Cisco’s Secure and Flexible Infrastructure for AI

DCN and Compute

Dave Prieto-CAI Account Executive

Modernizing data centers

Operational Simplicity

Centralize and simplify hyper-diverse and hyper-distributed data center operations

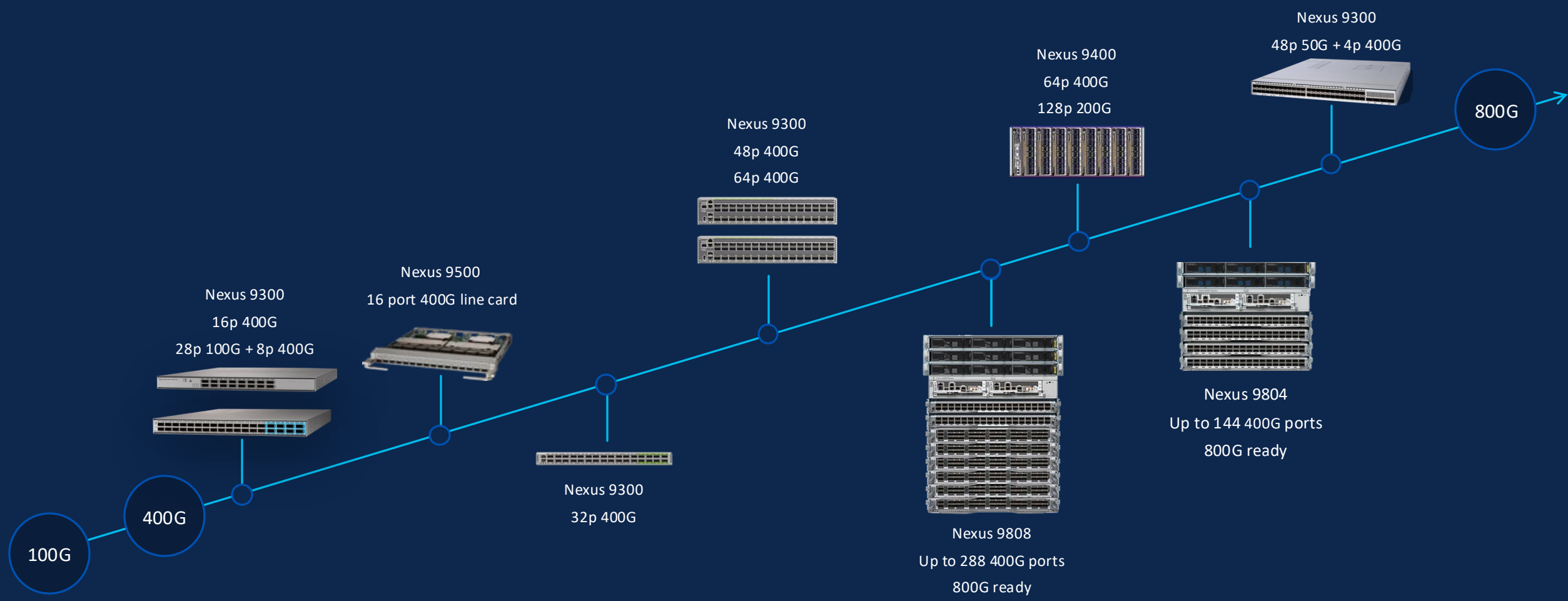
Artificial Intelligence

Empower AI/ML with 400/800Gbps and Flexible Compute Options

Platform Sustainability

Achieve net-zero goals faster while optimizing cost and power consumption

Power of Nexus 9000 – Cloud Scale and Silicon One



< 6.4T 25G SERDES | **6.4T to 25.6T 50G SERDES** | **>= 25.6T 100G SERDES**

Cloud service providers

Telco service providers

Enterprise

AI/ML networks

Cisco's 2-Fold AI Strategy

Using AI to maximize YOUR experience with **Cisco products**

In

Develop AI tools across the Cisco portfolio that help manage networks more effectively

- *Delivering better results*
- *Providing intelligent guidance*
- *Providing better security*
- *Solving day-to-day challenges*

Enabling **YOUR infrastructure** to support adoption of AI applications

On

Develop products that help accelerate YOUR adoption of AI for your business solutions

- *High-speed networking for AI training and inference clusters*
- *Flexible compute building blocks to build AI compute clusters*

Nexus Dashboard

Simple to integrate, Simple to Consume

Nexus Dashboard Insights

Visibility & Monitoring

- Topology
- Capacity utilization
- Control plane statistics
- Endpoint Visibility
- Multi-fabric support
- Custom Dashboards

Analytics & Correlation

- Flow Analytics – FTE
- AppDynamics integration
- Firmware Management
- Microburst detection
- Overlay traffic
- Anomaly analysis

Advisories & Tools

- Kafka messaging bus
- PSIRT notification
- TAC Assist
- Email notifications
- Field Notices

Cisco's 2-Fold AI Strategy

Using AI to maximize YOUR experience with **Cisco products**

In

Develop AI tools across the Cisco portfolio that help manage networks more effectively

- *Delivering better results*
- *Providing intelligent guidance*
- *Providing better security*
- *Solving day-to-day challenges*

Enabling **YOUR infrastructure** to support adoption of AI applications

On

Develop products that help accelerate YOUR adoption of AI for your business solutions

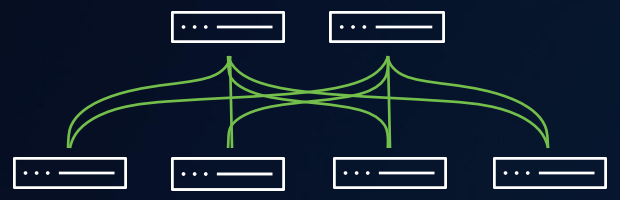
- *High-speed networking for AI training and inference clusters*
- *Flexible compute building blocks to build AI compute clusters*

Real Projects in PA

- **Predictive Policing:** Law Enforcement Agencies are using Palantir's Gotham software to analyze crime data and predict future crime hotspots.
- **Smart City Initiatives:** Pittsburgh is involved in several smart city projects that leverage AI for traffic management, public safety, and environmental monitoring
- **Fraud Detection:** The Commonwealth of Pennsylvania is using AI to detect fraud detection across various sectors. Ex)Dept of Labor(fraudulent unemployment claims), Dept of Revenue(tax evasion or fraud)
- **Chatbots for Citizen Services:** PA DHS introduced Olivia using a chatbot to answer frequently asked questions, help them apply for benefits, and provide information about services.
- **Predictive Maintenance:** The Pennsylvania Department of Transportation is exploring AI technologies to improve traffic flow, ehance road safety, and optimize public transportation systems.
- **Automated Permit Processing:** The City of Philadelphia is using AI to automate the permit processing process, which can save businesses time and money.
- **Early Disease Detection:** Healthcare Organizations are using AI to develop algorithms that can detect diseases like cancer at an early stage, when they are most treatable.

Cisco Data Center Networking

Cisco Nexus Dashboard



Private cloud-managed, flexible solution

General purpose data center solution

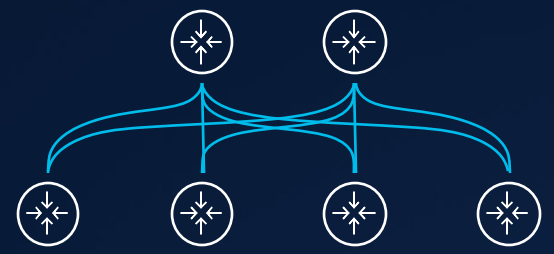
Greenfield and brownfield deployments

Any size spine-leaf data center (NX/ACI)

Nexus Dashboard for simplified operations

CloudScale and Silicon One based switches

Cisco Nexus Hyperfabric



Cisco cloud-managed, fully integrated

Easily design, order, deploy, monitor and upgrade fabrics

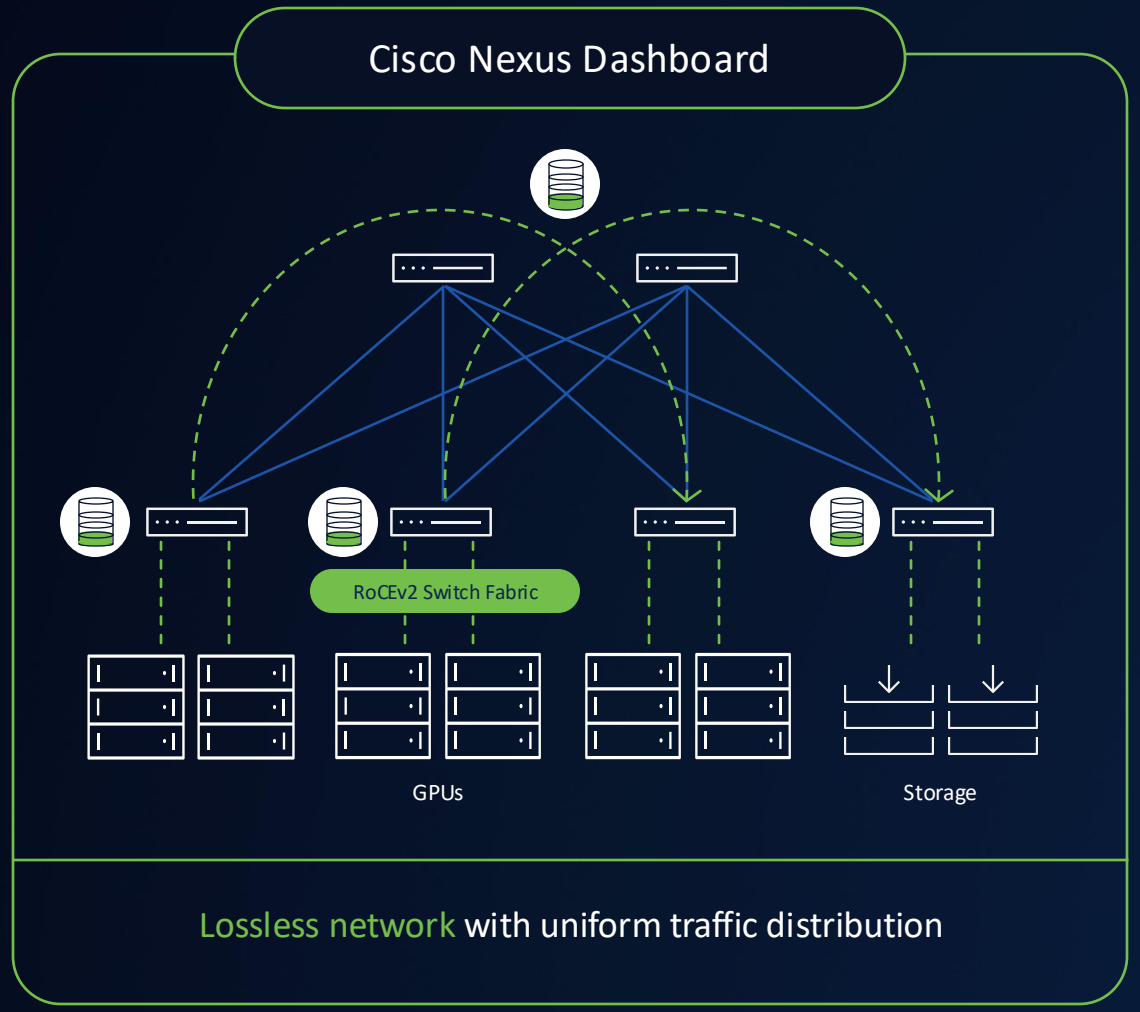
Purpose built vertical stack

Greenfield deployments only

Cisco 6000 Series switches (Silicon One)

Cisco GSX

Nexus Series Innovation



RDMA over Ethernet (RoCEv2)
Lossless network (PFC + ECN)

Non-blocking network

Low latency

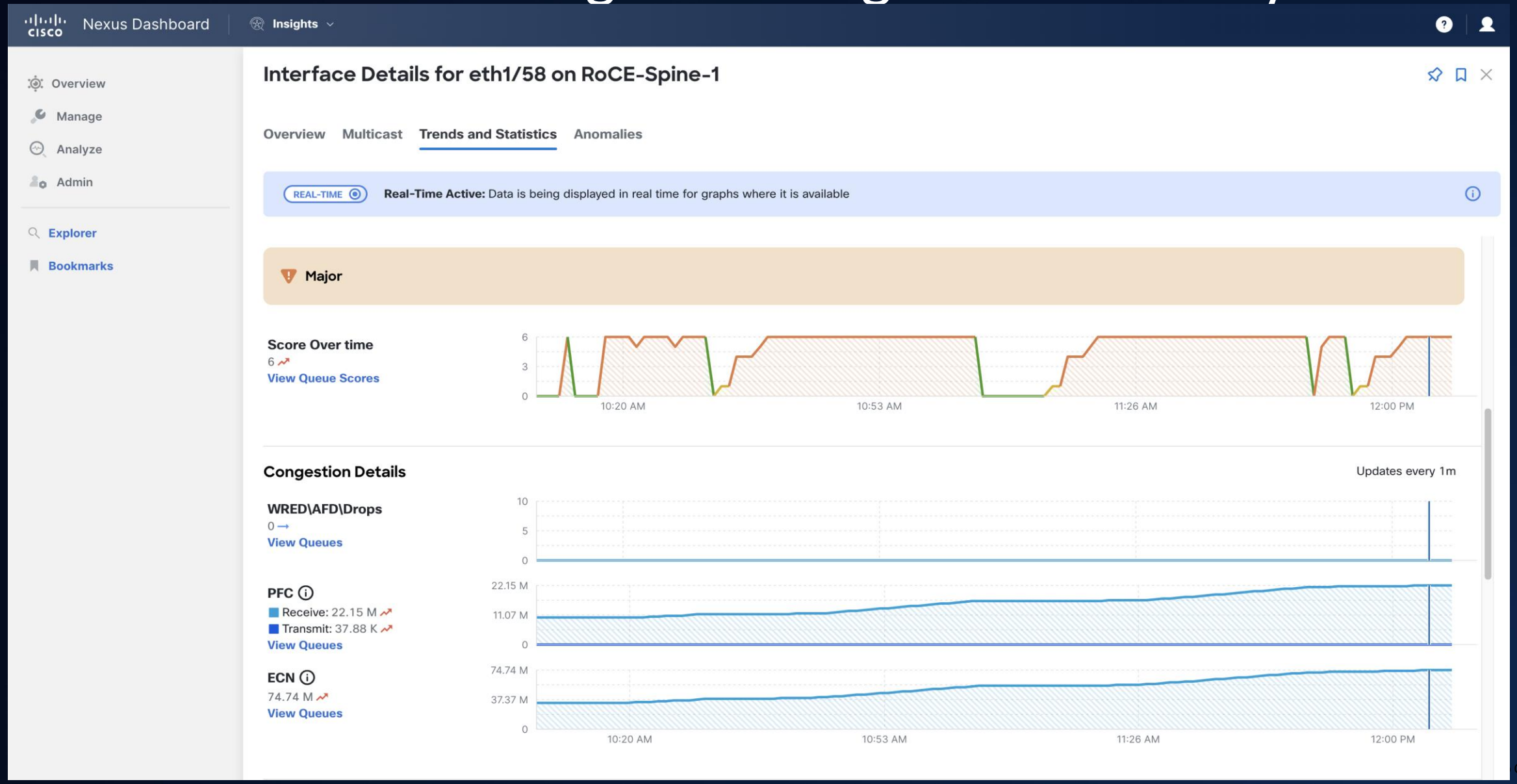
Congestion management

PFC: Priority Flow Control

ECN: Explicit Congestion Notification

Cisco GSX

Nexus Dashboard Insights – Congestion Visibility



Cisco Nexus Hyperfabric

AI Cluster

High-performance Ethernet

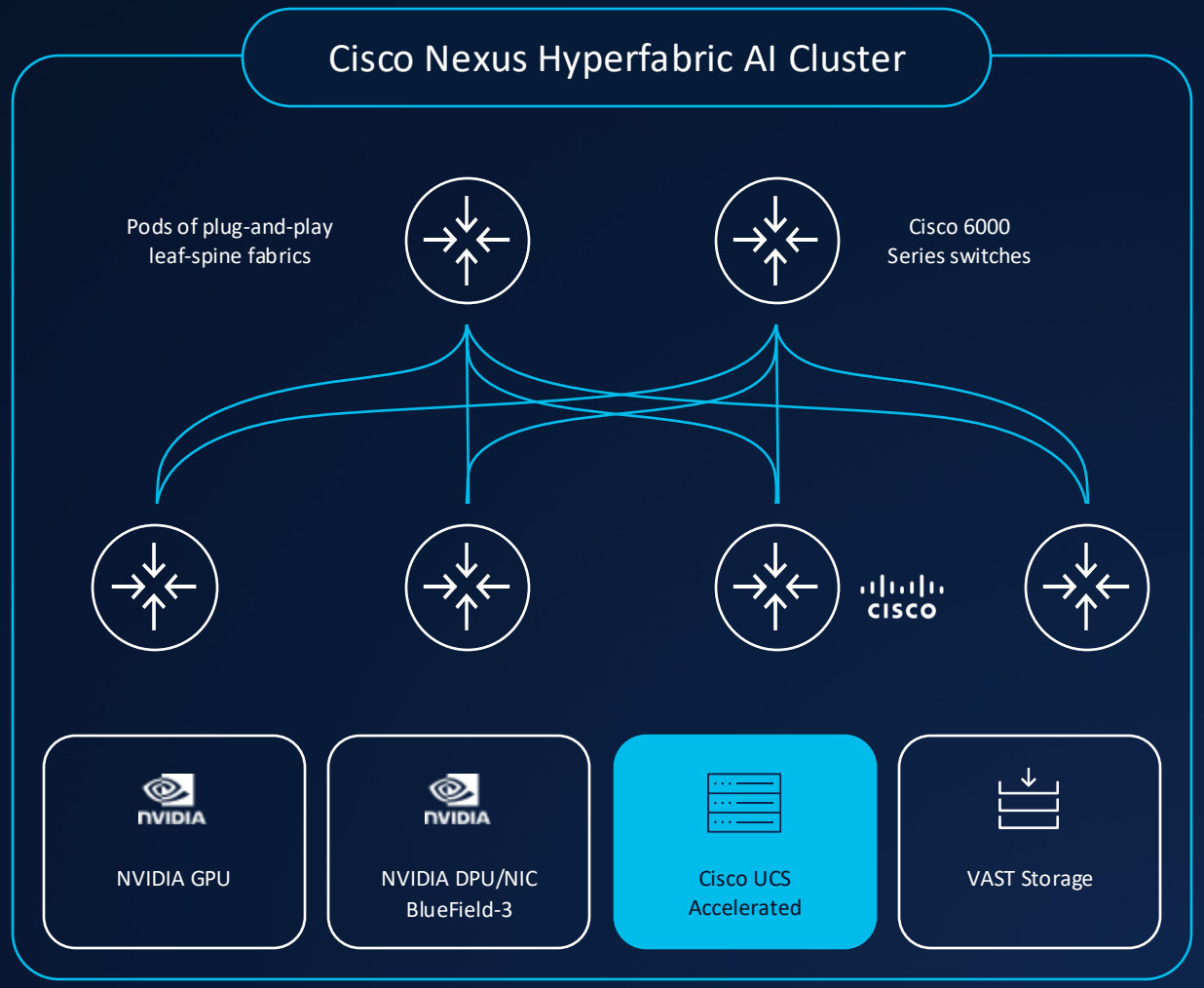
Cloud-managed operations

Unified stack including NVAIE

AI-native operational model

Democratize AI infrastructure

Visibility into full stack AI



Cisco GSX

The Blueprint For Today

Products and Services | Solutions | Support | Learn | Partners | Explore Cisco | Search

Products & Services / Cloud and Systems Management / Cisco Nexus Dashboard Fabric Controller / White Papers /

Cisco Data Center Networking Blueprint for AI/ML Applications

Updated: May 24, 2023

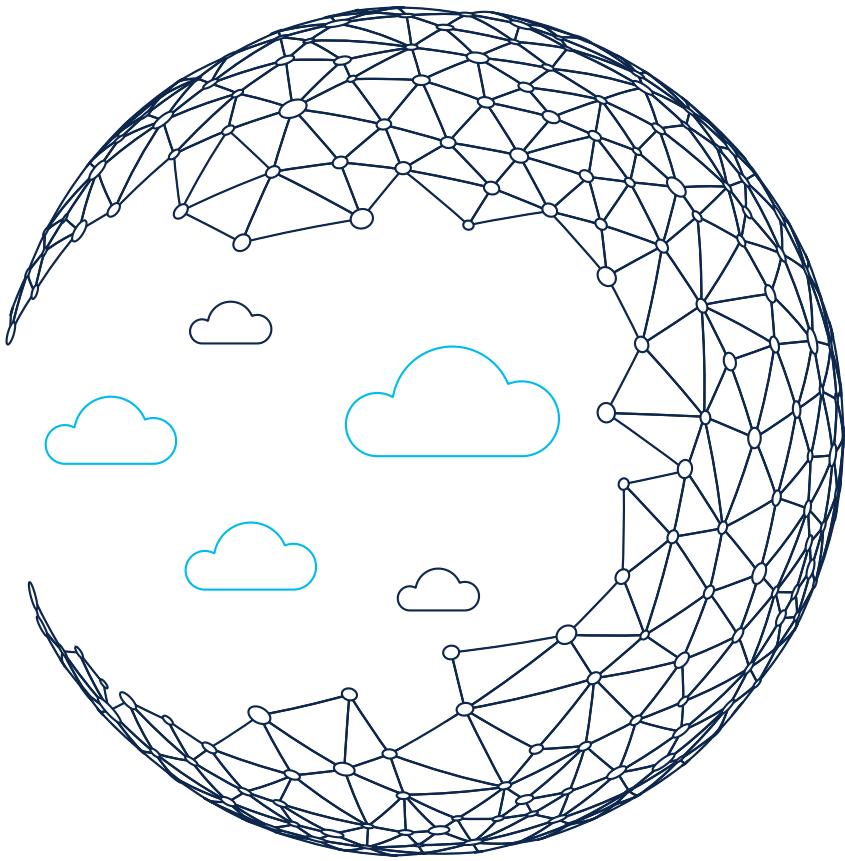
Bias-Free Language | Contact Cisco

Save | Download | Print

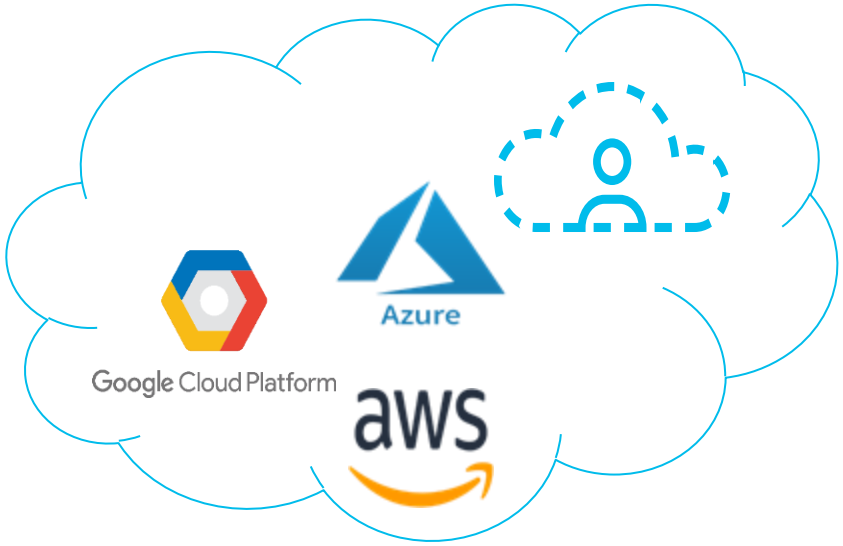
Introduction

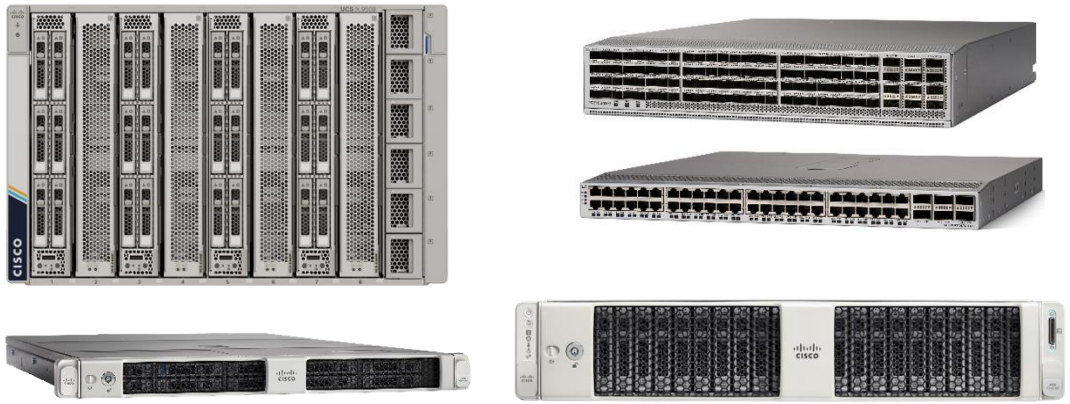
Table of Contents

Cisco UCS Compute



How will you run your AI/ML workloads?





Spark · hadoop · FlexPod A Cisco and NetApp Solution · FlashStack CI · PyTorch · SCALITY

kafka · NVIDIA NEMO · docker · kubernetes · TensorFlow · DL4J

AI/ML with UCS and UCS X-Fabric Technology

Model pre-training

I/O intensive

NVIDIA H100, A100, L40, L4

Fine-tuning pre-trained model

Compute intensive

NVIDIA H100, A100, L40, L4

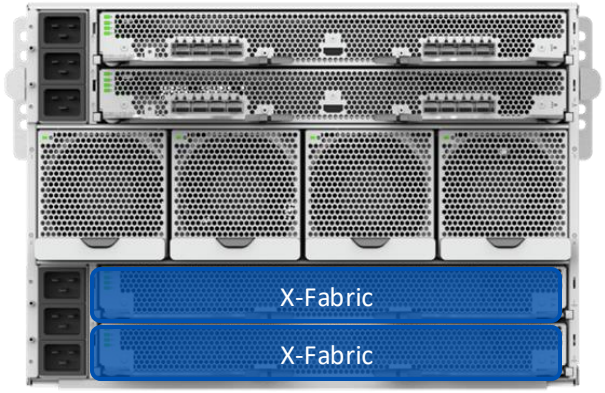
Inference / deploy

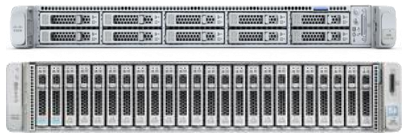
Latency sensitive

NVIDIA H100, A100, L40, L4

C240 M7

X210c M7 X440p





C2x0 M7

UCS® X-Fabric Technology

NVIDIA AI Enterprise

- Based on native PCIe Gen 4, upgradable to CXL in future
- Provides GPU acceleration to enterprise applications

Cisco Intersight®

Bringing High-Density GPU Servers to the Cisco UCS Family

Built for data-intensive use cases like model training and deep learning

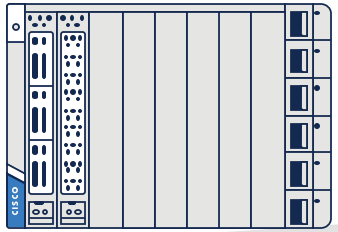
UCS Accelerated | UCS C885A M8 | NVIDIA HGX with 8 * H100 / H200 / MI300X GPUs | 2 AMD EPYC™ Processors

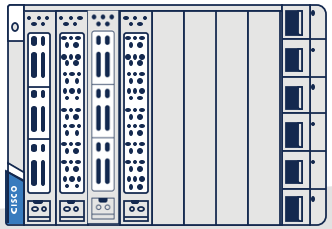
NEW

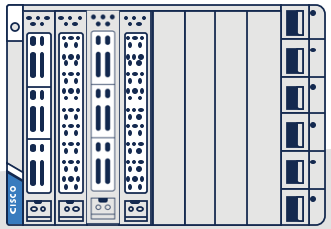
ORDERABLE OCTOBER 2024

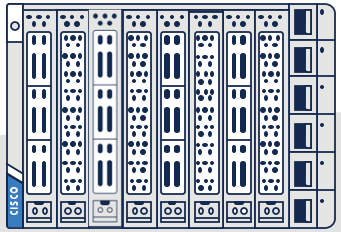
AI PODs for Inferencing

	Edge Inferencing (7B-13B Parameter)	RAG Augmented Inferencing (13B-40B+ Parameter)	Large-Scale RAG Augmented Inferencing (70B+ Parameter)	Scale-Out Inferencing Cluster (Inferencing Multiple Models)
Typical use case				
Hardware specification	**Small** **1x X210C compute node** • 2x Intel 5th Gen 6548Y+ • 512 GB System Memory • 5x 1.6 TB NVMe drives • 1x X440p PCIe • 1x NVIDIA L40S	**Medium** **2x X210C compute nodes** • 4x Intel 5th Gen 6548Y+ • 1 TB System Memory • 10x 1.6 TB NVMe drives • 2x X440p PCIe • 4x NVIDIA L40S	**Medium** **2x X210C compute nodes** • 4x Intel 5th Gen 6548Y+ • 1 TB System Memory • 10x 1.6 TB NVMe drives • 2x X440p PCIe • 4x NVIDIA H100 NVL	**Large** **4x X210C compute nodes** • 8x Intel 5th Gen 6548Y+ • 1.5 TB System Memory • 12x 1.9 TB NVMe drives • 4x X440p PCIe • 8x NVIDIA L40S









Performance and Scale

Inferencing Suite

X-Series-Direct

Edge infrastructure that is radically simplified

+

UCS X-Series Direct powered by Cisco Intersight

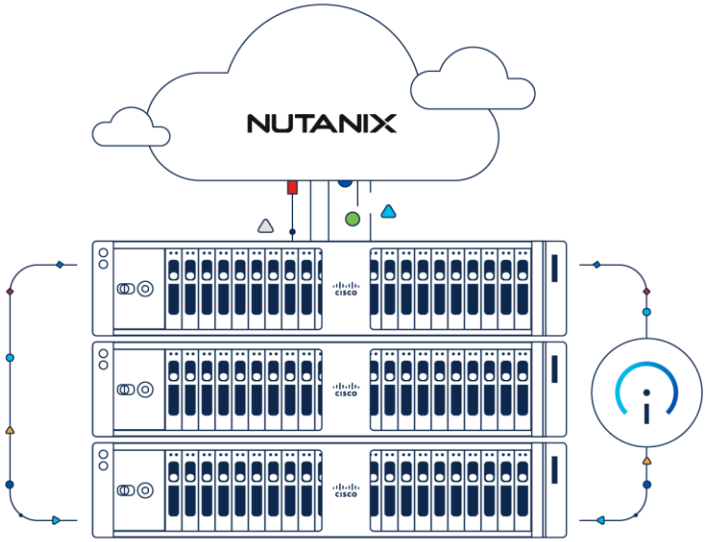
Cloud Operated

Unparalleled Flexibility

Future-ready

Cisco Compute Hyperconverged with Nutanix

Holistically built, managed, and supported by Cisco and Nutanix to eliminate complexity



Cisco Compute
and Networking

Simplify with cloud operations

Eliminate complexity with better visibility, control, and consistency across highly distributed environments

Accelerate IT transformation with more choice

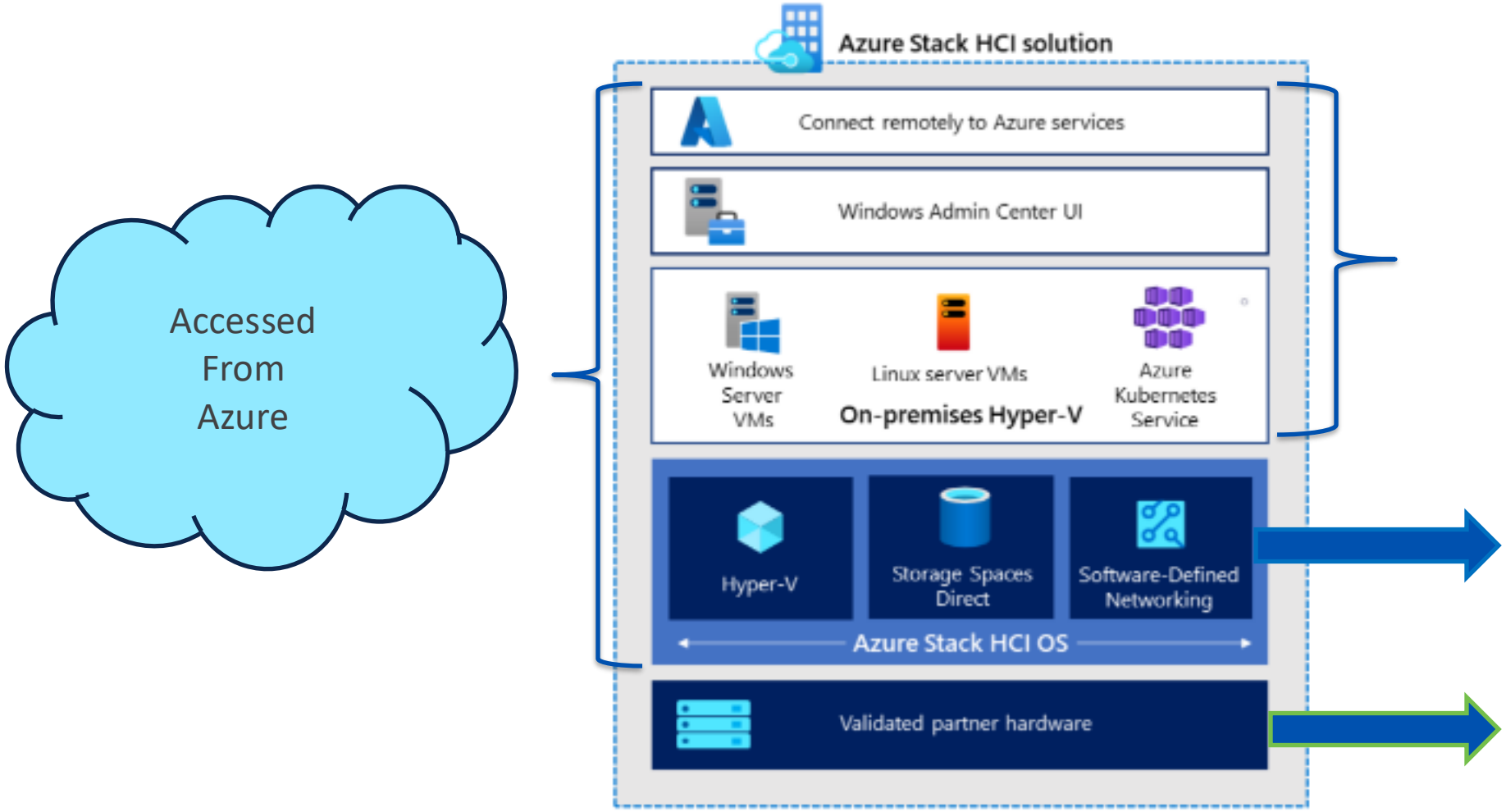
Effortlessly address modern apps and use cases with flexible deployment options, latest technologies, and multicloud connectivity

Resilient hyperconverged solution

Keep systems running with augmented support, resiliency, and security capabilities

Operate at scale	More choice and flexibility	End-to-end security	Comprehensive support	Easy to buy

Azure Stack HCI Software Architecture



Range of services become available after HCI OS:

- Azure Backup
- Azure Site Recovery
- Azure File Sync
- Azure Monitor
- Azure Update Management

- Includes product support
- Automatic patching & security updates
- Requires connection to Azure, every 30 days.

FlexPod for AI

How to buy | Partners | Log in | EN US

Products and Services | Solutions | Support | Learn | Explore Cisco | Search

Solutions / Design Zone /

Application 3: HPGMG (M...

Table of Contents

Conclusion

The amalgamation of High-Performance Computing (HPC) and Artificial Intelligence (AI) represents a powerful synergy that unleashes unprecedented computational capabilities, enabling us to tackle complex and data-intensive challenges with greater speed, accuracy, and efficiency. The combination of CPUs and GPUs with high-speed data fabric with end-end 100GbE network is essential for achieving optimal performance, scalability, and responsiveness.

Here's the importance of each component because it allows for the best of HPC and AI worlds:

- **Diverse Workload Support:** CPUs are essential for handling diverse tasks, including system management, control flow, and sequential processes, making them crucial for both HPC and AI infrastructure.
- **Parallel Processing:** GPUs are vital for parallelizable workloads, such as deep learning and scientific simulations, where their massive parallel processing power accelerates complex calculations.
- **Task Offloading:** Combining CPUs and GPUs allows for intelligent task distribution, enabling CPUs to offload parallel workloads to GPUs for enhanced efficiency and speed.
- **Optimal Performance:** Together, CPUs and GPUs offer a balanced and high-performance computing environment, capable of handling a wide range of workloads seamlessly.
- **Energy Efficiency:** CPUs are generally more power-efficient for certain tasks and are important for overall system management. GPUs, on the other hand, excel in computational throughput. Combining the two can lead to better energy efficiency and performance.
- **Fast data pipeline:** Data intensive workloads of HPC and AI often involve large datasets. A 100GbE network provides an extensive data pipeline, ensuring efficient data exchange between CPUs, GPUs, storage rapidly and without bottlenecks, improving overall performance.
- **Low Latency:** Low-latency communication is crucial for HPC and AI workloads, as it reduces the time spent waiting for data transfers and results in faster overall processing.
- **Scalability:** High-speed networking supports the scaling of resources, enabling you to add more compute nodes, GPUs, or storage as needed for growing workloads.
- **Resource Utilization:** CPUs and GPUs are fully utilized as data moves quickly between them, minimizing idle times and maximizing the overall system efficiency.

In this solution study, we tested various application (use cases) targeted for weather simulation (miniWeather), Nuclear Engineering - Radiation Transport (Minisweep), and Cosmology, Astrophysics, Combustion (HPGMG).

We documented recommended tunable parameters to achieve balanced configuration amongst compute, network and storage components and proved near linear scalability of the solution as the size of the cluster grew from 1 to 8 node.

CISCO

The bridge to possible