Cracking the Code of AI in the Data Center

Jeff Kreis – AE, Cisco Networking

April 29th, 2025

AI Changes **Everything**

$15.7T

Potential contribution to global economy by 2030

$300B

Global spending on AI by 2026

75%

Of large enterprises will rely on AI-infused processes by 2026

Healthcare and Life Sciences

Diagnosis
Drug discovery
Personalized medicine

Financial Services

Fraud detection
Risk assessment
Trading

Retail

Personalization
Inventory optimization
Virtual agents

Manufacturing

Predictive maintenance
Quality control
Demand forecasting

Agriculture

Yield optimization
Automated irrigation
Pest prediction & prevention

Transportation

Route optimization
Autonomous vehicles
Predictive maintenance

Energy

Distribution optimization
Fault prediction
Demand forecasting

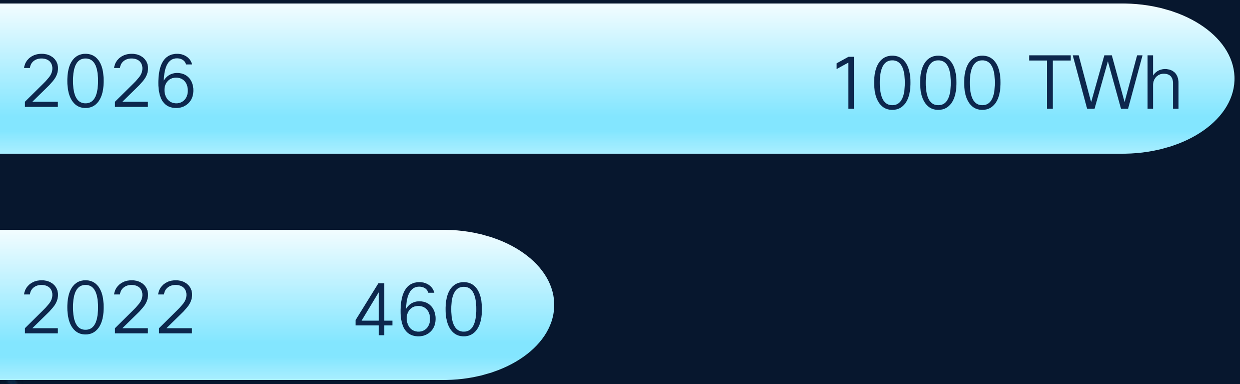
Public Sector

Smart cities
Security
Services improvement

Sources: PWC, IDC

Impact of AI Demand on Data Centers

AI impact on energy consumption could double by 2026



Growth will be led by power and the expansion of the data center sector, where U.S accounts for more than 1/3 of additional demand through 2026.

Updated regulations and technology improvements will be crucial to moderate the surge in energy consumption from data centers.

Source: IEA Electricity Report 2024

Impact of AI Demand on Data Centers

Efficient Data Centers are an important sustainability opportunity.

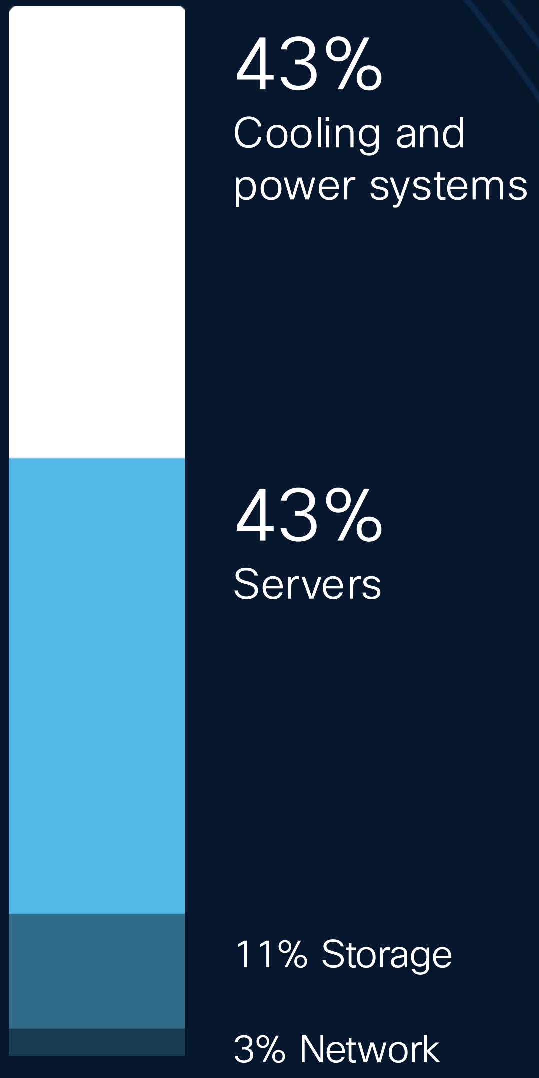
Today’s data center accounts for:

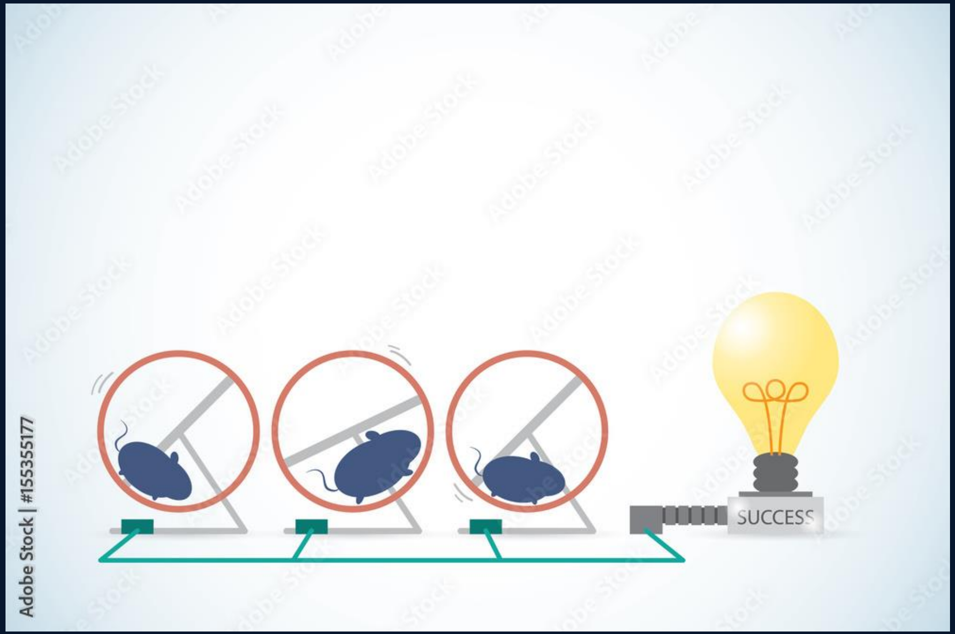
1-2% of global electricity demand

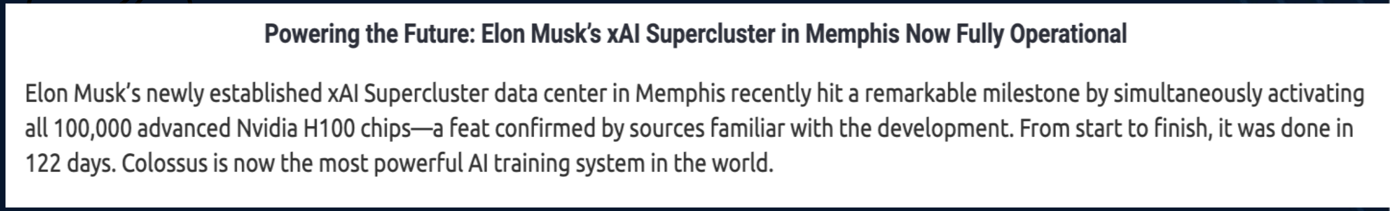
50X the power of a typical commercial office building

Every watt saved on computing results in:

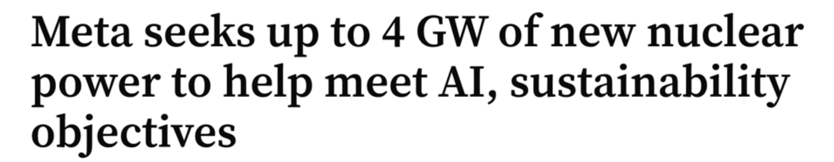
1.55 watts saved at the facility level

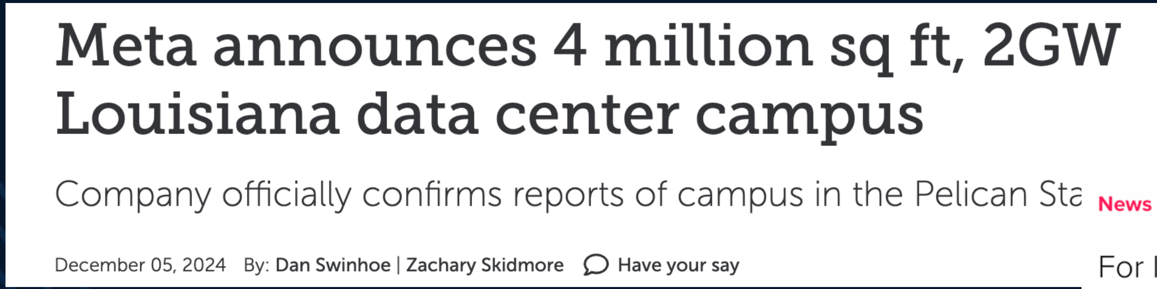





Powering the Future: Elon Musk's xAI Supercluster in Memphis Now Fully Operational

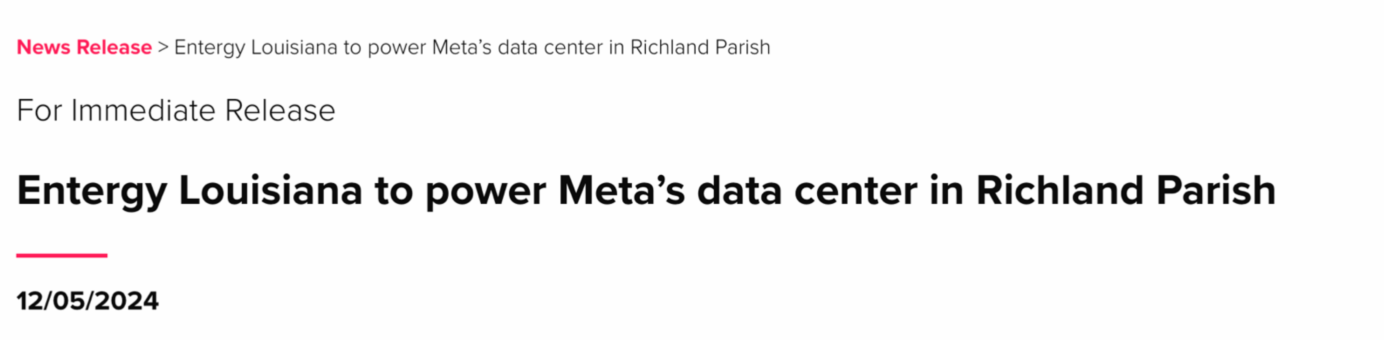
Elon Musk's newly established xAI Supercluster data center in Memphis recently hit a remarkable milestone by simultaneously activating all 100,000 advanced Nvidia H100 chips—a feat confirmed by sources familiar with the development. From start to finish, it was done in 122 days. Colossus is now the most powerful AI training system in the world.


Meta seeks up to 4 GW of new nuclear power to help meet AI, sustainability objectives


Meta announces 4 million sq ft, 2GW Louisiana data center campus

Company officially confirms reports of campus in the Pelican Sta

December 05, 2024 By: Dan Swinhoe | Zachary Skidmore Have your say


News Release > Entergy Louisiana to power Meta's data center in Richland Parish

For Immediate Release

Entergy Louisiana to power Meta's data center in Richland Parish

12/05/2024

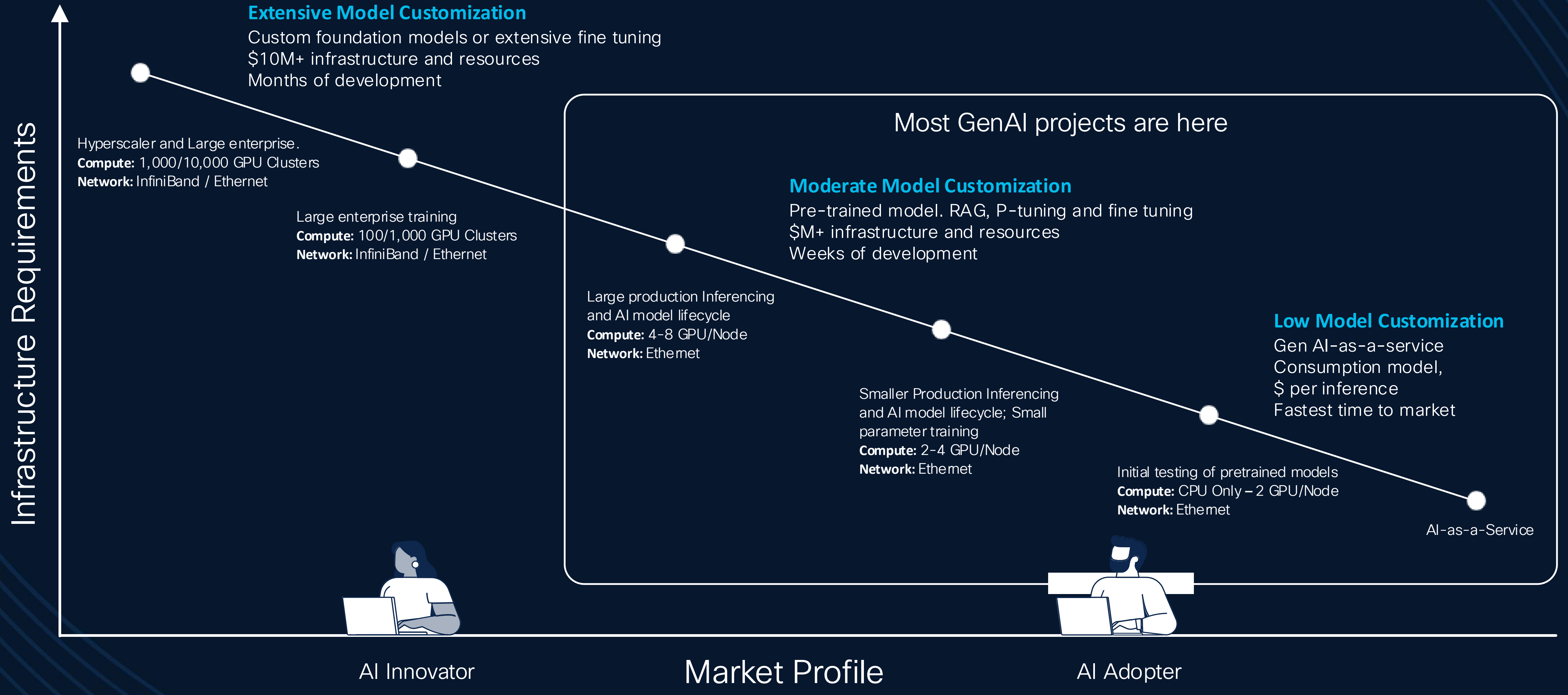
Traditional Data Center Challenges

Traditional DC Attributes	AI Workload Challenges
CPU-focused Compute	Inefficient for Parallel Processing
Lossy Ethernet	Lossless Network
Fixed & Inflexible Infrastructure	Difficulty Scaling & Adapting to Dynamic Workloads
Conventional Power & Cooling	Power Hungry Accelerators
Low Visibility, Siloed Management	Complex Orchestration of AI Resources

AI Compute Considerations

1. **Parallel Processing**: uses GPUs to handle 1000's of threads simultaneously.
2. **Deep Learning**: frameworks are optimized to utilize GPUs for efficiently training neural networks, involving matrix multiplications.
3. **Speed**: can significantly be reduced when training large neural networks with big datasets.
4. **Energy Efficiency**: is improved since GPUs can deliver more computational power per watt than CPUs.
5. **Specialized Hardware**: such as tensor cores in NVIDIA's GPUs are optimized for specific operations used in ML.
6. **Frameworks & Libraries**: like TensorFlow, PyTorch and CUDA libraries have extensive support for GPU acceleration.

AI Infrastructure Requirements



Extensive Model Customization

Custom foundation models or extensive fine tuning
$10M+ infrastructure and resources
Months of development

Hyperscaler and Large enterprise.
Compute: 1,000/10,000 GPU Clusters
Network: InfiniBand / Ethernet

Large enterprise training
Compute: 100/1,000 GPU Clusters
Network: InfiniBand / Ethernet

Most GenAI projects are here

Moderate Model Customization

Pre-trained model. RAG, P-tuning and fine tuning
$M+ infrastructure and resources
Weeks of development

Large production Inferencing and AI model lifecycle
Compute: 4-8 GPU/Node
Network: Ethernet

Smaller Production Inferencing and AI model lifecycle; Small parameter training
Compute: 2-4 GPU/Node
Network: Ethernet

Low Model Customization

Gen AI-as-a-service
Consumption model,
$ per inference
Fastest time to market

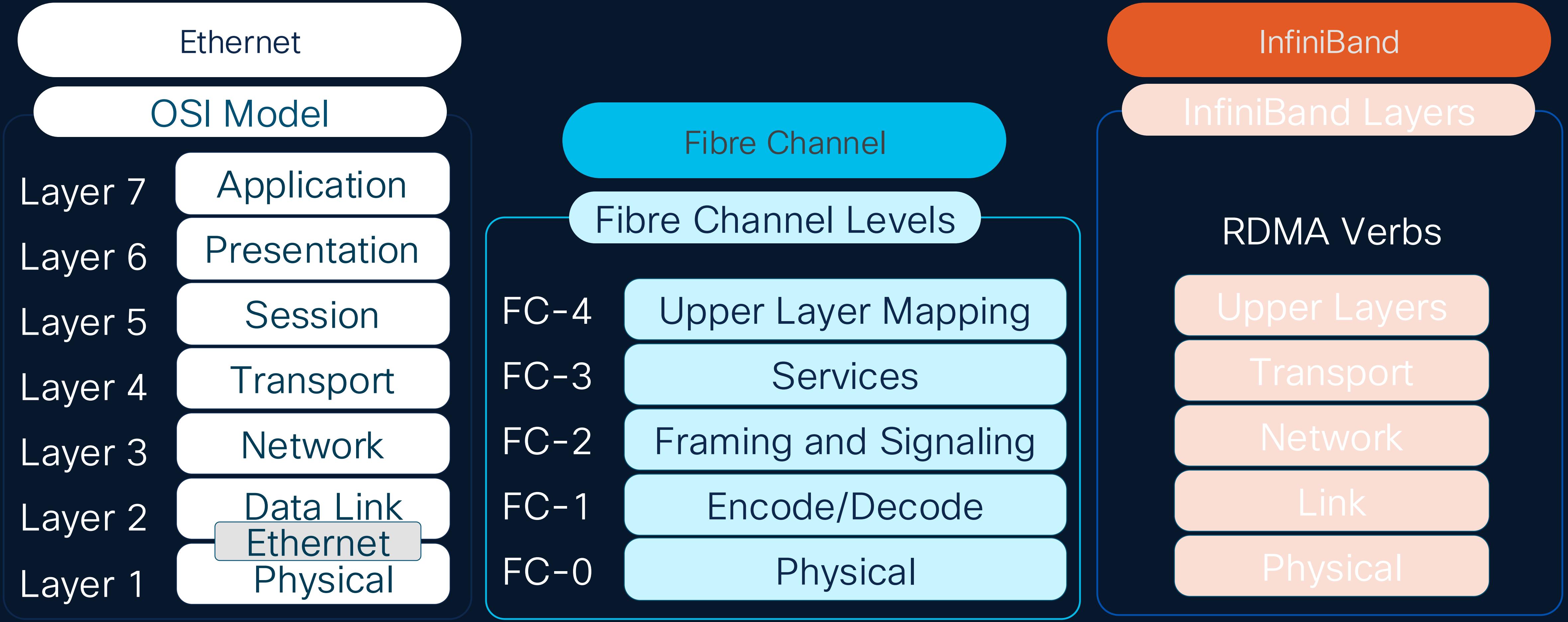
Initial testing of pretrained models
Compute: CPU Only – 2 GPU/Node
Network: Ethernet

AI-as-a-Service

Source: IDC: Enterprise Solutions for AI precedenceresearch - artificial-intelligence-market

Type of Networks in a Data Center

By Framing and Encoding



Ethernet

OSI Model

- Layer 7 Application
- Layer 6 Presentation
- Layer 5 Session
- Layer 4 Transport
- Layer 3 Network
- Layer 2 Data Link
- Ethernet
- Layer 1 Physical

Fibre Channel

Fibre Channel Levels

- FC-4 Upper Layer Mapping
- FC-3 Services
- FC-2 Framing and Signaling
- FC-1 Encode/Decode
- FC-0 Physical

InfiniBand

InfiniBand Layers

RDMA Verbs

- Upper Layers
- Transport
- Network
- Link
- Physical

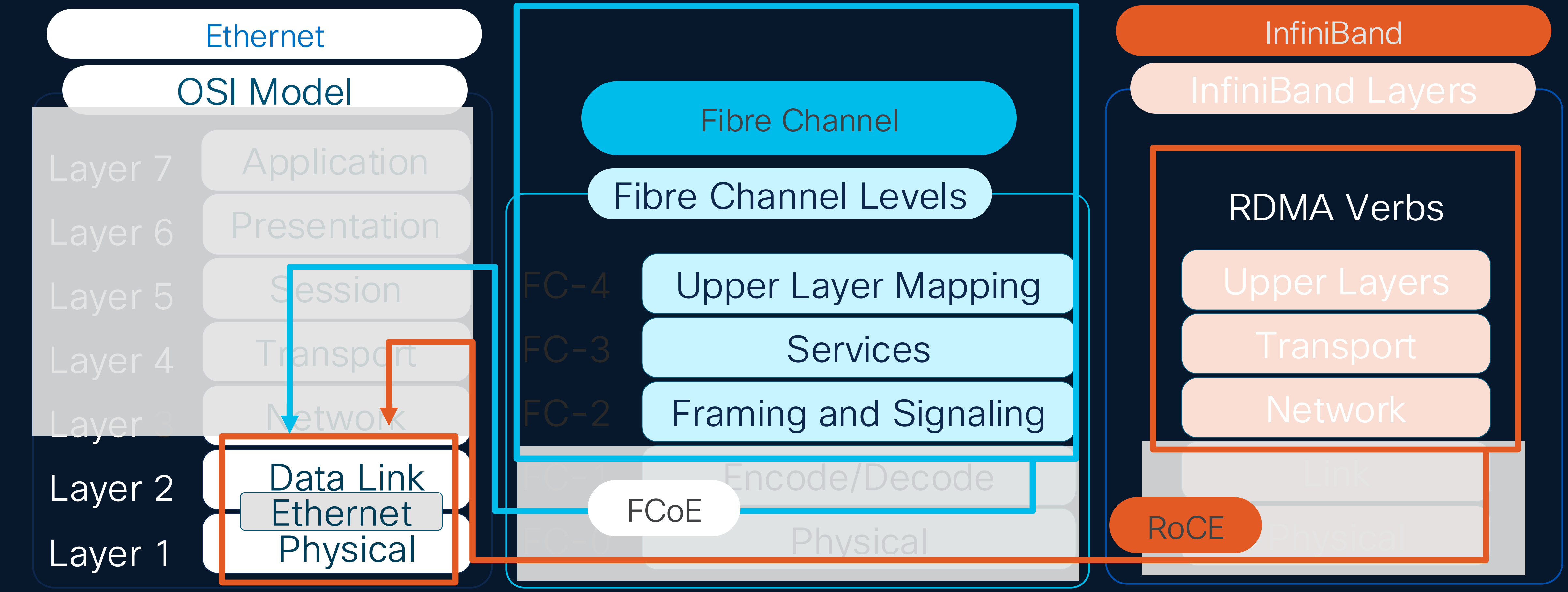
Optional Priority-based Flow Control (PFC). Pause Frames, etc.

B2B flow control. R_RDY, Credits, etc.

Credit-based flow control

Crossing The Boundaries of Network Types

What Fibre Channel did with FCoE, InfiniBand did with RoCE. Instead of IBoE, called it RoCE



Ethernet

OSI Model

Layer 7 Application

Layer 6 Presentation

Layer 5 Session

Layer 4 Transport

Layer 3 Network

Layer 2 Data Link

Ethernet

Layer 1 Physical

Fibre Channel

Fibre Channel Levels

FC-4 Upper Layer Mapping

FC-3 Services

FC-2 Framing and Signaling

FC-1 Encode/Decode

FC-0 Physical

FCoE

InfiniBand

InfiniBand Layers

RDMA Verbs

Upper Layers

Transport

Network

Link

Physical

RoCE

Optional Priority-based Flow Control (PFC). Pause Frames, etc.

B2B flow control. R_RDY, Credits, etc.

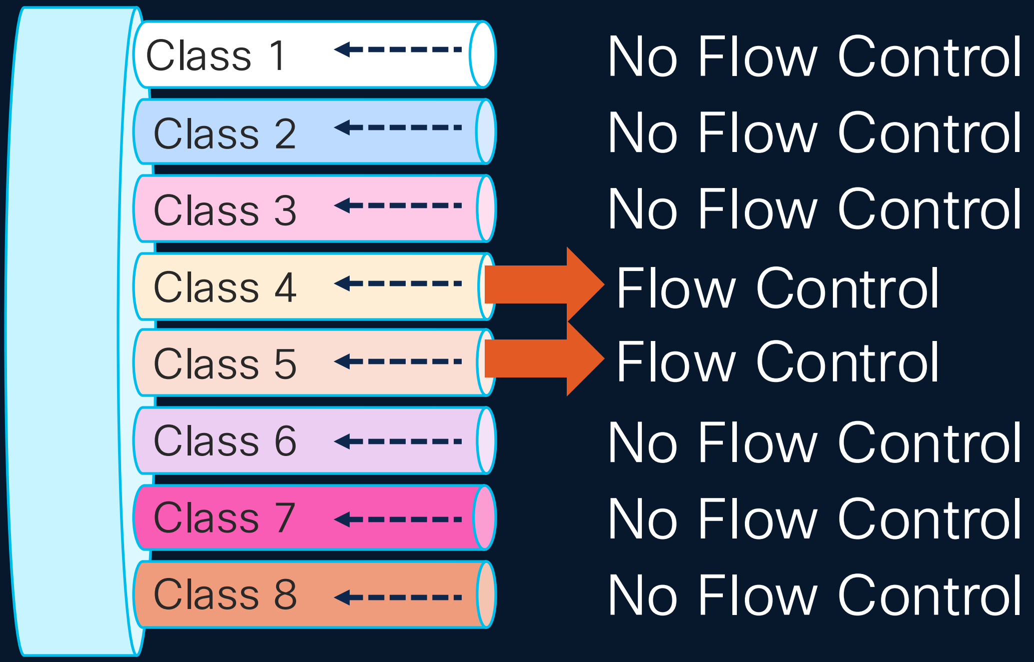
Credit-based flow control

Ethernet Flow Control

Traffic

Paces traffic in specific classes from directly-connected device while other classes are not flow controlled (IEEE 802.1Qbb).

Priority-based Flow Control (PFC)



Class 1 No Flow Control

Class 2 No Flow Control

Class 3 No Flow Control

Class 4 Flow Control

Class 5 Flow Control

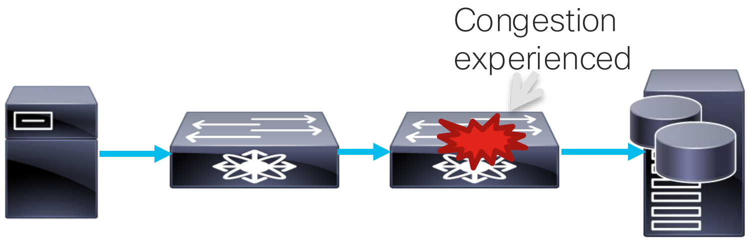
Class 6 No Flow Control

Class 7 No Flow Control

Class 8 No Flow Control

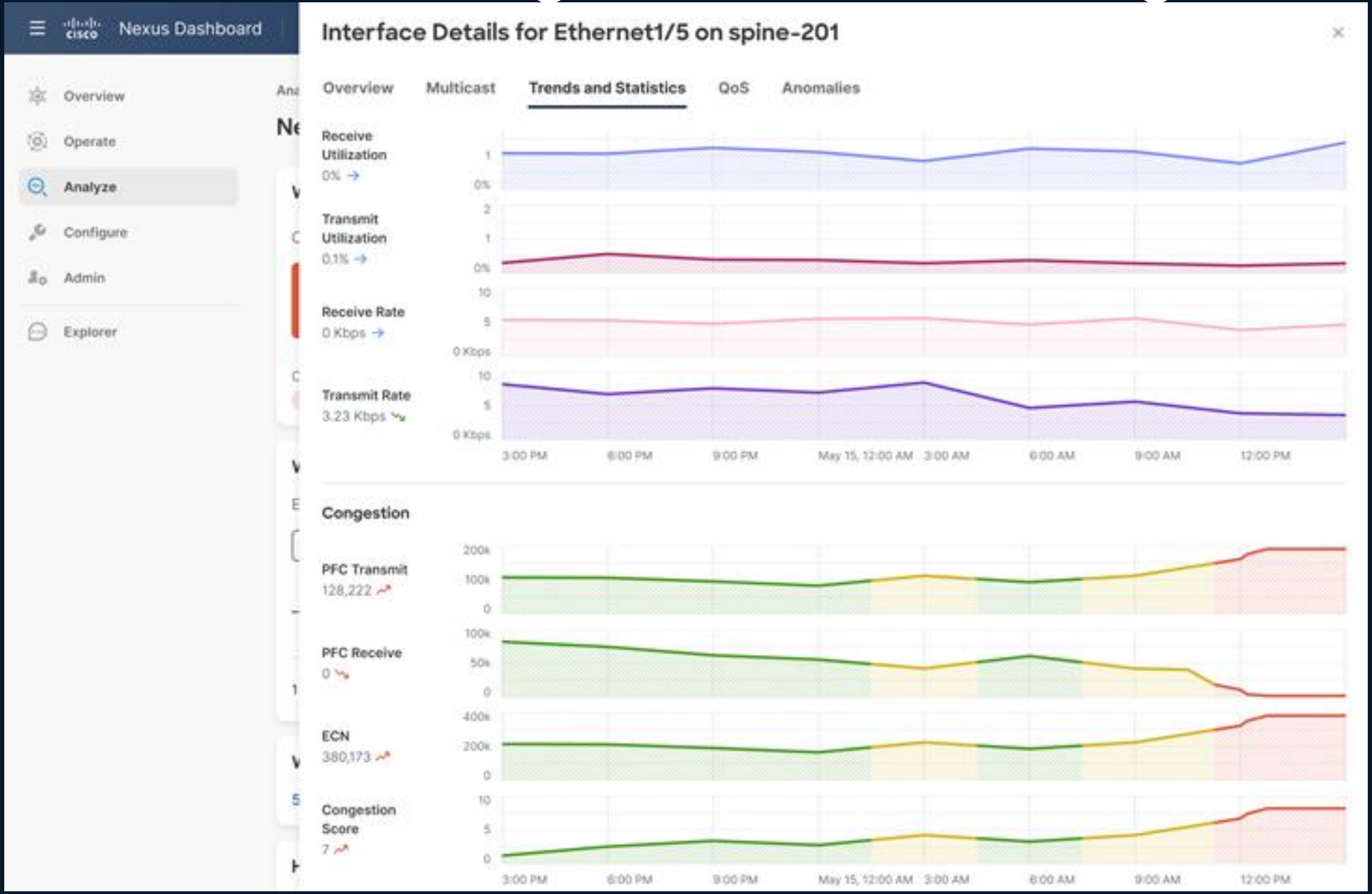
Explicit Congestion Notification

- IP Explicit Congestion Notification (ECN) is used for congestion notification.
- ECN enables end-to-end congestion notification between two endpoints on IP network
- ECN uses 2 LSB of Type of Service field in IP header



ECN	ECN Behavior
00	Non ECN Capable
10	ECN Capable Transport (0)
01	ECN Capable Transport (1)
11	Congestion Encountered

Nexus Dashboard Insights for Monitoring PFC & ECN

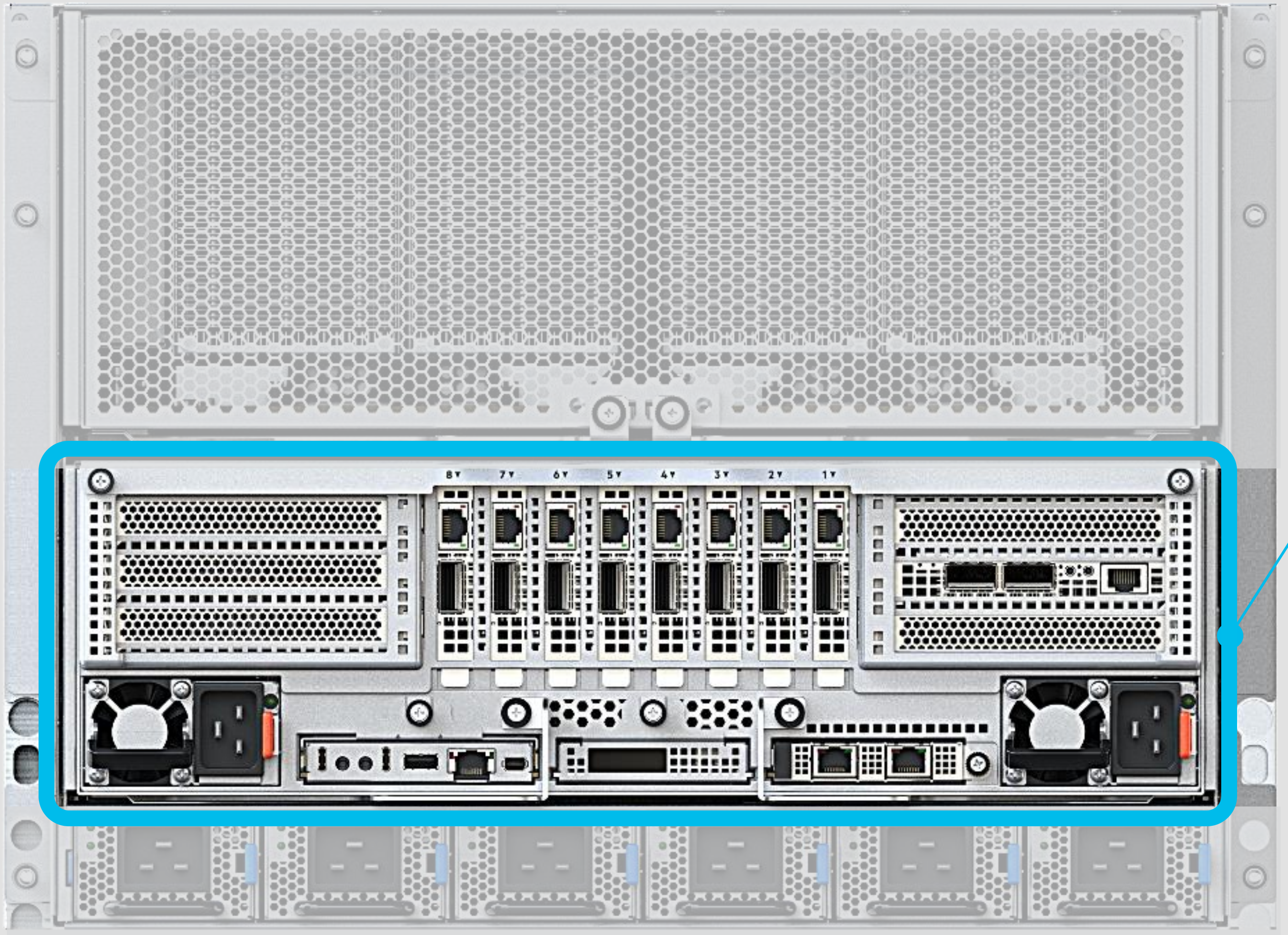


Bringing high-density GPU servers to the Cisco UCS family and to Cisco's AI solution portfolio

Discover data-intensive use cases
like model training and deep learning

Nvidia HGX with
8 Nvidia H100, H200 or AMD
Mi300X GPUs

2 AMD 4th Gen
EPYC™ Processors



Server Rear View

CPU & Memory

2x

AMD 9554 (Genoa) CPUs

64 cores & up to 3.75GHz

360W/CPU

or

2x

AMD 9575F (Turin) CPUs

64 cores & up to 5GHz

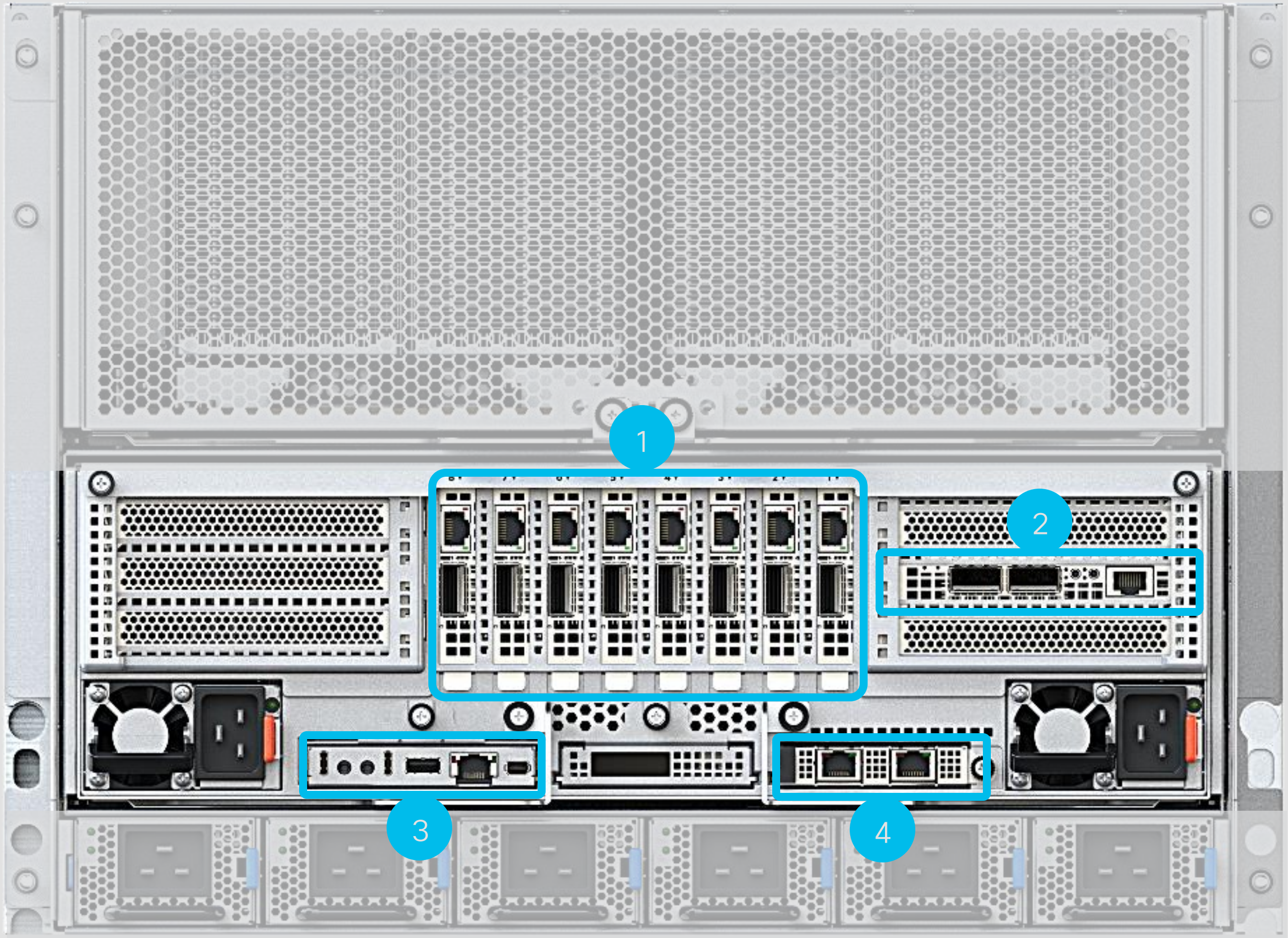
400W/CPU

24x

DDR5 RDIMMs

Up to 6,000 MT/S

128GB DIMM option for some fixed configs coming soon



Server Rear View

I/O & Other Components

1. 8x PCIe Gen5 x16 HHHL for east-west GPU-to-GPU traffic
2. 1x PCIe Gen5 x16 FHHL for north-south traffic
3. 1x Data Center Secure Control Module (DC-SCM)
4. 1x 1x OCP 3.0 PCIe Gen5 x8 for X710 2 x 10G RJ45 NIC for additional north-south or host management traffic

New MGX Server

UCS C845A M8

Highly Scalable 2-8 GPU MGX Server Designed to Drive a Multitude of AI Workloads. This MGX design will allow Cisco to utilize existing designs in implementing next-generation GPUs without costly redesigns.

MGX 4RU 19” EIA Rack

2, 4 or 8x Nvidia H100 NVL/H200 NVL/ B300A NVL/L40S/B40 GPUs, AMD MI210, Intel Guadi 3

AMD Turin CPUs @ 400W TDP

5x PCIe x16 FHHL Slots and 8 x PCIe x16 GPU Slots

AI Training and Inference, HPC, Data Analytics, Visualization and Hyperscale Cloud Applications, Large Language Models, Design and Simulation

- Service Providers
- Financial Services
- Manufacturing

- Healthcare and Life Sciences
- Automotive

UCS 845A M8 Product Overview

2-8-5

2 CPU / 8 GPU / 5 NICs

Front

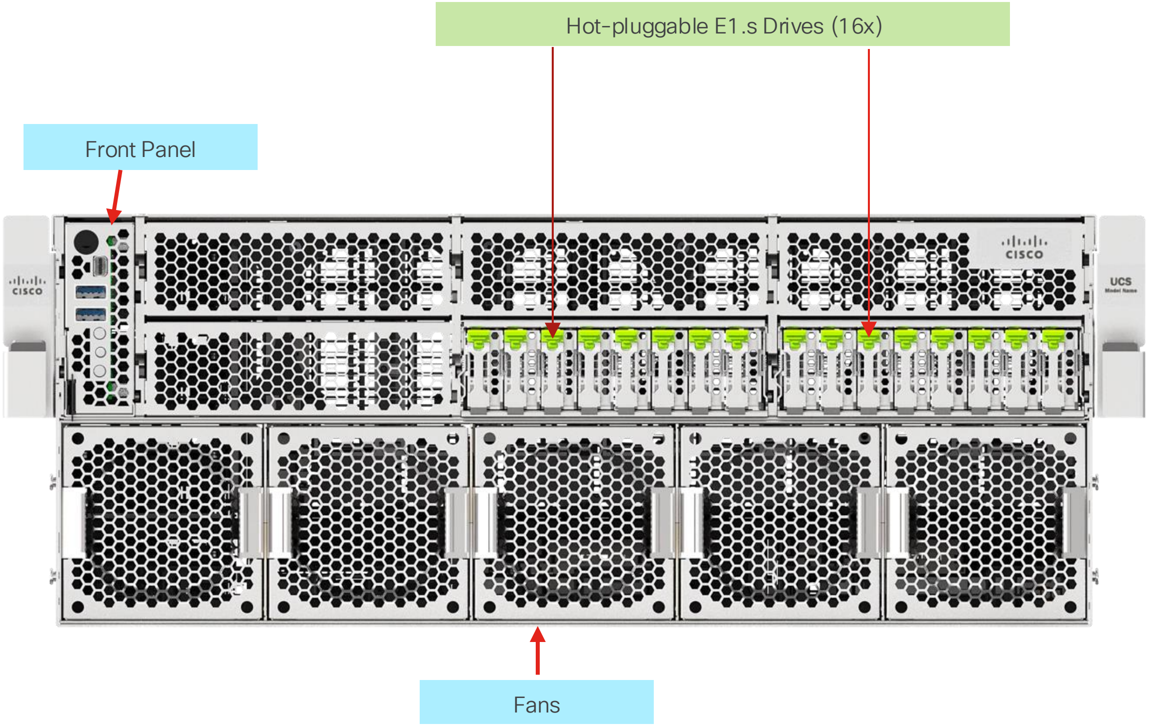
Back

A Flexible / Field Serviceable MGX AI Server

Product Specifications	
Form Factor	• 4RU Air-cooled Chassis
Compute + Memory	• Dual AMD Turin CPUs – Up to 400W TDP each • 32 DDR5 DIMMs - 5200MT/s 1DPC & 4400MT/s 2DPC
Storage	• 16x E1.S NVMe PCIe Gen5 Drives • Support for Boot RAID using Noe Valley and 2 x M.2 SATA Boot Drives
GPU	• Supports **up to** 8x Nvidia, AMD and Intel PCIe GPUs - 600W TDP
Network Cards	• 5 PCIe x16 – FHHL Slots for single slot NICs / DPUs • NVIDIA BF3 SuperNIC for N-S Traffic • Mellanox CX-7 Or BF3 (single slot) for E-W Traffic
Chassis Mngt.	• Driven via DC-SCM card with AST2600 BMC; TPM
BMC	• Network addressable through dedicated RJ-45 Ethernet port
Platform Root of Trust	• AST1060 (Similar PRoT as Bronco)
Firmware	• Cisco firmware enabled and Intersight Managed
GPU Switching	• RDMA Enabled PCIe Switches for GPU Direct
Cooling	• 10 x 80mm Fans
Front IO	• 1 USB 3.0, 1mdp, 1 ID button (w/ID LED), 1 Power Button (w/ Power & Status LED), 1 Reset button
Rear IO	• 1 USB 3.0, 1mDP, 1 ID button (w/ID LED), 1 Reset Button, 1 RJ45 (Mgmt)
Power Supply	• Up to 4x 3.2KW MCRPs PSU with N+1 Redundancy

Items above in blue are Cisco customizations

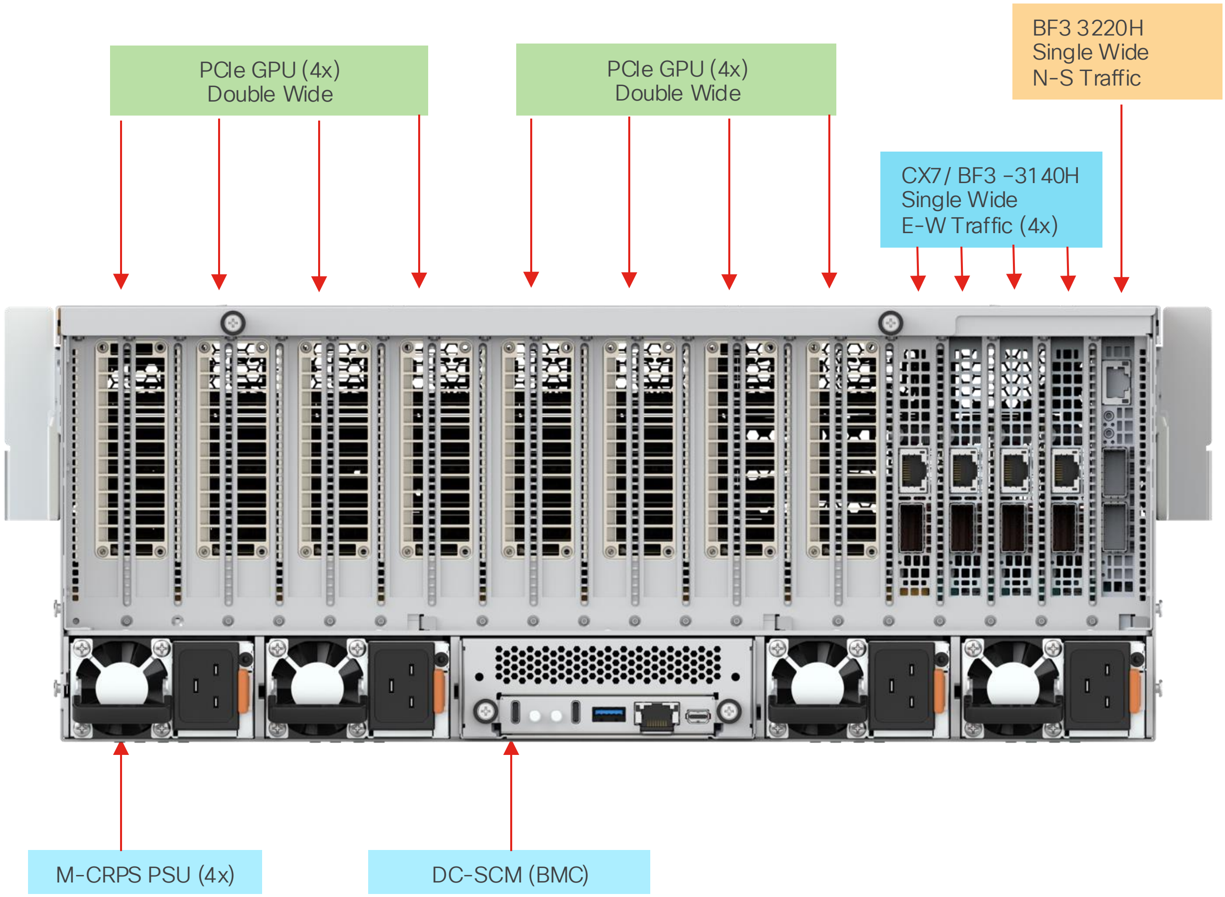
Front and Back Views



Hot-pluggable E1.s Drives (16x)

Front Panel

Fans



PCIe GPU (4x)
Double Wide

PCIe GPU (4x)
Double Wide

BF3 3220H
Single Wide
N-S Traffic

CX7 / BF3 -3140H
Single Wide
E-W Traffic (4x)

M-CRPS PSU (4x)

DC-SCM (BMC)

Network Definitions

Multiple networks of an AI/ML Infrastructure…

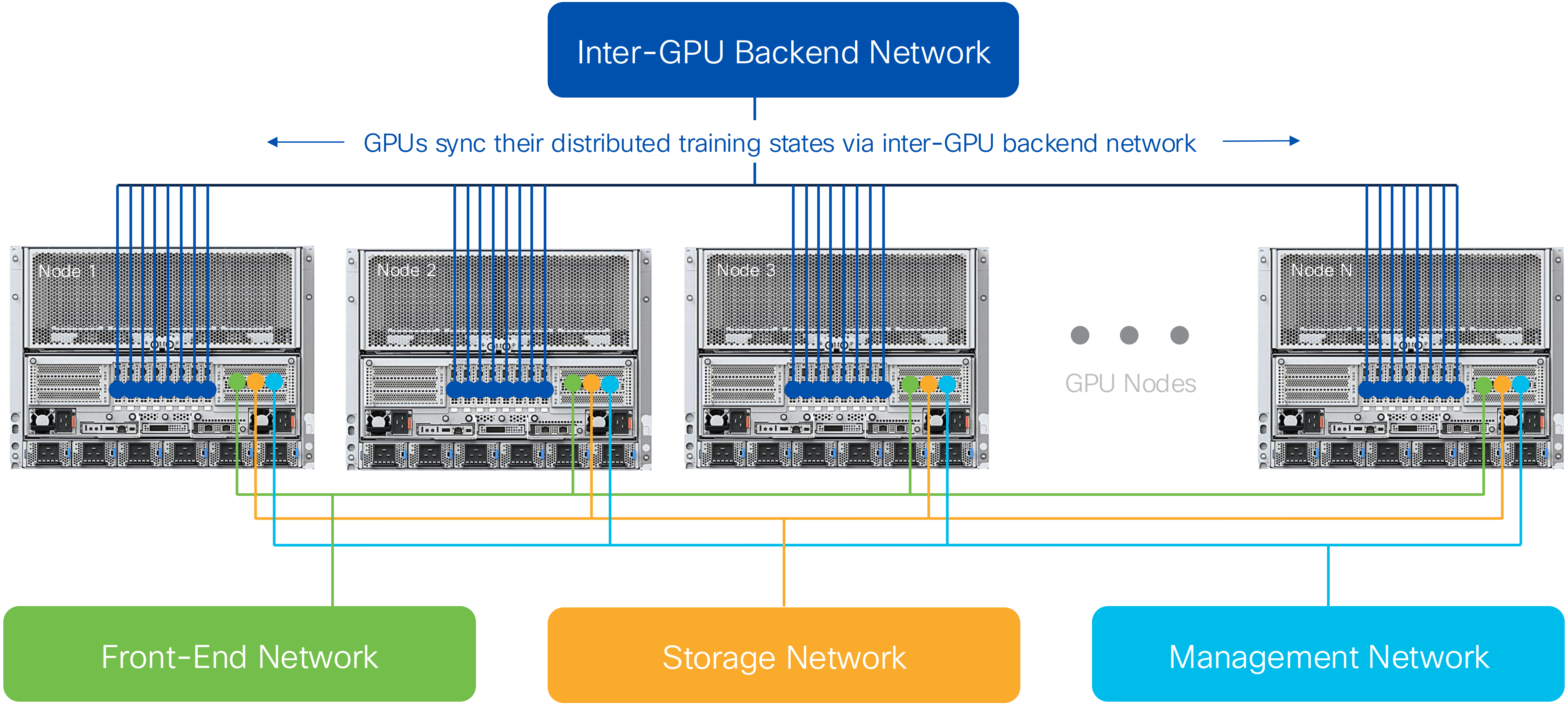
- Inter-GPU backend network: An Inter-GPU backend network connects the dedicated GPU ports for running distributed training. This network is also known as the back-end network, compute fabric, or scale-out network.

- Front-end network: A front-end network connects the GPU nodes to the data center network for inferencing, logging, managing in-band devices, and so on.

- Storage network: A storage network connects the GPU nodes to the shared storage devices providing parallel file system access to all the nodes for loading (reading) the data sets for training, and checkpointing (writing) the model parameters as they are learned. Some users may share the front-end network to connect storage devices, eliminating a dedicated storage network.

- Management network: A management network provides out-of-band connectivity to the devices of the AI/ML infrastructure, such as GPU nodes, network switches, and storage devices.

Networking Blueprint



Inter-GPU Backend Network

GPUs sync their distributed training states via inter-GPU backend network

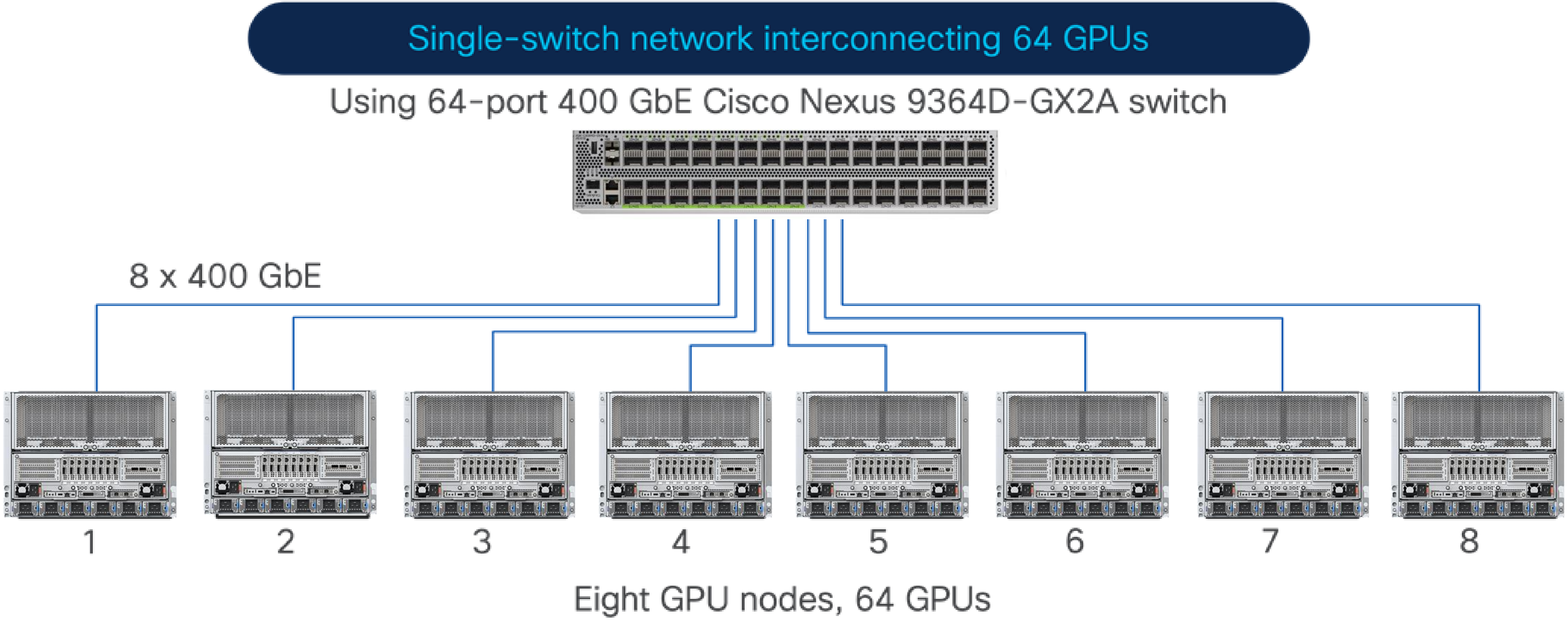
GPU Nodes

Front-End Network

Storage Network

Management Network

Designing a Smaller Inter-GPU Backend Network

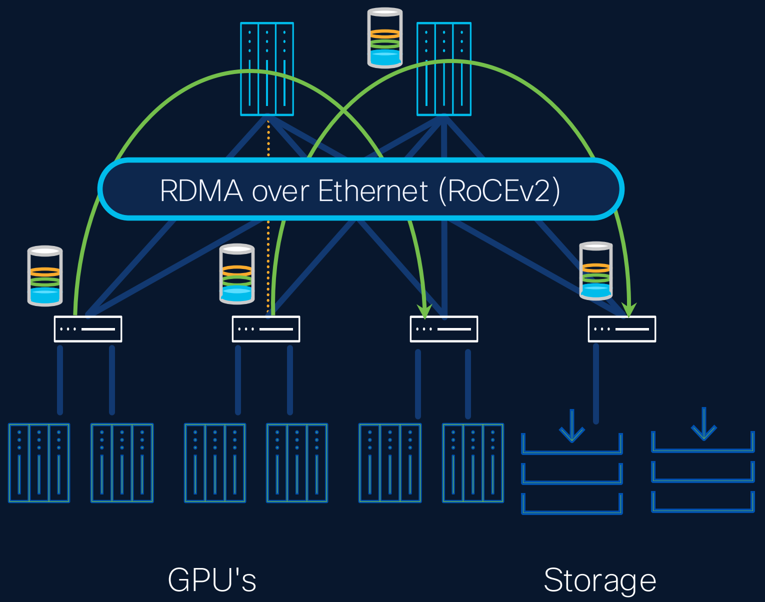


Single-switch network interconnecting 64 GPUs
Using 64-port 400 GbE Cisco Nexus 9364D-GX2A switch
8 x 400 GbE
1
2
3
4
5
6
7
8
Eight GPU nodes, 64 GPUs

- Smaller GPU clusters can use a single-switch network. For example, up to 64 GPUs can be interconnected using the 2 RU, 64-port 400 GbE, Cisco Nexus 9364D-GX2A switch (see above).

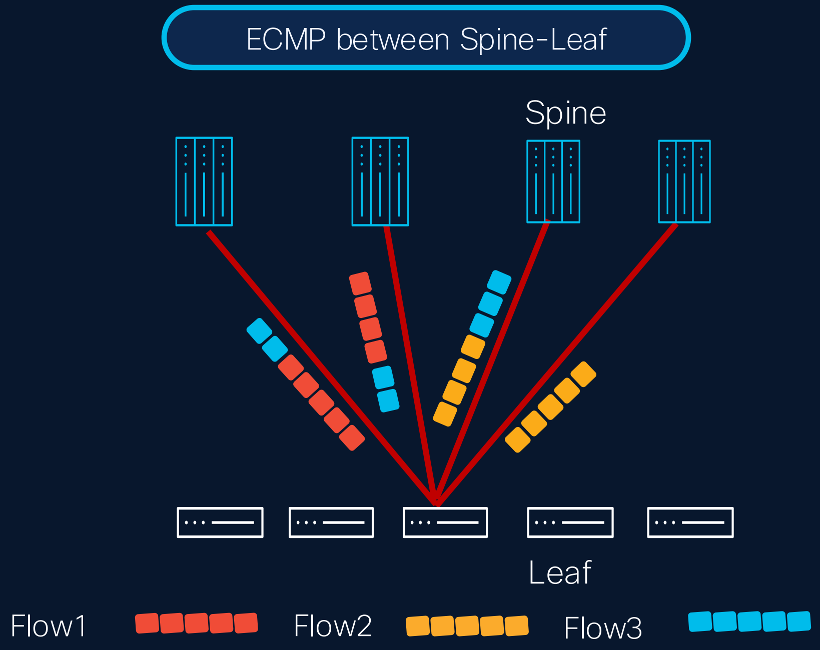
Nexus Dashboard

Automate your AI/ML network configurations



Manage network congestion

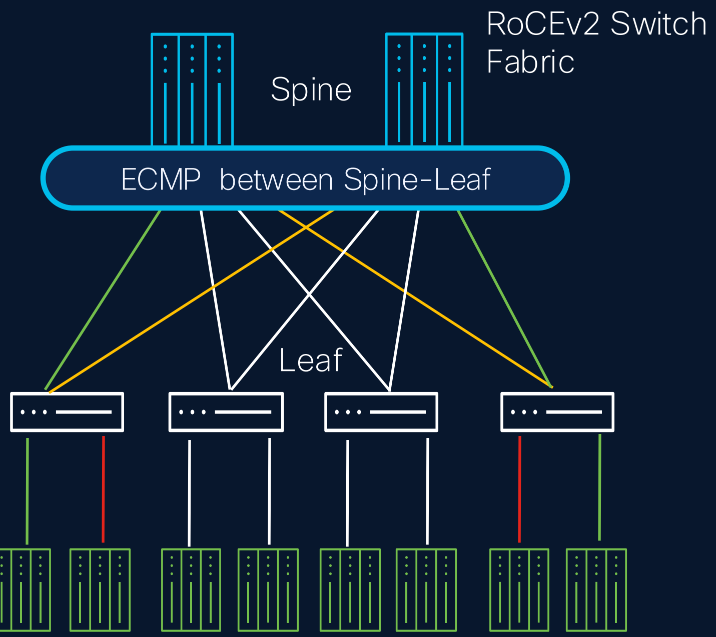
with Lossless Network (PFC + ECN)



Load balance flows/flowlets

based on link utilization

Better hashing results in AI fabrics
with uniform flow size and header information



Traffic efficiency through pinning rules

Map traffic from each downlink to the desired uplink

Allows efficient selection of Spines for communication between leaf and spines

Cisco AI Networking and Compute

Nexus Series with Nexus Dashboard

Minimize lock-in via an **open standards** RoCEv2 Ethernet fabric with intelligent buffering and streaming telemetry

Optimize training and inference network performance through deep visibility and actionable Insights

Accelerate and deliver deployments through automation with ready made AI templates

Unified Computing System (UCS)

Programmable modular system decoupling CPU, GPU, memory, storage and fabrics to deliver an AI perpetual architecture

Align AI sustainability targets to the compute platform that is sustainable by design

Accelerate and deliver AI infrastructure to the DC or Edge within minutes, not hours

Deploy AI anywhere with a full portfolio of AI-native infrastructure and software for the data center and the edge

AI Can be Fun

CISCO