AIDC地端化與智慧校園

以RoCE v2 實現高速資料交換

Prado Yang

Oct. 2025

Climate of AI DC

AI 資料中心的網路需求

AI 大語言模型啟動第四次工業革命

Hugging Face

Models
1,484,462

Chatbot Arena
https://lmarena.ai

繁中LLM 聊天機器人競技場
https://arena.twllm.com

Llama 3
From Meta

Microsoft
Phi-4

Grok

Gemini

Gemini

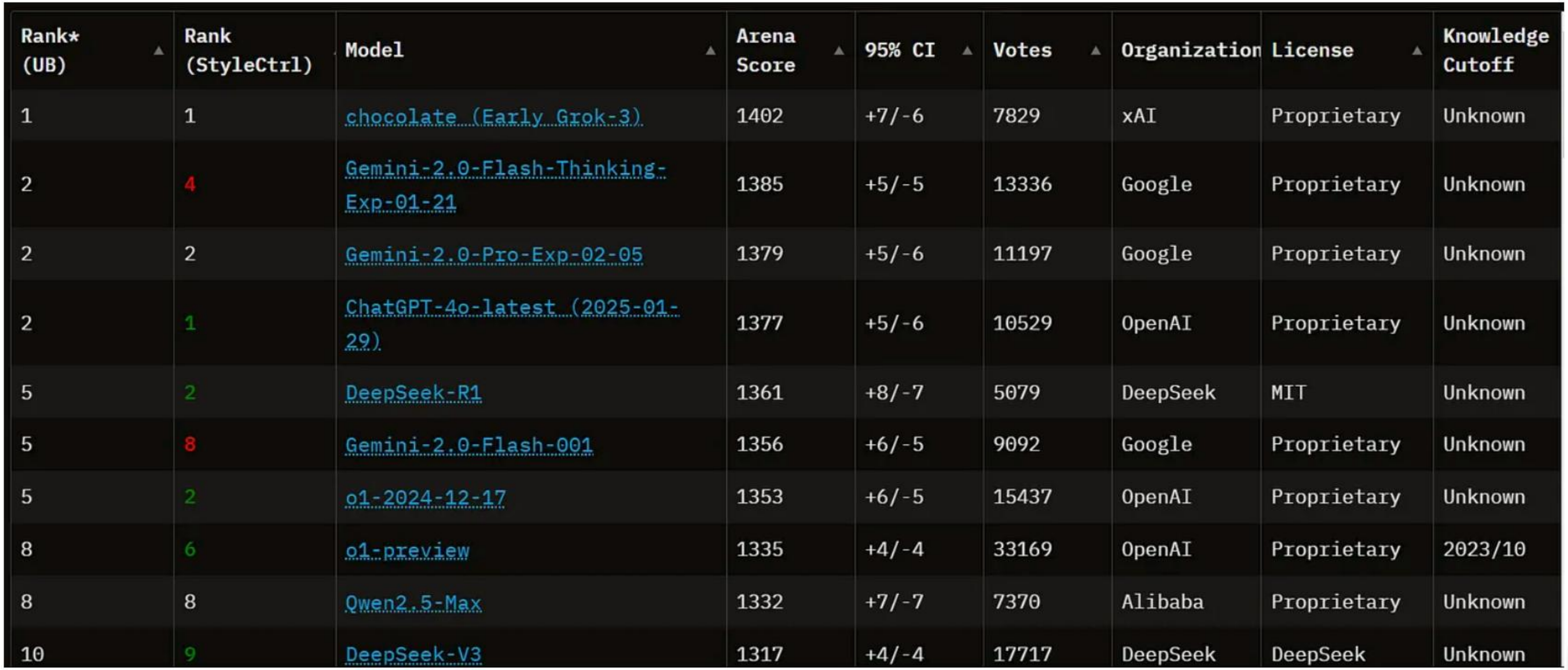
ChatGPT

deepseek

Gemini

ChatGPT

deepseek

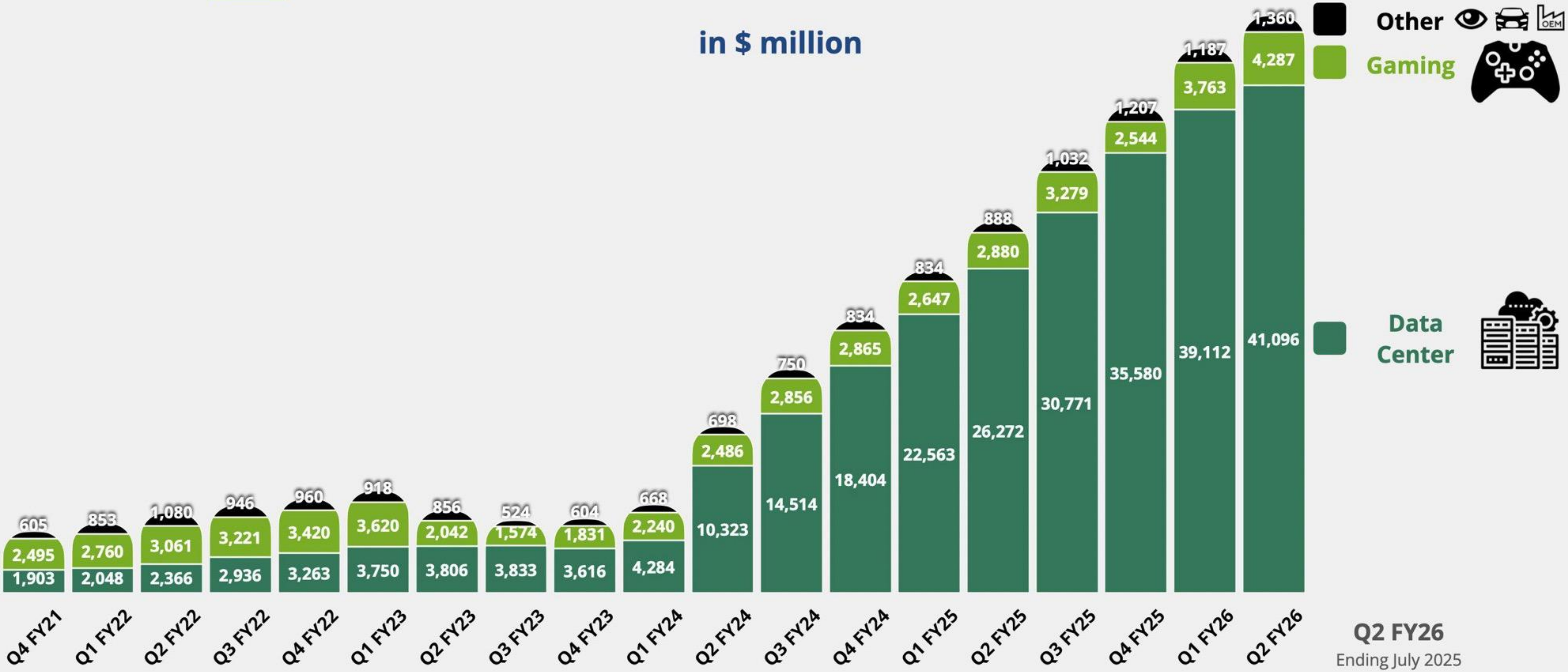


Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License	Knowledge Cutoff
1	1	chocolate (Early Grok-3)	1402	+7/-6	7829	xAI	Proprietary	Unknown
2	4	Gemini-2.0-Flash-Thinking-Exp-01-21	1385	+5/-5	13336	Google	Proprietary	Unknown
2	2	Gemini-2.0-Pro-Exp-02-05	1379	+5/-6	11197	Google	Proprietary	Unknown
2	1	ChatGPT-4o-latest (2025-01-29)	1377	+5/-6	10529	OpenAI	Proprietary	Unknown
5	2	DeepSeek-R1	1361	+8/-7	5079	DeepSeek	MIT	Unknown
5	8	Gemini-2.0-Flash-001	1356	+6/-5	9092	Google	Proprietary	Unknown
5	2	o1-2024-12-17	1353	+6/-5	15437	OpenAI	Proprietary	Unknown
8	6	o1-preview	1335	+4/-4	33169	OpenAI	Proprietary	2023/10
8	8	Qwen2.5-Max	1332	+7/-7	7370	Alibaba	Proprietary	Unknown
10	9	DeepSeek-V3	1317	+4/-4	17717	DeepSeek	DeepSeek	Unknown

DC IS THE NEW UNIT OF COMPUTING

NVIDIA Revenue Breakdown

in $ million



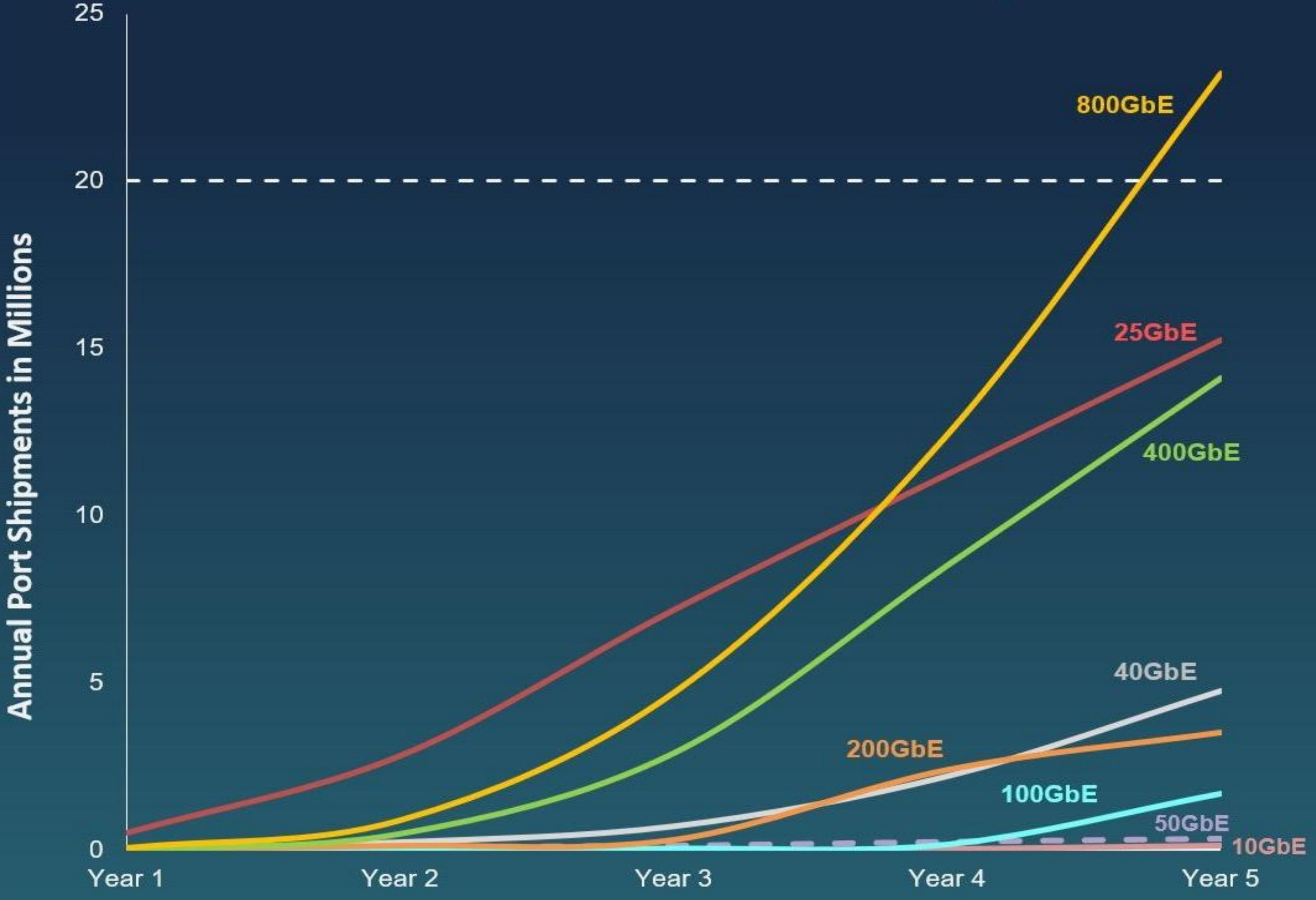
Q2 FY26
Ending July 2025

HOW THEY MAKE MONEY

appeconomyinsights.com

APP ECONOMY **INSIGHTS**

Data Center Ethernet Switch Annual Shipments



CREHAN RESEARCH Inc.

<u>Year 1 Equals:</u>

2023 for 800GbE

2016 for 25GbE

2018 for 400GbE

2011 for 40GbE

2019 for 200GbE

2012 for 100GbE

2020 for 50GbE

2001 for 10GbE

AI Data Center: Market view

InfiniBand is fading away, Industry is pivoting to Ethernet

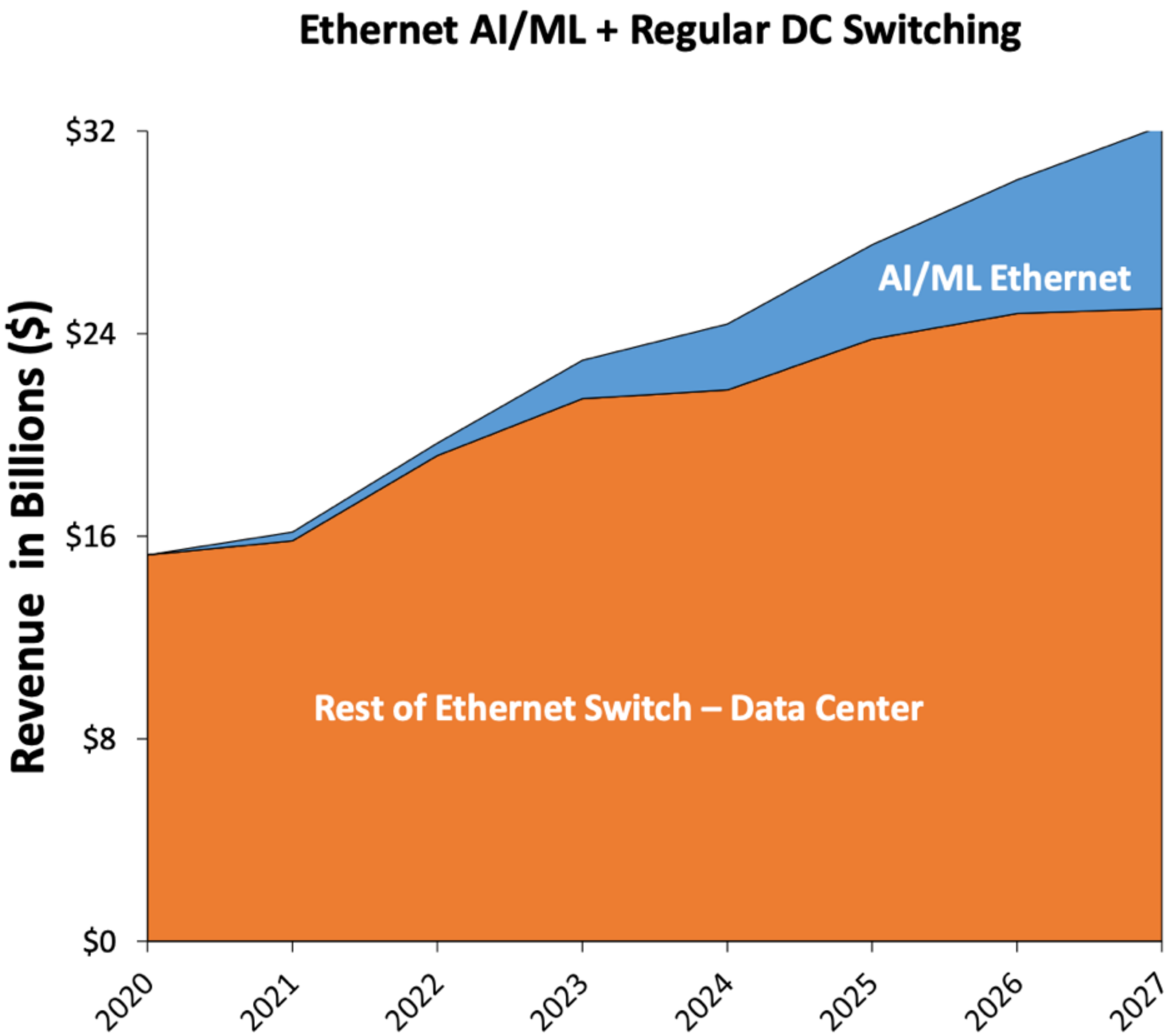
Data Center Networking - AI/HPC - Market and Forecast – 650 Group

	2024	2025	2026	2027	2028
Share Revenue (%)					
Ethernet - Frontend	23%	30%	34%	30%	34%
Ethernet - Backend	23%	41%	44%	51%	50%
InfiniBand (Switching Only)	54%	29%	22%	19%	17%

DC Network CapEx: Market Transition with AI

Investments in some enterprise DCs are being paused…

to make way for investment in AI DCs



Source: 650 Group, 2023

Avoid an Infrastructure Bottleneck

If the network is a bottleneck that delays training job completion, expensive GPU time is wasted, and training becomes network-bound instead of compute-bound

Juniper Mission for AI Data Centers

Unlocking the full potential of AI with unparalleled network performance and ease of operations

8 x GPU server

$350K +

Small AI cluster

AI frontend network
Inference

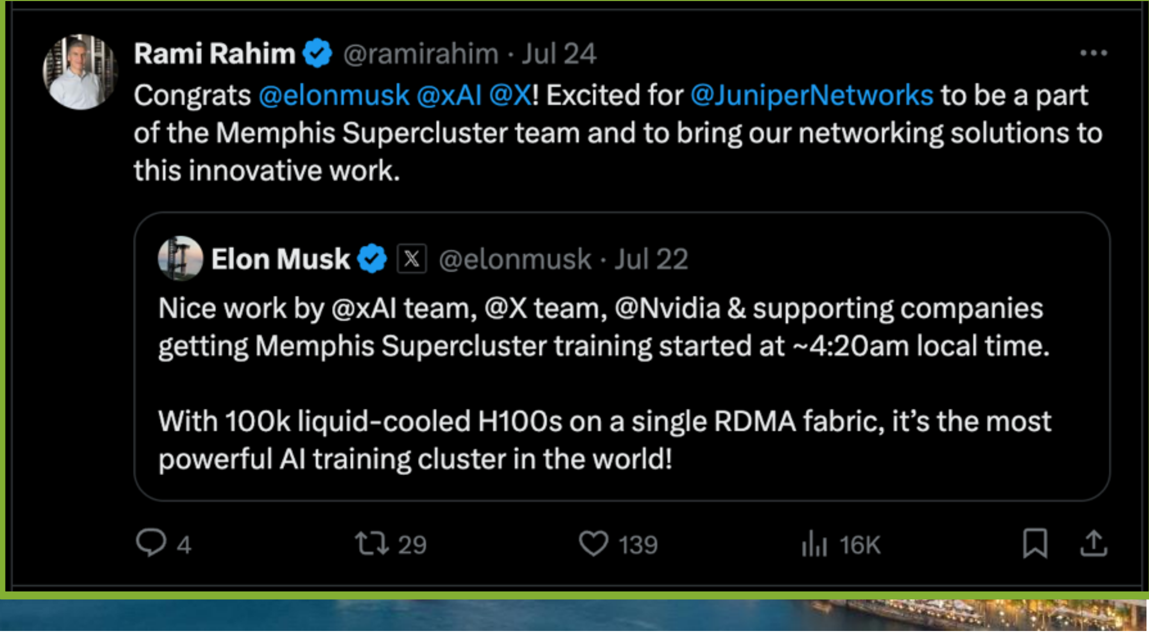
AI backend network
Training

$5M-10M

GPUs represent the bulk of the CapEx investment

The network connects GPUs in distributed training

xAI



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!

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

AI Technology on Juniper

SambaNova SYSTEMS

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

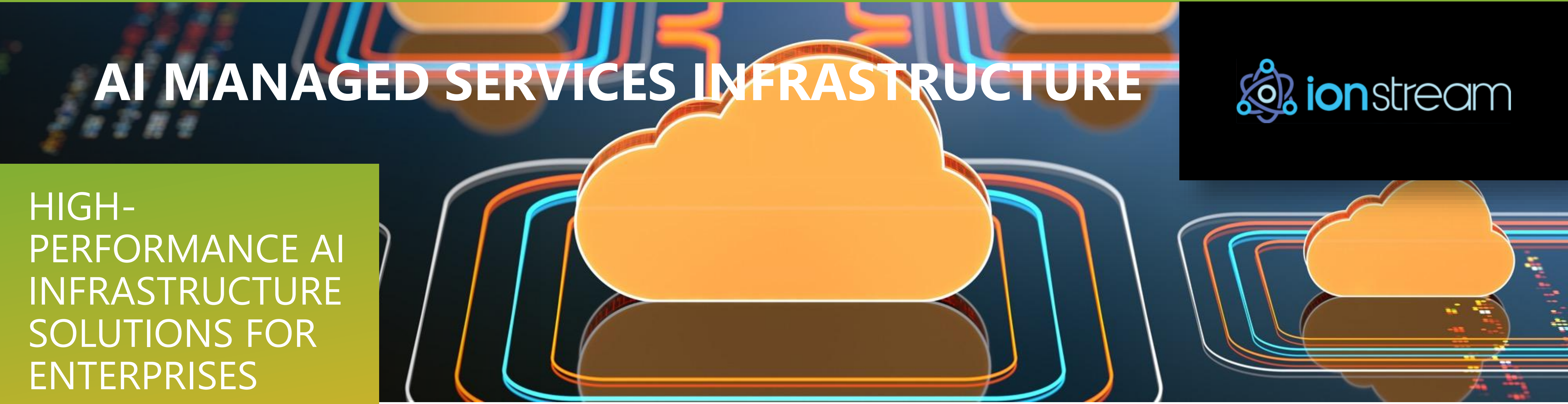
THE RESULTS

- **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

AI MANAGED SERVICES INFRASTRUCTURE



ionstream

HIGH-PERFORMANCE AI INFRASTRUCTURE SOLUTIONS FOR ENTERPRISES

THE CHALLENGE

- Unfamiliarity with NVIDIA InfiniBand
- Lengthy lead times from Cisco, Arista and NVIDIA
- Pricing from Arista/Cisco was high and slow to respond to ionstream’s questions

THE SOLUTION

- AI GPU pod features Juniper’s cutting-edge QFX 5240 powering 800GbE infrastructure
- Data center networking foundation empowers flexible GPU-as-a-service offerings to full service on-site deployments.
- Apstra Premium (future order)

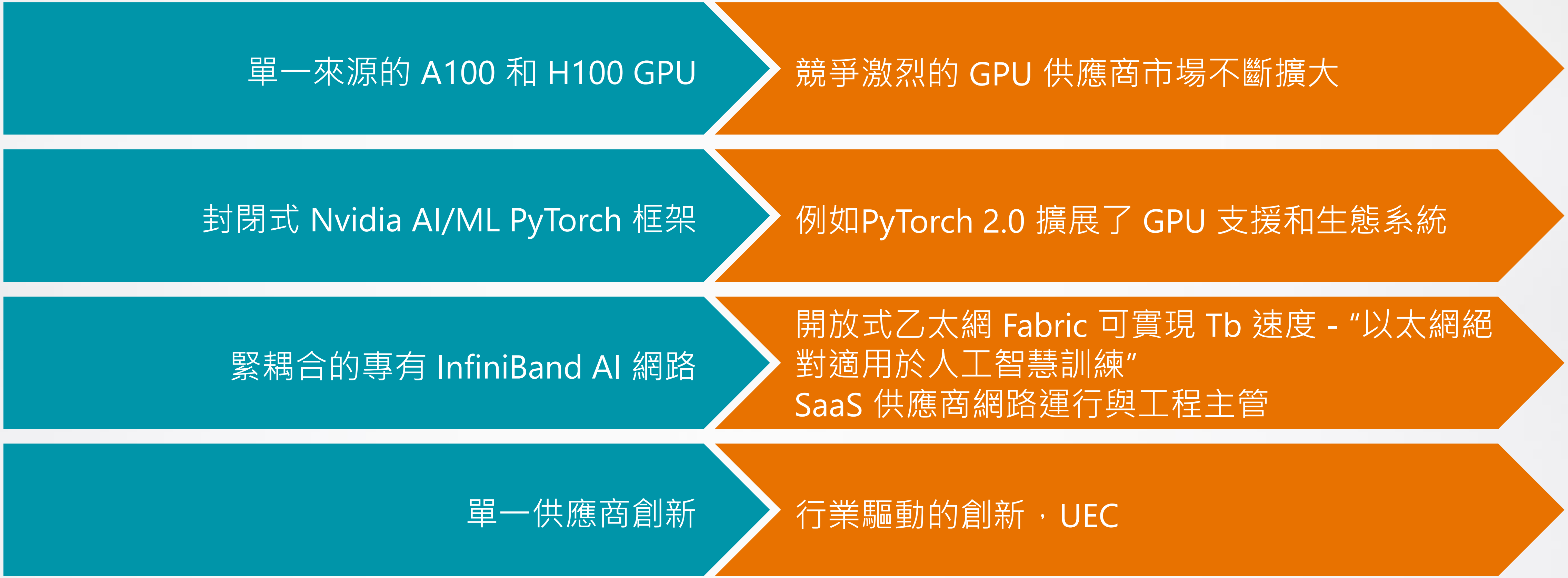
THE RESULTS

- Insights provided by specialist team on current and future designs **established instant credibility**
- Extensive competitive evaluation proved **Juniper’s superior ability to deliver** easy-to-deploy/operate, scalable networking driving faster time-to-value for AI clusters
- JVDs enable **quick spin up of AI-as-a-Service** to attract new clients

We’re able to deliver accessible, first-class AI transformation for our customers”

Next Gen of AIDC

從早期採用者. .到大眾市場

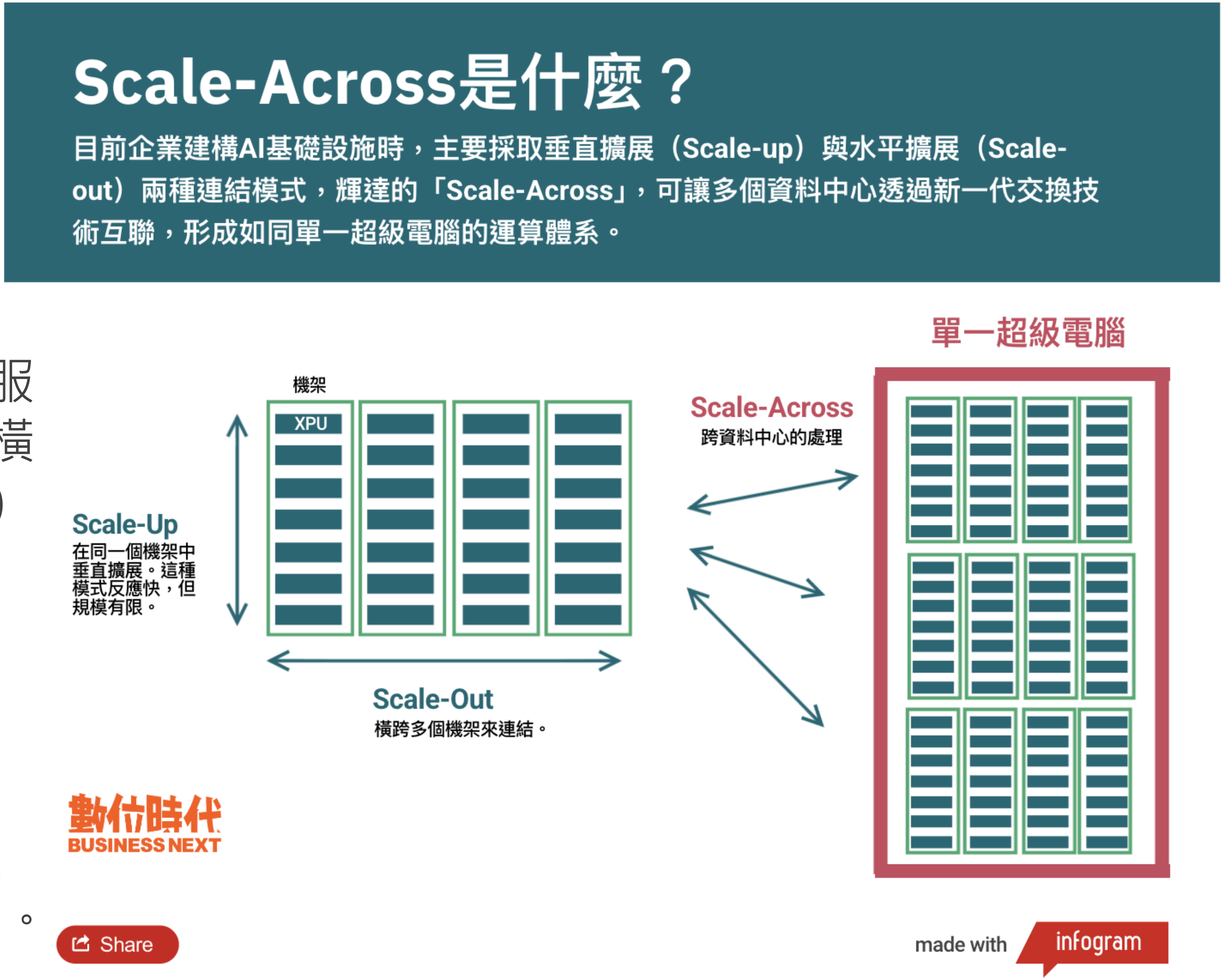


專有人工智慧	不斷演進的人工智慧解決方案
單一來源的 A100 和 H100 GPU	競爭激烈的 GPU 供應商市場不斷擴大
封閉式 Nvidia AI/ML PyTorch 框架	例如PyTorch 2.0 擴展了 GPU 支援和生態系統
緊耦合的專有 InfiniBand AI 網路	開放式乙太網 Fabric 可實現 Tb 速度 - “以太網絕對適用於人工智慧訓練” SaaS 供應商網路運行與工程主管
單一供應商創新	行業驅動的創新，UEC

What is Scale Across

- Scale Across」與傳統的「Scale-Up」（單一機櫃/伺服器擴容）和「Scale-Out」（橫向擴展更多機櫃/增加伺服器）不同，強調的是跨資料中心的協同運算能力。

- 突破了單一資料中心的規模及電力限制，能把多個城市、甚至不同國家的運算資源整合起來共同執行AI和大型運算任務。



Scale-Across是什麼？

目前企業建構AI基礎設施時，主要採取垂直擴展（Scale-up）與水平擴展（Scale-out）兩種連結模式，輝達的「Scale-Across」，可讓多個資料中心透過新一代交換技術互聯，形成如同單一超級電腦的運算體系。

Scale-Up
在同一個機架中垂直擴展。這種模式反應快，但規模有限。

Scale-Out
橫跨多個機架來連結。

Scale-Across
跨資料中心的處理

Shaping the Next Step Forward

GPUs are expensive and scarce

Single GPU

$33K +

8 x GPU server

$350K +

Small AI cluster

AI frontend network
Inference

AI backend network
Training

$5M-10M

Large AI clusters

$100s M

Data center CapEx comparison

	Traditional DC	AI Training DC
Compute	55%	80%
Storage	35%	14%
Network	10%	6%

Backend AI DC switching TAM:
$3B (2023)
65% CAGR (2022-2027)

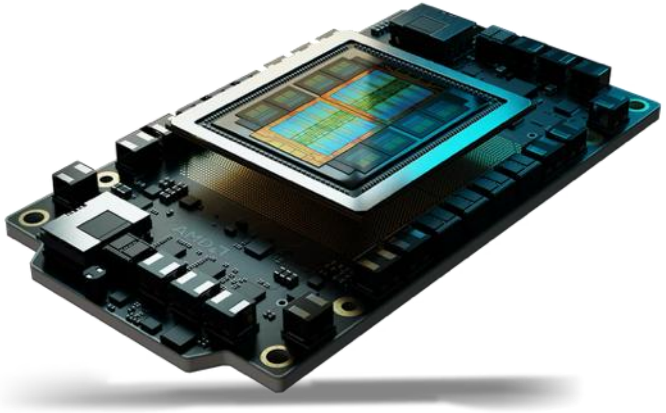
...the network is critical

Meta claimed last year that 33% of elapsed time in AI/ML is spent waiting for the network

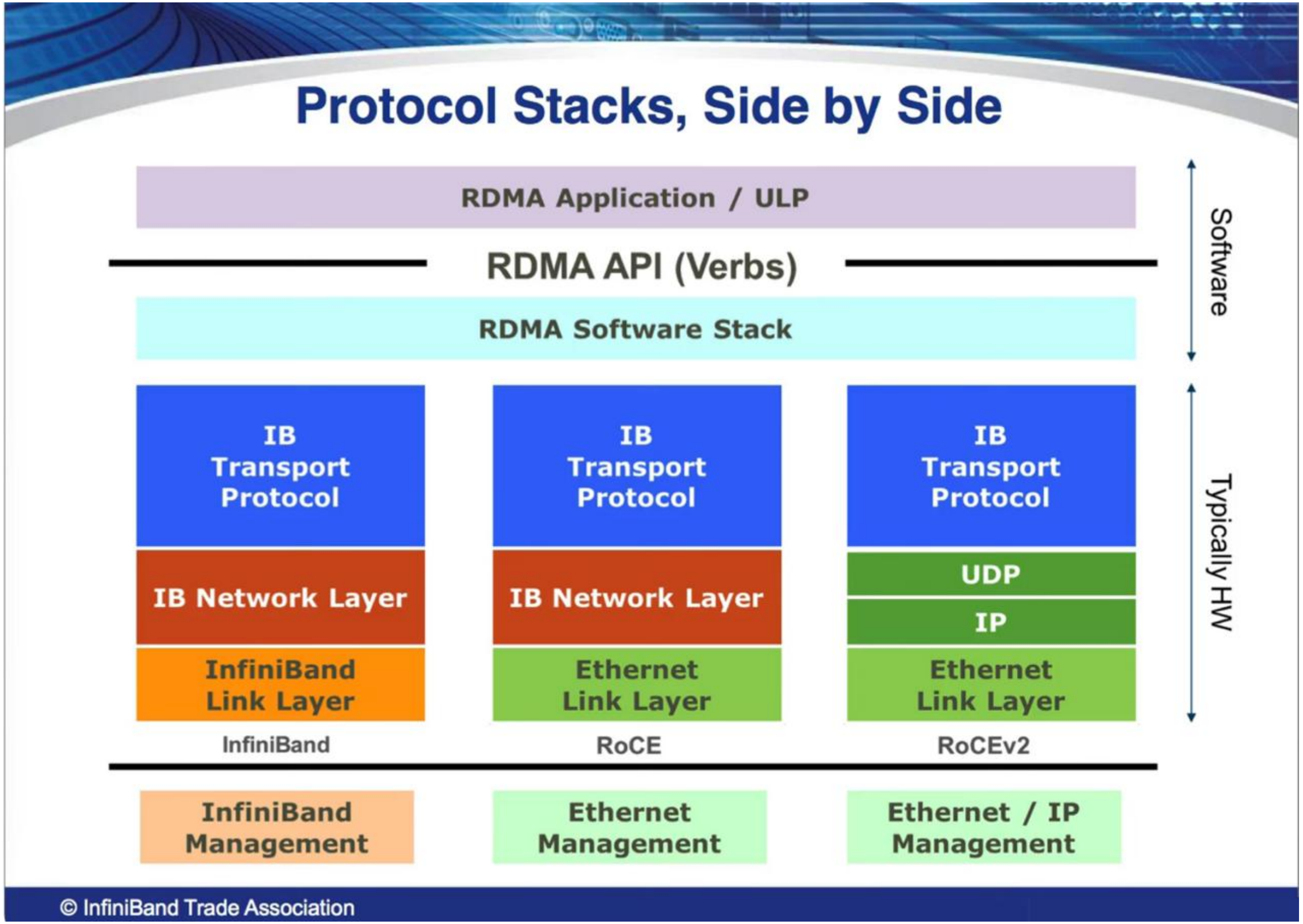
Source: Dell'Oro

Step Forward

- Market Trend of GPU/CPU/Memory resource
  - NVIDIA is priced beyond the reach of many users.
  - Most models aren’t tied to AMD or Intel GPUs; in fact, multi-vendor deployments are already common.
  - The majority of AIDCs now use multi-vendor computing architectures



What is RoCEv2 (RDMA over Converged Ethernet v2)



Protocol Stacks, Side by Side

RDMA Application / ULP

RDMA API (Verbs)

RDMA Software Stack

Software

InfiniBand	RoCE	RoCEv2
IB Transport Protocol	IB Transport Protocol	IB Transport Protocol
IB Network Layer	IB Network Layer	UDP IP
InfiniBand Link Layer	Ethernet Link Layer	Ethernet Link Layer
InfiniBand Management	Ethernet Management	Ethernet / IP Management

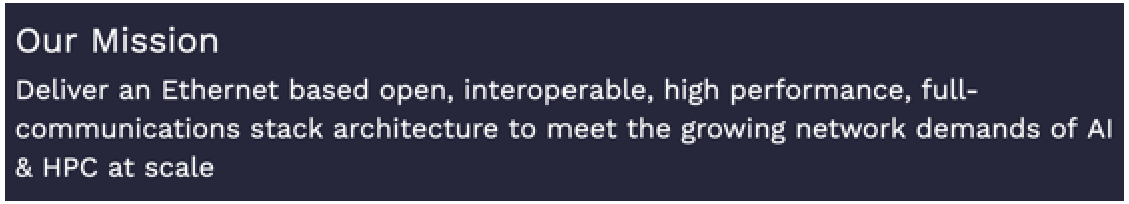
Typically HW

© InfiniBand Trade Association

https://www.youtube.com/watch?v=8kTAXhujn08&t=143s

Ultra Ethernet Consortium

Formed to create a new communication stack for Ethernet that is high-performance and open:



Our Mission

Deliver an Ethernet based open, interoperable, high performance, full-communications stack architecture to meet the growing network demands of AI & HPC at scale

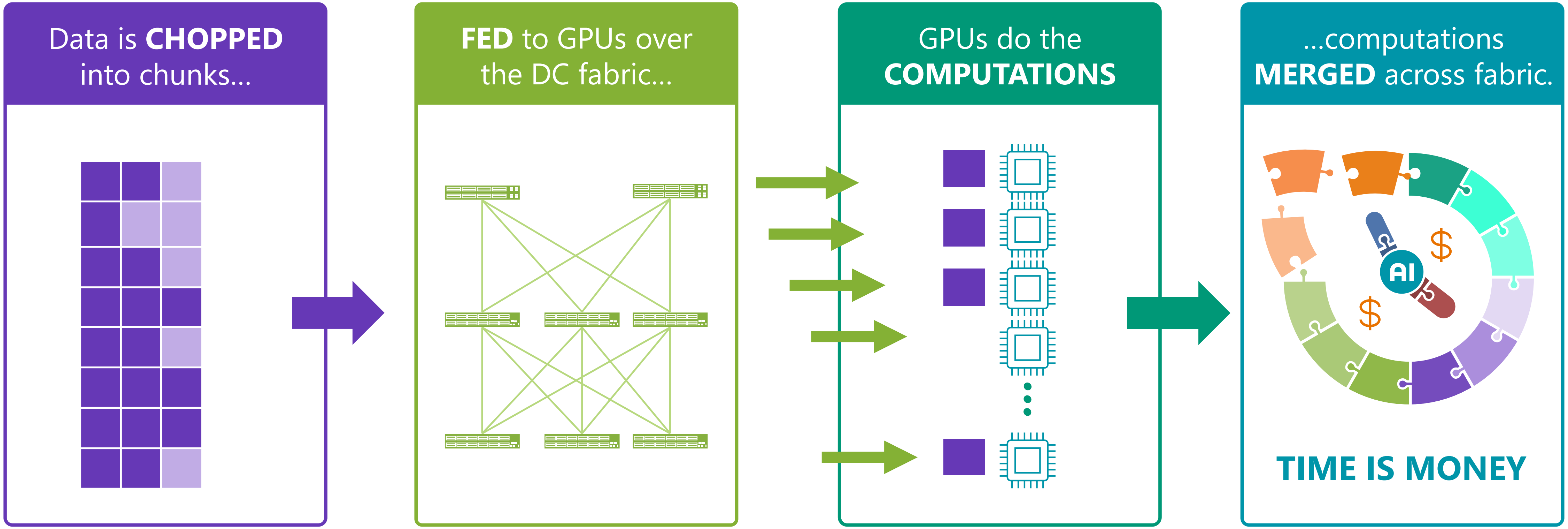
Adopting a clean slate approach to developing a complete communications stack for AI and HPC. A 1.0 Spec is planned for 2025 publication. UET is more than just a RoCEv2 replacement.

Juniper's view:

- Juniper is a General Member of UEC since Nov 2023
- A full spec for UET will take time to develop and even more time for adoption by hardware vendors
- A full communication stack will take significant time to develop so while we intend to participate in UEC, we will also be focused on tuning and optimizations for RoCEv2 which is heavily used in AI/ML today

https://www.juniper.net/us/en/the-feed/topics/ai-and-machine-learning/juniper-networks-ai-data-center-and-ultra-ethernet-consortium-uec.html

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



Data is **CHOPPED** into chunks…

FED to GPUs over the DC fabric…

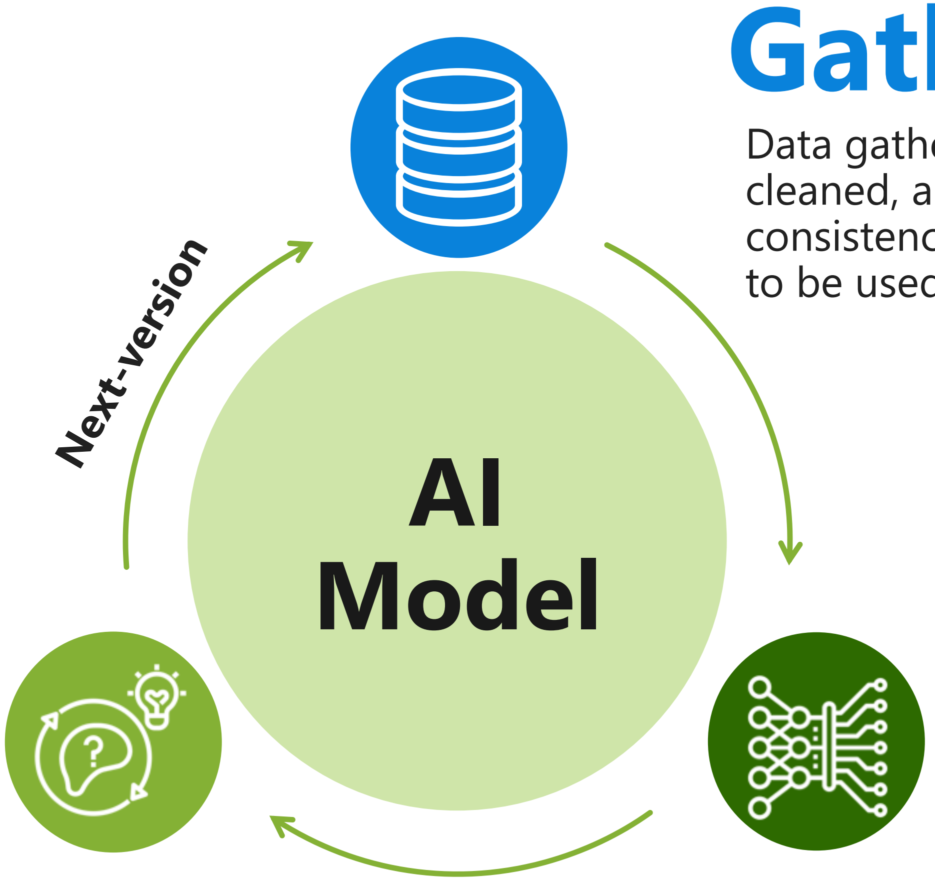
GPUs do the **COMPUTATIONS**

…computations **MERGED** across fabric.

TIME IS MONEY

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

Lifecycle of an AI Model



Next-version

AI Model

Gather data

Data gathered from various sources is cleaned, and verified for reliability and consistency. It is then prepared and curated to be used by the training model.

Training

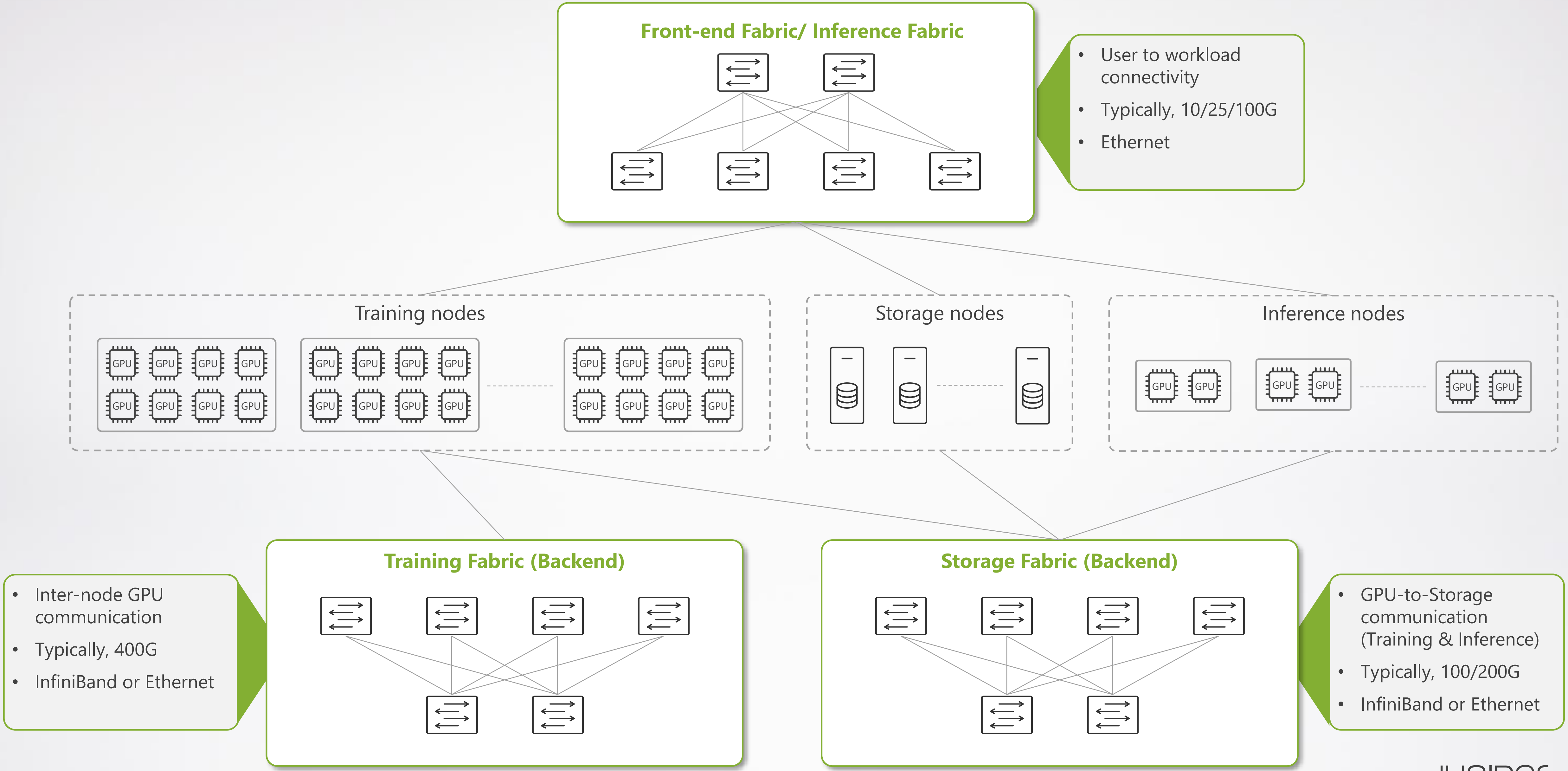
AI model is trained with the curated dataset and deep learning framework on GPU clusters.

Inference

The trained model is deployed on inference clusters to provide actionable outcomes from user inputs.

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 Data Center fabrics



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

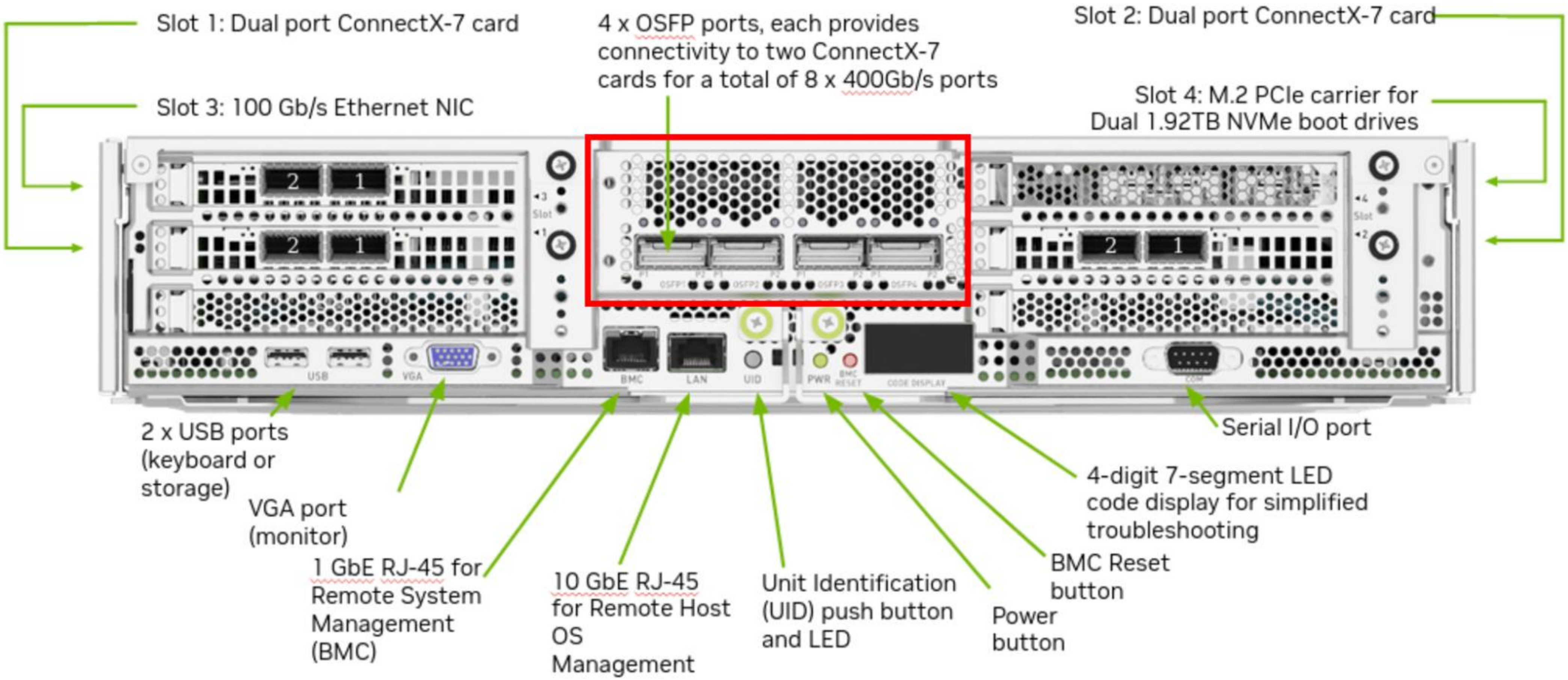
- Inter-node GPU communication
- Typically, 400G
- InfiniBand or Ethernet

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

[NVIDIA – DGX-H100 Doc](#)

DGX H100 – Network Ports

DGX H100 - NVIDIA Server with GPUs to accelerate deep learning applications



Slot 1: Dual port ConnectX-7 card

4 x OSFP ports, each provides connectivity to two ConnectX-7 cards for a total of 8 x 400Gb/s ports

Slot 2: Dual port ConnectX-7 card

Slot 3: 100 Gb/s Ethernet NIC

Slot 4: M.2 PCIe carrier for Dual 1.92TB NVMe boot drives

2 x USB ports (keyboard or storage)

VGA port (monitor)

1 GbE RJ-45 for Remote System Management (BMC)

10 GbE RJ-45 for Remote Host OS Management

Unit Identification (UID) push button and LED

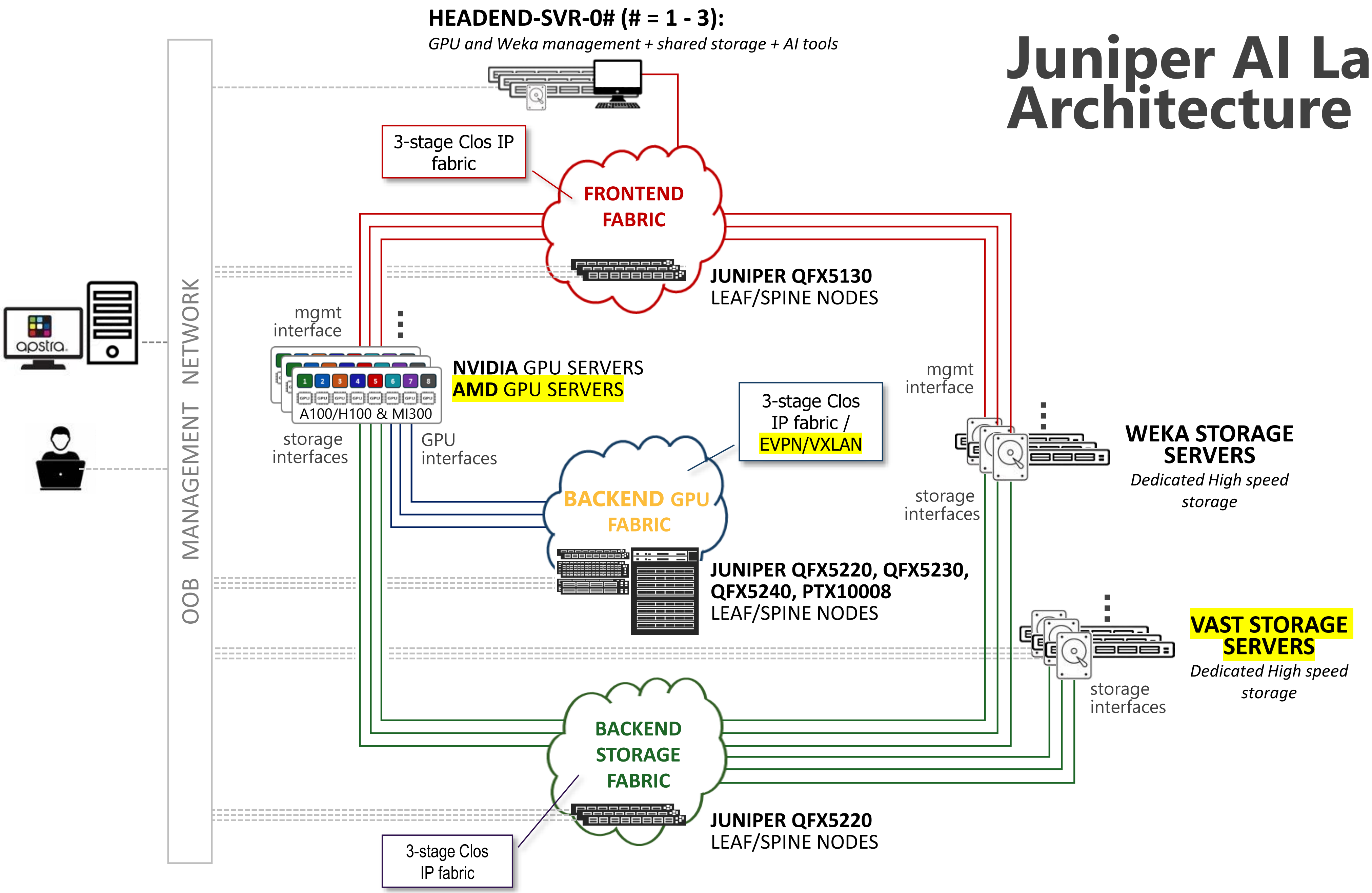
Power button

BMC Reset button

4-digit 7-segment LED code display for simplified troubleshooting

Serial I/O port

Juniper AI Lab Architecture



HEADEND-SVR-0# (# = 1 - 3):

GPU and Weka management + shared storage + AI tools

3-stage Clos IP fabric

FRONTEND FABRIC

JUNIPER QFX5130
LEAF/SPINE NODES

mgmt interface

NVIDIA GPU SERVERS
AMD GPU SERVERS

A100/H100 & MI300

storage interfaces

GPU interfaces

3-stage Clos IP fabric / EVPN/VXLAN

BACKEND GPU FABRIC

JUNIPER QFX5220, QFX5230, QFX5240, PTX10008
LEAF/SPINE NODES

mgmt interface

storage interfaces

WEKA STORAGE SERVERS

Dedicated High speed storage

VAST STORAGE SERVERS

Dedicated High speed storage

storage interfaces

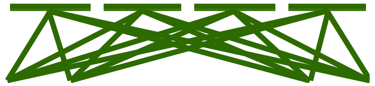
BACKEND STORAGE FABRIC

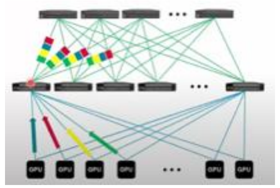
JUNIPER QFX5220
LEAF/SPINE NODES

3-stage Clos IP fabric

OOB MANAGEMENT NETWORK

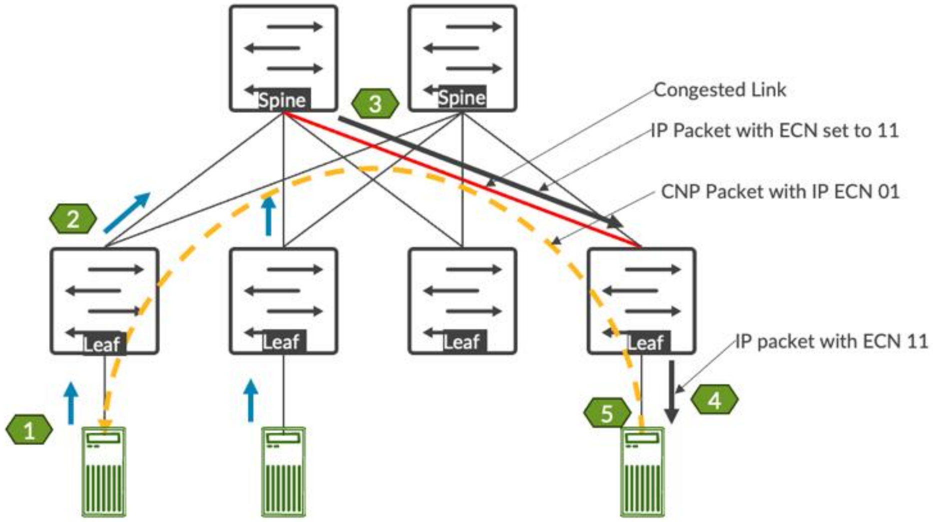
Open Ethernet offer best TCO with performance





	InfiniBand *Propriety Mellanox NICs and Switches*	**AI Optimized Ethernet** *Shallow Buffer, Deep Buffer across leaf and spine*	**Scheduled Fabric** *Cell or Ethernet-based disaggregated chassis*
Vendor Lock-In	Yes	No	Yes
Operational Consistency	No	Yes	No
Cost	Higher	Lower	Higher
Scale	Central/Limited	Distributed/Higher	Central/Limited
Performance for AI workloads	Yes	Yes	Yes

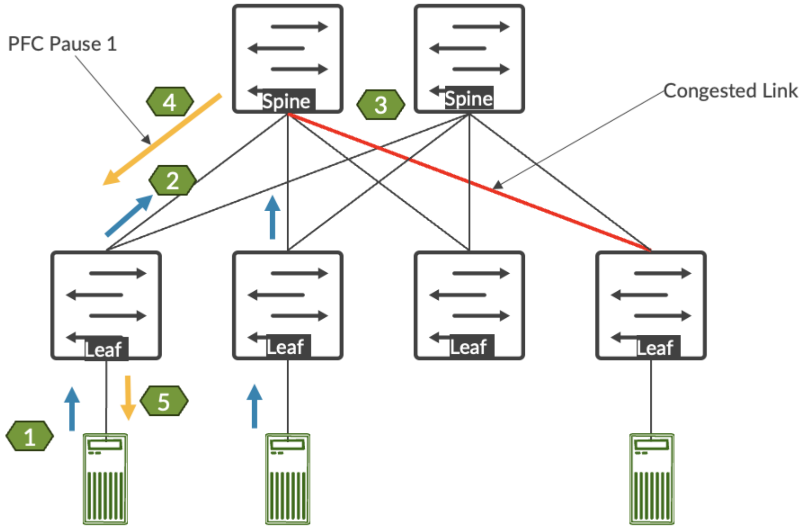
Congestion Control with DCQCN (ECN + PFC)



Explicit Congestion Notification (ECN)

- Node on detecting congestion, sets ECN bit to the packets
- Destination finds ECN is set and sends a congestion notification packet to source

- Source slows down the traffic



Priority Flow Control (PFC)

- Node on detecting congestion, sends PFC pause frame towards the previous node

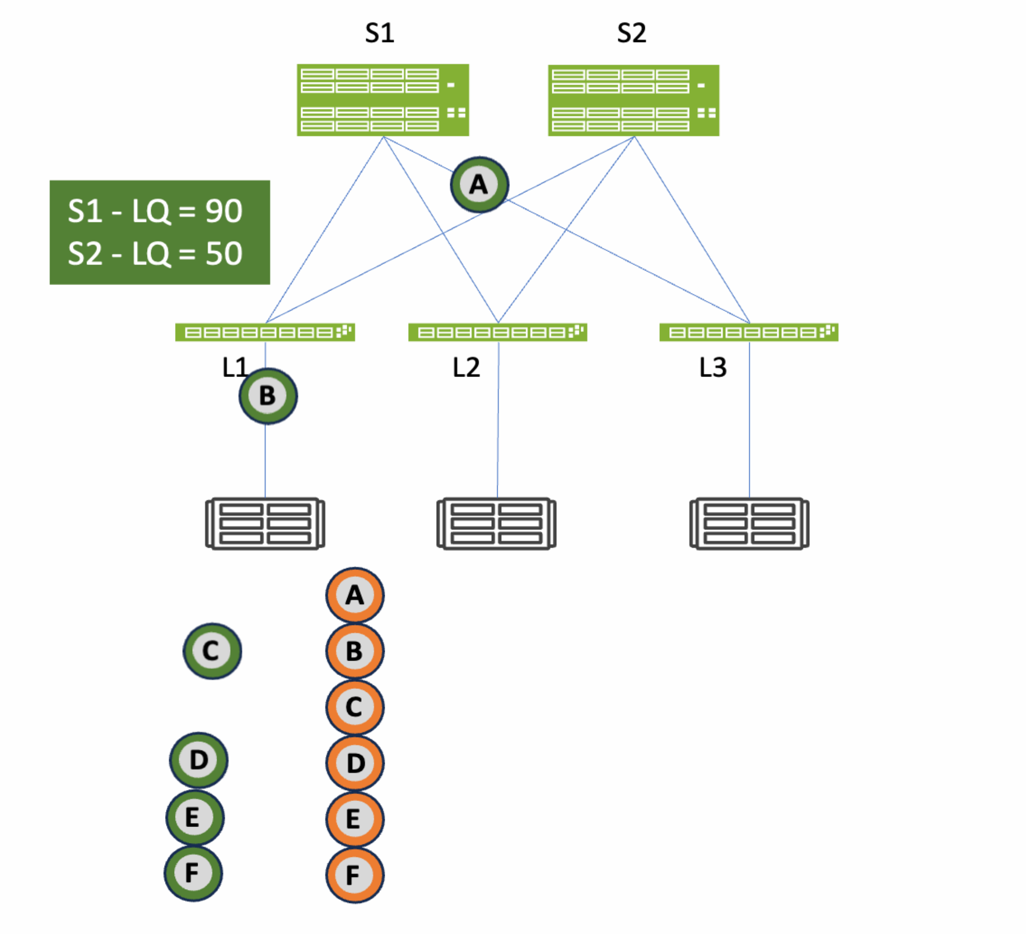
- Previous node slows down the traffic in that queue and sends update to the previous node towards the source

- Source slows down the traffic

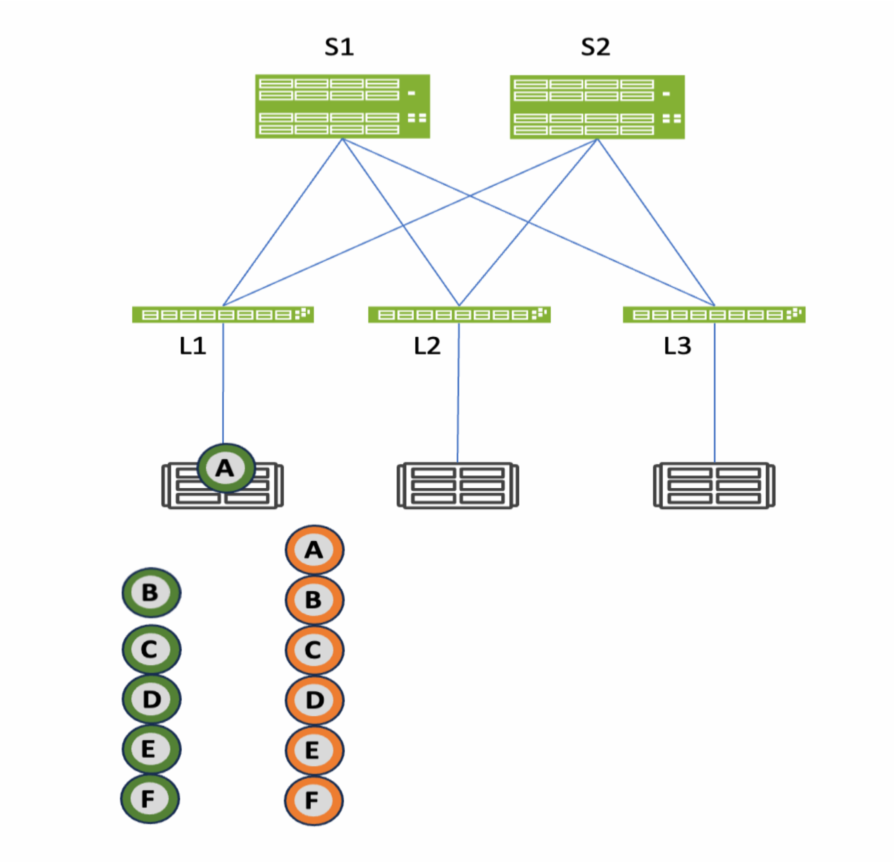
🟥 Congested Link

Network-based Dynamic Load Balancing (DLB)

DLB flowlet mode



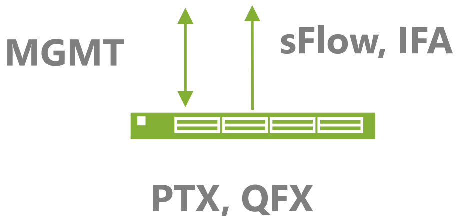
DLB packet mode

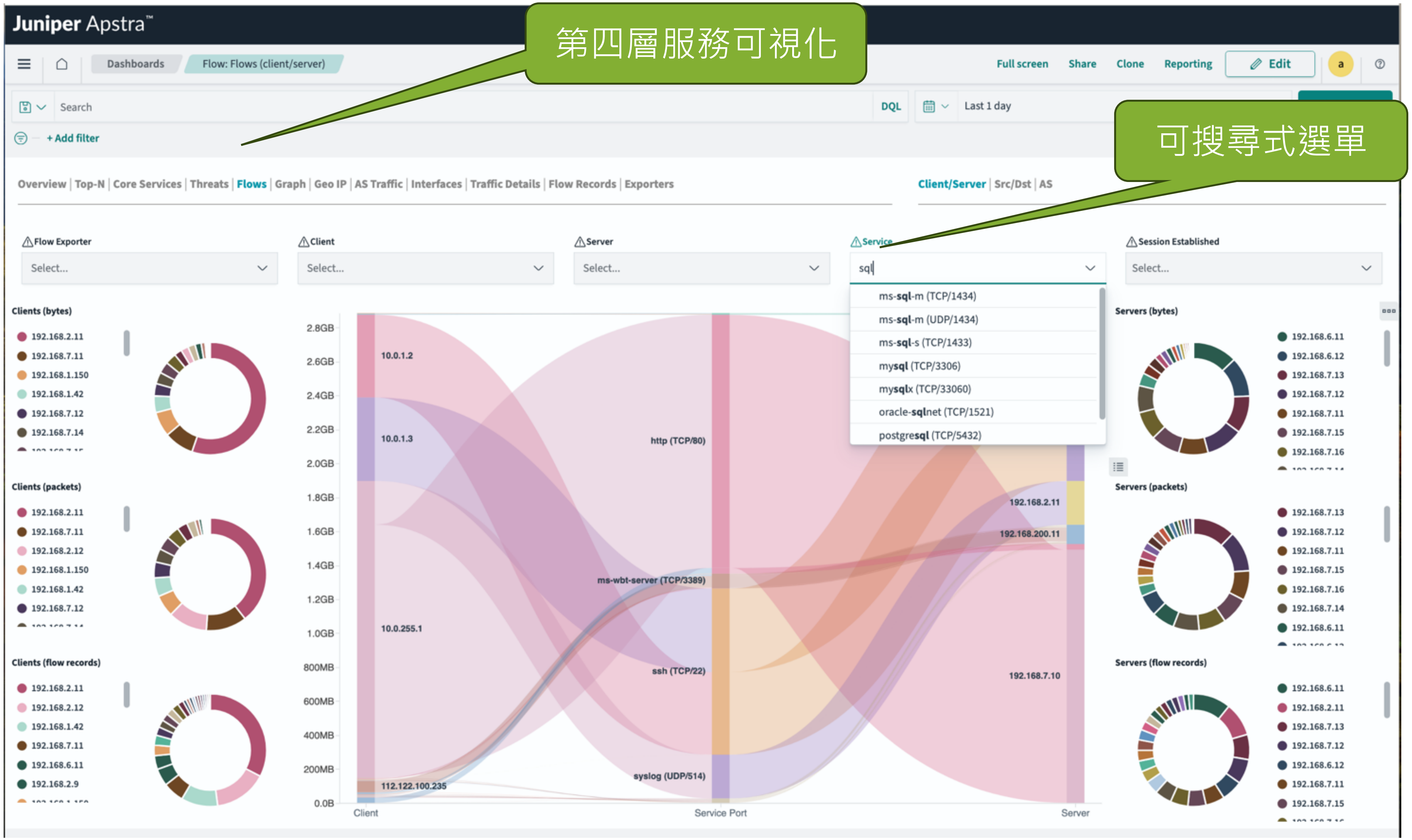


Apstra 服務可視化的效益

Benefits
- planning/ bandwidth
- verify firewall rule
- Geo IP
- DDoS Attack
- troubleshooting
- Service awareness

Juniper Apstra™

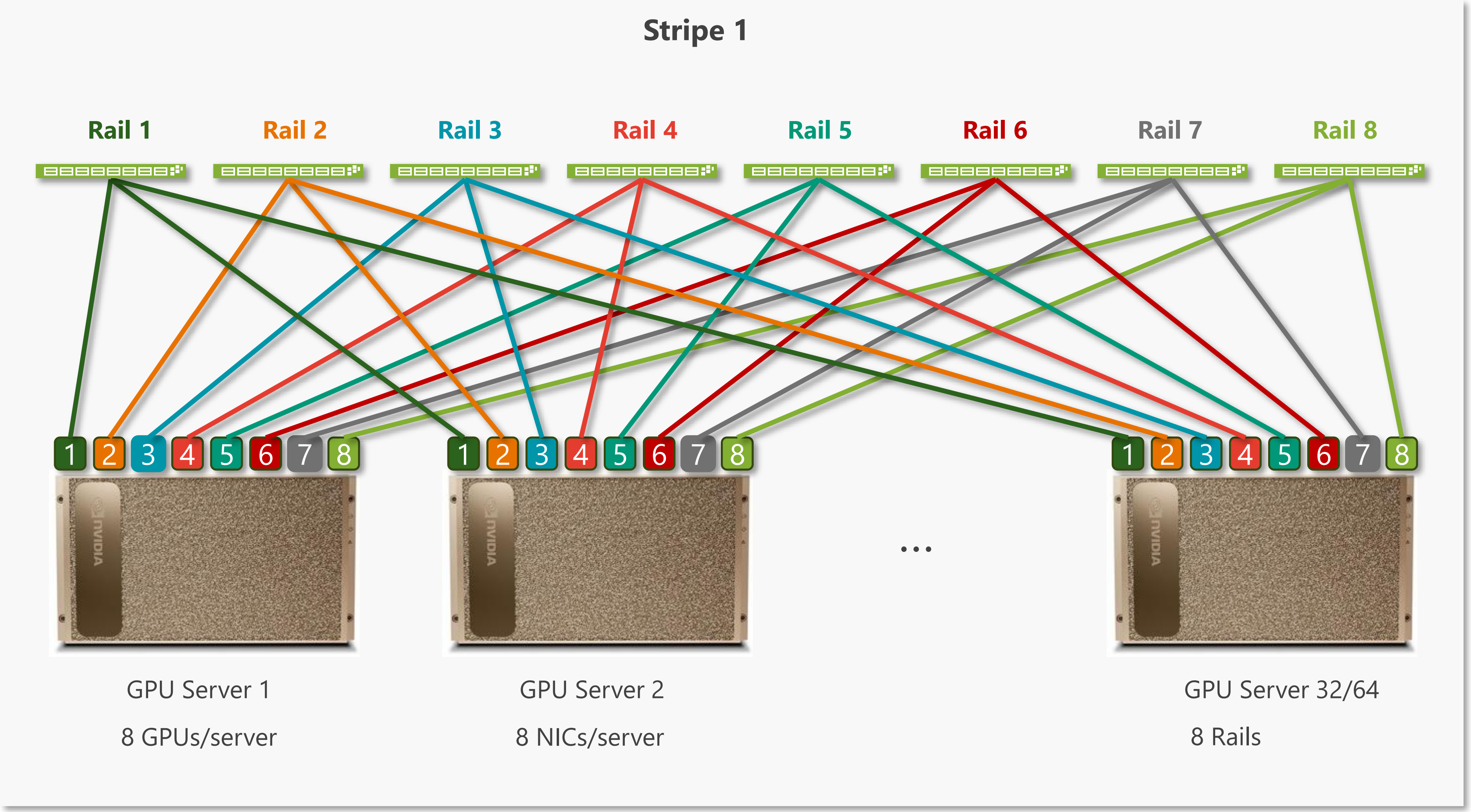



第四層服務可視化

可搜尋式選單

GPU Fabric Rail-Optimized Design

- 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/RCCL manages traffic among rails to provide 8 independent high BW channels avoiding collision at leaf.



支援NG-AI的網路設備應為考量重點

維運

ANSIBLE

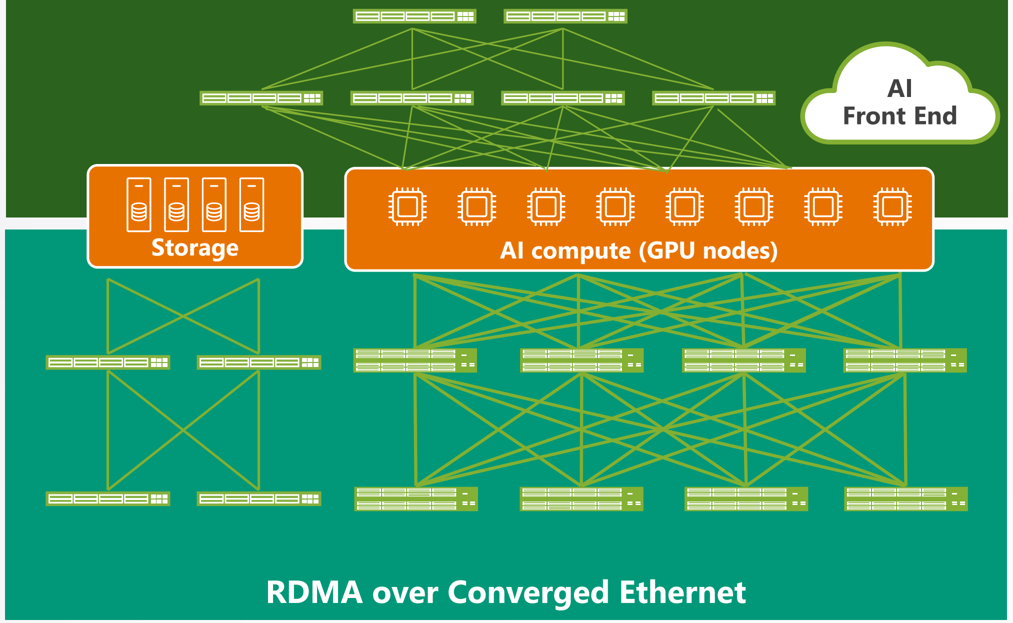
Terraform

Juniper Apstra™

融合 AI NetOps

一致的人工智慧平臺 NetOps 工作流程和自動化，提供操作簡單性、速度和可靠性

前端



100G/400G/800G 乙太網矩陣

QFX5120, QFX5130 交換機為開放式乙太網 Fabric 提供最佳性價比

後端

GPU 高效人工智慧基礎設施

新型高密度 400G/800G PTX, QFX 交換機為開放式乙太網 Fabric 提供最高容量和規模

IBN + AIOps 配有人工智慧擴展和先進的流量管理，可提供靈活性和更高的經濟性

Thank you

JUNIPER NETWORKS | Driven by Experience™