

Cisco Full Stack AI DC

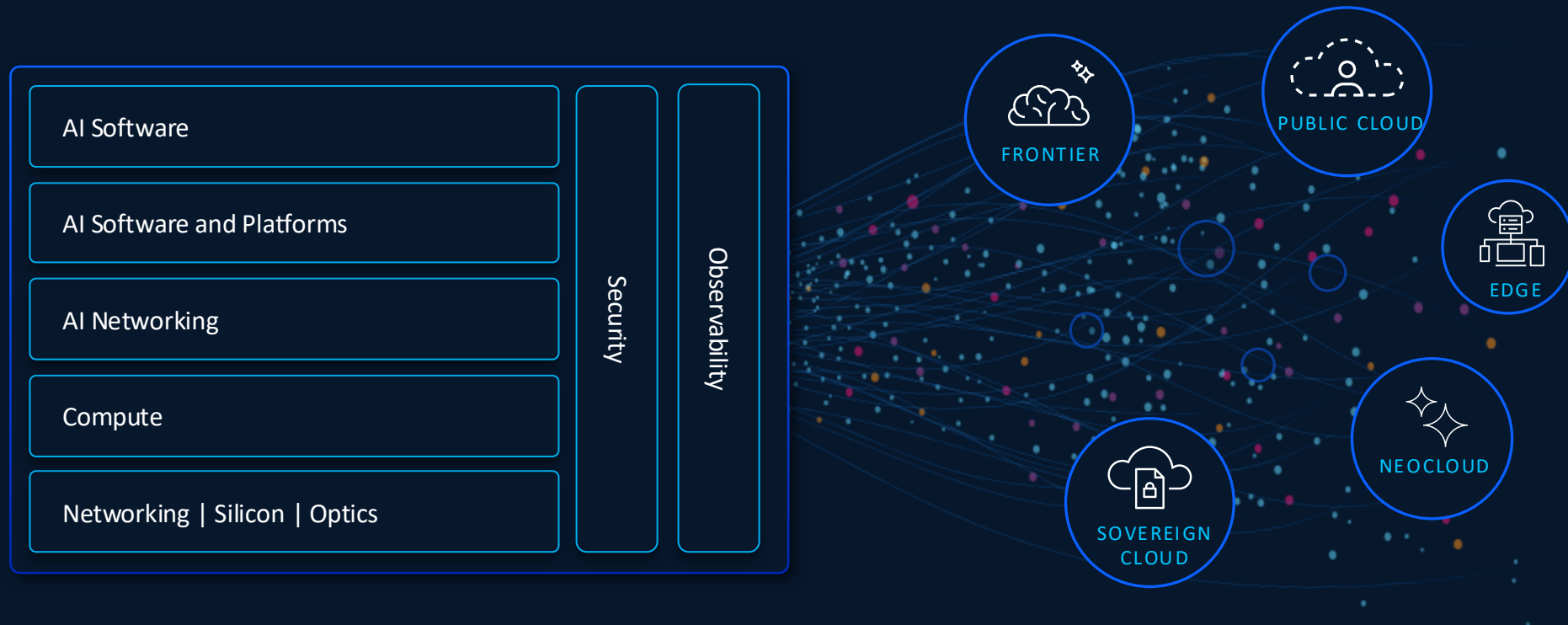
Cisco Tech Club

AI is accelerating faster than traditional infrastructure and operations can adapt

As AI moves from experimentation to execution, every
agentic action is a routing challenge, a trust decision and a
telemetry event

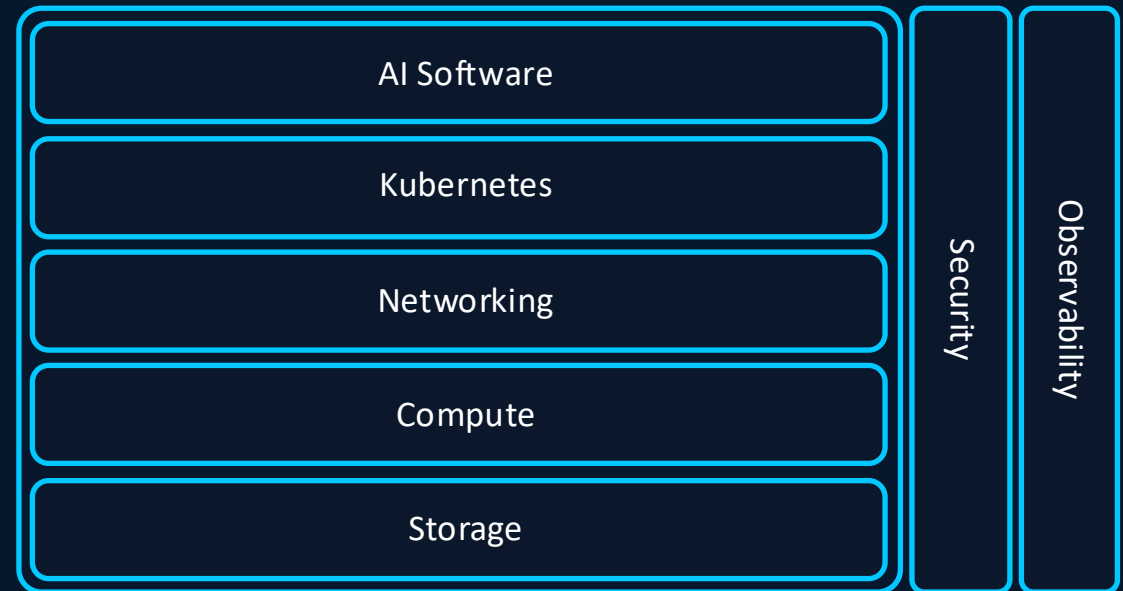


Production AI needs Full Stack AI, Everywhere AI Runs



Cisco Full Stack AI

Cisco AI Pod
Cisco Secure AI Factory with NVIDIA



Cisco AI PODs

Scalable, repeatable full stack to power AI workloads

Cisco Validated Designs

Professional Services

AI Software

Kubernetes Platform

Cisco AI Networking

Cisco Accelerated Computing

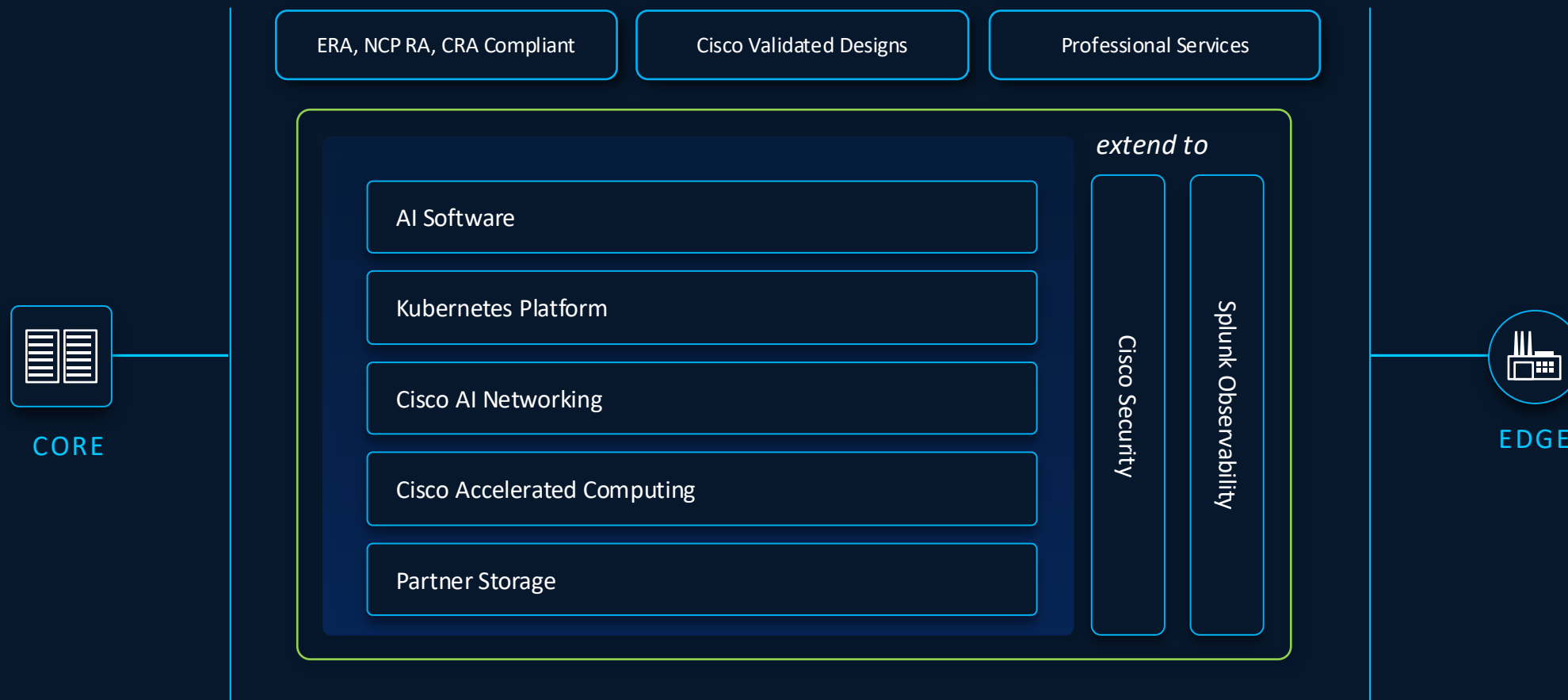
Partner Storage

extend to

Cisco Security

Splunk Observability

Cisco Secure AI Factory with NVIDIA



Accelerate the delivery of trusted agentic and physical AI applications with a secure, simple, and scalable AI infrastructure

Cisco Secure AI Factory with NVIDIA

ERA, NCP RA, CRA Compliant

Cisco Validated Designs

Professional Services

Compliant
with NVIDIA
ERA and
NCP RA



ERA: 1– 1K GPUs

HGX – NVL8 | MGX PCIe GPUs

Certified Partner Storage

Cisco N9100 switches w/ NVIDIA Spectrum-X silicon

Cisco N9300 switches w/ Cisco Silicon One

NCP RA and CRA: 1k – 100K+ GPUs

NVL72 – Vera Rubin | Grace Blackwell

HGX – NVL8 | MGX PCIe GPUs

Certified Partner Storage

Cisco N9100 switches w/ NVIDIA Spectrum-X silicon

Cisco N9300 switches w/ Cisco Silicon One

Cisco 8000 switches w/ Cisco Silicon One

Rack-to-fabric air and liquid cooled systems



Cisco Confidential

Cisco Secure AI Factory with NVIDIA

ERA, NCP RA, CRA Compliant

Cisco Validated Designs

Professional Services

Cisco validated designs

Design, deploy, operate, secure, and extend with documented best practices from Cisco and ecosystem partners



Training



Optimization



Inferencing

Faster deployments

Less risk

Predictability

Cisco accelerated compute portfolio

Inferencing, Agentic AI
at the edge

Optimization,
Inferencing, Agentic AI

Model Training

Large scale Inference, Agentic AI,
Trillion Parameter Model Training

PCIe GPU: Edge Platform

Cisco Unified Edge
NVIDIA RTX PRO



PCIe GPU Servers

Supermicro MGX
NVIDIA RTX PRO



Cisco UCS®
C845A M8
NVIDIA RTX PRO &
AMD Instinct MI



Cisco UCS® X-Series
Modular Servers
NVIDIA RTX PRO



Cisco UCS® C245 M8
NVIDIA RTX PRO



Cisco UCS®
C240 M8
NVIDIA RTX PRO



Dense GPU Servers

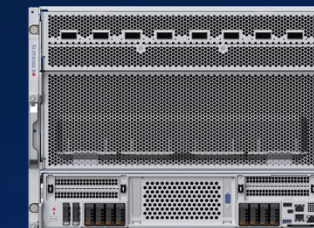
Supermicro
NVIDIA HGX
Rubin NVL8
liquid-cooled



Supermicro
NVIDIA HGX
B300 NVL8
liquid-cooled



Supermicro
NVIDIA HGX
B300 NVL8
air-cooled



Rack-Scale Systems

Supermicro
NVIDIA
Vera Rubin
NVL72



Supermicro NVIDIA
Grace Blackwell 300
NVL72

Storage Server

Cisco UCS® C225 M8N



Liquid Cooled System Power Usage

NVIDIA HGX B300 Servers



SYS-422GS-LCC
Liquid-cooled

- 1x HGX B300 8-GPU Board
- 2x Intel Xeon 6 6768P CPUs
- 16x or 32x 128GB DDR5 RDIMMs
- 2x 1.9TB M.2 boot
- 8x ConnectX-8 for the backend network
- 2x BF-3 B3240 DPUs for the frontend network

~10% less than AC

~13 kW

Up to 100+ kW Rack

NVIDIA HGX Rubin LC8 Server



SYS-222GS-R2BT-ALC
Liquid-cooled

- 1x HGX Rubin 8-GPU Board LC
- 2x Intel Xeon GNR-SP 6776P
- 32x 64GB DDR5 RDIMMs
- 2x 1.9TB M.2 boot drives
- 8x 3.84TB data drives
- 8x NVIDIA CX-9 E-W + BF-4 B4240 N-S

~24 kW

Up to 200+ kW Rack

Rack-Scale Systems



SRS-GB300-NVL720-Mx

- 2x Grace CPUs (480GB LPDDR5X per CPU)
- 4x B300 GPUs (288GB HBM3e per GPU)
- 2x CX-8 NICs, w/ 2x 800G OSFP
- 1x BF3 B3240
- 1x 1.92 TB M.2 boot
- 4x E1.S 3.84TB

~140 kW



SRS-VRRBN-NVL720-Mx

- 2x Vera CPUs (720GB SOCAMM per CPU)
- 4x Rubin GPUs (294GB HBM4 per GPU)
- 4x CX-9 NICs, w/ 2x 800G OSFP
- 1x BF4 B4240
- 1x 1.92 TB M.2 boot
- 4x E1.S 7.68TB

~230 kW

Liquid cooling portfolio options

Scale cooling density for any workload type



Cooling
capacity

50 kW

80 kW

200 kW

250 kW

1.8 MW

Rear-door heat exchanger

- Cooling capacity: up to 80 kW per rack
- Air-to-liquid deployment

Side-car CDU

- Liquid-to-air cooling
- Fast deployment
- Cooling capacity up to 200 kW
- *LCS-SCDU-200AR001*

In-rack CDU

- EIA and OCP19 rack versions

In-row CDU

- Low PUE* – warm-water handling
- Redundant pumps and power

*PUE: Power Usage Effectiveness: total facilities power / IT equipment power; closer to 1 is best.

Liquid cooling portfolio options

Scale cooling density for any workload type



Cooling capacity
50 kW 80 kW

Rear-door heat exchanger

- Cooling capacity: up to 80 kW per rack
- Air-to-liquid deployment



200 kW

Side-car CDU

- Liquid-to-air cooling
- Fast deployment
- Cooling capacity up to 200 kW
- *LCS-SCDU-200AR001*

Requires *uncommon* air cooling
in the datacenter

Requires a Facility Water System



250 kW

In-rack CDU

- EIA and OCP19 rack versions



1.8 MW

In-row CDU

- Low PUE* – warm-water handling
- Redundant pumps and power

*PUE: Power Usage Effectiveness: total facilities power / IT equipment power; closer to 1 is best.

Secure AI Factory Dense GPU Server Management Options

Option 1: Cisco Software Direct to Device APIs

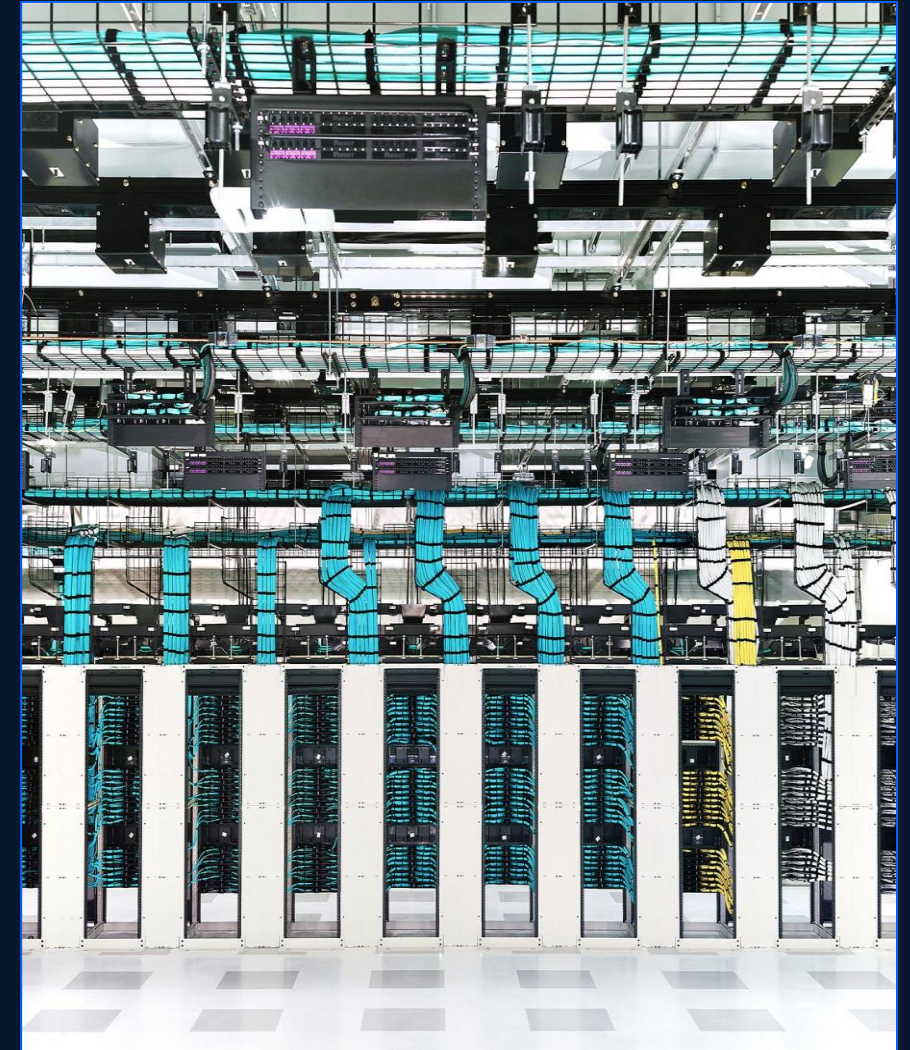
- No dependency on Supermicro software
- Cisco software directly manages:
 - ✓ Supermicro servers
 - ✓ Supermicro power infrastructure
 - ✓ Cisco networking
- CDU monitoring via API integration
- Facility infrastructure / DCIM management - targeted future phase

Note: Scope of CDU management is under investigation

Option 2: 3rd-Party Management Direct to Device APIs (on-prem)

- 3rd-party software directly manages Supermicro infrastructure
- Options may include:
 - ✓ Custom automation using Redfish APIs
 - ✓ NVIDIA Base Command Manager (BCM)
 - ✓ NVIDIA Mission Control (includes BCM)
 - ✓ Other management tools, such as Supermicro SCC
- Cisco software manages Cisco networking

Note: TAC provides L0/L1 support. For L2/L3 support, requests are routed to SMCI after initial triage.



Cisco AI Software Options

Optimization,
Inferencing, Agentic AI

Optimization,
Inferencing, Agentic AI

and

Model Training

Large scale Inference, Agentic AI,
Trillion Parameter Model Training

NVIDIA MGX/PCIe GPU Server



SYS-422GL-FNR2

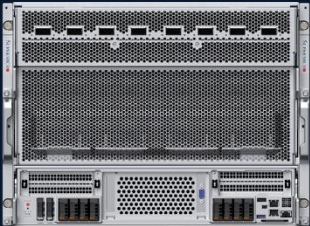
Cisco Management Platform

SaaS	On-prem
Intersight and Hyperfabric	Intersight + ND

Customer Options for compute: Hyperfabric, Intersight, 3rd party server mgmt.

Customer Options for network: Hyperfabric, Nexus Dashboard

NVIDIA HGX B300 Servers



SYS-822GS-NB3RT
Air-cooled



SYS-422GS-LCC
Liquid-cooled

Cisco Management Platform

SaaS	On-prem
Intersight and Hyperfabric	Intersight + ND

Customer Options for compute: Hyperfabric, Intersight, 3rd party server mgmt.

Customer Options for network: Hyperfabric, Nexus Dashboard

NVIDIA HGX Rubin LC8 Server



SYS-222GS-R2BT-ALC
Liquid-cooled

Cisco Management Platform

SaaS	On-prem
Hyperfabric	Cisco S/W - under planning

Customer Options for compute: Hyperfabric, Cisco S/W*, 3rd party server mgmt.

Customer Options for network: Hyperfabric, Cisco S/W

Rack-Scale Systems



SRS-GB300-NVL720-Mx

Cisco Management Platform

SaaS	On-prem
Hyperfabric	Third-party S/W + ND

Customer Options for compute: Hyperfabric, 3rd party server mgmt.

Customer Options for network: Hyperfabric, Nexus Dashboard



SRS-VRRBN-NVL720-Mx

Cisco Management Platform

SaaS	On-prem
Hyperfabric	Cisco S/W – under planning

Customer Options for compute: Hyperfabric, Cisco S/W, 3rd party server mgmt.

Customer Options for network: Hyperfabric, Cisco S/W

Cisco Secure Networking

Cisco Cloud Control

Unified Login | Unified Inventory | Unified Topology

FULLY INTEGRATED DATA CENTER NETWORK STACK

Nexus Dashboard

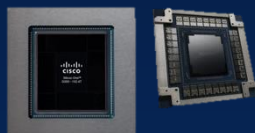
On premises

Nexus Hyperfabric

Cloud management

SILICON

Cisco Silicon One
NVIDIA Spectrum-X Ethernet



SYSTEMS

Cisco N9000
Systems



OPTICS

Cisco Optics



SOFTWARE

Cisco NX-OS
Cisco ACI, SONiC



SECURITY AND OBSERVABILITY

Ultra Ethernet

MEMBER

Cisco Secure Networking



NUTANIX

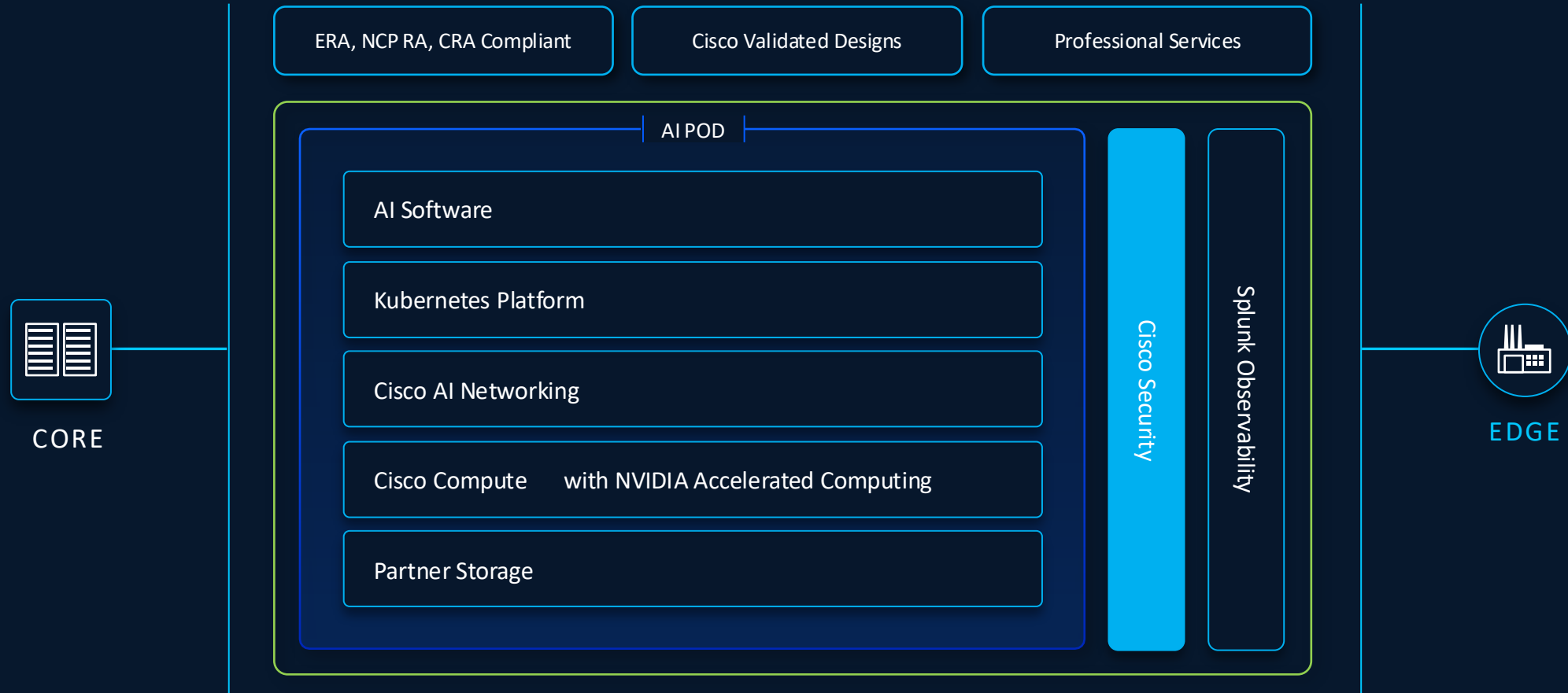


ISOVALENT

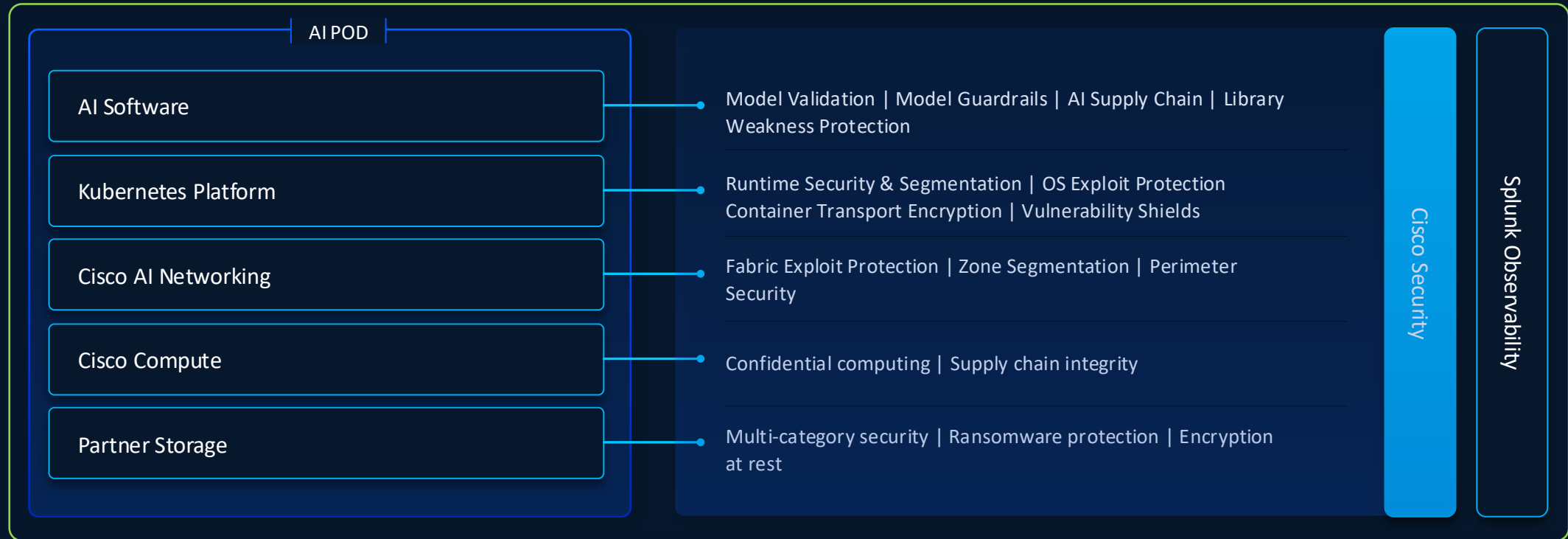
High performance, advanced Kubernetes Networking, load
balancing, and runtime security

SECURITY AND OBSERVABILITY

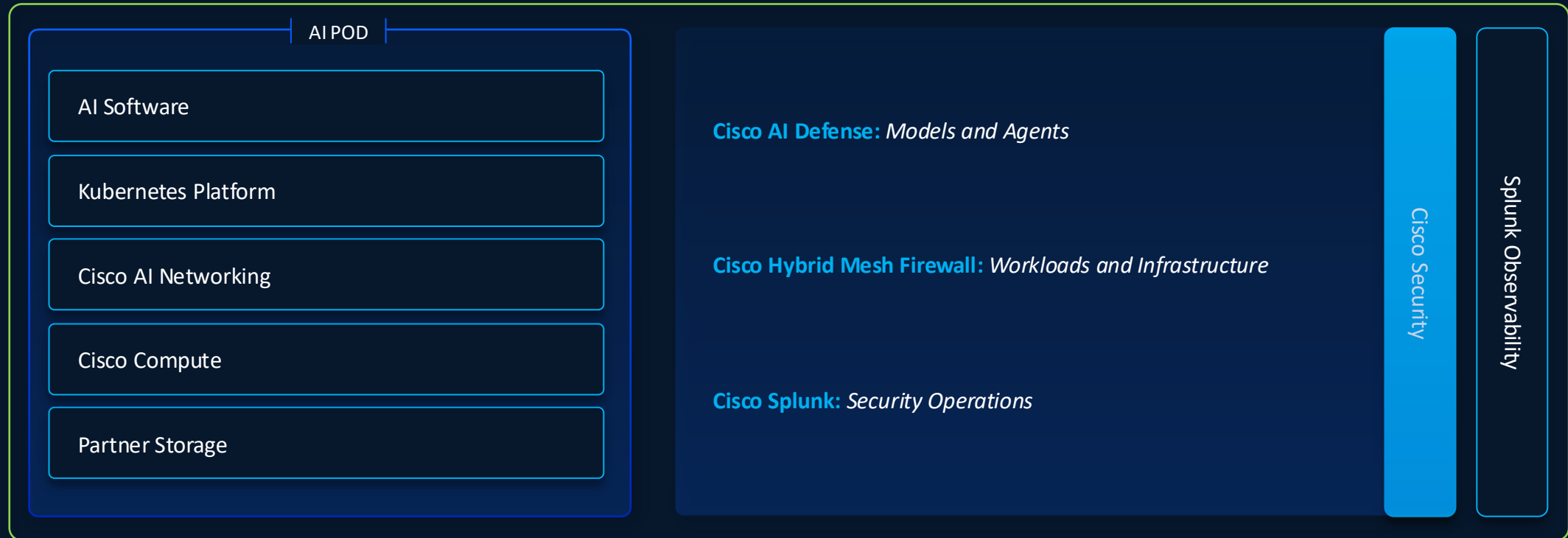
Cisco Full Stack Security



Cisco Full Stack Security

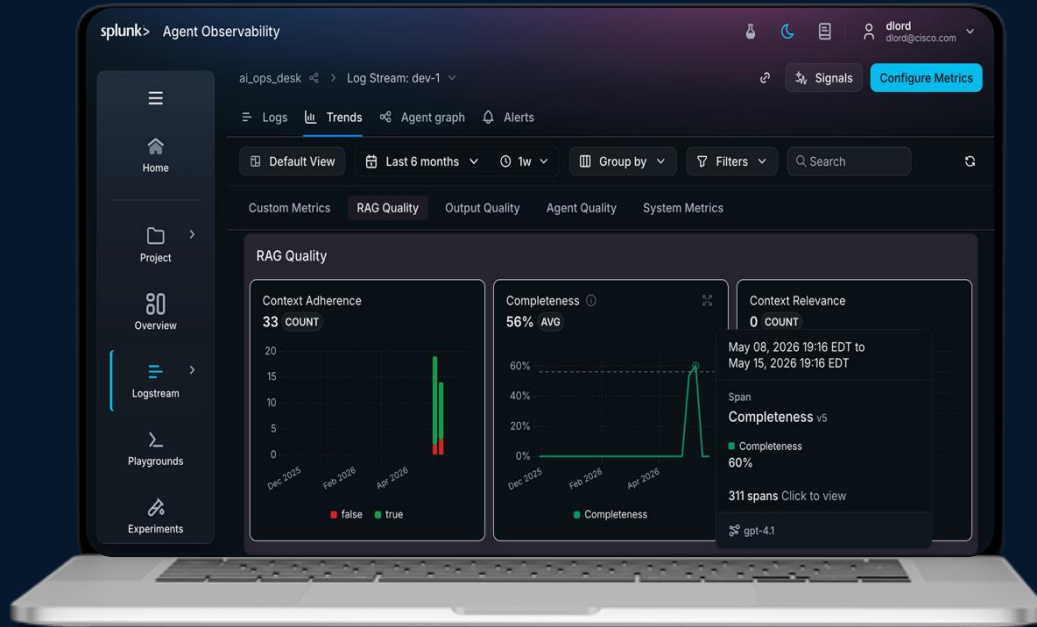


Cisco Full Stack Security



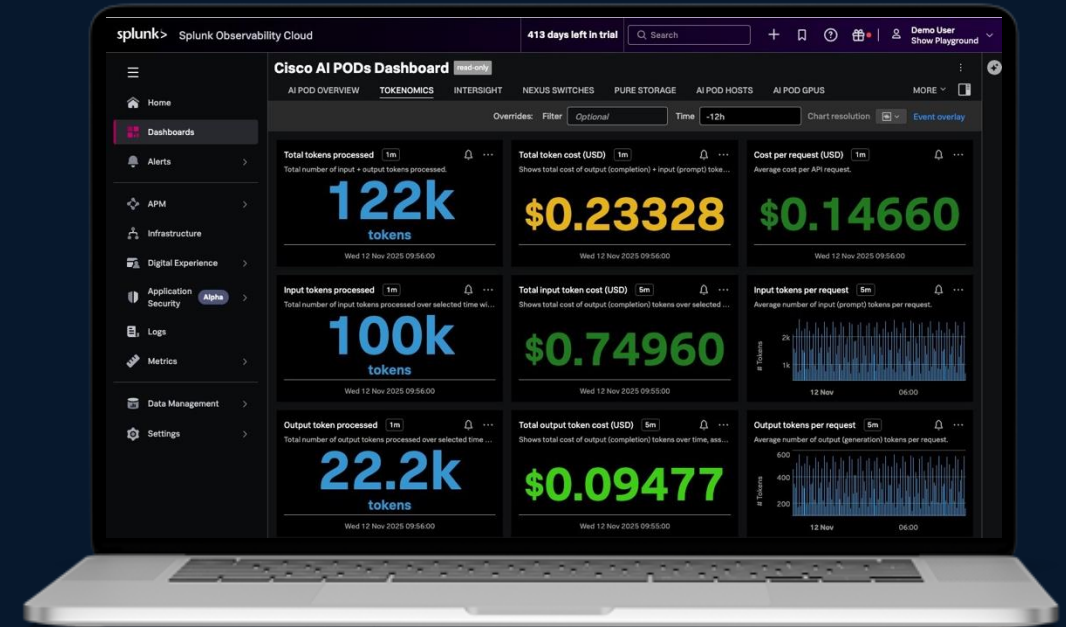
Splunk Full Stack AI Observability

Monitor the performance, quality, security, and cost of your AI stack



AI Agent Monitoring

Monitor the performance, quality, security, and cost of LLM and agentic applications



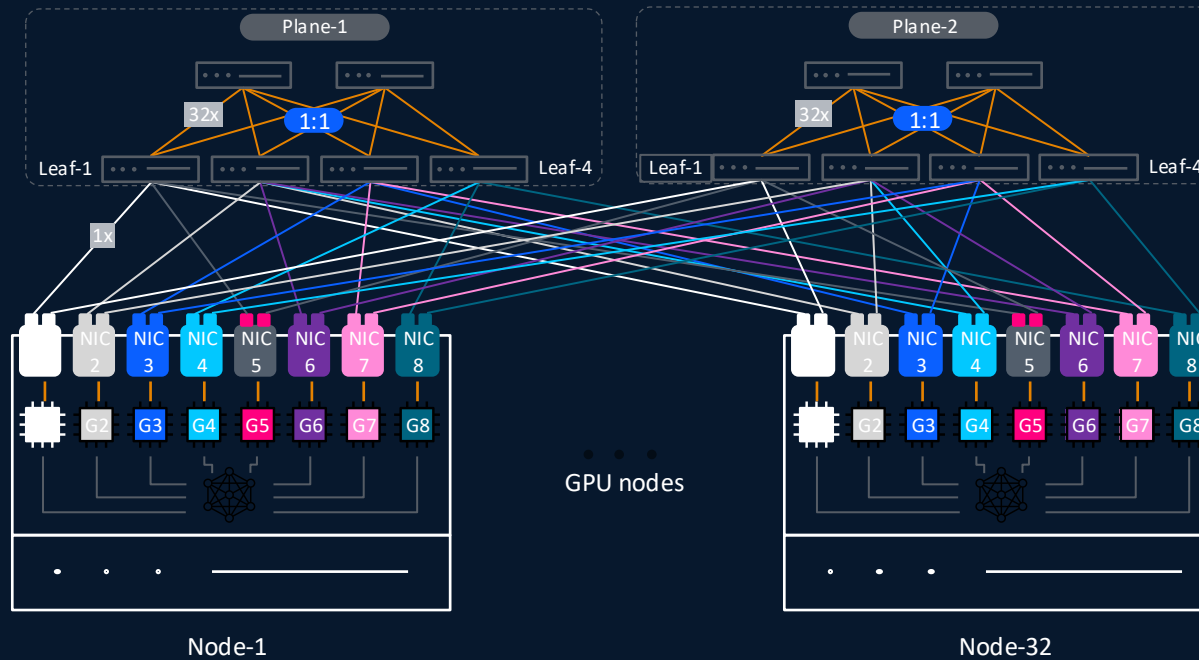
AI Infrastructure Monitoring

Monitor the health, availability, and consumption (or usage) of AI infrastructure

Network Topology examples

HGX B300 - 256 GPU Cluster

Using Cisco Nexus 9364E-SG2 switches (128 port x 400 GE)

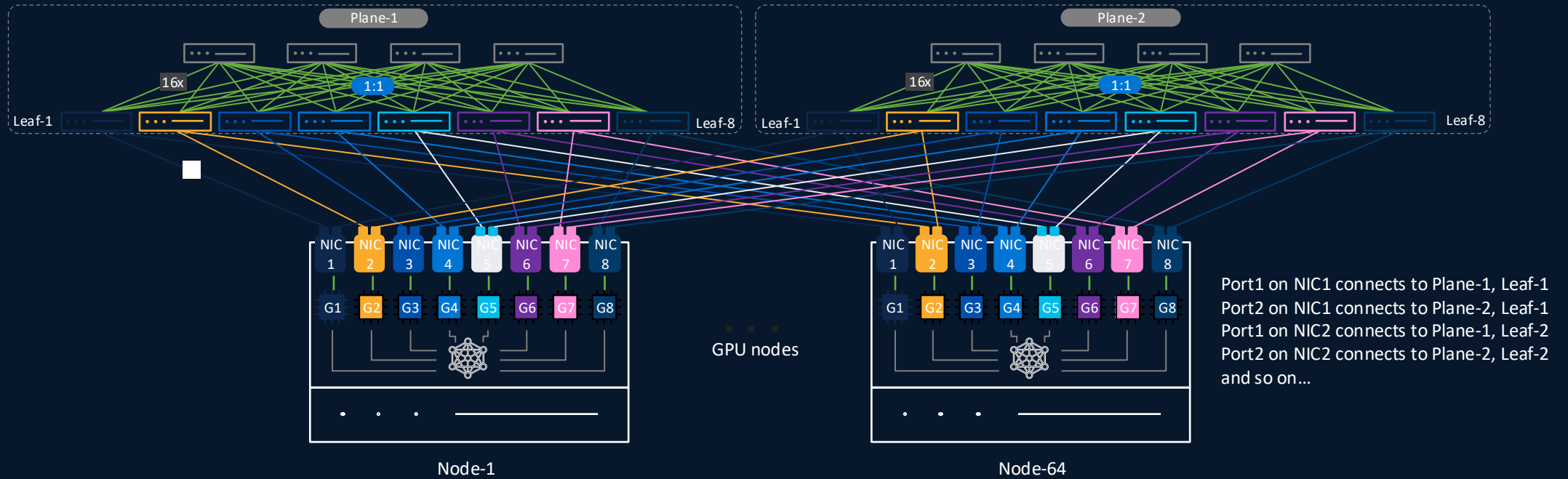


Port1 on NIC1 connects to Plane-1, Leaf-1
Port2 on NIC1 connects to Plane-2, Leaf-1
Port1 on NIC2 connects to Plane-1, Leaf-2
Port2 on NIC2 connects to Plane-2, Leaf-2
and so on...
Port1 on NIC5 connects to Plane-1, Leaf-1
Port2 on NIC5 connects to Plane-2, Leaf-1
Port1 on NIC6 connects to Plane-1, Leaf-2
Port2 on NIC6 connects to Plane-2, Leaf-2
and so on...

1. Total GPUs = 8 GPUs in a node x 32 nodes = 256 GPUs
2. Leaf switches = 4 per plane x 2 planes = 8
3. Spine switches = 2 per plane x 2 planes = 4
4. Total switches = 12
5. Leaf-node links = 64 per leaf x 4 leaf x 2 planes = 512 @ 400G
6. Leaf-spine links = 32 per leaf-spine pair x 2 spines x 4 leafs x 2 planes = 512 @ 400G

HGX B300 - 512 GPU Cluster

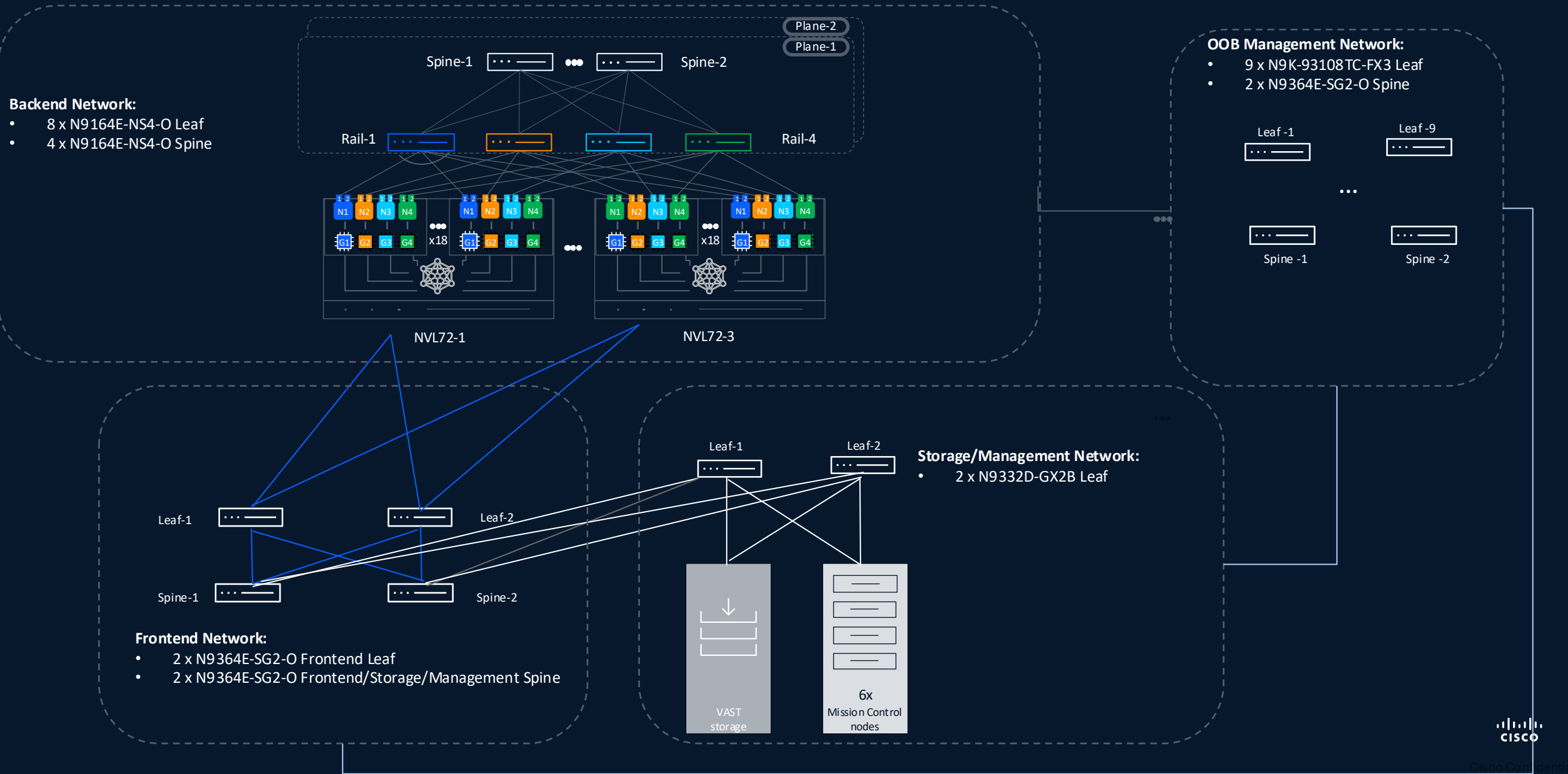
Using Cisco Nexus 9364E-SG2 switches (128 port x 400 GE)



1. Total GPUs = 8 GPUs in a node x 64 nodes = 512 GPUs
2. Leaf switches = 8 per plane x 2 planes = 16
3. Spine switches = 4 per plane x 2 planes = 8
4. Total switches = 24
5. Leaf-node links = 64 per leaf x 8 leaf x 2 planes = 1024 @ 400G
6. Leaf-spine links = 16 per leaf-spine pair x 4 spines x 8 leafs x 2 planes = 1024 @ 400G

GB300 NVL72: 216 GPU: AI Cluster Network using NXOS

Logical Topology





Děkujeme za Vaši pozornost

Následující Tech Club webinář:

20.10.2026 Bezpečnostní koncept “Zero Trust” v době AI agentů

Přednášející: Jiří Tesař



Následující webinář



Tech Club portál



Tech Club Webex Space



Tech Club Spotify