

Modernizing the DC to Support AI

Building, Observing, and Securing the Next Gen Infrastructure

Cloud and AI Infrastructure Specialist SE

Jason Powell Eliot Ngwa



Cisco powers how people and technology work together across the physical and digital worlds



AI-Ready Data Centers



Future-Proofed Workplaces

Secure Global Connectivity



Digital Resilience

Accelerated by Cisco AI

Every organization's AI approach and needs are different

Build the Model | Training

Optimize the Model | Fine-tuning and RAG

Use the Model | Inferencing



And applications have evolved

From simple and deterministic to complex and adaptive

Pre-AI Period

- Deterministic logic
- Predictable behavior
- Static query / response

Large Language Models (LLMs)

- Natural language input / output
- No workflows / ~1 span

Retrieval-Augmented Generation (RAG) with LLMs

- Hybrid and multi-cloud environments / Microservices
- Linear workflows / ~10s of spans

Agentic Applications

- Autonomous task completion
- Unpredictable behavior
- Multi-step actions

Past

Present

Future

The emphasis in AI infrastructure is shifting from training massive models to optimizing inference

Brad Lightcap



“The industry is realizing that inference—not just training—is where the economic value of AI is created.”

Sarah Guo



“We’re seeing a fundamental forking of AI scaling: the old model was ‘bigger is better’ for training, but now inference efficiency is becoming the real battleground.”

Jonathan Ross
groq

“The shift to inference marks a new phase of AI’s power—moving from potential to real-world impact at scale.”

Aidan Gomez
cohere

“Training pushed AI forward, but inference is where the real-world performance gaps are being solved now.”

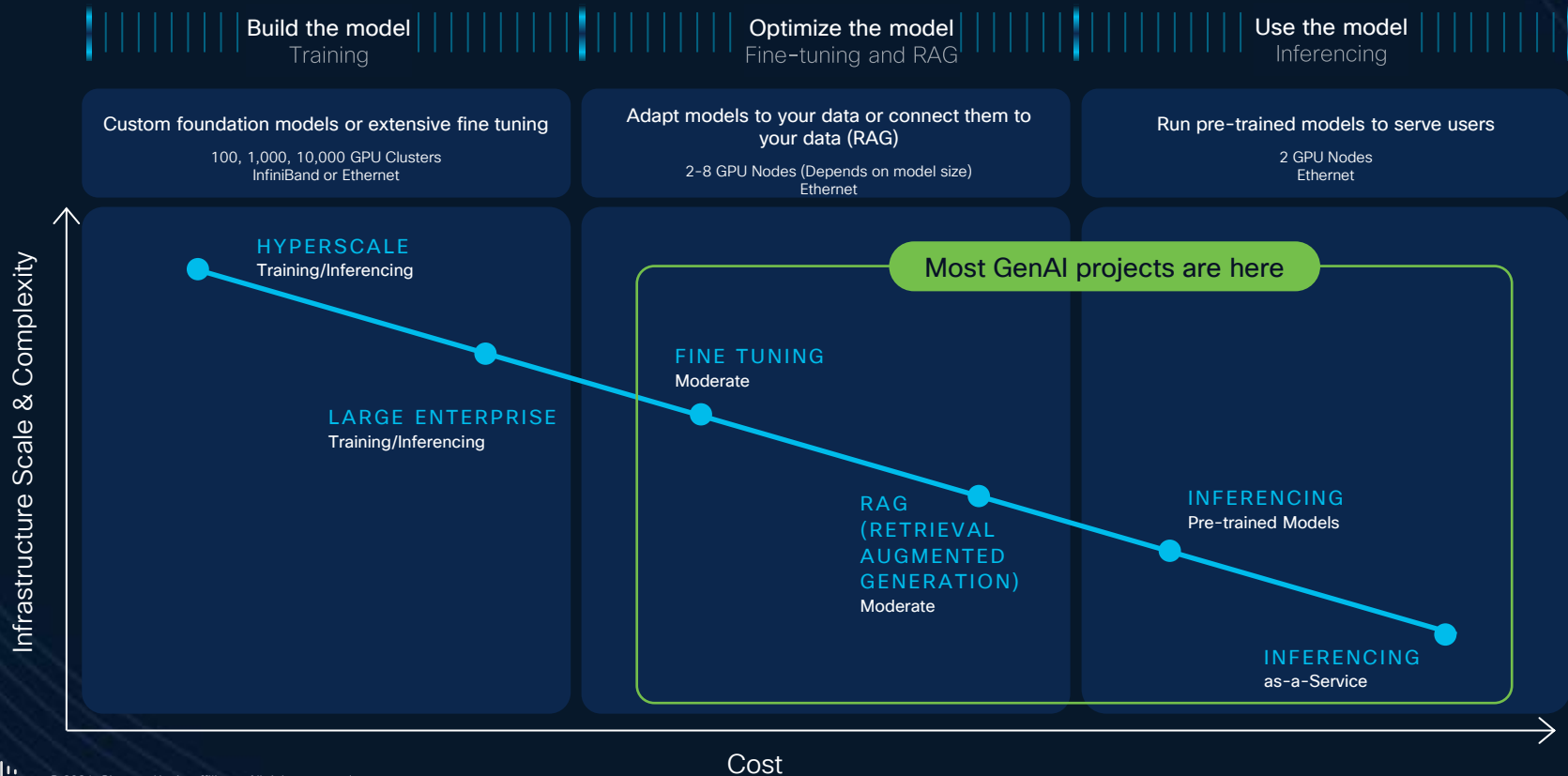
Arvind Jain
glean

“For enterprise AI, it’s all about inference—we’re not training new models, we’re optimizing the ones we already have to deliver results instantly.”

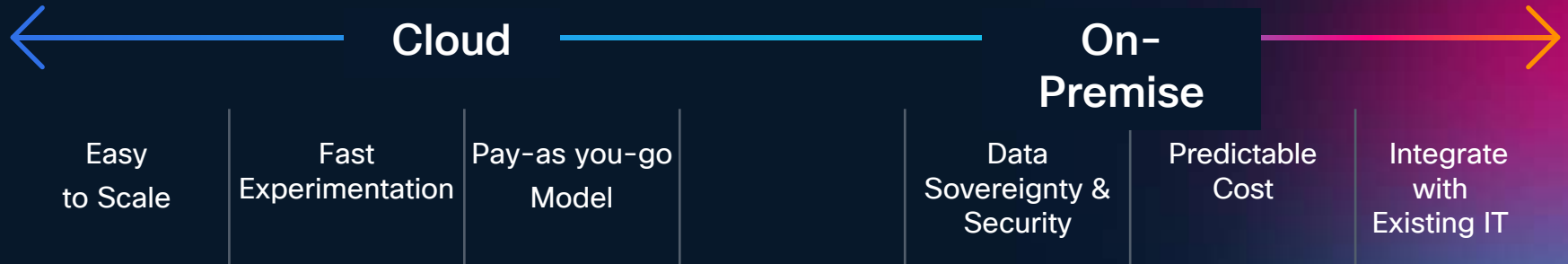
Jonathan Ross
groq

“We’re shifting from ‘Can AI work?’ to ‘How can we scale it effectively?’—it’s all about inference now.”

AI Workload Spectrum: Training, Fine-Tuning, or Inference



Where It Runs: Cloud vs On-Premise



Cisco Supports AI Everywhere



Infrastructure to power AI



Data to drive insights and context



Partners to accelerate the value of AI



Security for AI, AI for security



Software to unlock productivity

AI is forcing a rethink of the tech stack

>80%

not AI ready – need hardware acceleration

Network

speed, fabric designs, and performance

20%

of CIOs have a mature data platform strategy

>48%

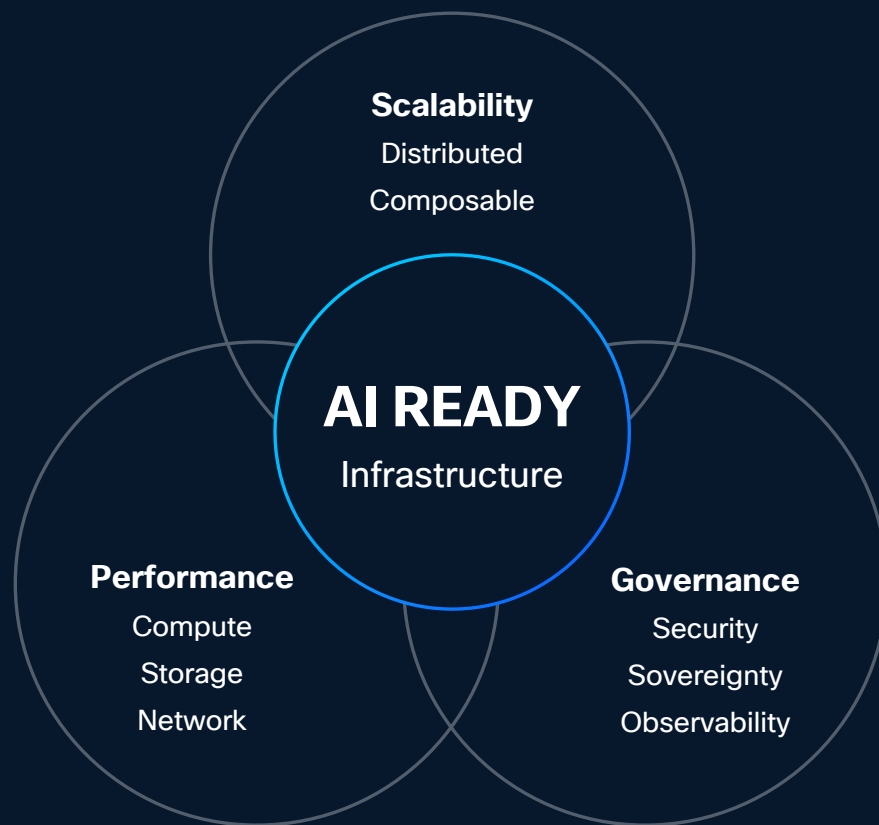
investigating VMware alternatives

>48%

of network, compute, and storage infrastructure is 10+ years old

40%

increase in security breaches year over year – due to old infrastructure



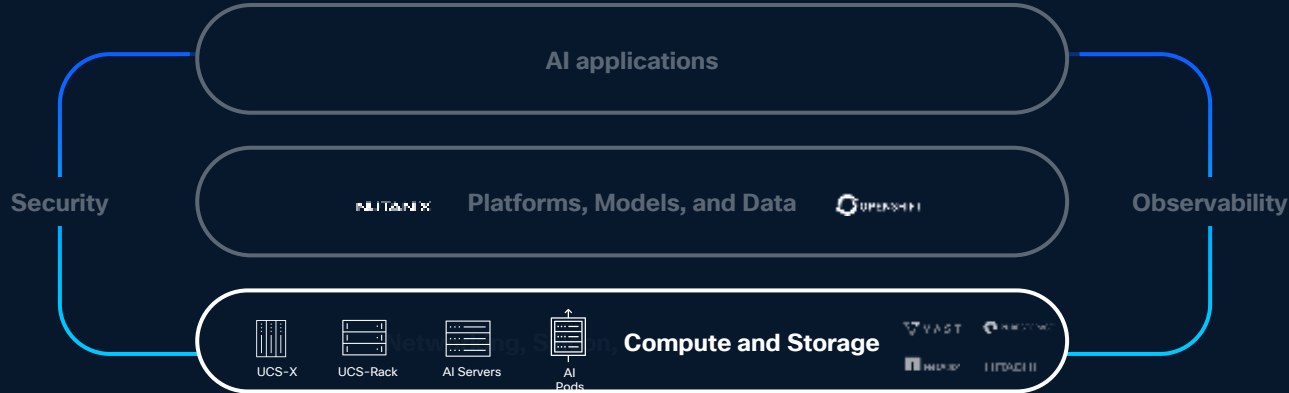
Sources: IDC; Gartner; Forrester; TechRadar; CIO.com, Accenture Tech Vision (2025)



Compute for AI

Unified Computing Systems

Deliver resources in any location from the cloud



Cloud Managed

AI-ready infrastructure

Train, fine tune and inference on accelerated servers and integrated full stack systems

Simplify at Scale

Unified infrastructure operations for faster time to value and easier lifecycle management

Hybrid Multicloud

Modernize with validated converged and hyperconverged platforms that support distributed applications

Cisco AI Compute Portfolio

Unified approach to accelerated AI compute

Validated solutions for AI with compute, network, storage, and software



GPU Accelerated

GPU Optimized

Edge

Build the model | Training

Optimize the model | Fine-tuning and RAG

Use the model | Inferencing

Cisco AI Compute Portfolio

Dense GPU AI Servers

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



Cisco UCS® C885A M8 Rack Server

NVIDIA HGX platform with
8 NVIDIA H100 NVL/H200 NVL and M300X GPUs
2 AMD 4th Gen/5th Gen EPYC processors

For fine tuning and scaleable
inference use cases



Cisco UCS® C845A M8 Rack Server

NVIDIA MGX platform with
2/4/6/8 NVIDIA H100 NVL/H200 NVL/L40S GPUs
2 AMD 5th Gen EPYC processors

Bringing High-Density GPU Servers to the Cisco UCS Family

Built for LLM training, deep learning, fine-tuning, and HPC

UCS Accelerated | UCS C880A M8

NEW

AVAILABLE NOW



2 CPUs

Intel Xeon 6th Gen Scalable Processor

NVIDIA HGX with 8 GPUs

NVIDIA B300 with NVL8 Air Cooled

Network

(8) NVIDIA ConnectX-8 GPU Board
Integrated (E-W)

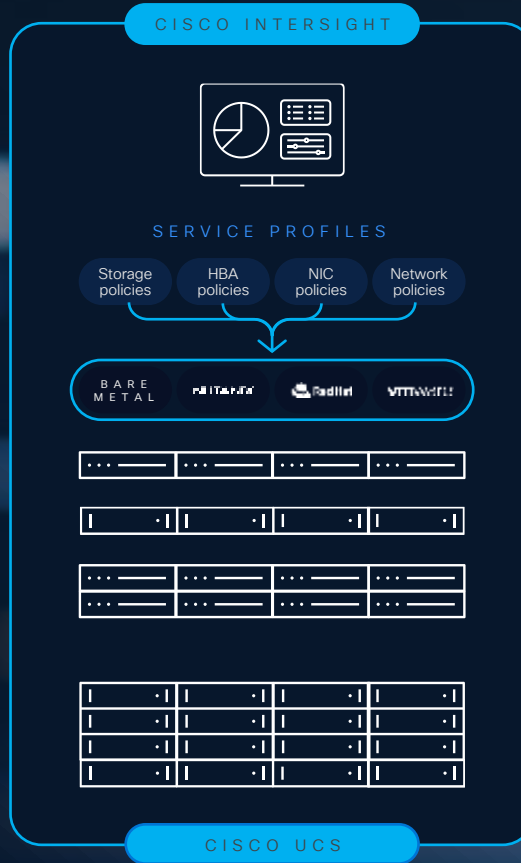
(2) NVIDIA BF3 B3220, NVIDIA BF3240,
NVIDIA ConnectX-7 (N-S)

Power

(12) 50V 3200W (N+N redundancy)

Computing that's elastic, responsive and intelligent

Delivering industry differentiation with Cisco's SaaS-managed, fabric-based data center



Elastic

Shape infrastructure to any workload, depending on business needs

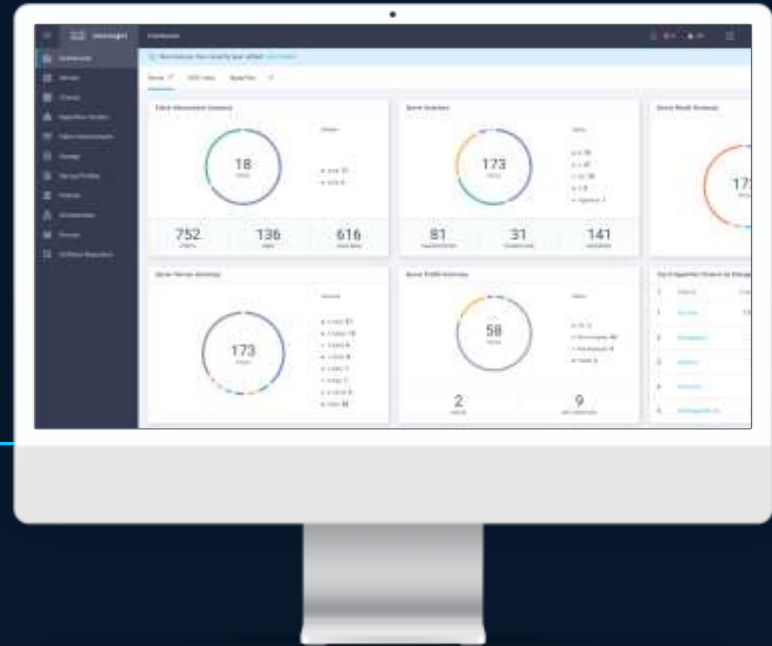
Intelligent

Connect and scale from the data center to the edge with 360° visibility

Responsive

Adapt to the future with unparalleled simplicity, performance and resiliency

Work smarter and faster with a simplified, unified operating model



Cisco Intersight

See your global on-premises, cloud, and edge environments

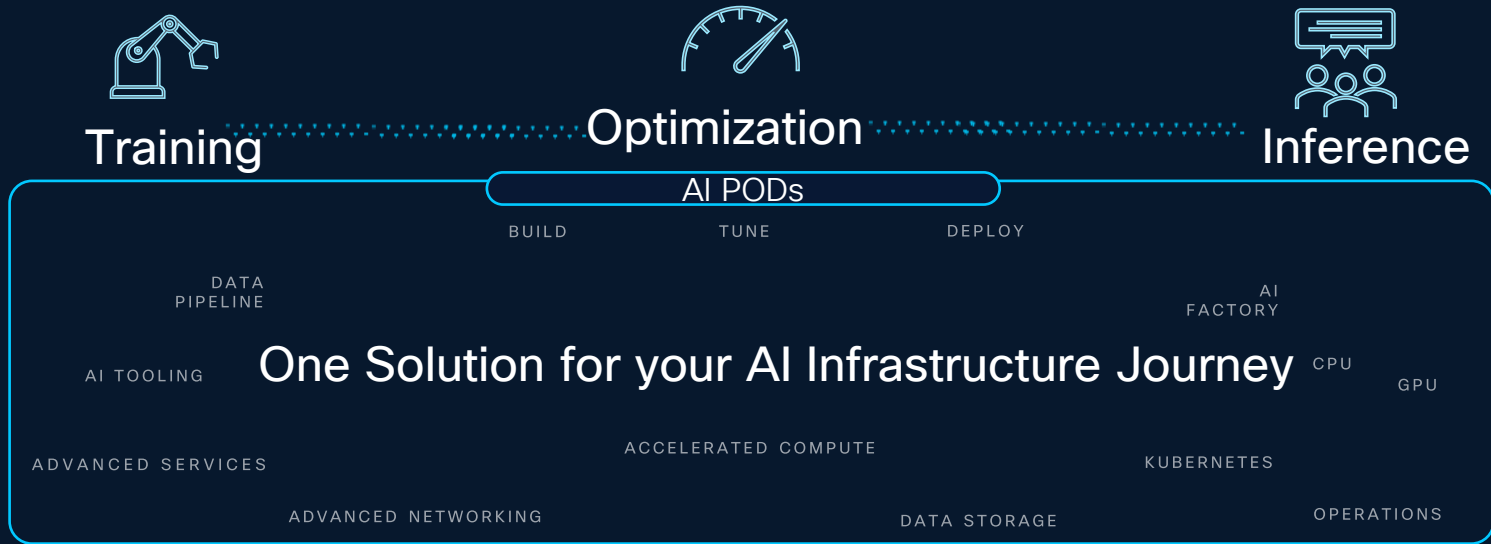
Connect your infrastructure operations across compute and storage

Secure operations with built-in advisories and continuous risk mitigation

Automate deployments, configuration, workflows, and day-0 to day-N tasks

Cisco AI Pods

A scalable architecture, built to support any AI workload simply & efficiently



Cisco AI PODs

A scalable architecture, built to support any AI workload simply & efficiently

Deploy AI with confidence

Cisco CVD, NVIDIA ERA

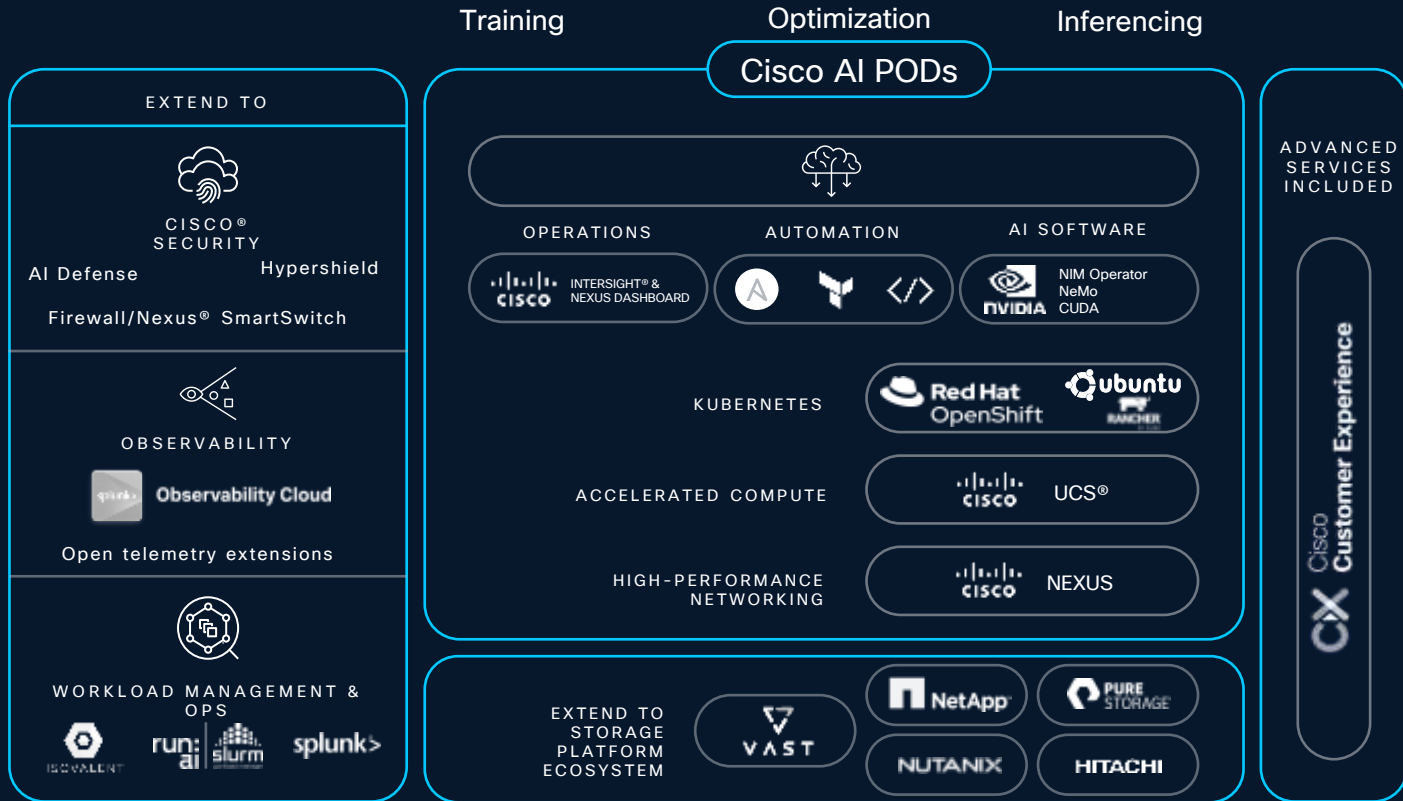
Fully supported stack including Cisco and 3rd party components

Cisco CX Success Track

Orderable, use case driven AI-ready infrastructure stacks

**Inferencing.
Optimization.
Training.**

Incremental, atomic-level –or– fabric-based cluster scale



Cisco Validated Designs – Reducing Risk

What a Validated Design Guide provides



Reliability

CVDs are extensively tested. You can confidently set performance expectations when you deploy your solution.



Less risk

Using a CVD reduces both the risk that products won't work together and the risk that they won't perform as promised.



Comprehensiveness

CVDs provide everything from system designs to configuration instructions to a bill of materials (BOM).



24-hour support

Because CVD solutions are guaranteed to work as specified, we offer 24-hour support options for any issues that might arise.

Easily Deploy New Systems With Expert Guidance
Consistent, Successful Technology Deployments Addressing Business Initiatives

CVDs to simplify and automate AI infrastructure

1

CVD blueprint for AI networks



Best performing AI/ML networks, focus on application performance



Intelligent buffer, low latency, telemetry and RoCEv2



Dynamic congestion avoidance



One IP network for both front-end and back-end



Automation for day-2 operations



Validated designs for network and ecosystem partners

2

EXPANDED ROADMAP

CVDs for simplified AI-ready infrastructure



NVIDIA AI Enterprise



Red Hat OpenShift AI

NUTANIX

GPT-in-a-box on Nutanix Hyperconverged

CLOUDERA

Gen-AI with Cloudera Data Platform



intel.

3

NEW

CVD playbooks for deploying common AI models



Large language models (GPT3, BERT, T5)



Computer vision models (ResNet, EfficientNet, YOLO)



Generative models (GANs, VAEs)



As your needs become more specific, so do our recommendations



Your Trusted AI Advisor

Helping you optimize your AI infrastructure

Workload sizing



GPUs recommendations



Multiple LLM model support



End-to-end guidance



Bot assistance and BOM generation

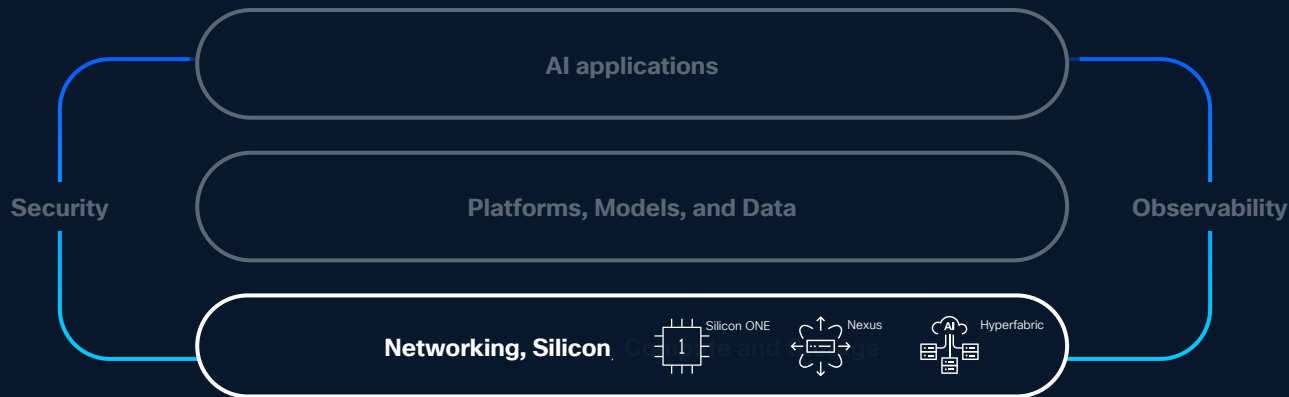


Expanded model selection



Data Center Networking

Industry leading data center fabrics coupled with operational simplicity



Fabric Options

Choice of Fabric

Cisco Nexus and Hyperfabric connect and protect the most demanding workloads, powered by Silicon One

Simplified Operations

Choose on-premise or cloud managed operational model that delivers operational insights, efficiency and sustainability

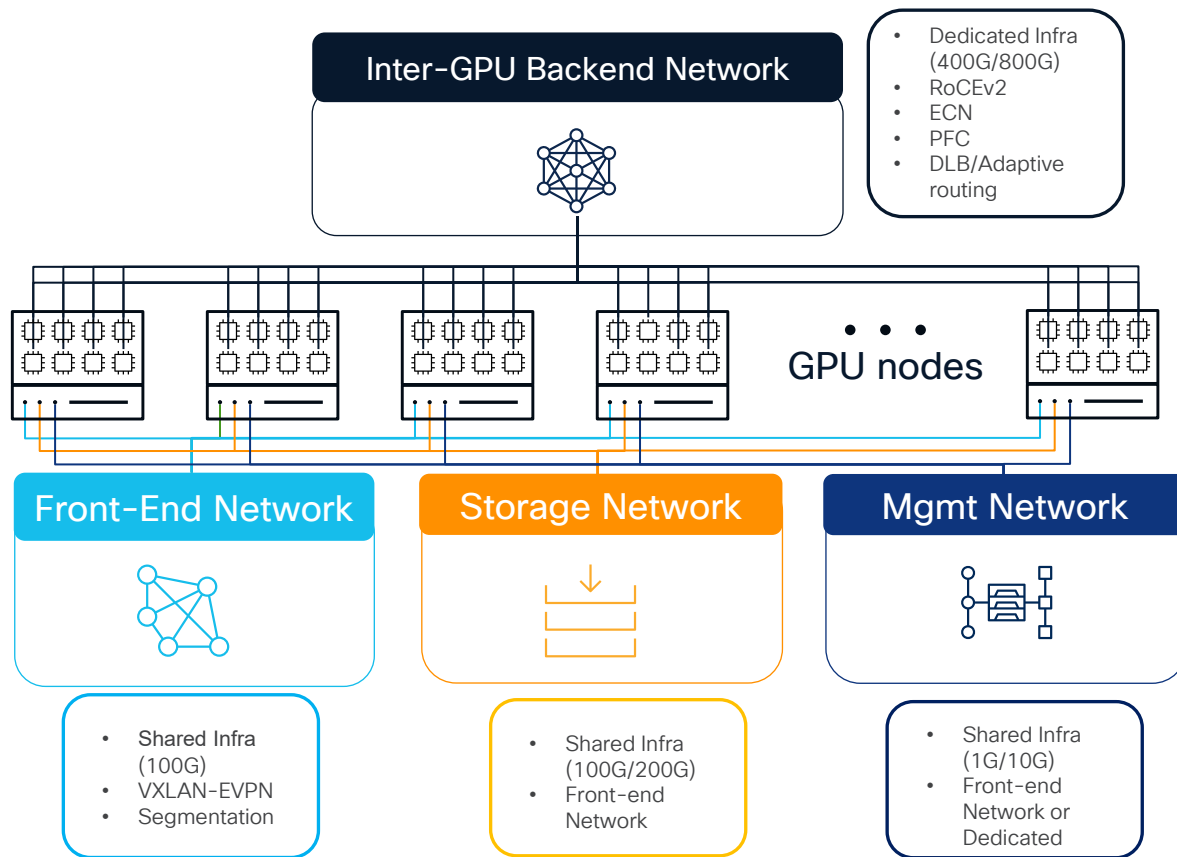
Validated Designs

Design, deploy and operate with co-developed reference architectures and best practice from Cisco and NVIDIA



Datacenter Networking for AI

Multiple networks for AI infrastructure



AI Infrastructure with
Cisco Nexus 9000 Switches



Cisco Data Center
Networking Blueprint for
AI/ML Applications



Cisco Nexus 9300 Series – 100G/400G/800G Fixed Switches

NXOS | ACI leaf

NXOS | ACI spine | ACI Leaf

NXOS L3 Only

2RU Switches

1RU Switches



Nexus 9364C-H1
Nexus 9364C-GX
64p 100G



Nexus 9348D-GX2A
48p 400G



Nexus 9364D-GX2A
64p 400G



Nexus 9364E-SG2
64p 800G



Nexus 9336C-FX2
36p 100G



Nexus 93600CD-GX
28p 40/100G & 8p 400G



Nexus 9316D-GX
16p 400G



Nexus 9332D-H2R
Nexus 9332D-GX2B
32p 400G



Bringing AI to the Enterprise.

Cisco and NVIDIA expand partnership to accelerate AI adoption in the enterprise

Partnership focus on
Private data centers

Cisco Full-stack
AI infrastructures

Accelerated Compute

High-performance Network
(backend and frontend)

NVIDIA-Certified
Storage Ecosystem

Cisco Security

How are we partnering?



Cisco will be part of NVIDIA's Enterprise Reference Architecture



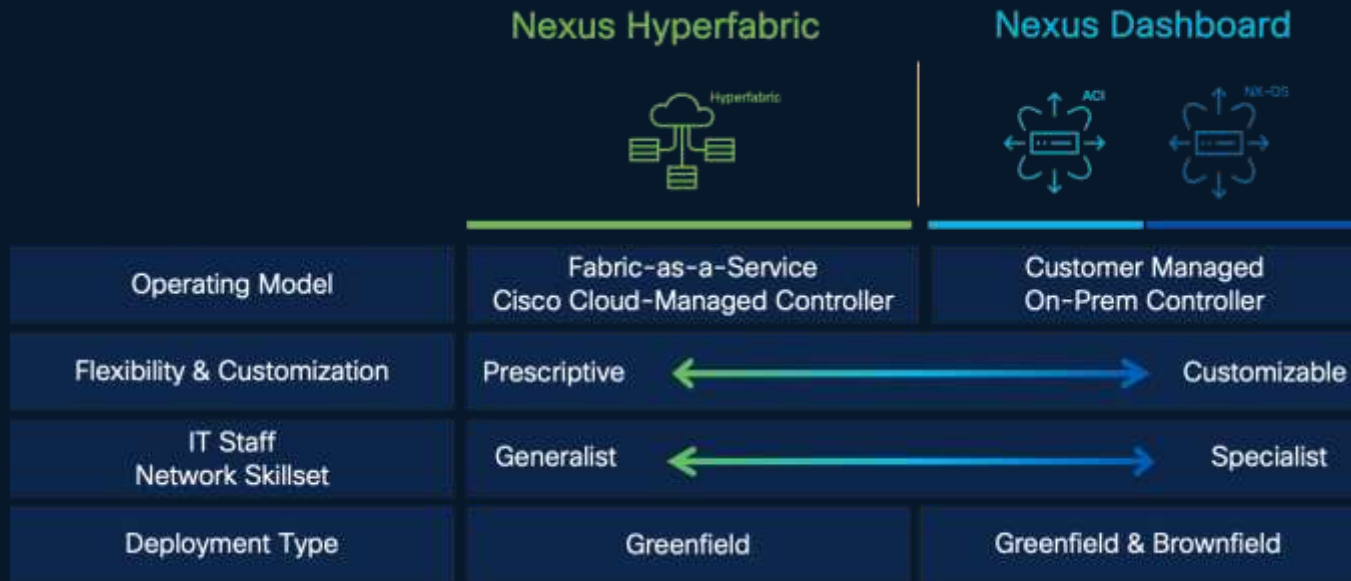
Cisco will be part of NVIDIA's Cloud Partner Reference Architecture



Deliver AI Factory with customer choice: Cisco Silicon One® or NVIDIA Spectrum-X Silicon



Managing AI Infrastructure



Greenfield: new fabrics not being managed
by Nexus Dashboard

Cisco Secure AI Factory with NVIDIA

