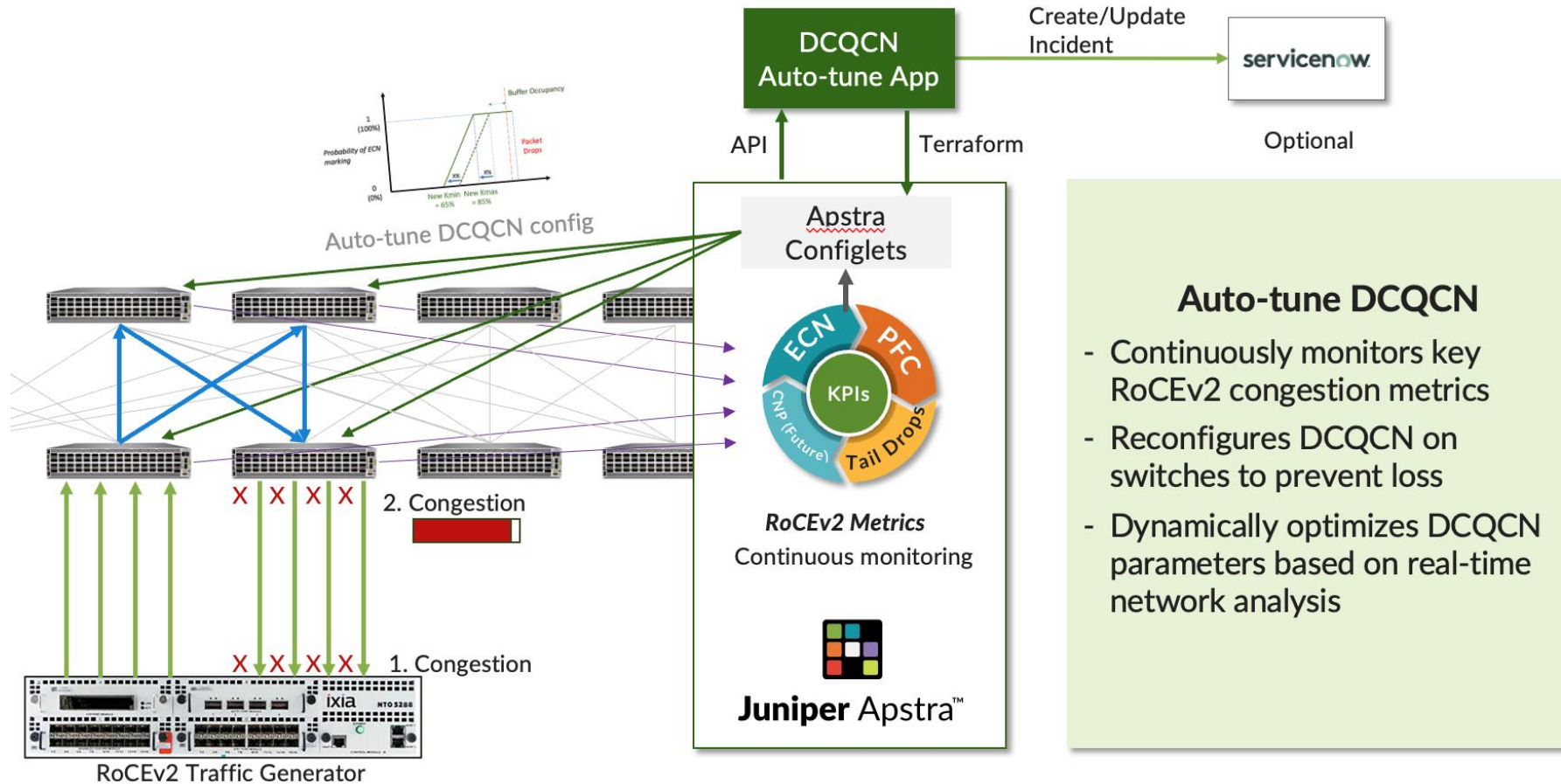


Enterprise DC use cases and solution mapping



Apstra AI DC 自動微調設定減低網路擁塞

Auto-tune DCQCN to maximize GPU performance



Goal – Achieve max network throughput without losing packets

Reducing tuning time from weeks to minutes!!

Deployment Status

Anomalies

Root Causes

Build Errors

Build Warnings

Uncommitted Changes

Create Blueprint

...

Q

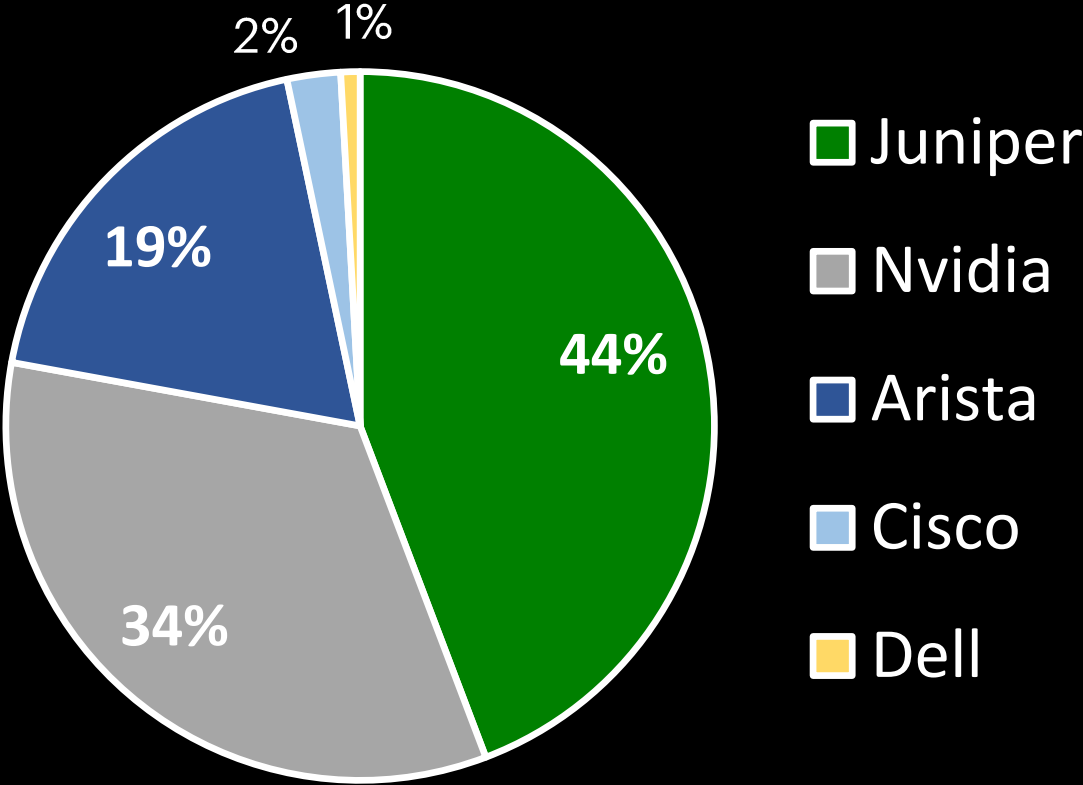
Chuntai_Hu_03010_Jun_SE_Demo_600_24706d3acd09 - evpn-ve
x-virtual
Datacenter

Physical Structure:	1 pod, 2 racks 2 spines, 3 leaves, 5 generic systems
Virtual Structure:	3 routing zones, 27 virtual networks
Analytics	
Service Deployment Status	5
Service Anomalies	0
Probe Anomalies	0
Root Causes:	0

Version 161
Total lines of config 5141
Last modified a day ago

HPE Networking is leading the way in the AI data center

800GbE OEM Market Share, 2024



Source: 650 Group, March 2024

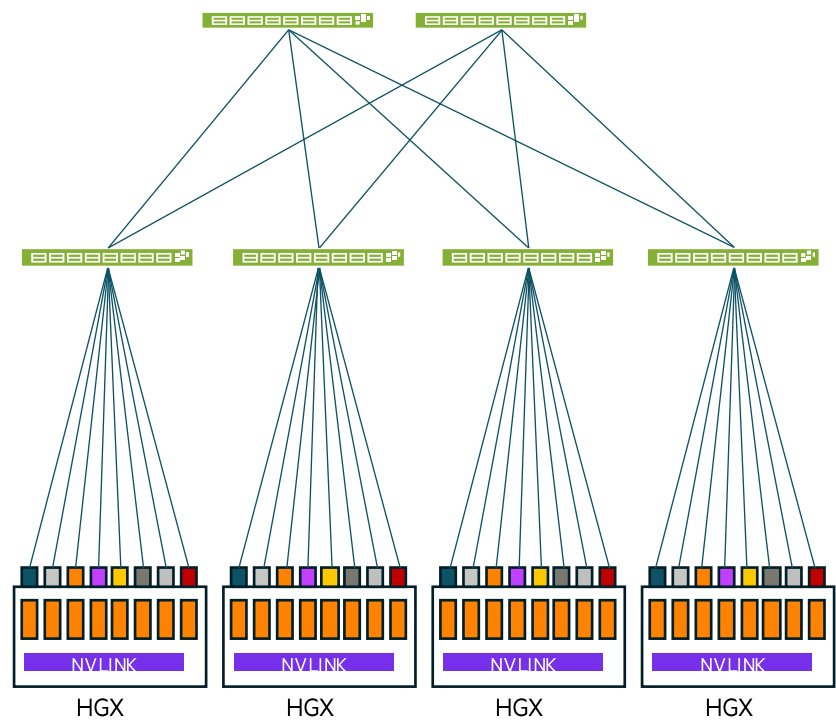


AI cluster (Back End Network) 的設計建議

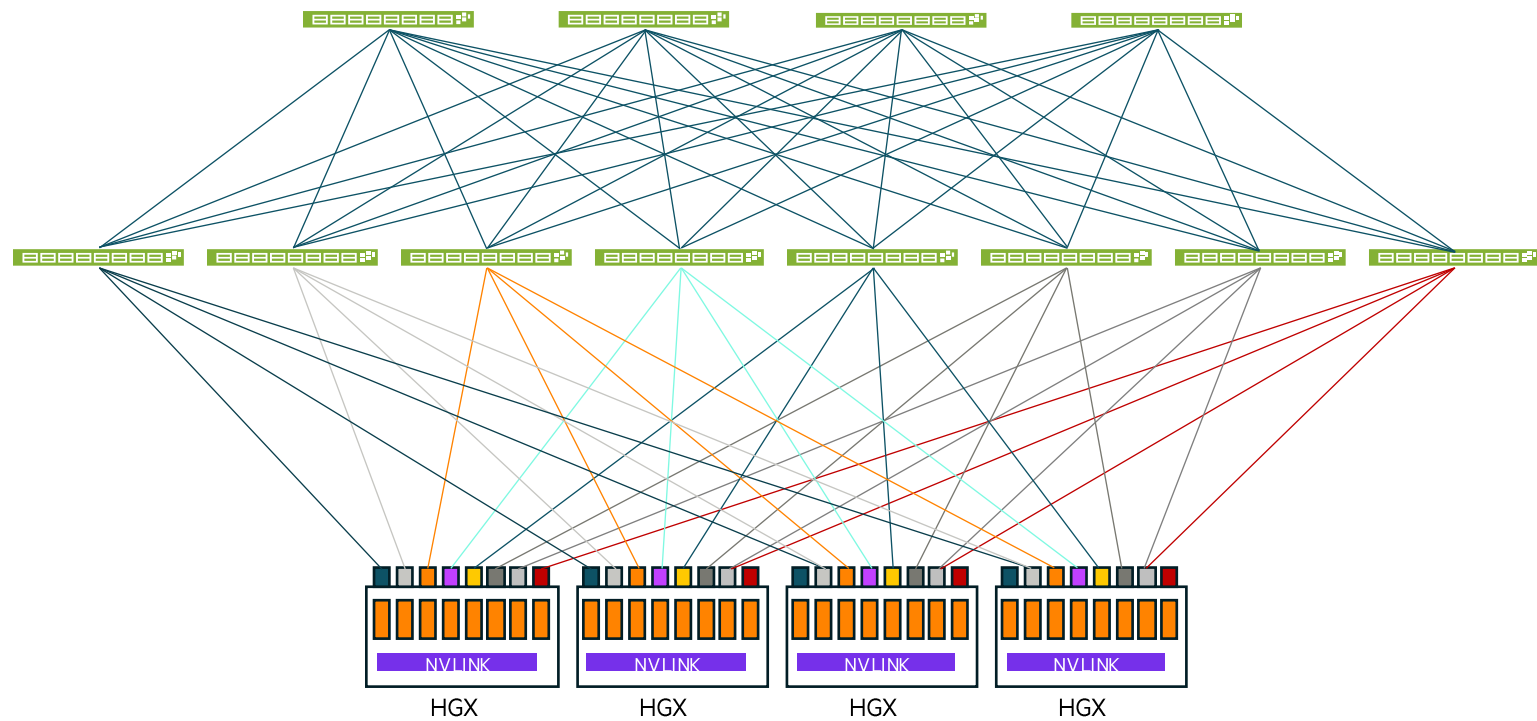


AI/ML Data center clos topologies

Standard Fat Tree



Rail Optimized FAT Tree Topology

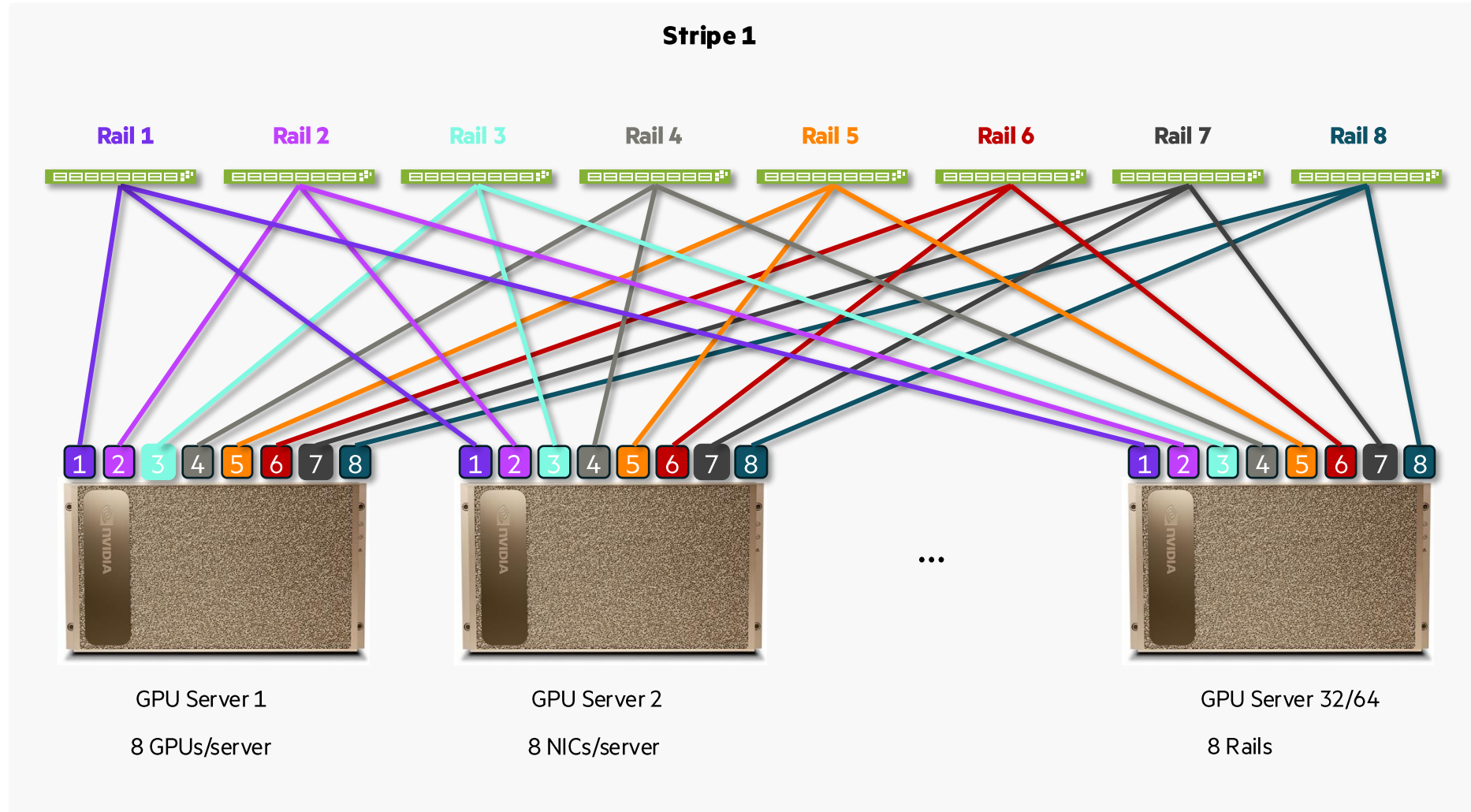


GPUS

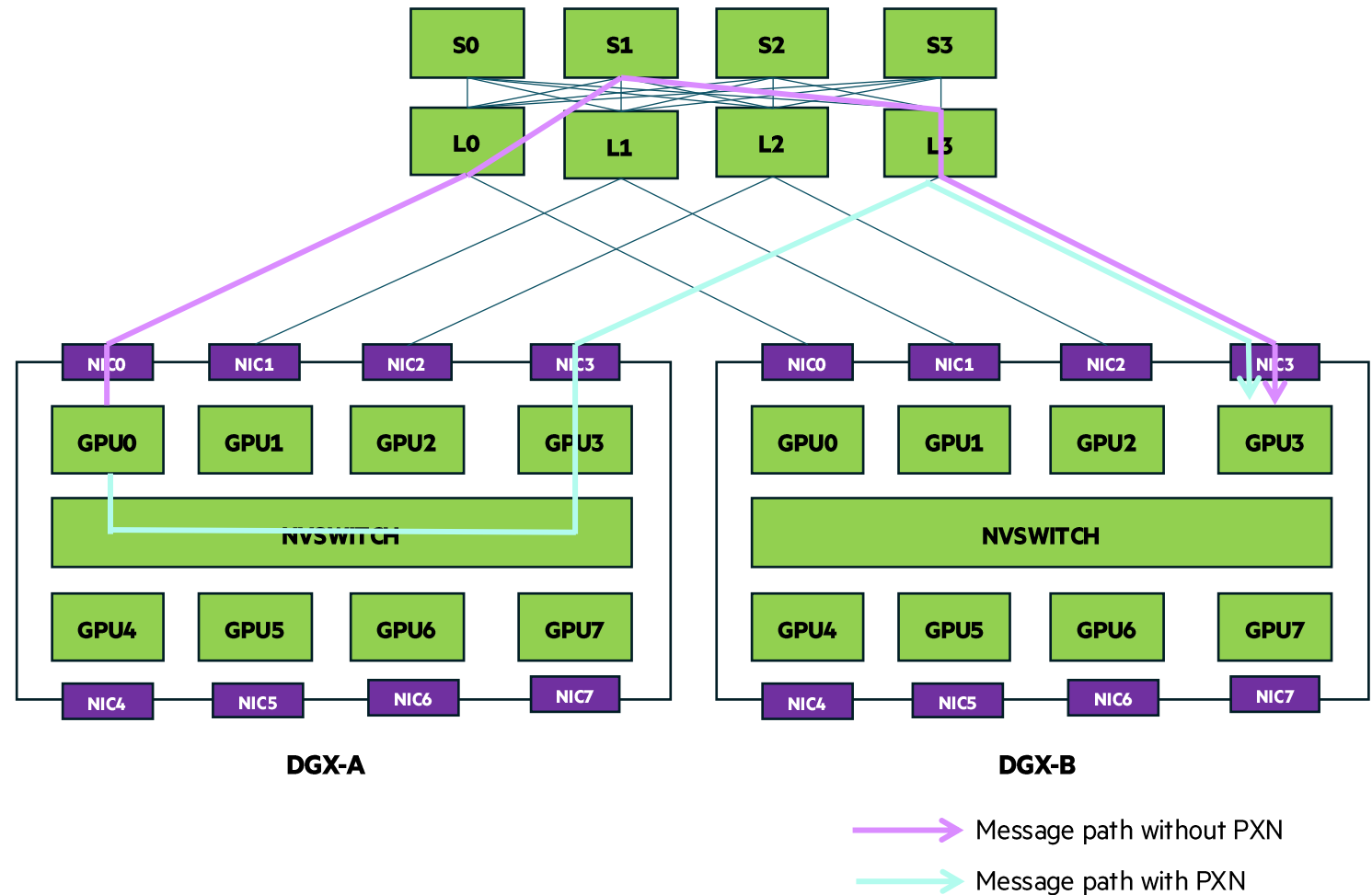


GPU Fabric Rail-Optimized Design (Nvidia / AMD)

- GPU Servers have
 - 8 GPUs
 - 8 NICs
- 8 Leaf switches create 8 Rails with GPUs 1 hop away
- **Rail:** NIC n of each server connects to Leaf n
- A group of 8 Leafs with 8 Rails form a **Stripe**
- NCCL manages traffic among rails to provide 8 independent high BW channels avoiding collision at leaf.



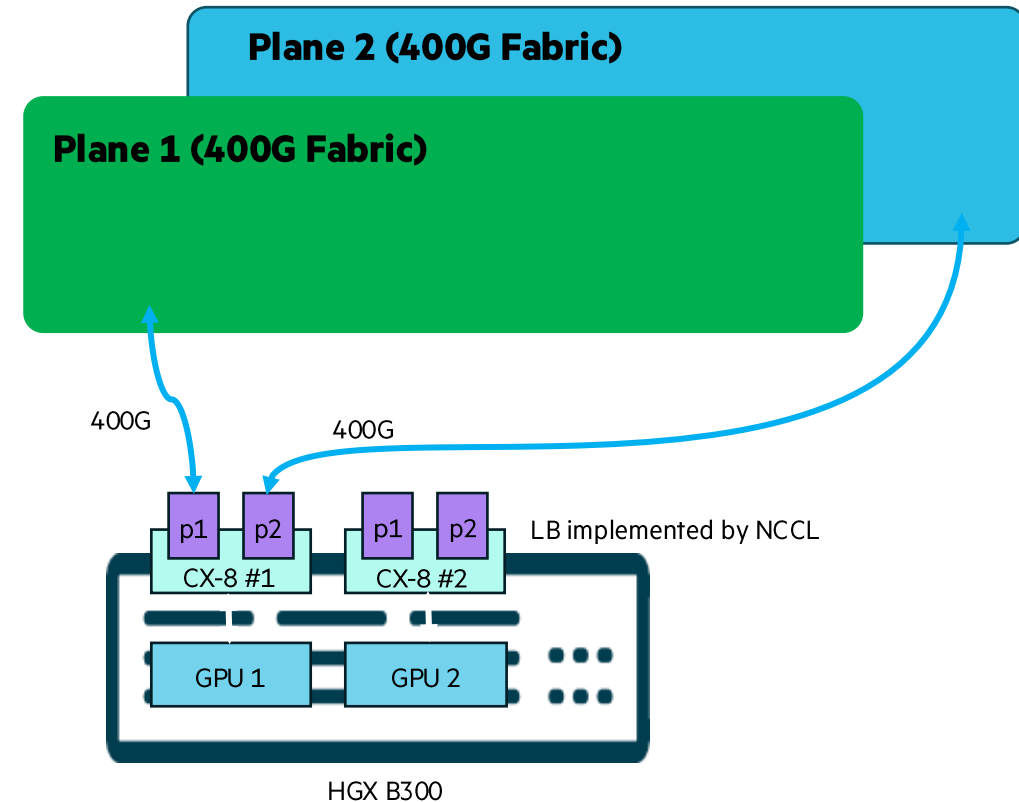
Rail-optimized Topologies – PXN Nvidia Hx00/Bx00



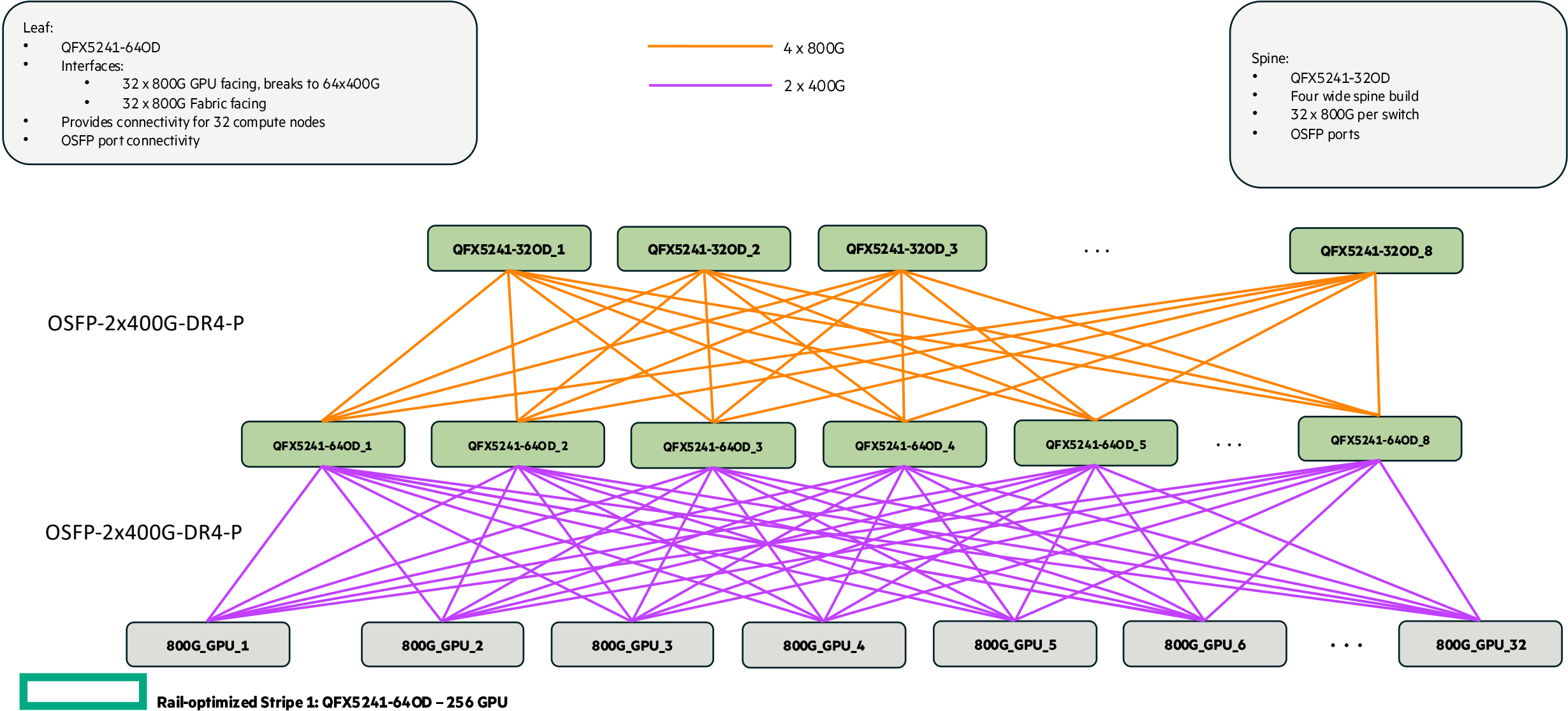
- PXN leverages NVIDIA NVLink Switch connectivity between GPUs with the node to first move data on a GPU to the same rail as the destination and then sends it to the destination without crossing rails.
- Before PXN the message in the diagram would have traversed through three hops of network switches (L0, S1, and L3)
- After PXN the message in the diagram will only travers one network switch (L3)
- With a QFX5230 leaf in a rail specific topology you can have 256 H100 GPU that are within a single network hop of each other.

We'll see dual-plane topologies in 2026

- NVIDIA Cloud Partner (NCP) Reference Architecture (RA)
- Starting from B300/GB300, Nvidia NCP RA recommends dual-plane topologies for E-W fabric to avoid single point of failure.
- Not every customer is expected to follow NCP RA – we expect some to deploy single plane.



256 GPU Cluster, 800G Connectivity, GPU Fabric (DR4) QFX5241-64OD leaf/ QFX5241-64OD spine, 1:1



256 GPU Cluster, 800G Connectivity, GPU Fabric (VR4)

QFX5241-64OD leaf/ QFX5241-64OD spine, 1:1

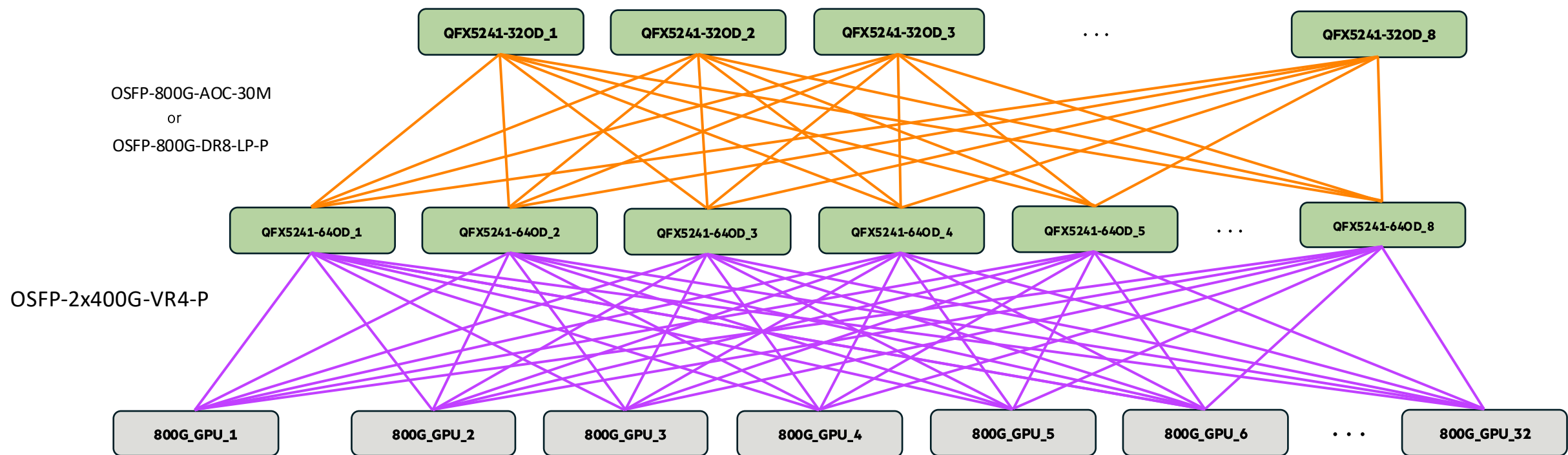
Leaf:

- QFX5241-64OD
- Interfaces:
 - 32 x 800G GPU facing, breaks to 64x400G
 - 32 x 800G Fabric facing
- Provides connectivity for 32 compute nodes
- OSFP port connectivity

— 4 x 800G
— 2 x 400G

Spine:

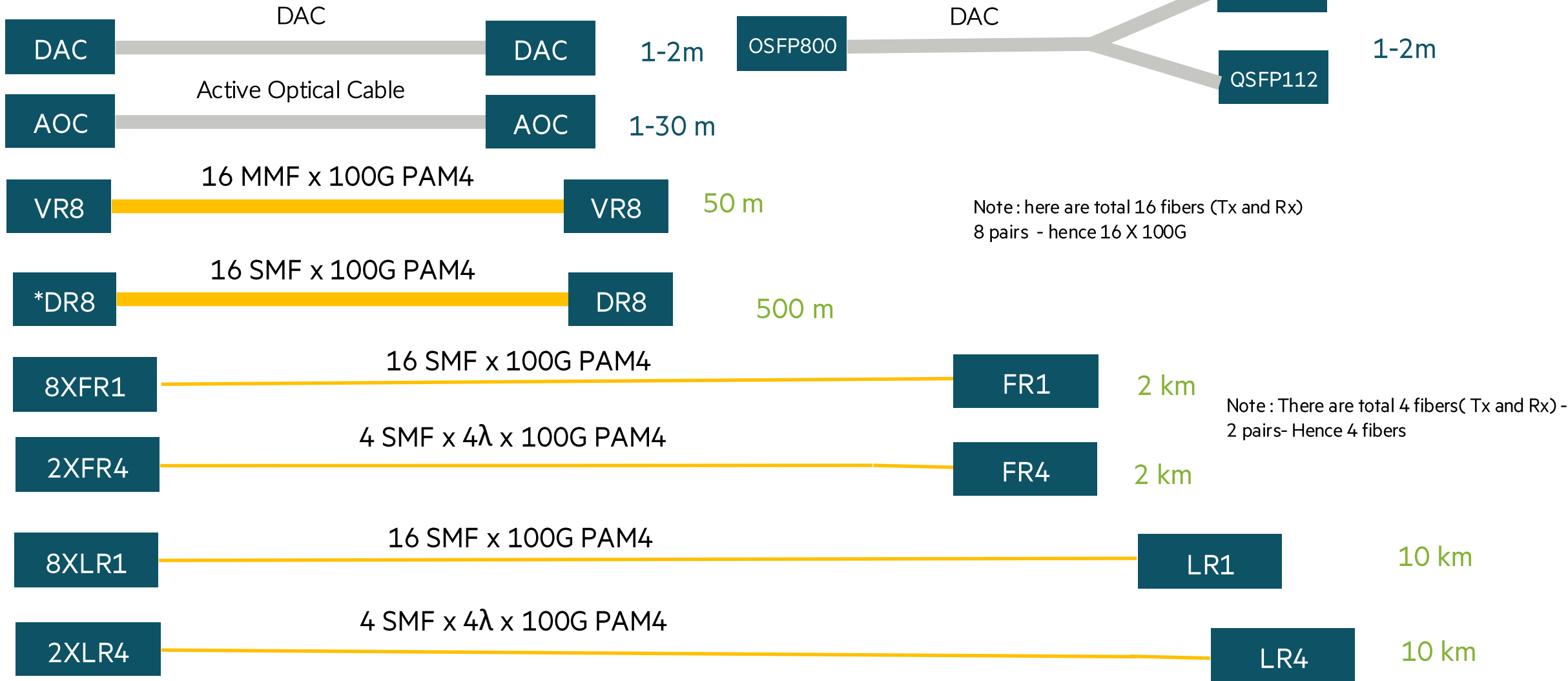
- QFX5241-32OD
- Four wide spine build
- 32 x 800G per switch
- OSFP ports



 Rail-optimized Stripe 1: QFX5241-64OD – 256 GPU

800G Optics & Cables

OSFP800 & QSFPDD-800 support for AI/ML clusters



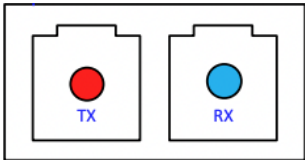
*offered in both Dual MPO-12 and MPO16 connector variants



Break-out options for 800G ports

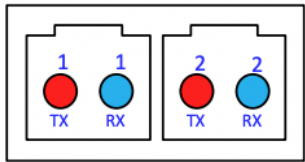
connector designs for break-out optics

Duplex LC



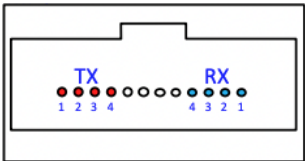
Single-lane duplex optics
e.g. 100G/400G FR4 and LR4

Duplex CS



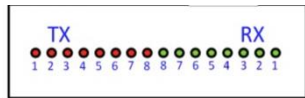
2-way break-out optics
e.g. 2x100G LR4/CWDM4 and 2X400G-FR4/LR4

MPO-12



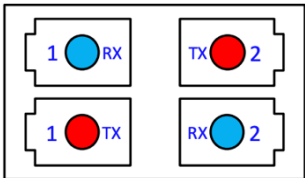
4-way break-out cable
e.g. 40G/100G SR4, 100G PSM4, 400G DR4/4x100G-FR/LR

MPO-16



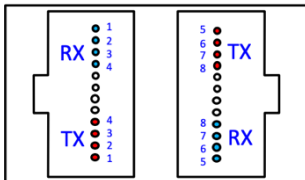
8-way break-out optics
Used for 400G SR8 and 800G-DR8/8x100G-FR/LR

Dual duplex LC



2-way break-out optics
Favored approach for 2 x 400G break-out optics

Dual MPO-12



8-way break-out optics
Favored approach for 8 x 100G break-out optics

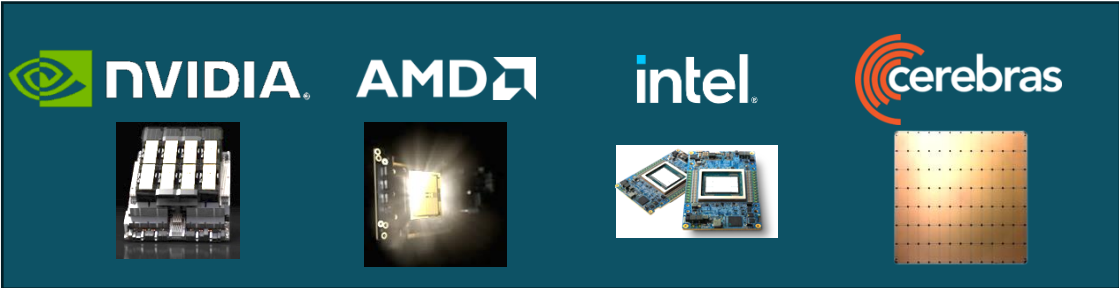
Connector options evolving for 800G optics as break-out becomes more widely deployed

後端網路 (GPU Cluster)一定要用InfiniBand嗎?

Ubiquitous Ecosystem

(Management & Operations Tools)

GPU Options

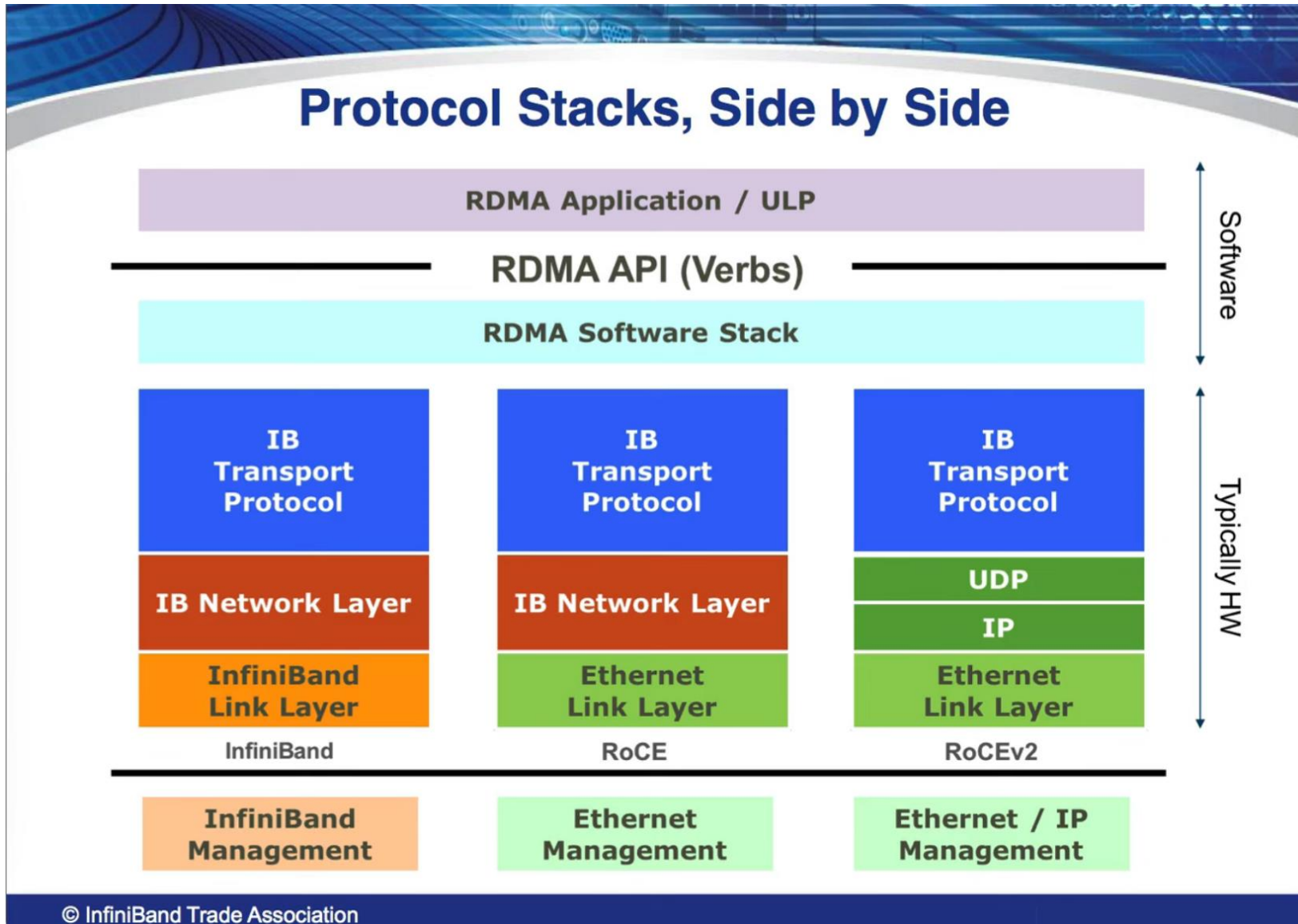


Ethernet	✓	✓	✓	✓	✓
Ultra Ethernet	✓	✓	✓	✓	✓
InfiniBand	✗	✓	✗	✗	✗

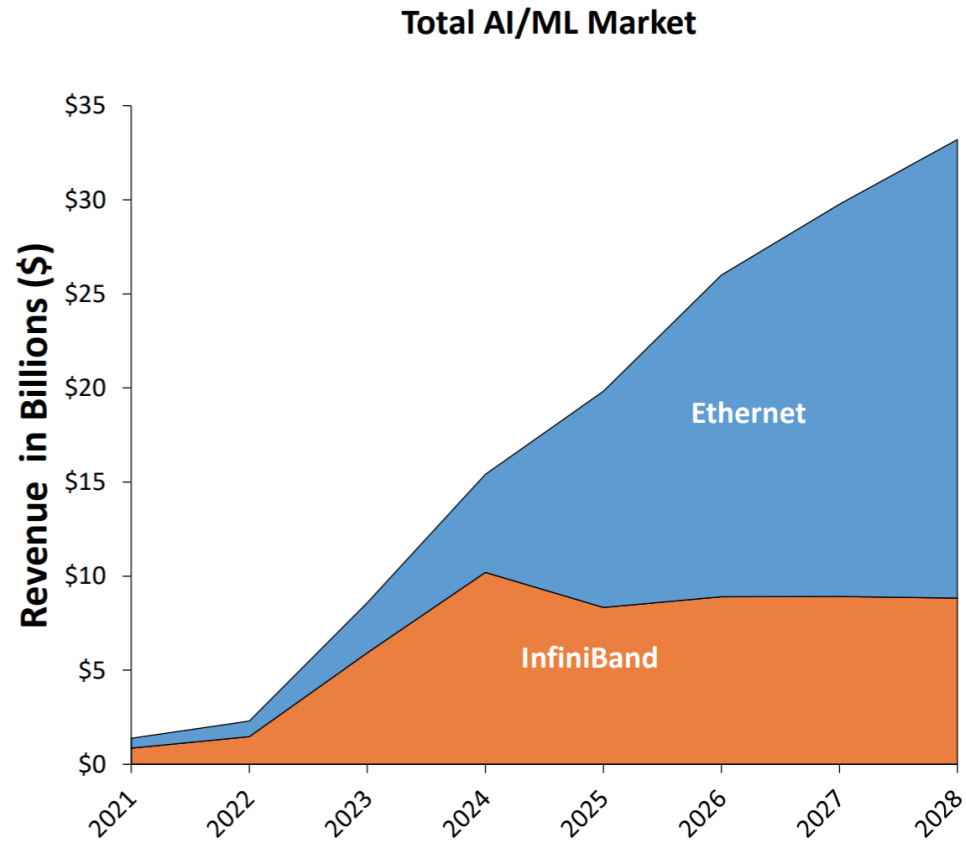
Ethernet: 55% lower TCO vs. NVIDIA InfiniBand

Source: ACG Research

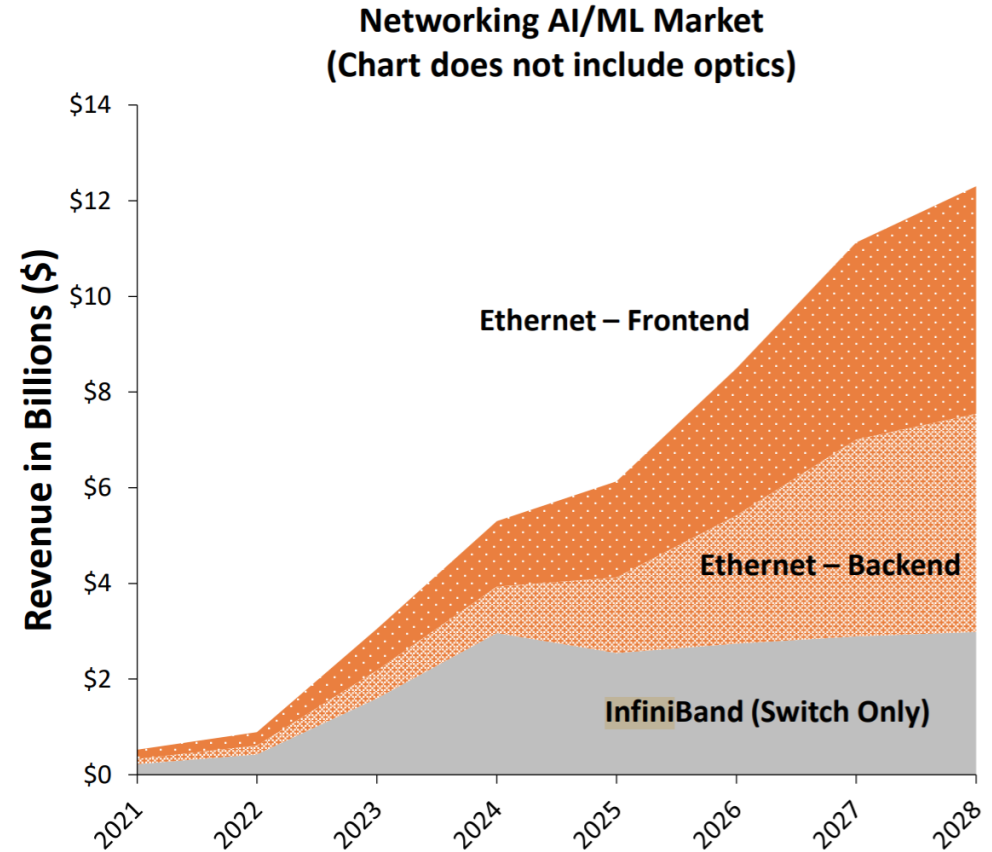
What is RoCEv2 (RDMA over Converged Ethernet v2)



Ethernet is the right choice



Ref: Cloud CAPEX, Data Center, and Ethernet Switch 1Q24 Results and June Forecast, 650 Group, Alan Weckel, 26 June 2024

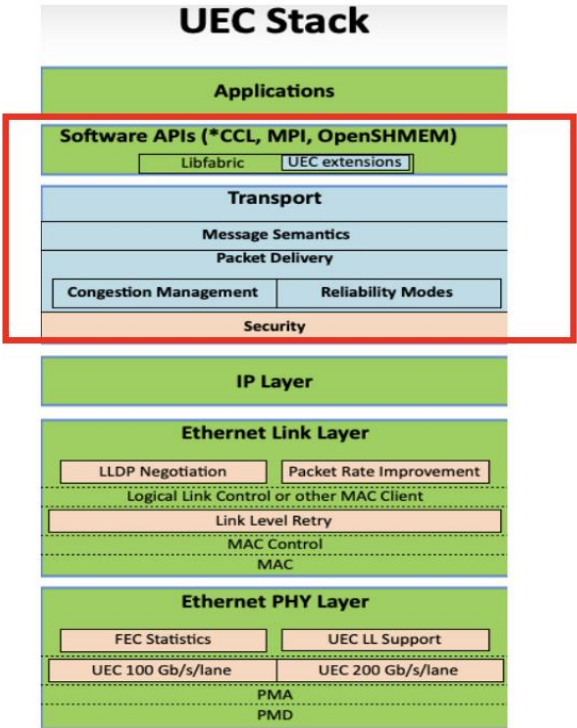


Ref: <https://www.fierce-network.com/cloud/data-center-ethernet-nvidias-next-multi-billion-dollar-business>, 11 Sept 2024

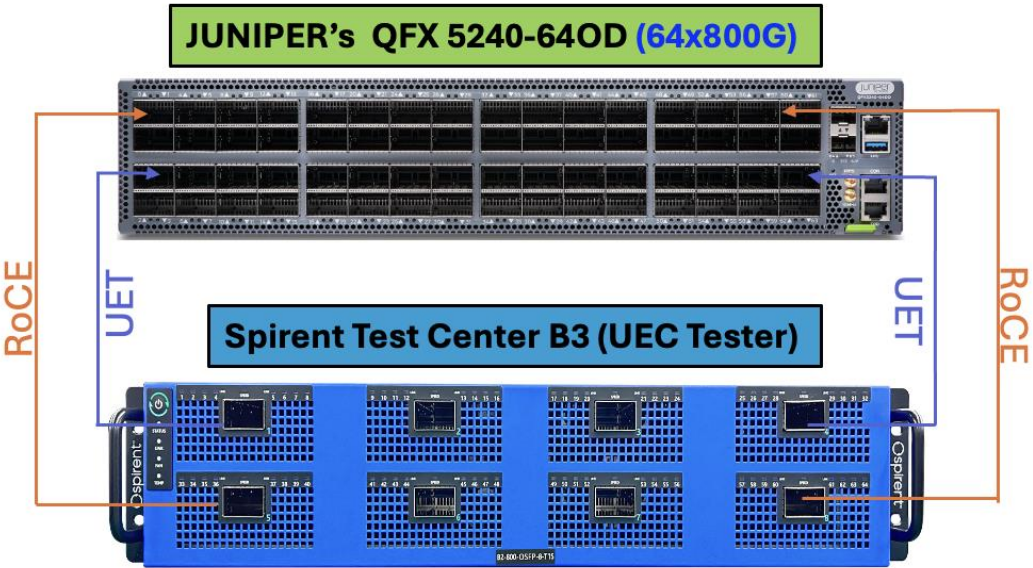


Tokyo Interop 2025 – AI Data Center

World's first Ultra Ethernet Transport (UET) Test Solution



Ultra ~~Ethernet~~

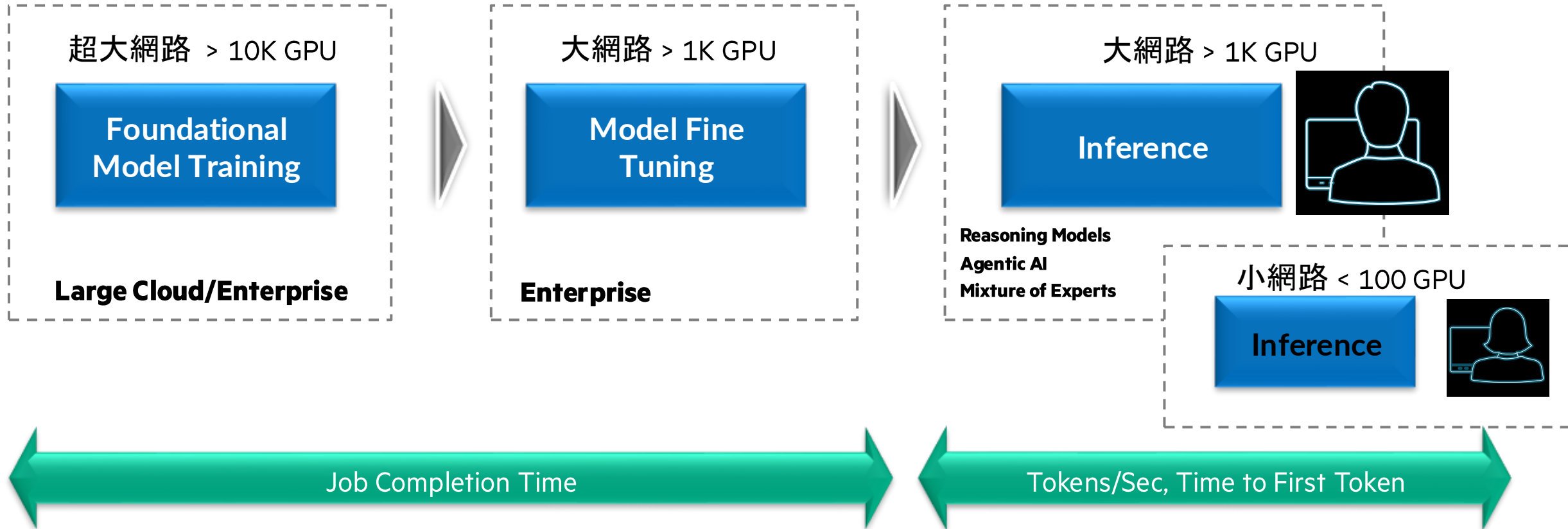


總結與行動



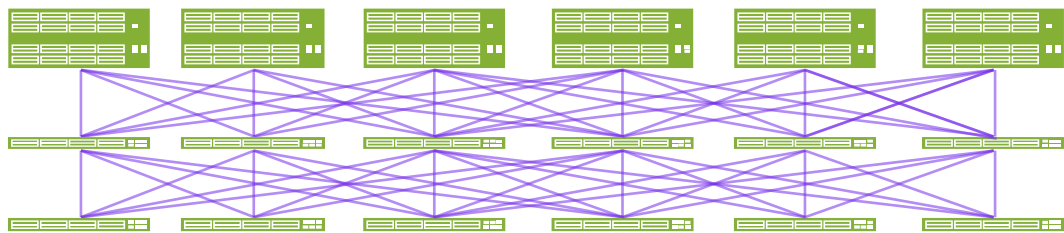
Networking: The Backbone of the AI Lifecycle

The network plays a critical role for optimizing key AI metrics



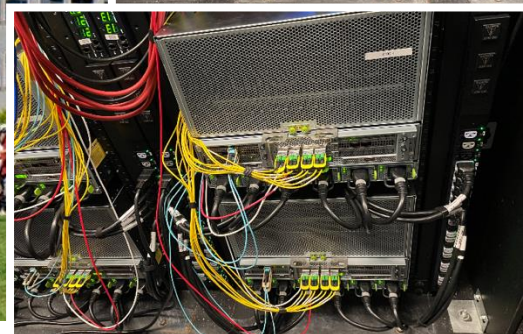
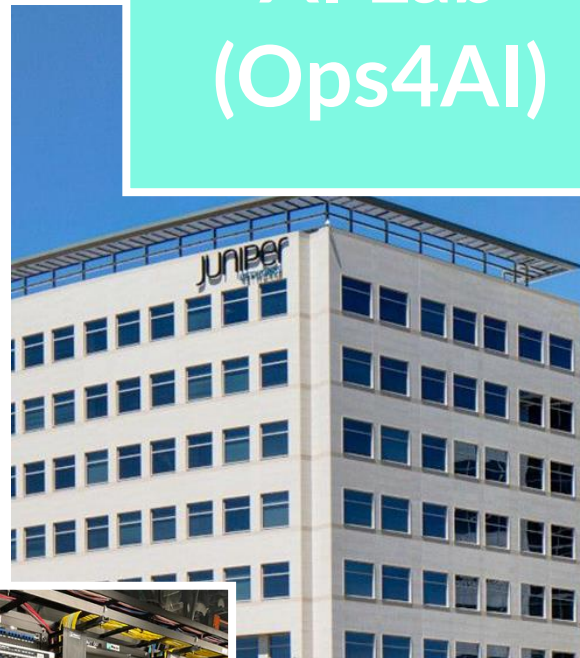
AI Performance Begins and Scales with the Network

AI Lab 驗證的端到端AI DC解決方案

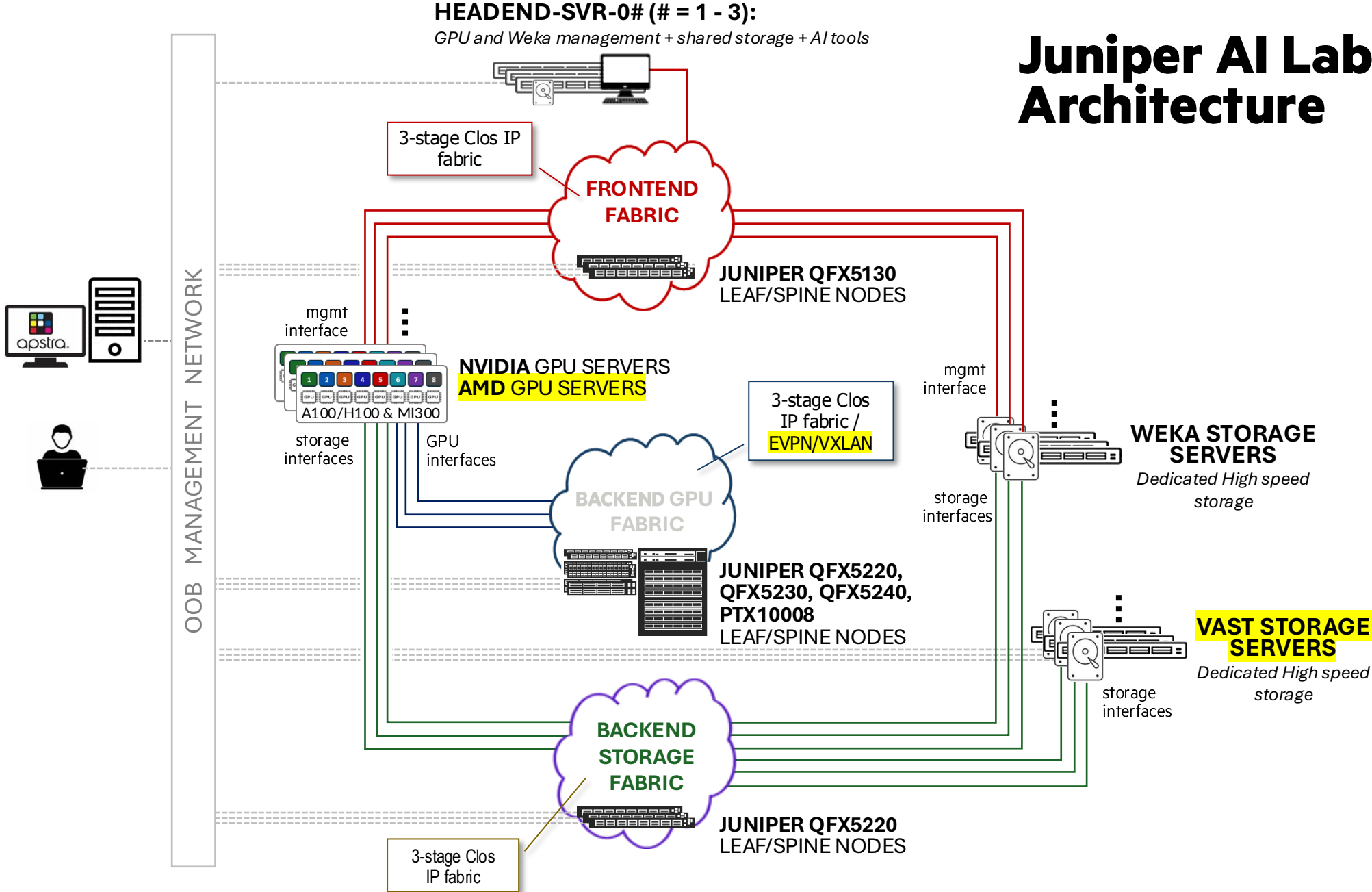


- 全面的端到端 JVD 範本，用於部署 AI 網路解決方案
- 經過嚴格的預先測試和驗證
- 降低複雜性和風險
- 適合大中小型 AI DC 的各種規格

Juniper
AI Lab
(Ops4AI)

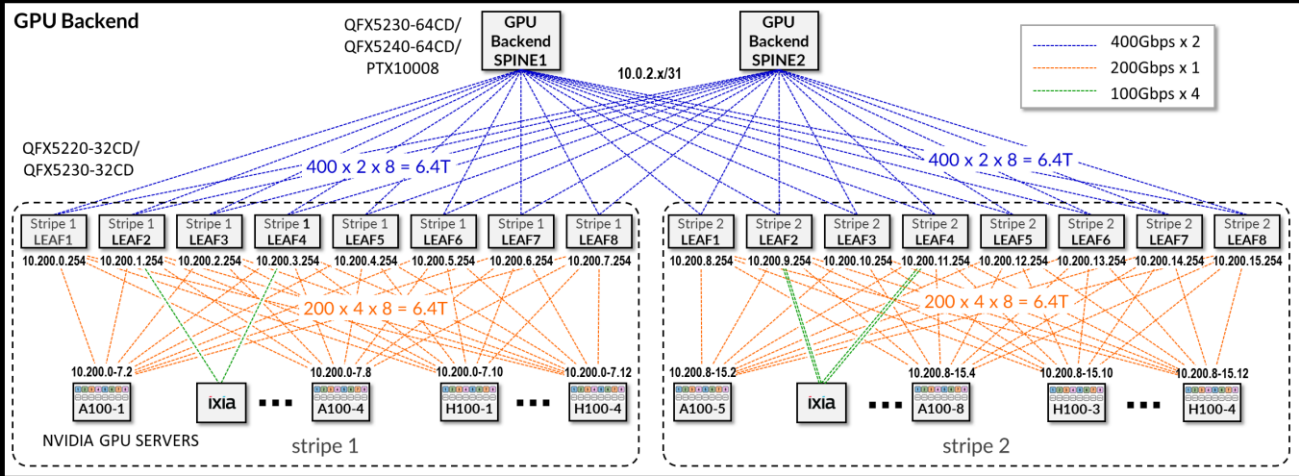


Juniper AI Lab Architecture



AI-optimized Ethernet: Put to test

We achieved similar results testing with Ethernet as compared with InfiniBand in MLPerf benchmarks



- Bert-Large, DLRM, Llama2, and GPT-3 training models (MLCommons)
- Customer models

Juniper Ethernet Training Time

InfiniBand Training Time

BERT-Large
(Training)

~2.6 min
(64 A100 GPU)

2.5-3.3 min
(64 A100 GPU)

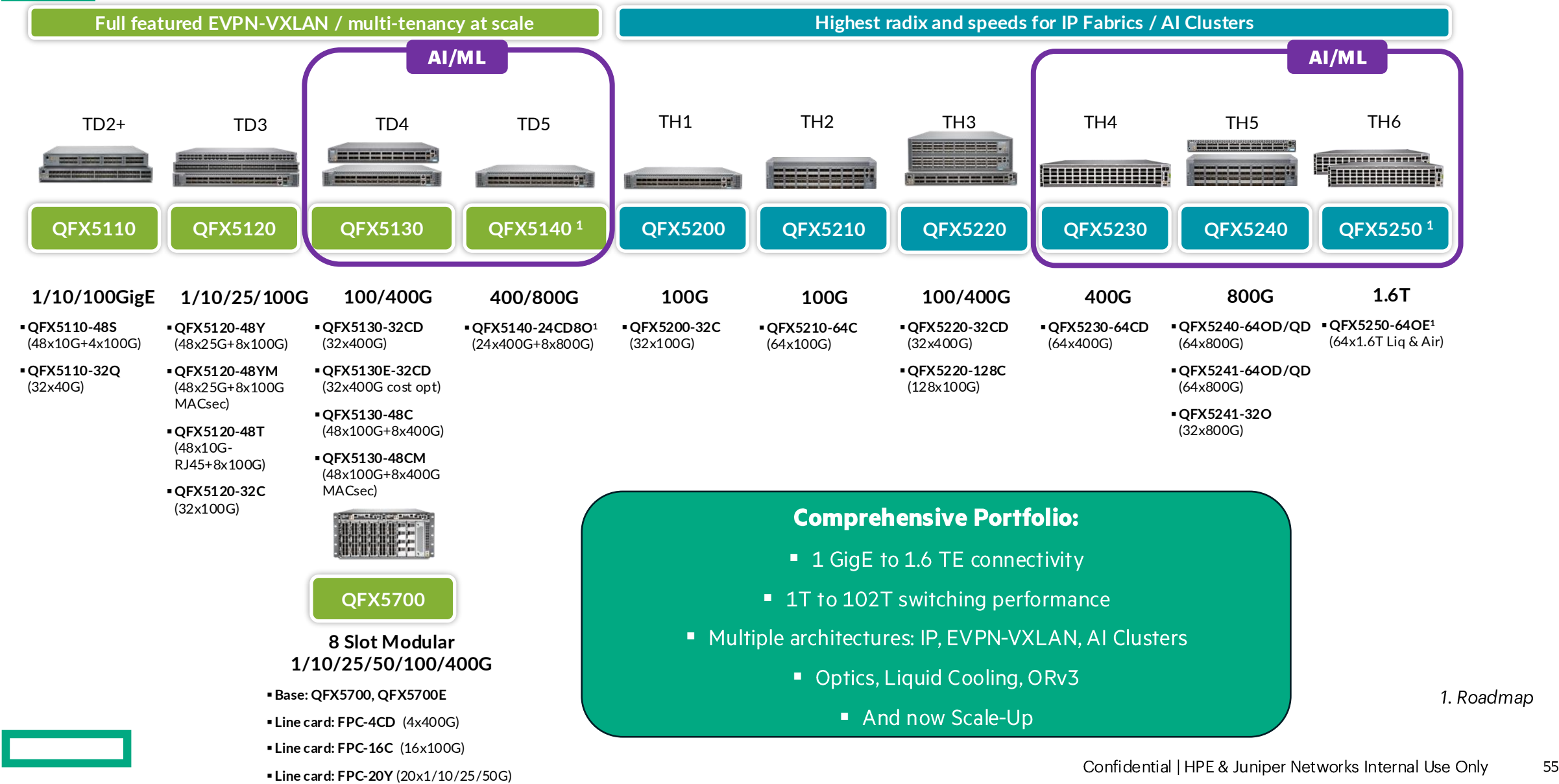
Llama2 NeMo
(Fine Tuning)

7.9 min*
(32 H100 GPU)

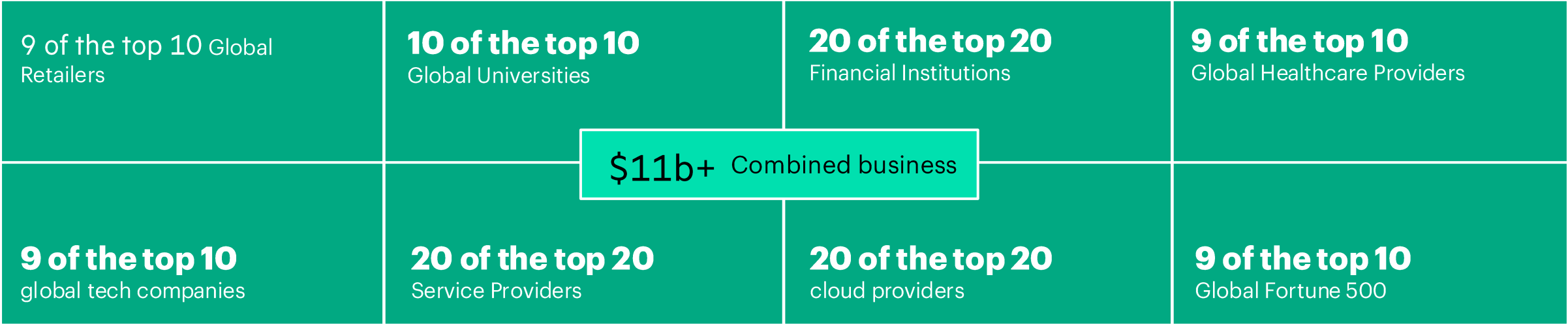
~5.1 min
(64 H100 GPU)

*Juniper Ethernet is using 32 H100 GPUs vs. 64 H100 GPUs for InfiniBand

HPE Juniper Datacenter Portfolio



Unlock networking with HPE



Thank You

