Cracking the Code of AI in the Data Center

Jeff Kreis – SE, Cloud & AI Infrastructure

Mastin Bailey – AE, Cloud & AI Infrastructure

December 3, 2024

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

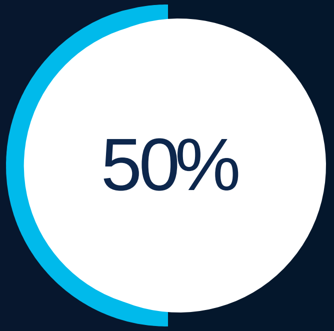
Sources: PWC, IDC

AI Deployment: Race Against Time

98% feel increased urgency over the past year.

CEOsand leadership are driving urgency for AI across ~50% of organizations.

85% say they have less than 18 months to deploy an AI strategy, or they will see negative business effects.



AI is a priority spend for IT budgets: 50% of companies say they've already dedicated 10-30% of their budget to AI.

AI Deployment: Balance Urgency and Readiness

98%

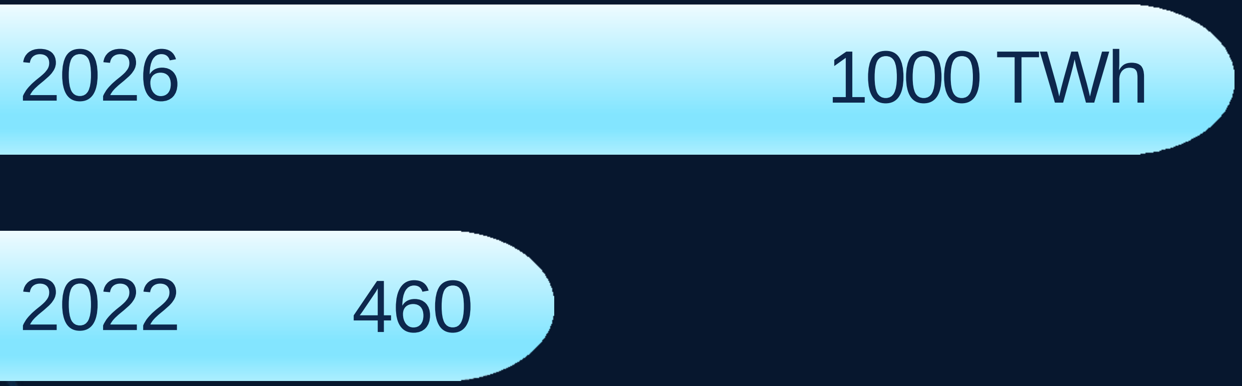
of global organizations reported an urgency to deploy AI powered technologies while only 14% are fully prepared to deploy and leverage AI*

*Cisco AI Readiness Report 2024

- Lines of business are driving AI infrastructure demand
- Significant GPU lead-times
- Nvidia dominates market and mindshare with AMD and Intel challengers
- InfiniBand and Ethernet compete for AI fabric

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.

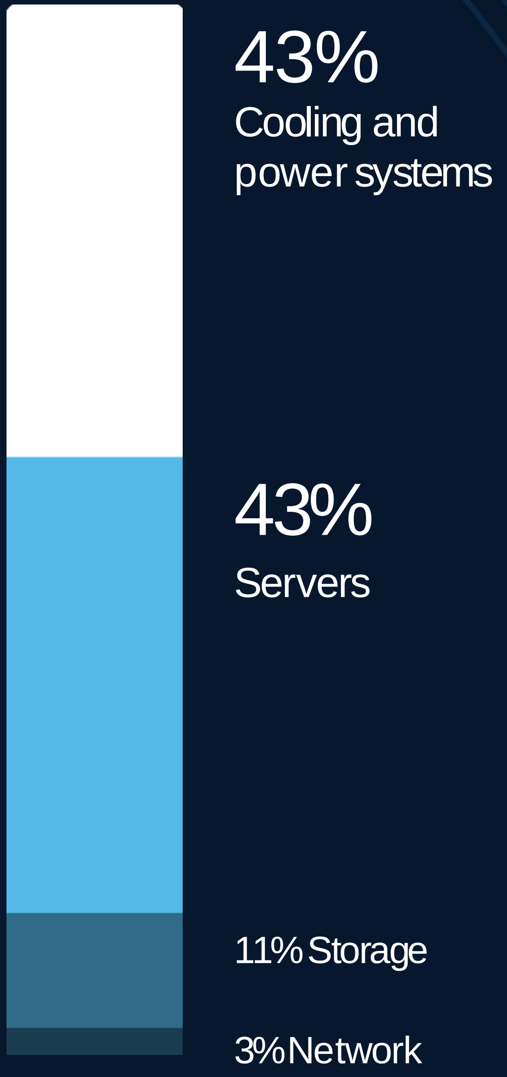
Today's data center accounts for:

1-2% of global electricity demand

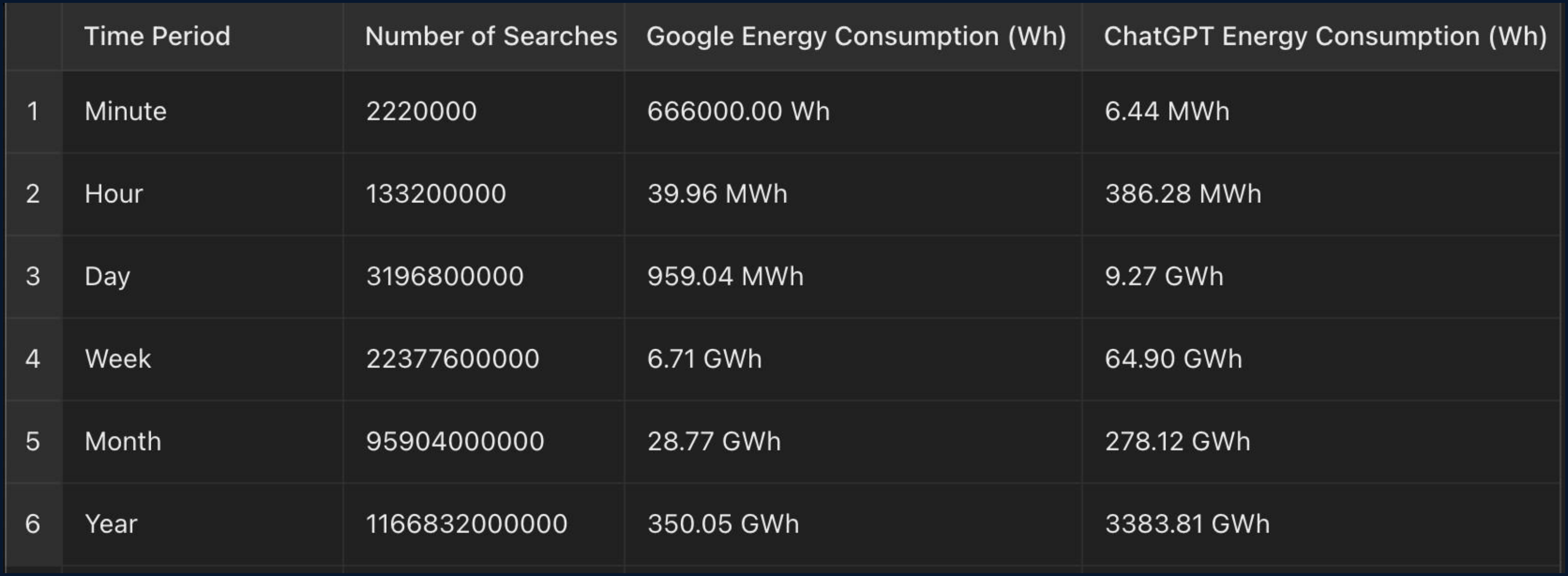
50X the power of atypical commercial office building

Every watt saved on computing results in:

1.55 watts saved at the facility level



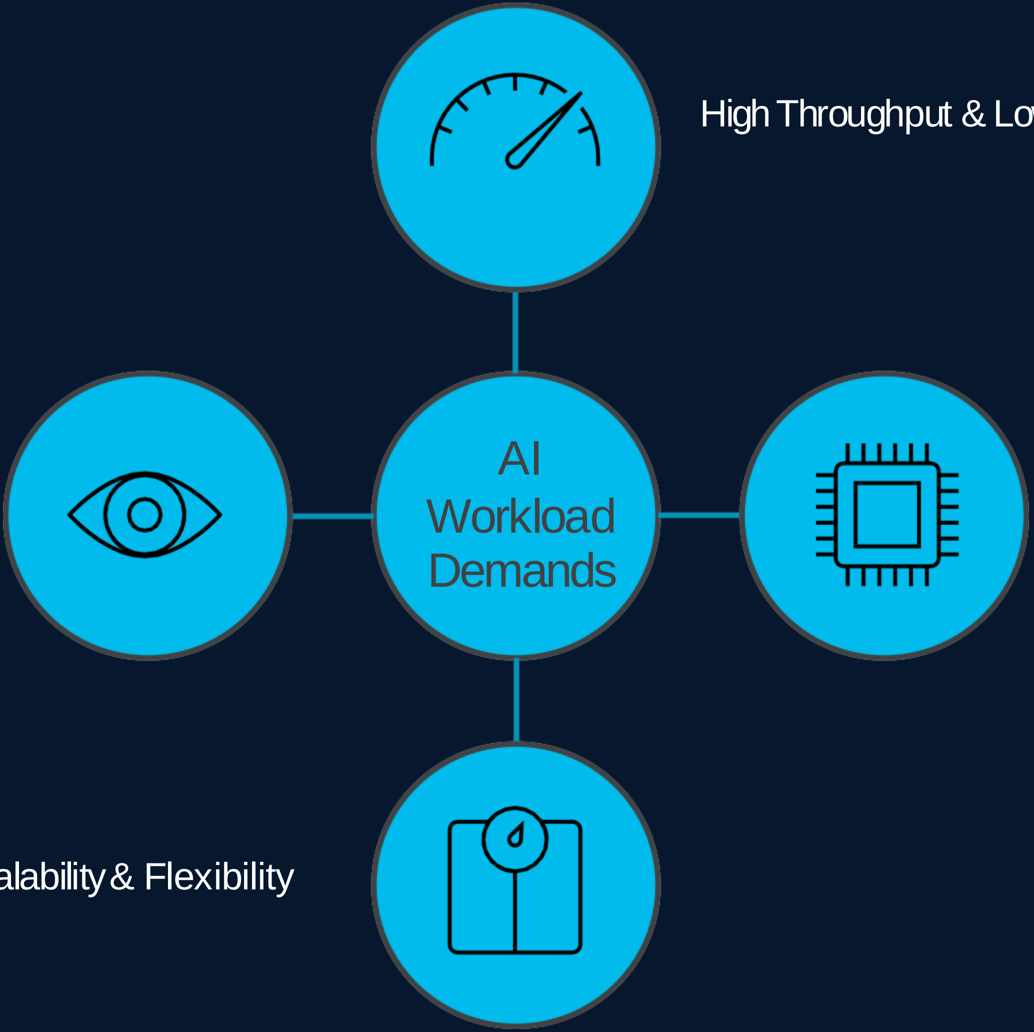
Thought Experiment



	Time Period	Number of Searches	Google Energy Consumption (Wh)	ChatGPT Energy Consumption (Wh)
1	Minute	2220000	666000.00 Wh	6.44 MWh
2	Hour	133200000	39.96 MWh	386.28 MWh
3	Day	3196800000	959.04 MWh	9.27 GWh
4	Week	22377600000	6.71 GWh	64.90 GWh
5	Month	95904000000	28.77 GWh	278.12 GWh
6	Year	1166832000000	350.05 GWh	3383.81 GWh

- The annual energy Google uses for its searches could power **63,936 Tesla Model 3s** to make a round trip across the USA.
- The annual energy ChatGPT uses for the same number of queries could power **618,048 Tesla Model 3s** for the same round trip.
- **ChatGPT would power nearly 10 times as many Tesla's on this journey compared to Google**.

Meeting the Demands of AI Workloads



High Throughput & Low Latency Network

Visibility & Adaptability

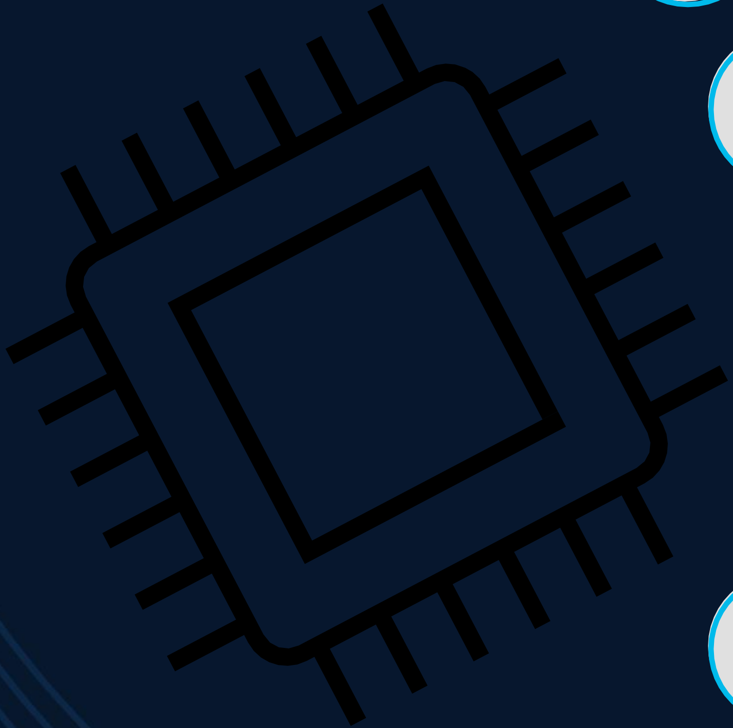
Accelerated Computing

Scalability & Flexibility

Why Traditional Data Centers Fail Short for AI

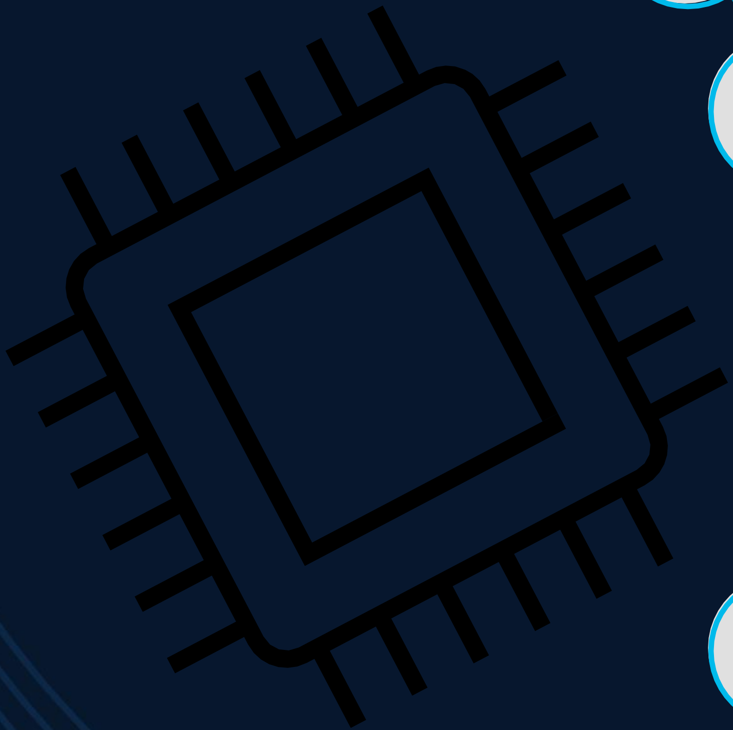
Traditional DC Attributes	AI Workload Challenges
CPU-focused Compute	Inefficient for Parallel Processing
Lossy Ethernet	Lossless Network
Fixed & Inflexible Infrastructure	Difficulty Scaling & Adapting to Dynamic Workloads
Conventional Power & Cooling	Power Hungry Accelerators
Low Visibility, Siloed Management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

AI Compute Solutions: The Case for GPUs



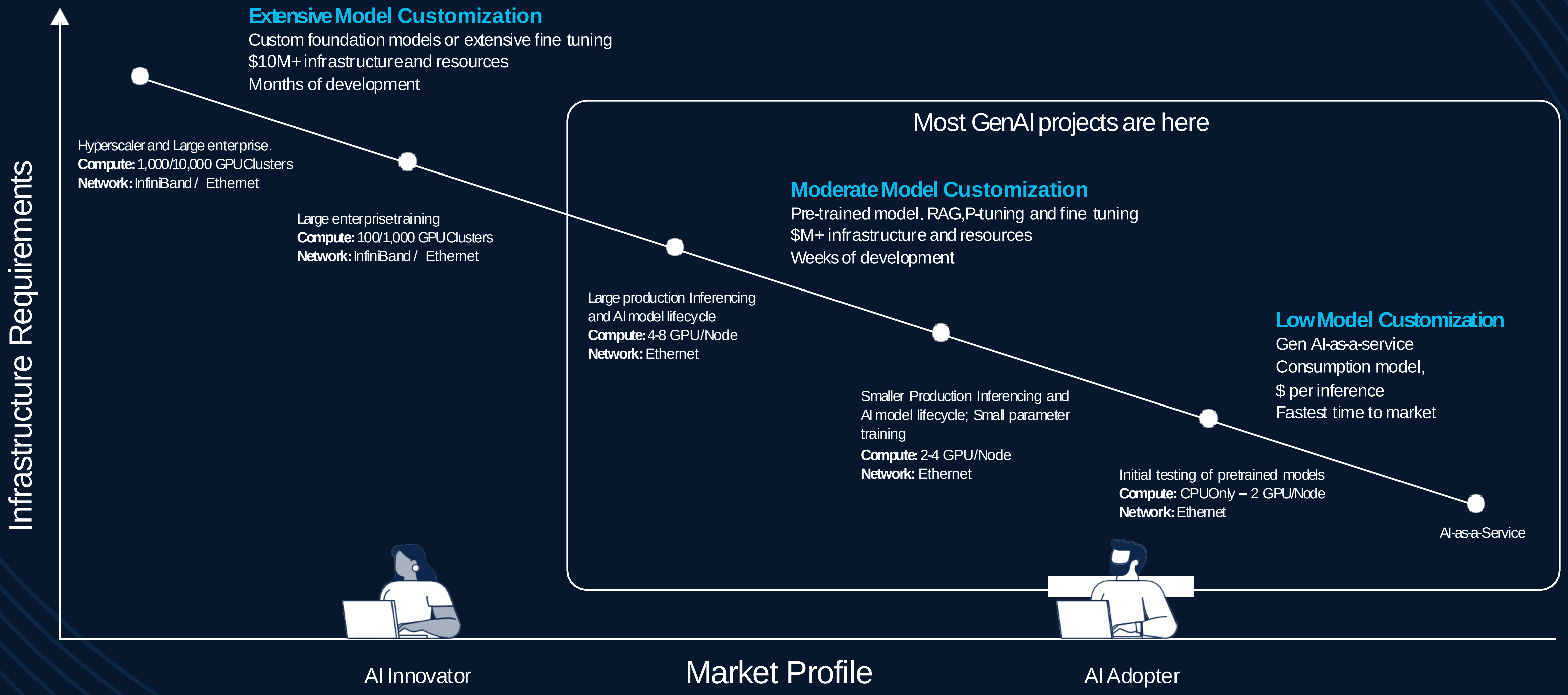
1. **Parallel Processing**: uses GPUs to handle 1000's of threads simultaneously.
2. **Deep Learning**: frameworks are optimized to utilize GPUs for efficiently training neural networks, involving matrix multiplications.
3. **Speed**: can significantly be reduced when training large neural networks with big data sets.
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.

CPUs in AI: Supporting the Heavy Lifting



1. **General-purpose computations**: like sequential processing, executing complex instructions, or moving data into memory are served well by CPUs.
2. **Data Preprocessing**: like data cleaning and feature extraction, can be efficiently handled by CPUs.
3. **Control Tasks**: used to manage the overall system, orchestrate the data flow, and control other components like GPUsare handled by CPUs.
4. **Training Smaller Models**: may not require GPUs, making CPUssufficient.
5. **Inferencing**: for some applications does not require intense parallel power, and CPUscan be used effectively.
6. **Cost-effectiveness**: for tasks that don’t benefit from parallelization, CPUsmight make more sense. They also allow you to get started without additional investment.

AI Infrastructure Requirements



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

AI Innovator

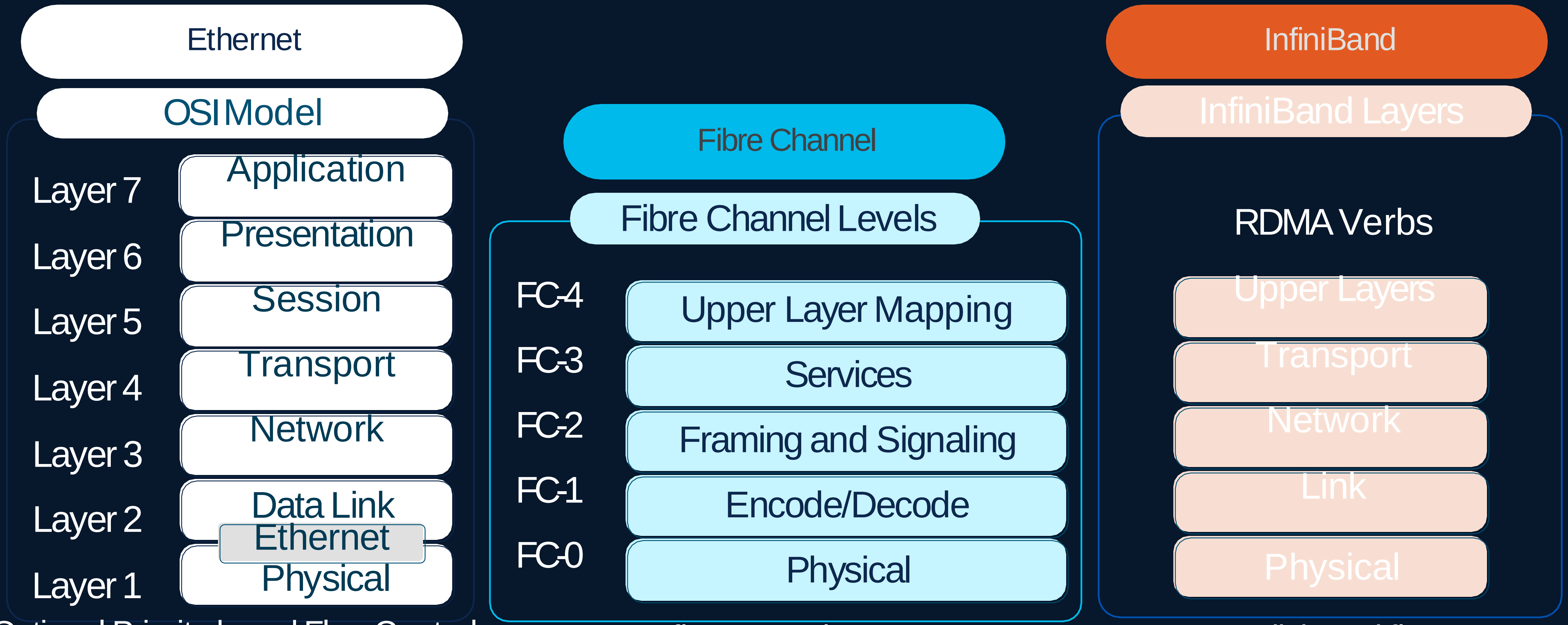
Market Profile

AI Adopter

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

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

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

B2B flow control. R_RDY, Credits, etc.

InfiniBand

InfiniBand Layers

RDMA Verbs

Upper Layers

Transport

Network

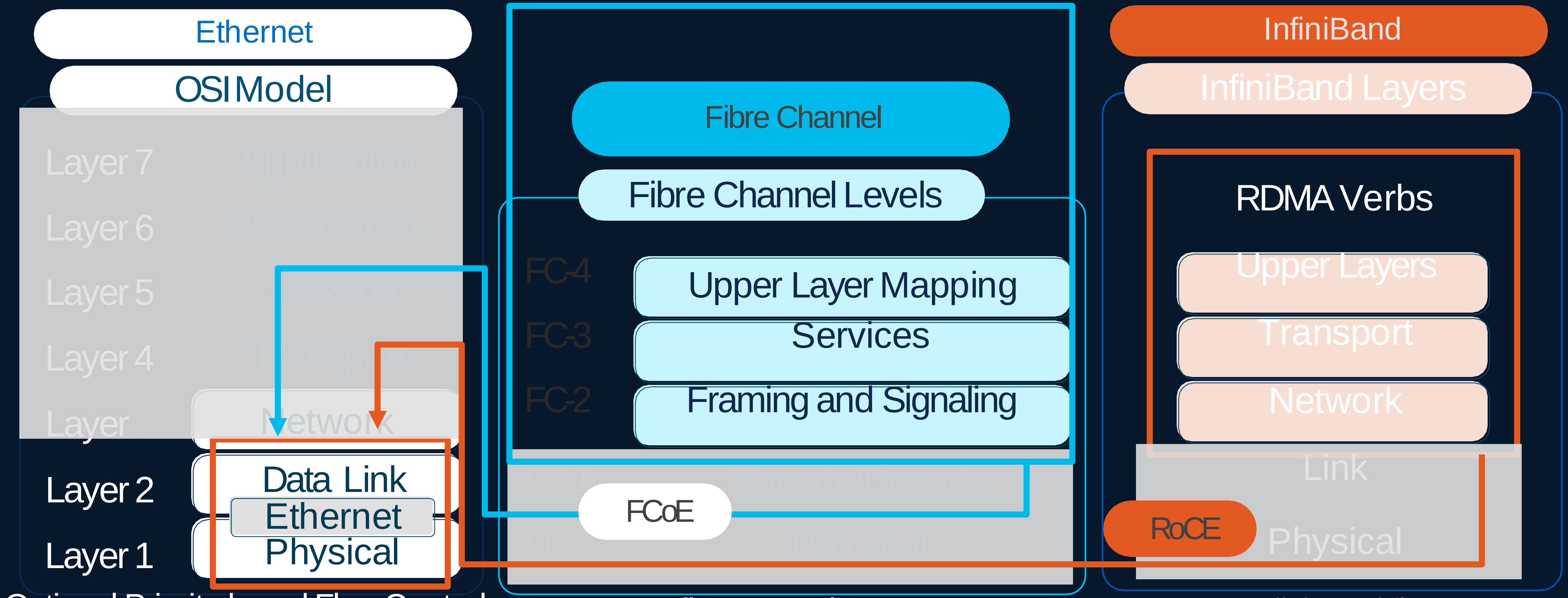
Link

Physical

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



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

B2B flow control. R_RDY, Credits, etc.

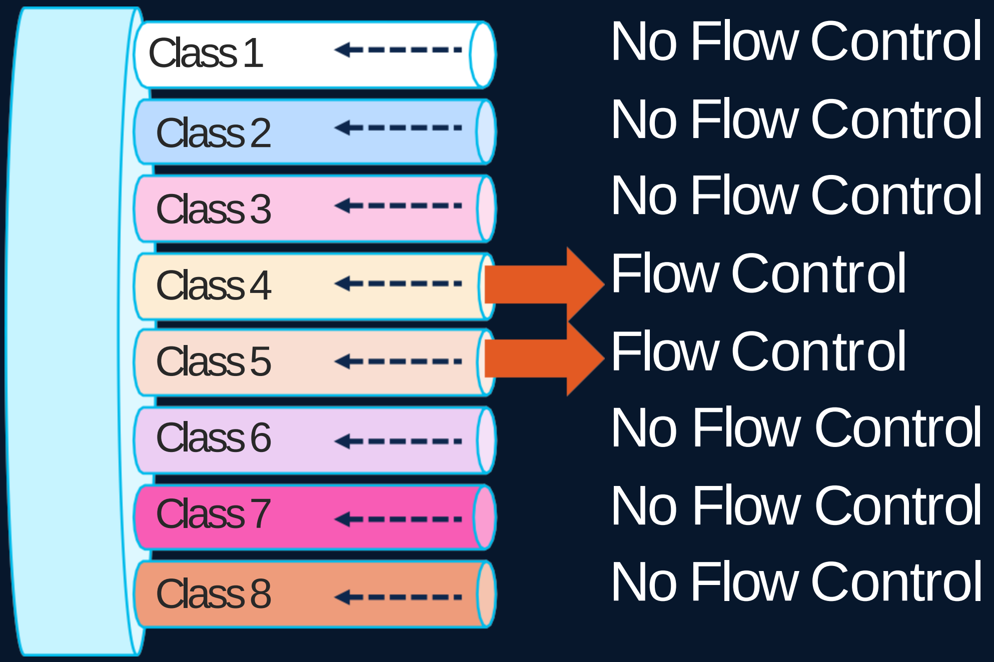
Credit-based flow control

Ethernet Flow Control

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

Traffic

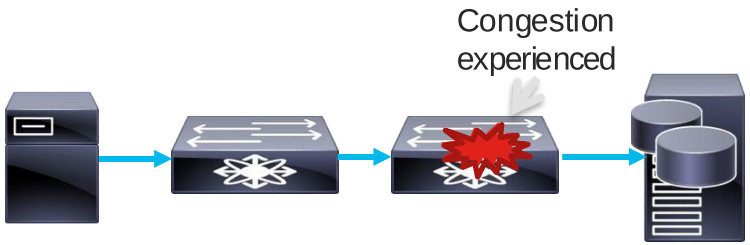
Priority-based Flow Control (PFC)



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

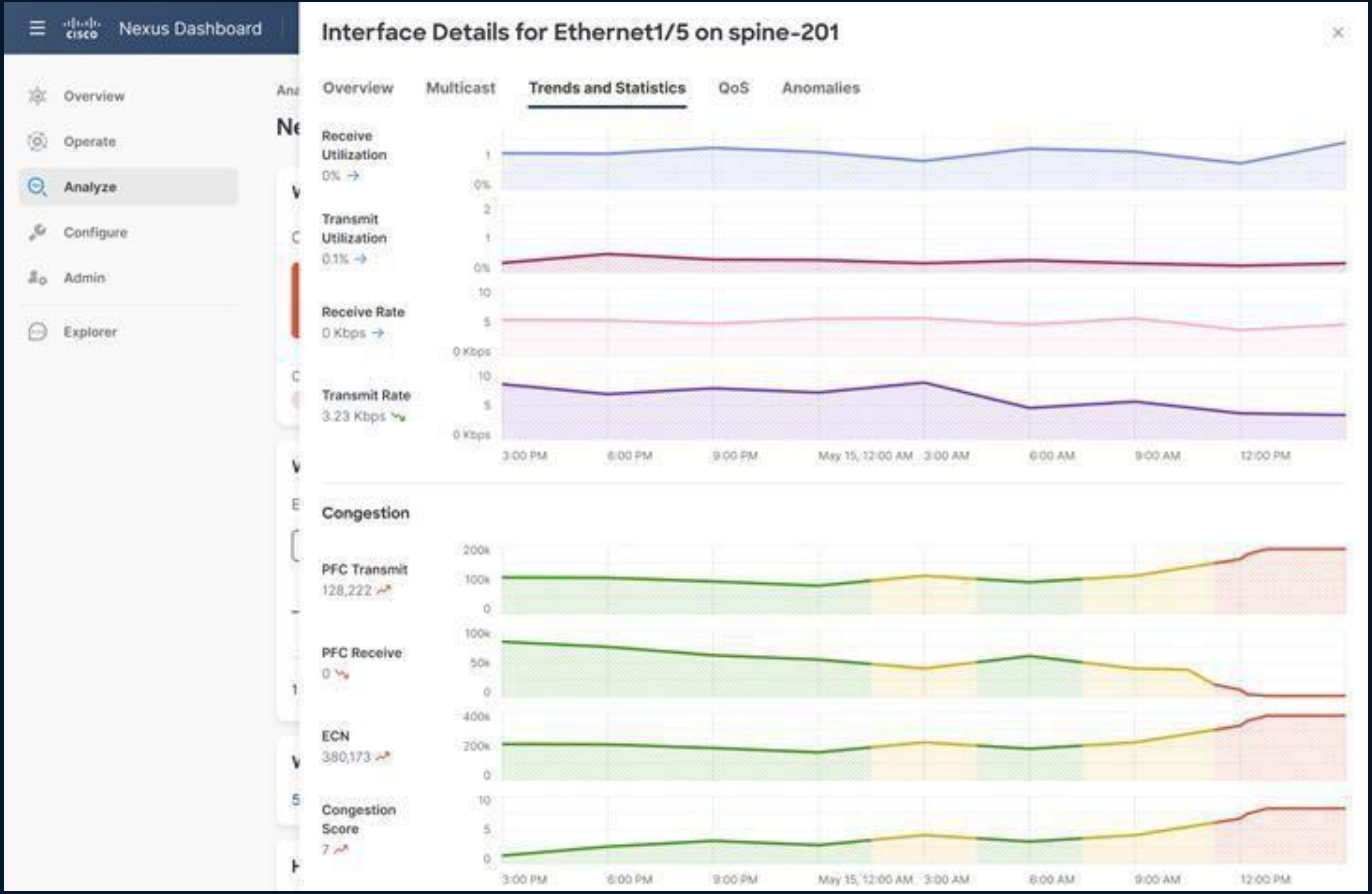
Explicit Congestion Notification

- 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



Interface Details for Ethernet1/5 on spine-201

Receive Utilization 0%

Transmit Utilization 0.1%

Receive Rate 0 Kbps

Transmit Rate 3.23 Kbps

Congestion

PFC Transmit 128,222

PFC Receive 0

ECN 380,173

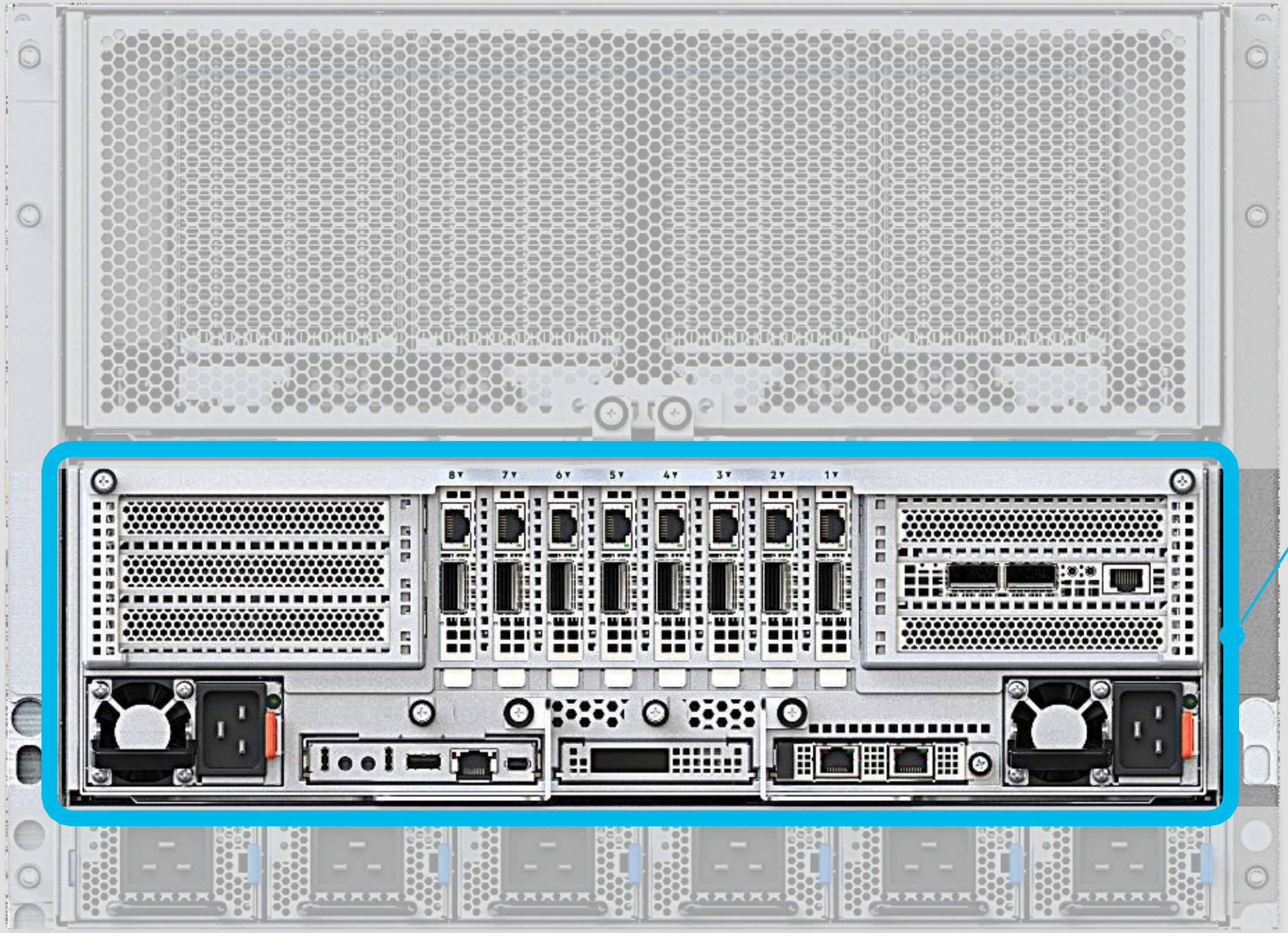
Congestion Score 7

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



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

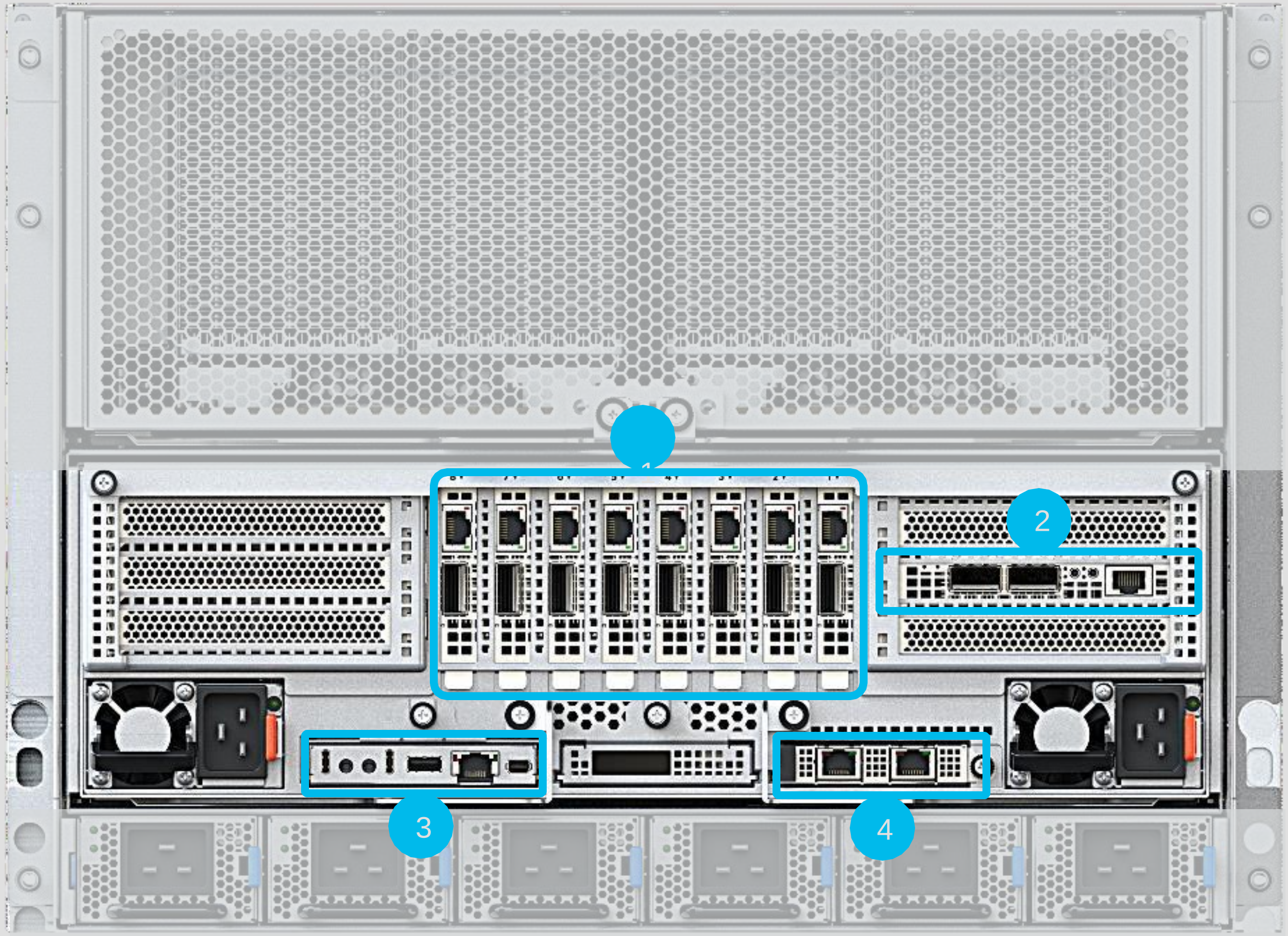
24x

DDR5RDIMMs

Up to 6,000 MT/S

128GBDIMMoption for some fixed configs coming soon

Server Rear View



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 OCP3.0 PCIe Gen5 x8 for X710 2 x 10G RJ45 NIC for additional north-south or host management traffic

Network Definitions

Multiple networks of an AI/ML Infrastructure…

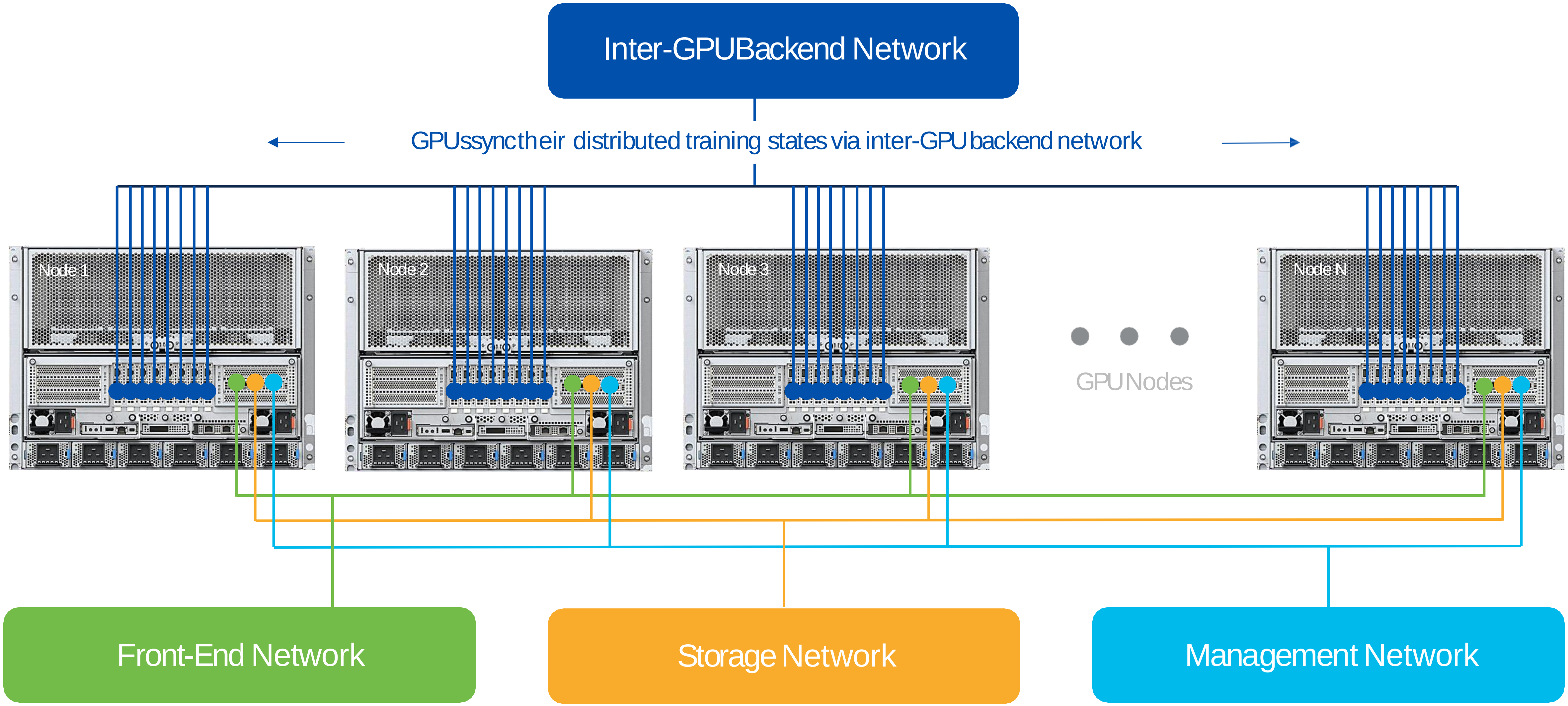
- Inter-GPU backend network: An Inter-GPU backend network connects the dedicated GPU ports for running distributed training. This network is also known as the back-end network, compute fabric, or scale-out network.

- 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-GPUBackend Network

GPUssynctheir distributed training states via inter-GPU backend network

Front-End Network

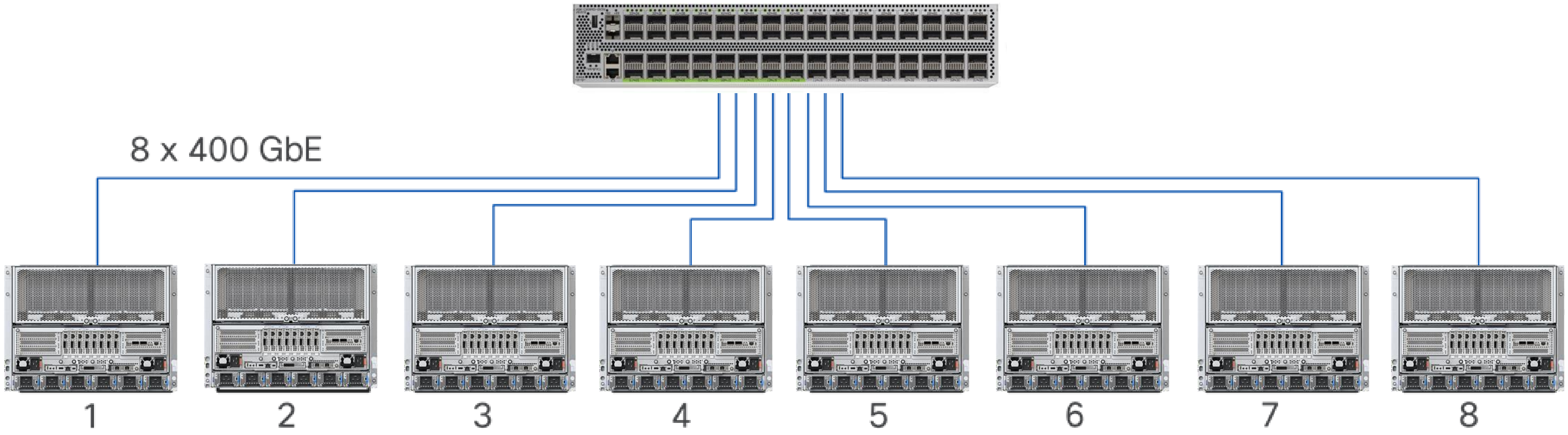
Storage Network

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



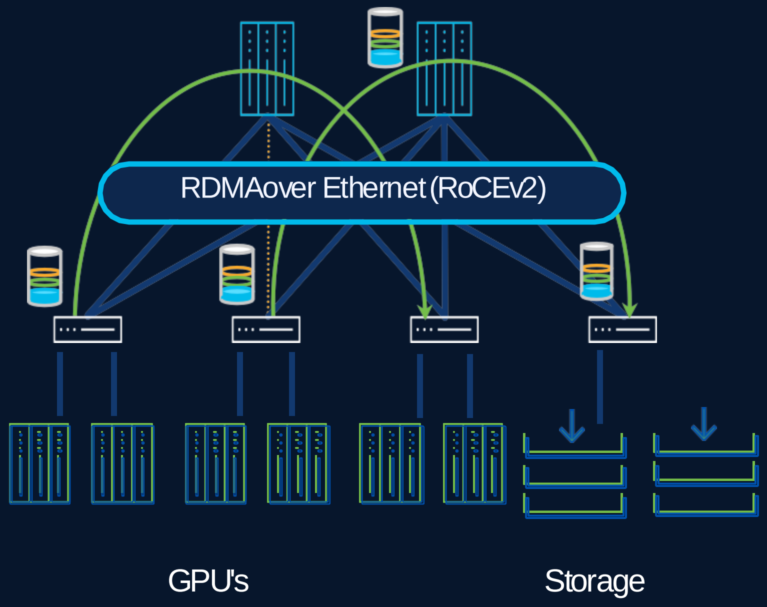
Eight GPU nodes, 64 GPUs

- 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

Dec ‘24 – Q1 ’25



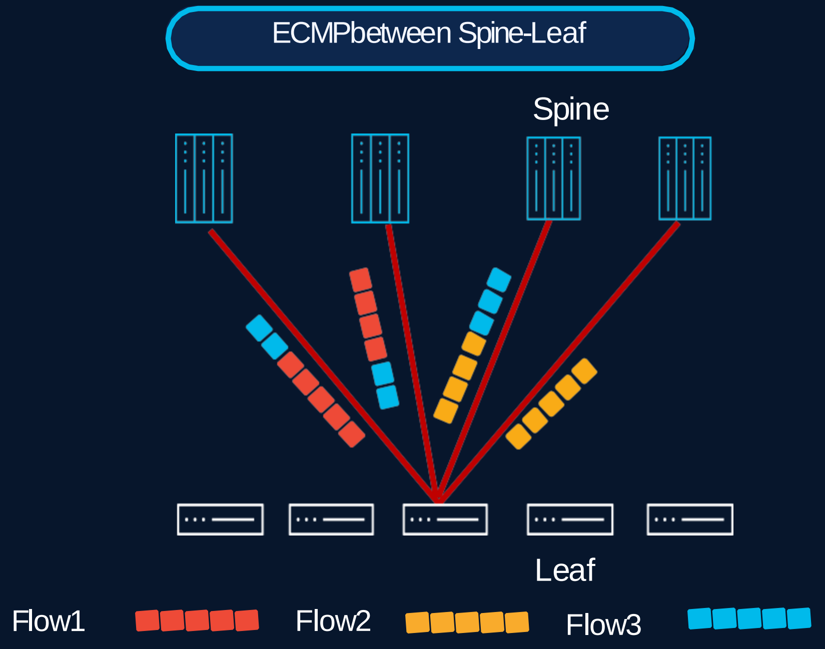
RDMAover Ethernet (RoCEv2)

GPU's

Storage

Manage network congestion

with LosslessNetwork (PFC+ECN)



ECMPbetween Spine-Leaf

Spine

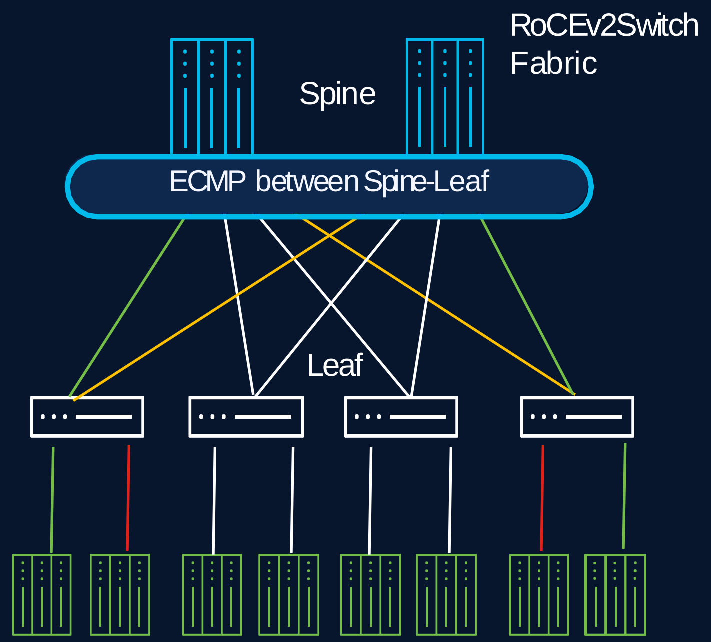
Leaf

Flow1 🟥🟥🟥🟥🟥 Flow2 🟧🟧🟧🟧🟧 Flow3 🟦🟦🟦🟦🟦

Load balance flows/flowlets

based on link utilization

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



RoCEv2Switch Fabric

Spine

ECMP between Spine-Leaf

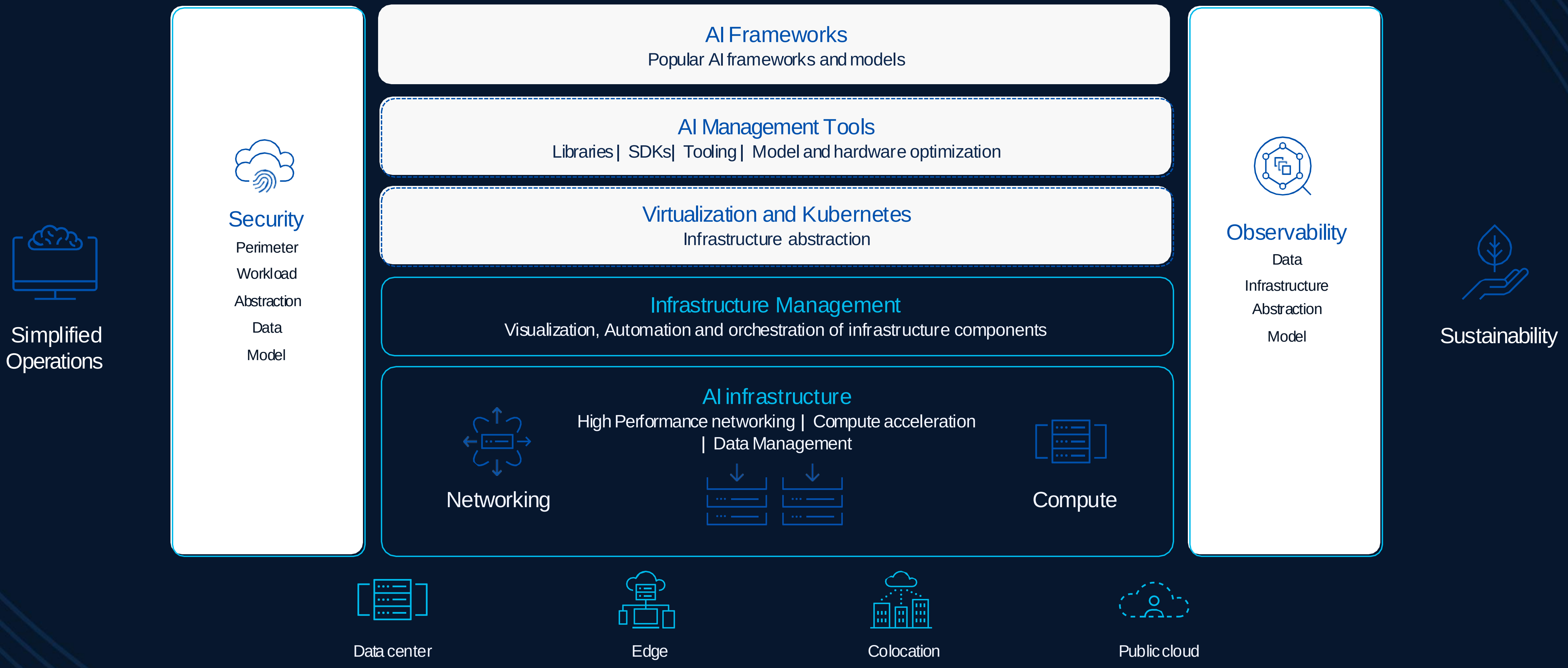
Leaf

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 Stack



Simplified Operations

Security

Perimeter

Workload

Abstraction

Data

Model

AI Frameworks

Popular AI frameworks and models

AI Management Tools

Libraries | SDKs| Tooling | Model and hardware optimization

Virtualization and Kubernetes

Infrastructure abstraction

Infrastructure Management

Visualization, Automation and orchestration of infrastructure components

AI infrastructure

High Performance networking | Compute acceleration | Data Management

Networking

Compute

Observability

Data

Infrastructure

Abstraction

Model

Sustainability

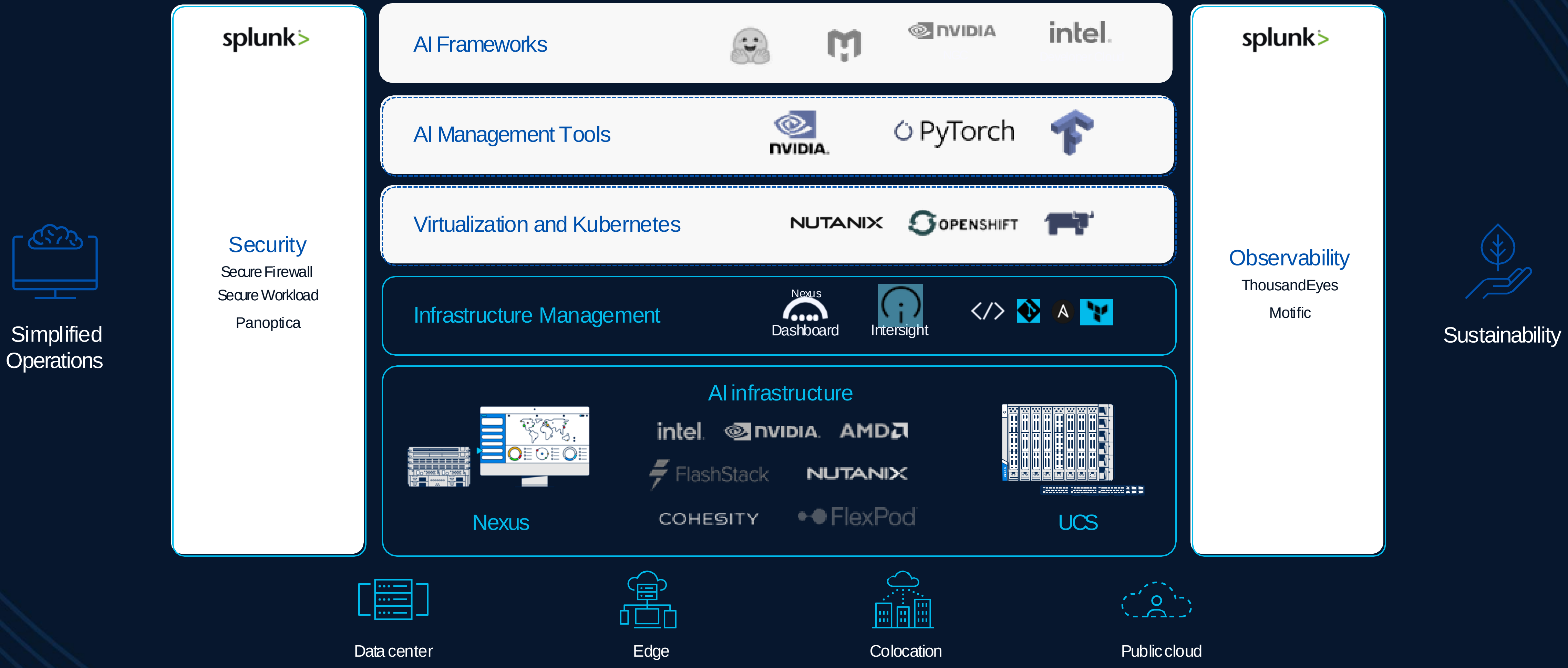
Data center

Edge

Colocation

Public cloud

Cisco AI Stack



Simplified Operations

Security
Secure Firewall
Secure Workload
Panoptica

AI Frameworks

AI Management Tools

Virtualization and Kubernetes

Infrastructure Management

Nexus Dashboard

Intersight

AI infrastructure

Nexus

UCS

Observability
ThousandEyes
Motific

Sustainability

Data center

Edge

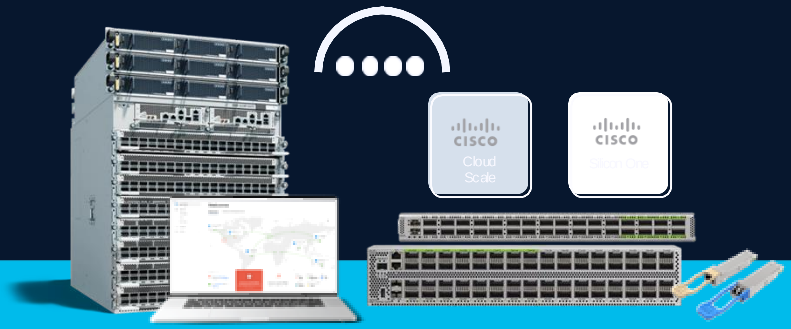
Colocation

Public cloud

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 DCor 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

Cisco AI Infrastructure **Simplified**

Enterprise grade AI solutions

Mainstream AI Infrastructure

Evolution not Revolution

Align AI initiatives with existing compute, network, storage and tooling investments

Accelerate AI Projects

Standardize and De-risk

Streamline AI deployments with validated reference architectures built upon best of breed hardware and software

Breadth and Scale of Data

Enabling AI Applications

Telemetry from 100s of millions of connected devices feed intelligence to the Cisco portfolio and your applications