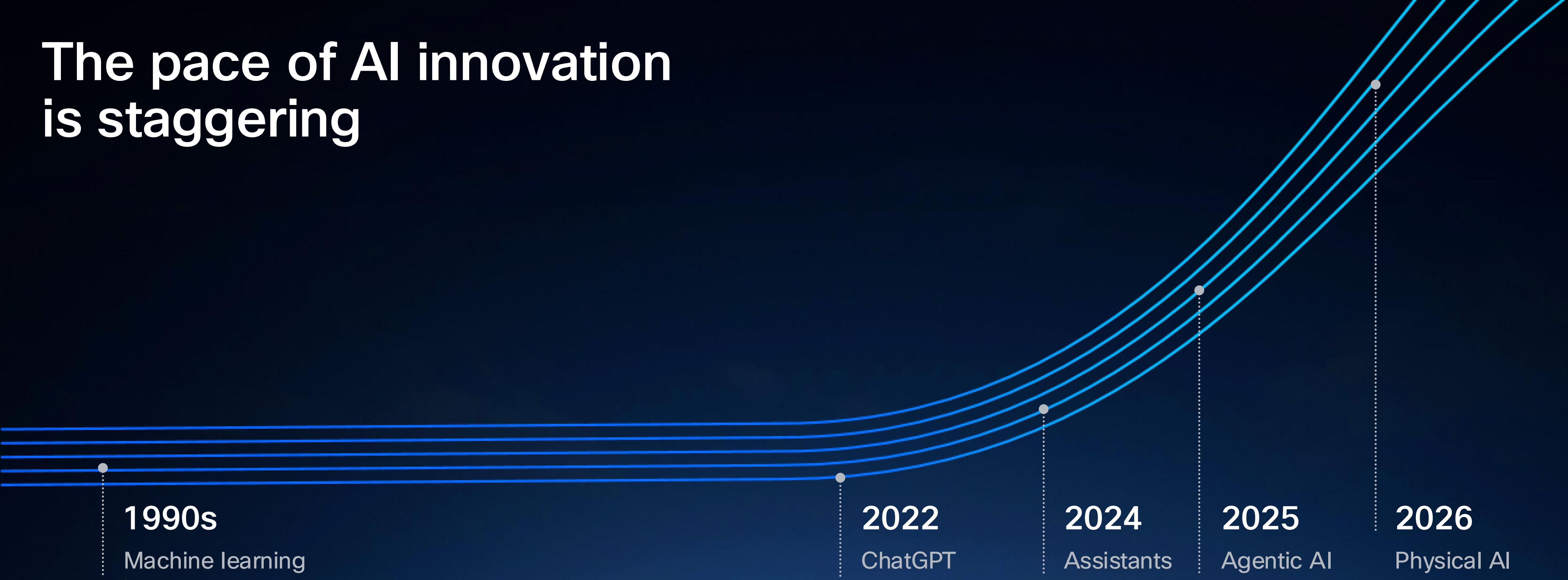
Cisco AI Strategy

Aleksandar Stepancev
Technical Lead, Cisco Balkans
Country Lead, Cisco Serbia & Montenegro

Novembar 2025

The pace of AI innovation is staggering



1990s
Machine learning

2022
ChatGPT

2024
Assistants

2025
Agentic AI

2026
Physical AI

Manufacturing

Predictive maintenance

Quality control

Demand forecasting

Public sector

Smart cities

Security and safety

Services improvement

Retail

Personalization

Inventory optimization

Sales forecasting

Financial services

Fraud detection

Risk assessment

Trading

Healthcare

Diagnosis

Drive-thru optimization

Patient support

Education

Learning & teaching experiences

Smart & secure facilities

Industry challenges

Surge in AI

Unimaginable opportunity, unprecedented threats

Evolving risks

Complex, distributed environments, new vulnerabilities

Digital Everything

Every physical experience will be digitized

Priorities

Modern infrastructure

Multicloud | App/ Infra Modernization | Ops Transformation

Cybersecurity

Reduce Attack surface | Full Stack visibility | Unified posture

AI and data

AI Infra buildout | Data Engineering | AI-Driven Solutions

Cisco is leading architectural shifts

AI is network bound

Bottlenecks and latency stall AI investments

Private data center "re-acceleration"

Surge in new AI workloads on-premises

Common foundation for AI safety

Responsible AI framework for safe, trustworthy AI

Hyper-distributed security

AI's unpredictable nature creates new challenges

6 months of unprecedented Cisco Innovation

2025 January

- AI Defense security
- AI Servers

2025 February

- Smart Switches with Hypershield, Hybrid Mesh Firewall
- Mistral AI alliance

2025 March

- Secure AI Factory with NVIDIA
- AGNTCY.Org: Internet of agents

2025 April

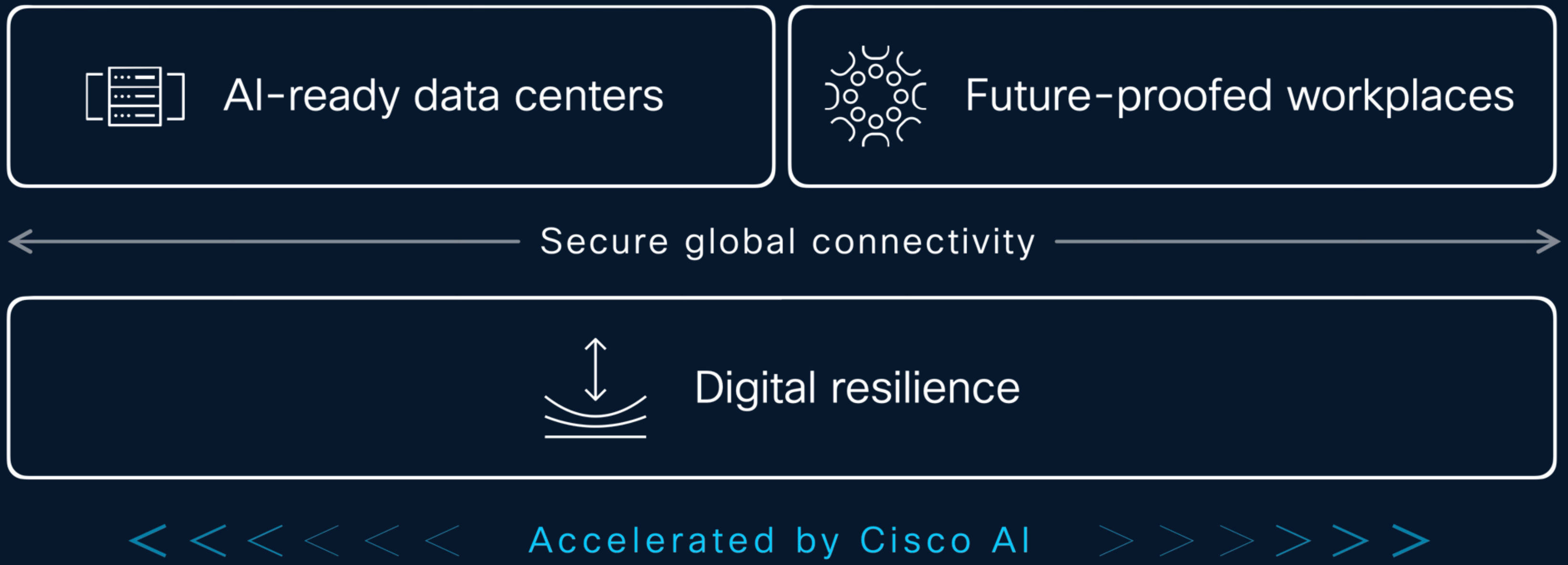
- Foundation AI security model
- ServiceNow partnership

2025 May

- AI infrastructure Partnership
- AI initiatives in the Middle East

2025 June

- Expanded AI Pods
- AI Canvas
- AgenticOps
- Deep Network Model



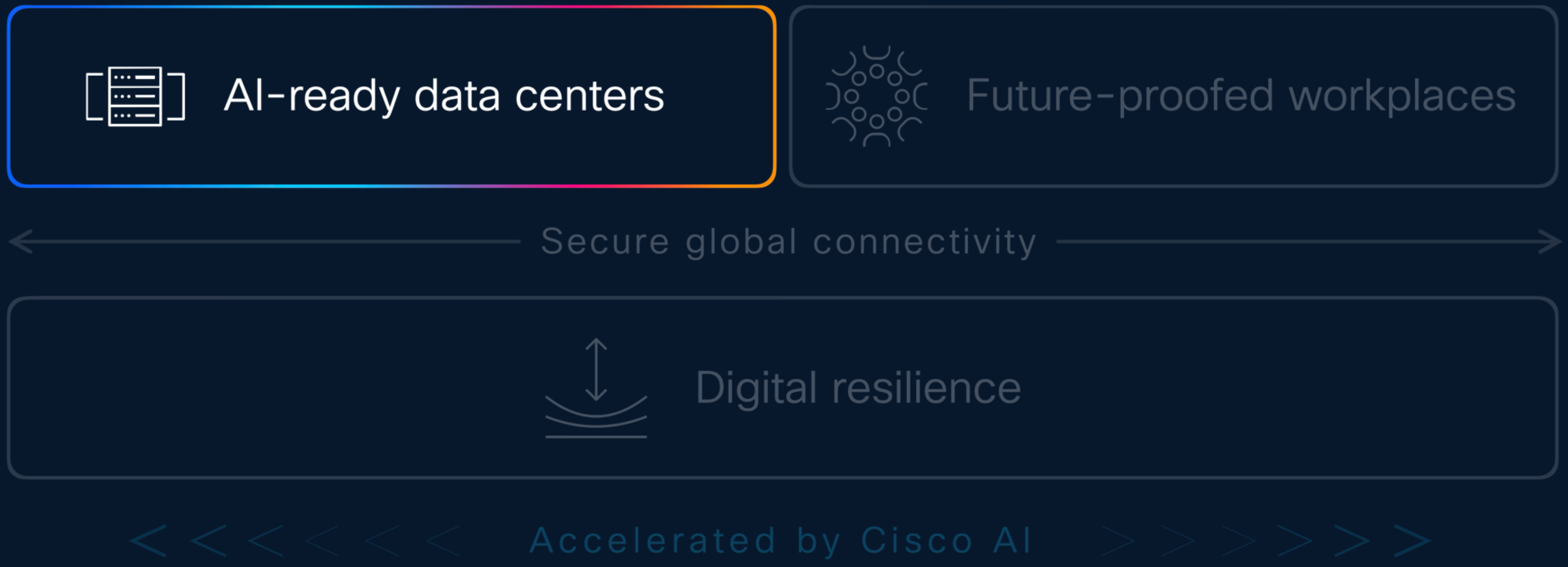
AI-ready data centers

Future-proofed workplaces

Secure global connectivity

Digital resilience

Accelerated by Cisco AI



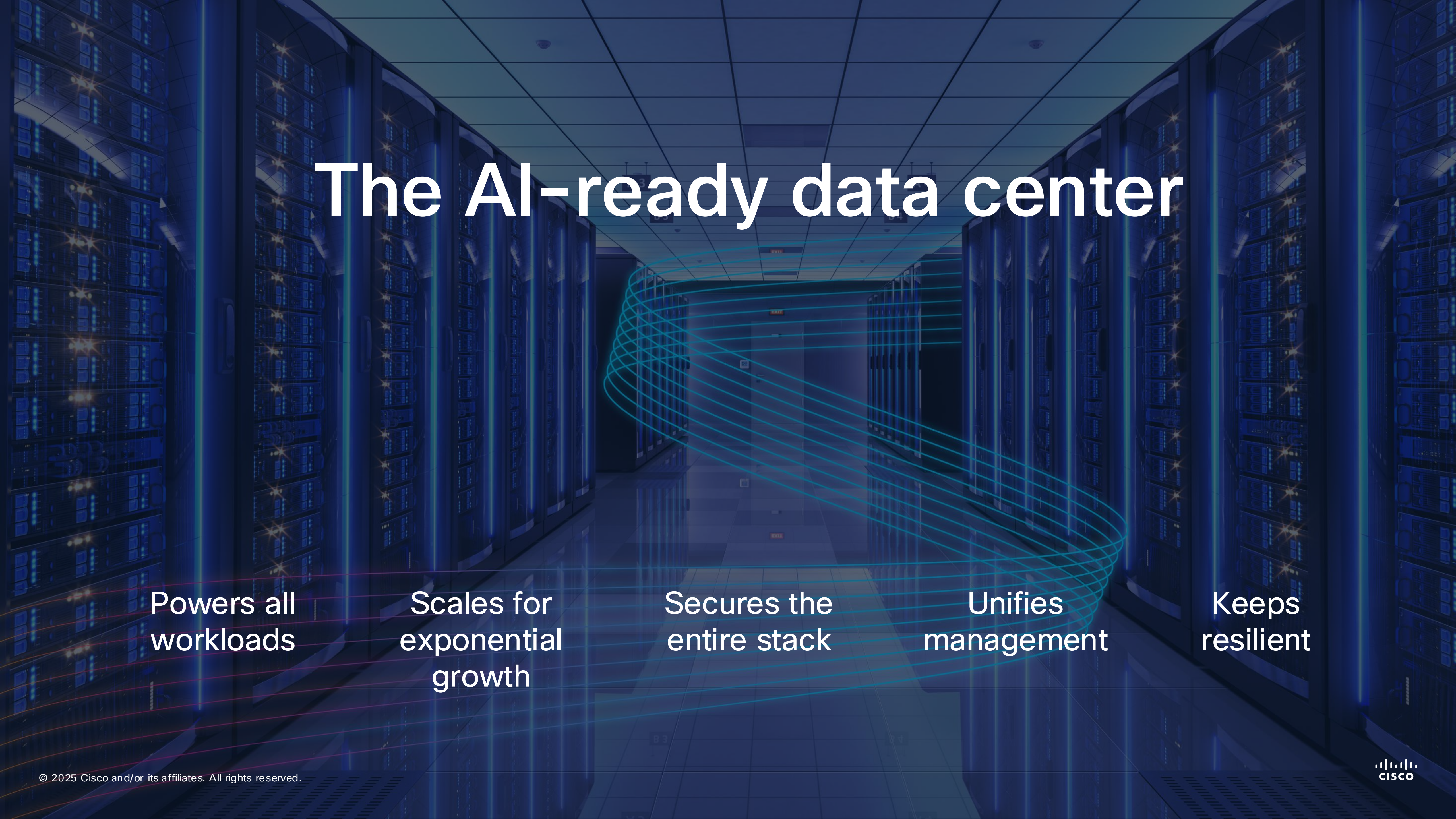
AI-ready data centers

Future-proofed workplaces

Secure global connectivity

Digital resilience

Accelerated by Cisco AI



The AI-ready data center

Powers all workloads

Scales for exponential growth

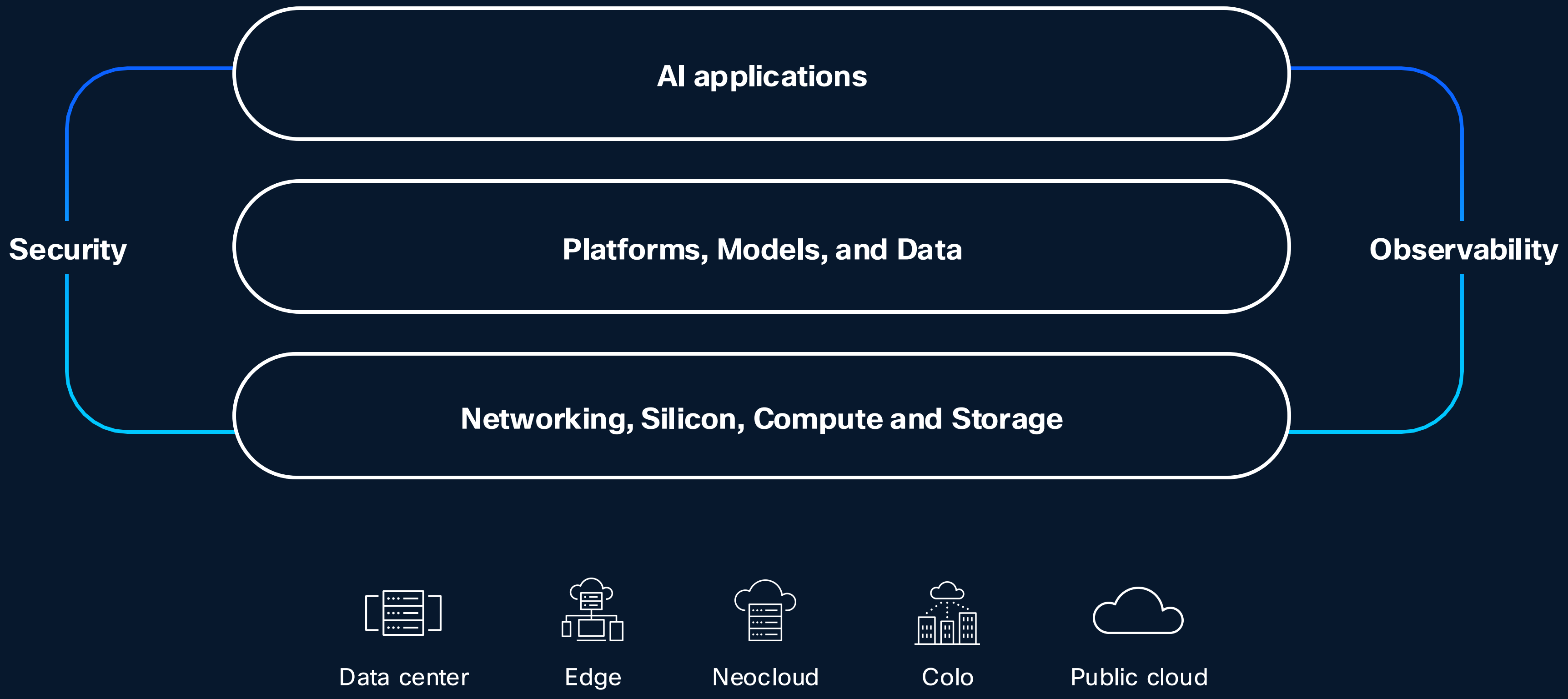
Secures the entire stack

Unifies management

Keeps resilient

Connecting, protecting and powering the entire stack

through our technology and with strategic partners



Security

AI applications

Platforms, Models, and Data

Networking, Silicon, Compute and Storage

Observability

Data center

Edge

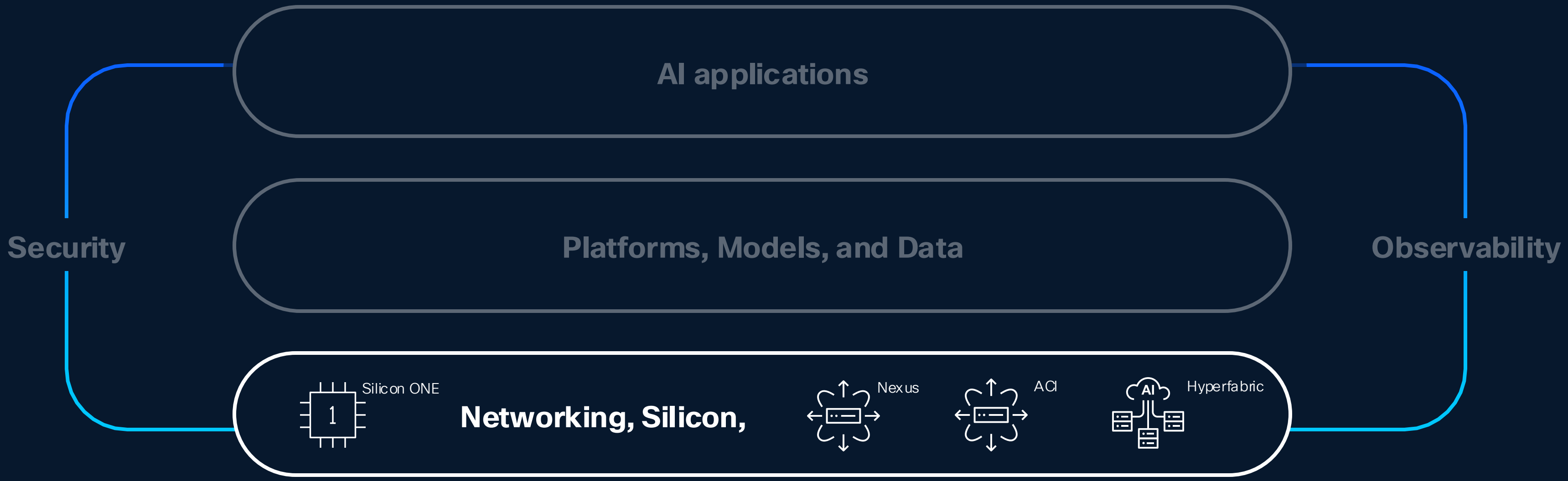
Neocloud

Colo

Public cloud

Data Center Networking

Industry leading data center fabrics coupled with operational simplicity



AI applications

Platforms, Models, and Data

Security

Observability

Networking, Silicon,

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

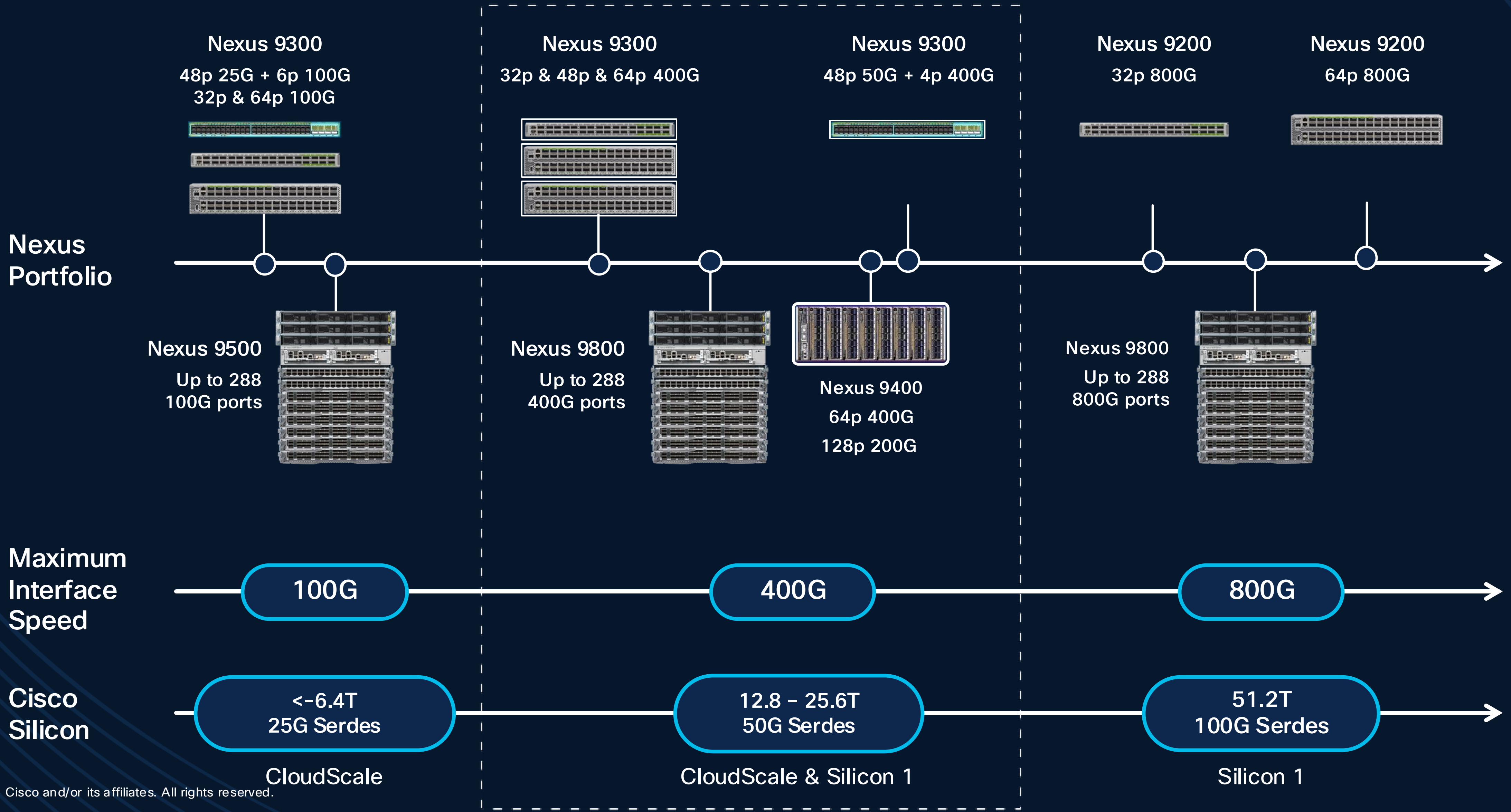
Design, deploy and operate with co-developed reference architectures and best practice from Cisco and NVIDIA

Cisco Data Center Networking Portfolio

Greenfield: new fabrics not being managed by Nexus Dashboard

	Nexus Hyperfabric	Nexus Dashboard
	Hyperfabric	ACI NX-OS
Operating Model	Fabric-as-a-Service Cisco Cloud-Managed Controller	Customer Managed On-Prem Controller
Flexibility & Customization	Prescriptive ↔ Customizable	
IT Staff Network Skillset	Generalist ↔ Specialist	
Deployment Type	Greenfield	Greenfield & Brownfield

Secure and Futureproof Infrastructure



Nexus Portfolio

Nexus 9300
48p 25G + 6p 100G
32p & 64p 100G

Nexus 9500
Up to 288
100G ports

Nexus 9300
32p & 48p & 64p 400G

Nexus 9800
Up to 288
400G ports

Nexus 9300
48p 50G + 4p 400G

Nexus 9400
64p 400G
128p 200G

Nexus 9200
32p 800G

Nexus 9800
Up to 288
800G ports

Nexus 9200
64p 800G

Maximum Interface Speed

100G

400G

800G

Cisco Silicon

<-6.4T
25G Serdes

CloudScale

12.8 – 25.6T
50G Serdes

CloudScale & Silicon 1

51.2T
100G Serdes

Silicon 1

Cisco 6000 series switches

Spine/Leaf: Cisco HF6100-64E

- OSFP form factor
- 64x 100/200/400/800GbE
- Breakouts supported

Spine/Leaf: Cisco HF6100-32D

- 1RU
- QSFP56-DD form factor
- 32x 100/200/400 Gbe
- Breakout supported

Leaf: Cisco HF6100-60L4D

- 4x 100/400GbE QSFP56-DD (16x 100G breakout)
- 60x 10/25/50GbE SFP56

Benefits

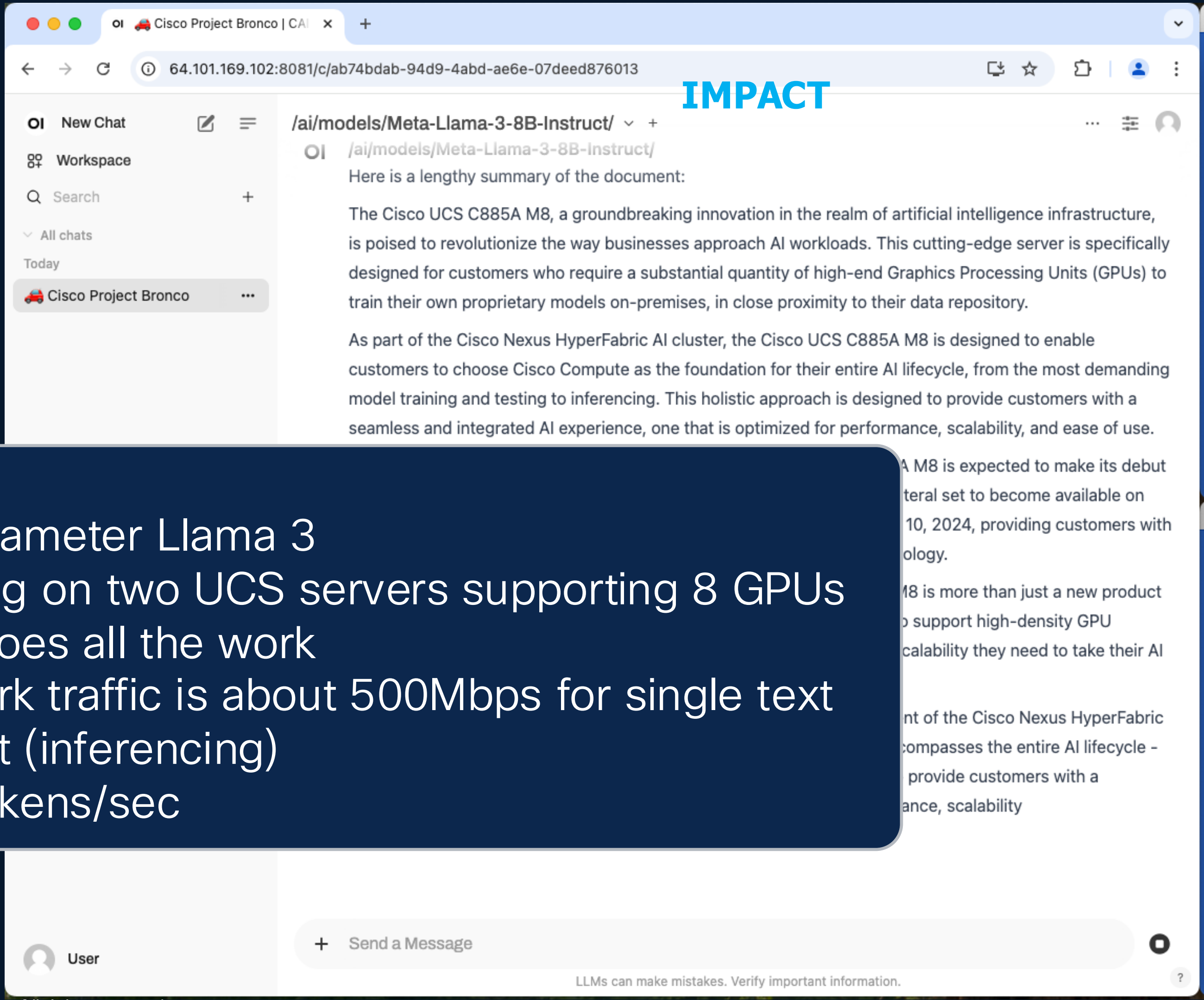
Cloud-managed | Open & extenseable | High performance Silicon One | Optimized for AI workloads

Inferencing Network Impact Example

Single text prompt

AI Monitor

https://github.com/pl247/ai-monitor


Cisco Project Bronco | CAI
64.101.169.102:8081/c/ab74bdab-94d9-4abd-ae6e-07deed876013
New Chat
Workspace
Search
All chats
Today
Cisco Project Bronco
User
/ai/models/Meta-Llama-3-8B-Instruct/
Send a Message
LLMs can make mistakes. Verify important information.

IMPACT

/ai/models/Meta-Llama-3-8B-Instruct/

Here is a lengthy summary of the document:

The Cisco UCS C885A M8, a groundbreaking innovation in the realm of artificial intelligence infrastructure, is poised to revolutionize the way businesses approach AI workloads. This cutting-edge server is specifically designed for customers who require a substantial quantity of high-end Graphics Processing Units (GPUs) to train their own proprietary models on-premises, in close proximity to their data repository.

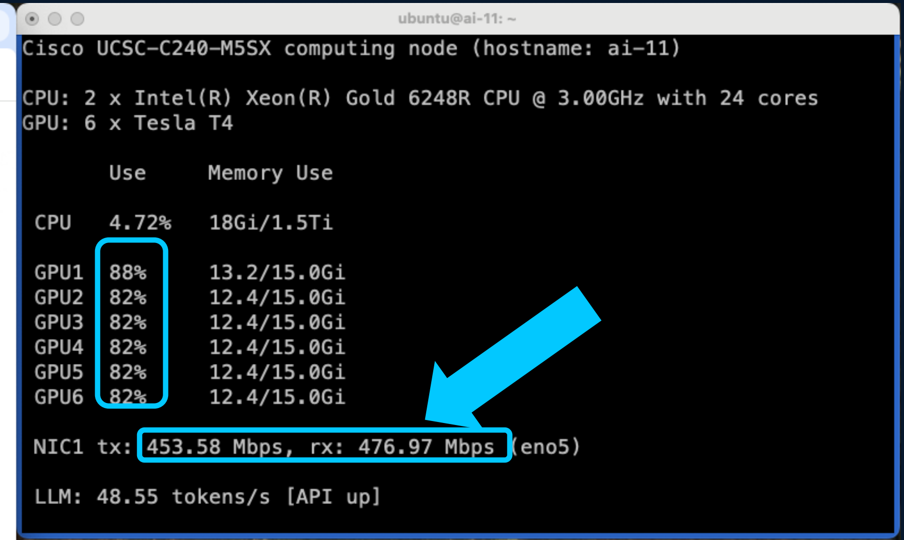
As part of the Cisco Nexus HyperFabric AI cluster, the Cisco UCS C885A M8 is designed to enable customers to choose Cisco Compute as the foundation for their entire AI lifecycle, from the most demanding model training and testing to inferencing. This holistic approach is designed to provide customers with a seamless and integrated AI experience, one that is optimized for performance, scalability, and ease of use.

A M8 is expected to make its debut
teral set to become available on
10, 2024, providing customers with
ology.

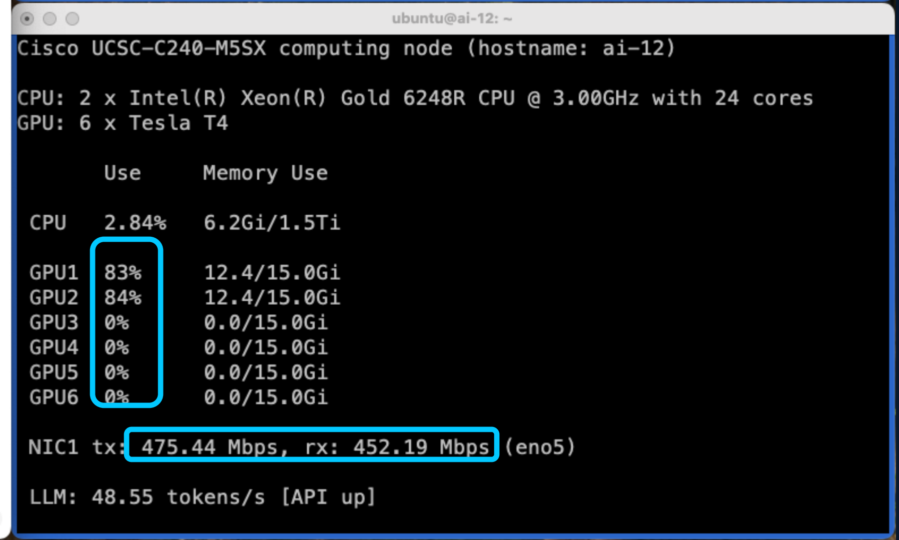
M8 is more than just a new product
support high-density GPU
calability they need to take their AI

nt of the Cisco Nexus HyperFabric
ompasses the entire AI lifecycle -
provide customers with a
ance, scalability

- 8B Parameter Llama 3
- Running on two UCS servers supporting 8 GPUs
- GPU does all the work
- Network traffic is about 500Mbps for single text prompt (inferencing)
- ~50 tokens/sec



```
Cisco UCSC-C240-M5SX computing node (hostname: ai-11)

CPU: 2 x Intel(R) Xeon(R) Gold 6248R CPU @ 3.00GHz with 24 cores
GPU: 6 x Tesla T4

       Use     Memory Use

 CPU   4.72%   18Gi/1.5Ti

 GPU1  88%     13.2/15.0Gi
 GPU2  82%     12.4/15.0Gi
 GPU3  82%     12.4/15.0Gi
 GPU4  82%     12.4/15.0Gi
 GPU5  82%     12.4/15.0Gi
 GPU6  82%     12.4/15.0Gi

 NIC1 tx: 453.58 Mbps, rx: 476.97 Mbps (eno5)

 LLM: 48.55 tokens/s [API up]
```

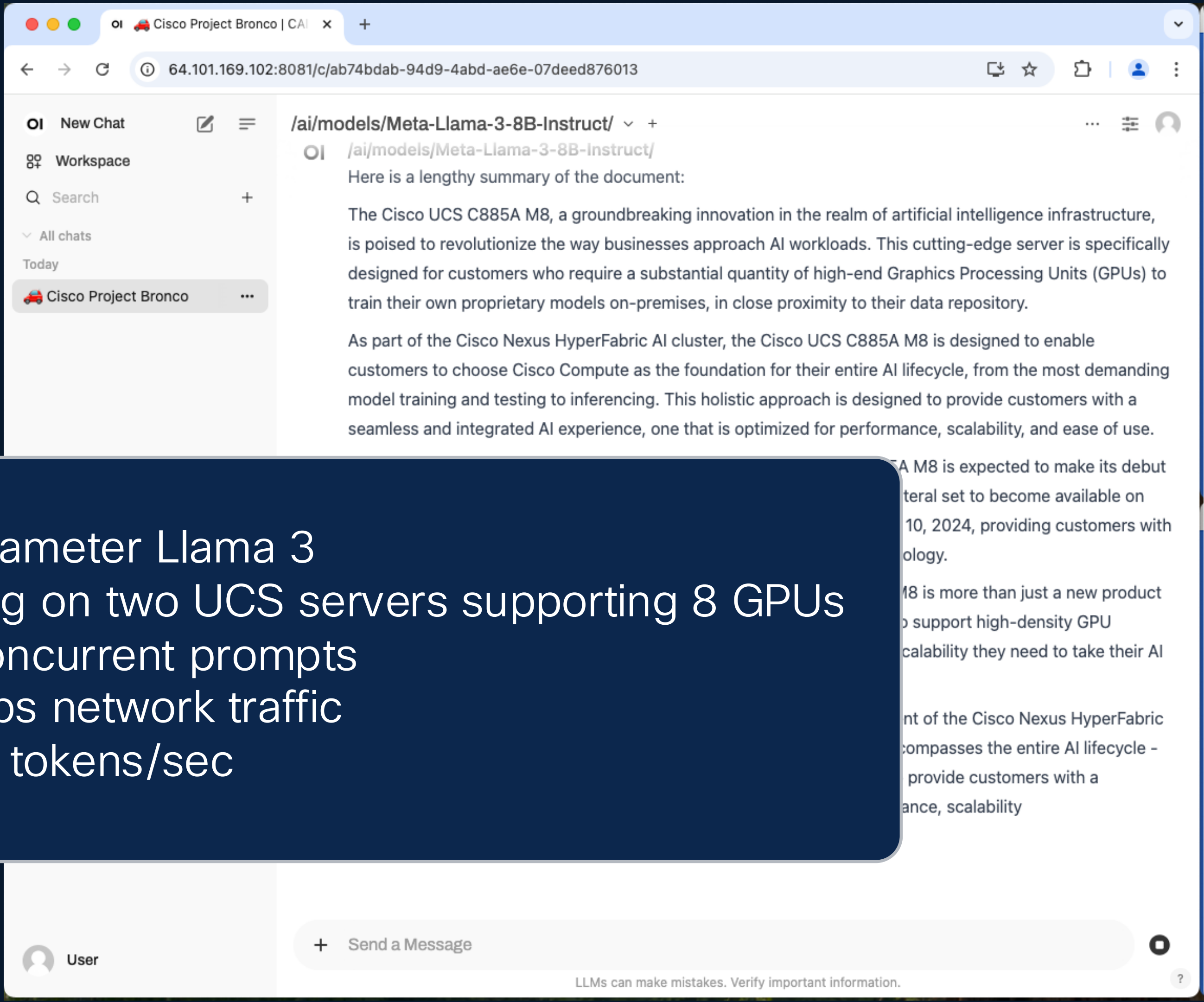


```
Cisco UCSC-C240-M5SX computing node (hostname: ai-12)

CPU: 2 x Intel(R) Xeon(R) Gold 6248R CPU @ 3.00GHz with 24 cores
GPU: 6 x Tesla T4

       Use     Memory Use

 CPU   2.84%   6.2Gi/1.5Ti

 GPU1  83%     12.4/15.0Gi
 GPU2  84%     12.4/15.0Gi
 GPU3  0%      0.0/15.0Gi
 GPU4  0%      0.0/15.0Gi
 GPU5  0%      0.0/15.0Gi
 GPU6  0%      0.0/15.0Gi

 NIC1 tx: 475.44 Mbps, rx: 452.19 Mbps (eno5)

 LLM: 48.55 tokens/s [API up]
```

Inferencing Network Impact Example

200 concurrent connections

https://github.com/pl247/ai-monitor



Cisco Project Bronco | CA
64.101.169.102:8081/c/ab74bdab-94d9-4abd-ae6e-07deed876013
New Chat
Workspace
Search
All chats
Today
Cisco Project Bronco
User
/ai/models/Meta-Llama-3-8B-Instruct/
Send a Message
LLMs can make mistakes. Verify important information.

/ai/models/Meta-Llama-3-8B-Instruct/

Here is a lengthy summary of the document:

The Cisco UCS C885A M8, a groundbreaking innovation in the realm of artificial intelligence infrastructure, is poised to revolutionize the way businesses approach AI workloads. This cutting-edge server is specifically designed for customers who require a substantial quantity of high-end Graphics Processing Units (GPUs) to train their own proprietary models on-premises, in close proximity to their data repository.

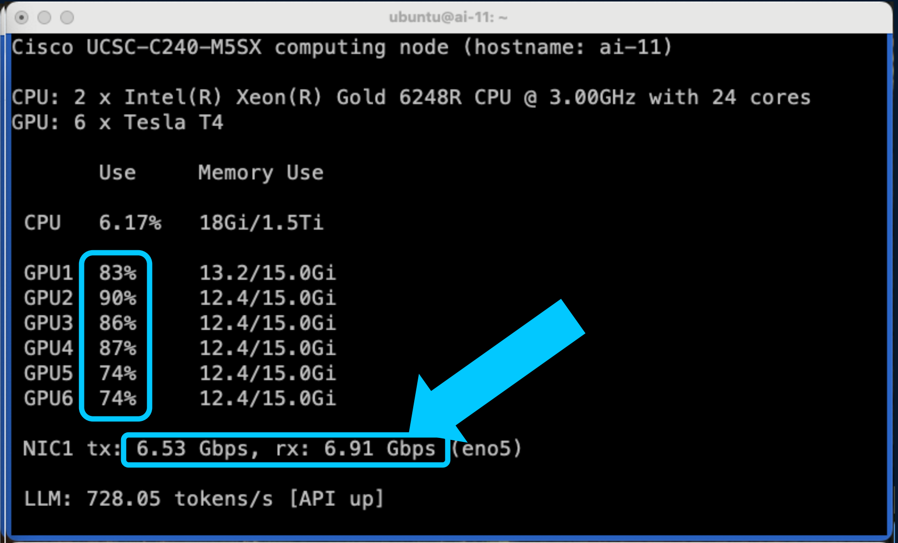
As part of the Cisco Nexus HyperFabric AI cluster, the Cisco UCS C885A M8 is designed to enable customers to choose Cisco Compute as the foundation for their entire AI lifecycle, from the most demanding model training and testing to inferencing. This holistic approach is designed to provide customers with a seamless and integrated AI experience, one that is optimized for performance, scalability, and ease of use.

A M8 is expected to make its debut teral set to become available on 10, 2024, providing customers with ology.

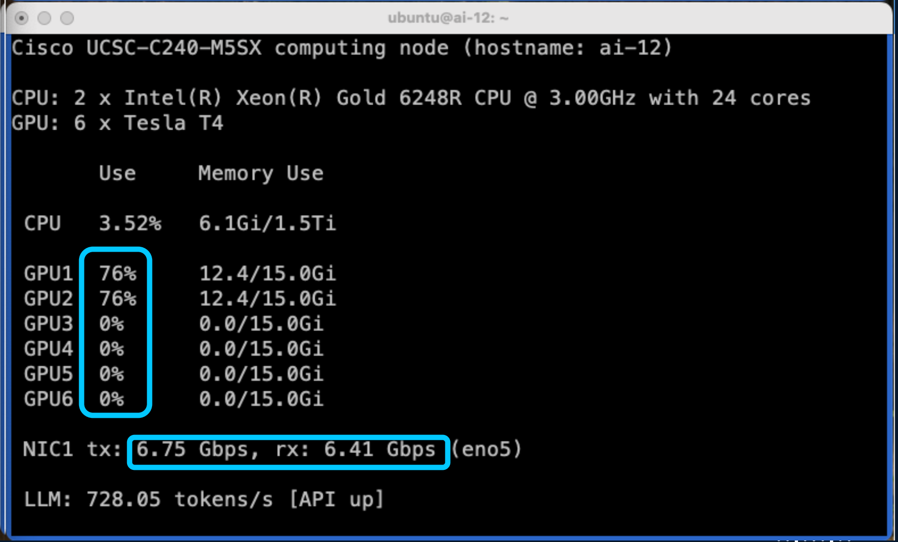
18 is more than just a new product support high-density GPU calability they need to take their AI

nt of the Cisco Nexus HyperFabric compasses the entire AI lifecycle - provide customers with a ance, scalability

- 8B Parameter Llama 3
- Running on two UCS servers supporting 8 GPUs
- 200 concurrent prompts
- 6+ Gbps network traffic
- ~700+ tokens/sec



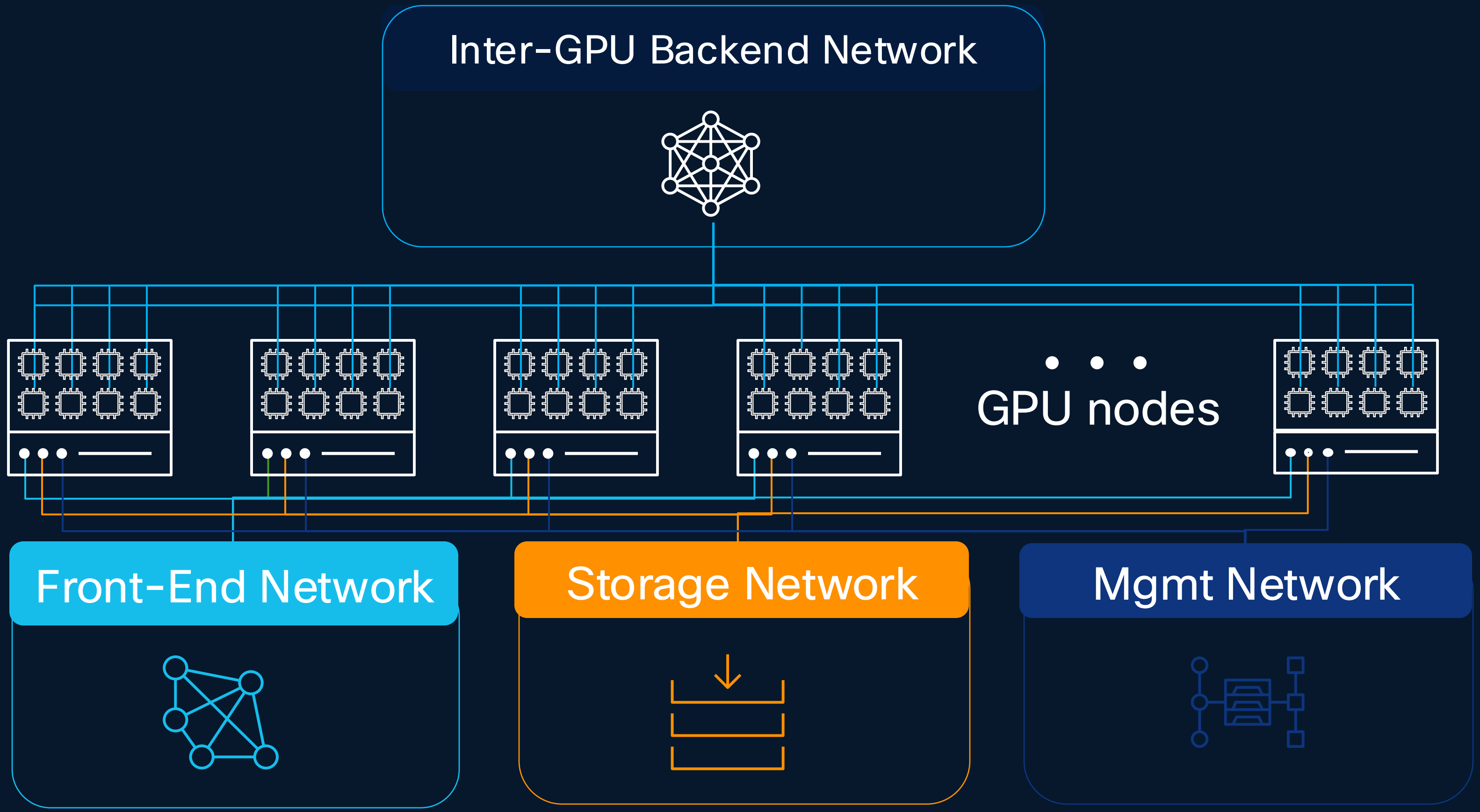
```
Cisco UCSC-C240-M5SX computing node (hostname: ai-11)

CPU: 2 x Intel(R) Xeon(R) Gold 6248R CPU @ 3.00GHz with 24 cores
GPU: 6 x Tesla T4

       Use     Memory Use

 CPU   6.17%   18Gi/1.5Ti

 GPU1  83%     13.2/15.0Gi
 GPU2  90%     12.4/15.0Gi
 GPU3  86%     12.4/15.0Gi
 GPU4  87%     12.4/15.0Gi
 GPU5  74%     12.4/15.0Gi
 GPU6  74%     12.4/15.0Gi

 NIC1 tx: 6.53 Gbps, rx: 6.91 Gbps (eno5)

 LLM: 728.05 tokens/s [API up]
```



```
Cisco UCSC-C240-M5SX computing node (hostname: ai-12)

CPU: 2 x Intel(R) Xeon(R) Gold 6248R CPU @ 3.00GHz with 24 cores
GPU: 6 x Tesla T4

       Use     Memory Use

 CPU   3.52%   6.1Gi/1.5Ti

 GPU1  76%     12.4/15.0Gi
 GPU2  76%     12.4/15.0Gi
 GPU3  0%      0.0/15.0Gi
 GPU4  0%      0.0/15.0Gi
 GPU5  0%      0.0/15.0Gi
 GPU6  0%      0.0/15.0Gi

 NIC1 tx: 6.75 Gbps, rx: 6.41 Gbps (eno5)

 LLM: 728.05 tokens/s [API up]
```

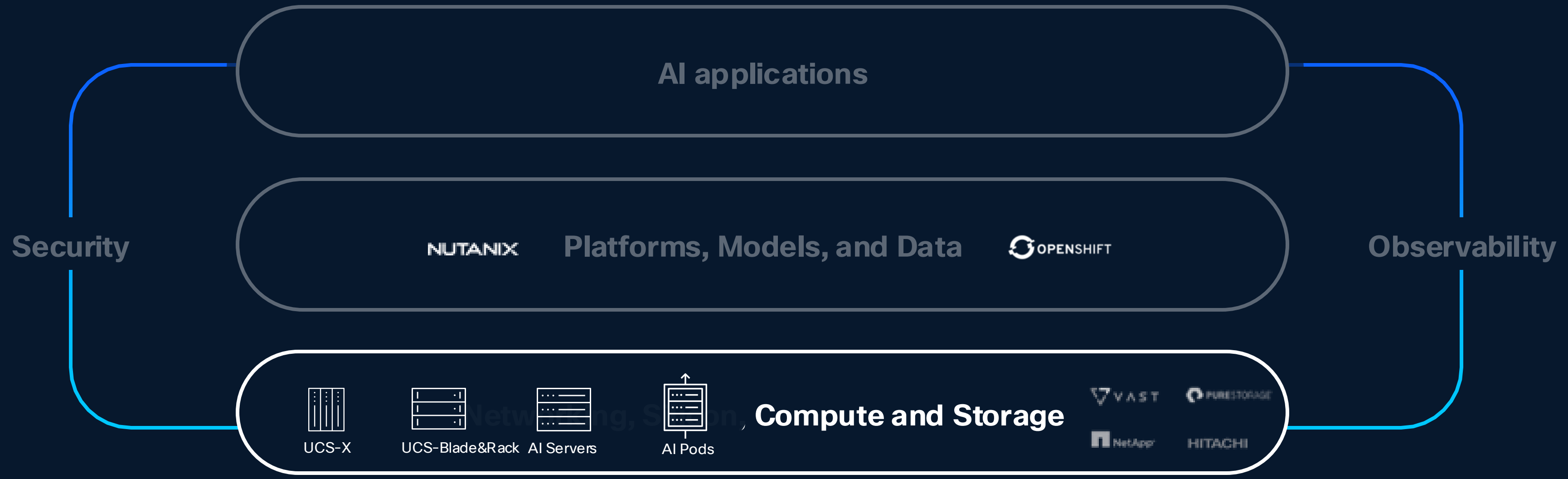
AI Networking

Dedicated Backend Networks



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 Multcloud

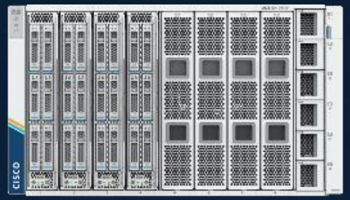
Modernize with validated converged and hyperconverged platforms that support distributed applications

Cisco UCS Compute Portfolio

MAINSTREAM ENTERPRISE SERVERS

UCS X-Series
X9508 Chassis

IFM Module



UCS X-Series Direct

UCS X210c M7

UCS X210c M8

UCS X410c M7

UCS B200 M6

UCS X215c M8

UCS C240 M8E3S
36 EDSFF E3.S 1T

UCS C240 M8SX
28 HDD/SDD/NVMe

UCS C240 M8L
16 LFF + 4 SFF

UCS C240 M7SN
28 NVMe

UCS C240 M6S
14 SSD/HDD Media drive

UCS C240 M6N
14 NVMe Media Drive

UCS C220 M8E3S
16 EDSFF E3.S 1T

UCS C220 M8S
10 HDD/SSD/NVMe

UCS C220 M7N
10 NVMe

UCS C245 M8SX
28 HDD/SDD

UCS C225 M8S
10 HDD/SSD

UCS C225 M8N
10 NVMe

AI SERVERS

UCS C885A M8
8RU Dense GPU Server

UCS C845A M8
4RU MGX Server

Cisco AI Compute Portfolio

Unified approach to accelerated AI compute

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

GPU Accelerated

GPU Optimized

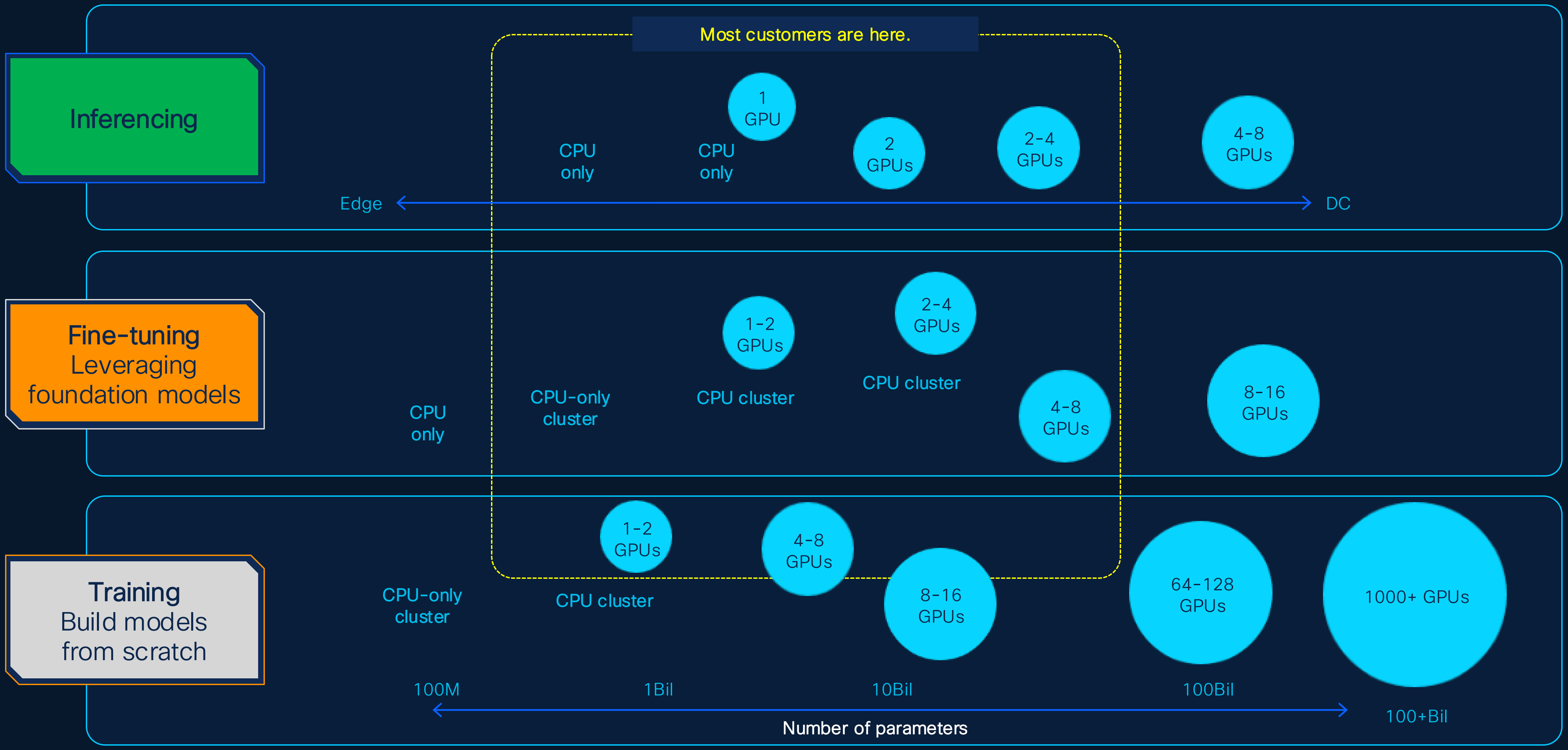
Unified Edge

Build the model | Training

Optimize the model | Fine-tuning and RAG

Use the model | Inferencing

Enterprise AI infrastructure mapping



Most customers are here.

Inferencing

CPU only

CPU only

1 GPU

2 GPUs

2-4 GPUs

4-8 GPUs

Edge ← → DC

Fine-tuning
Leveraging foundation models

CPU only

CPU-only cluster

1-2 GPUs

CPU cluster

2-4 GPUs

CPU cluster

4-8 GPUs

8-16 GPUs

Training
Build models from scratch

CPU-only cluster

1-2 GPUs

CPU cluster

4-8 GPUs

8-16 GPUs

64-128 GPUs

1000+ GPUs

100M

1Bil

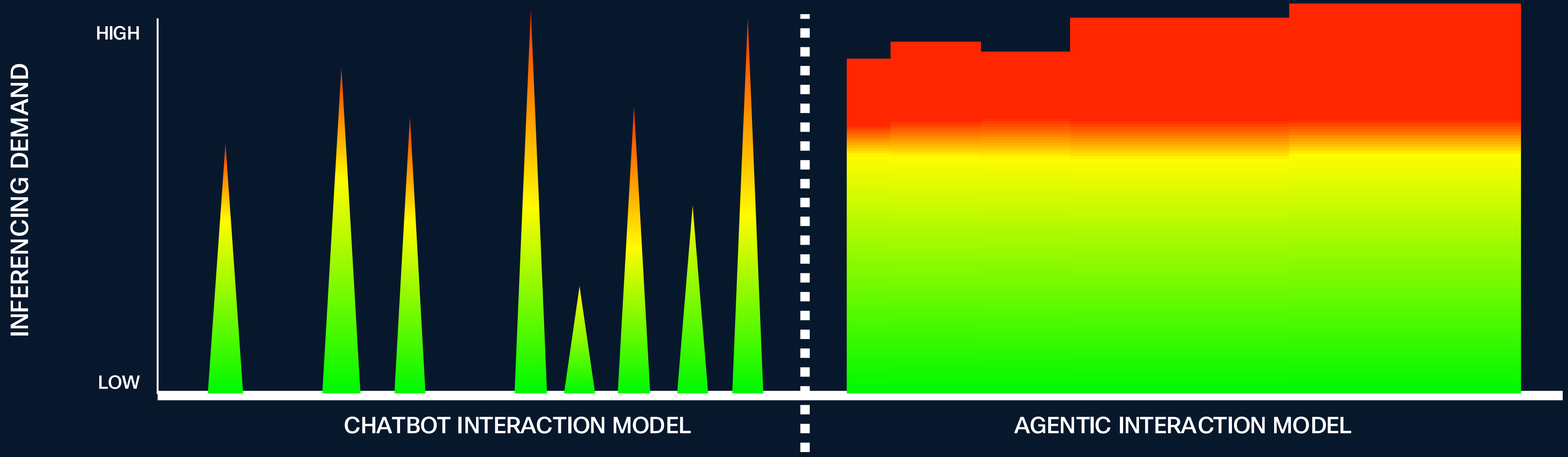
10Bil

100Bil

100+Bil

Number of parameters

A new level of inferencing demand with Agentic AI



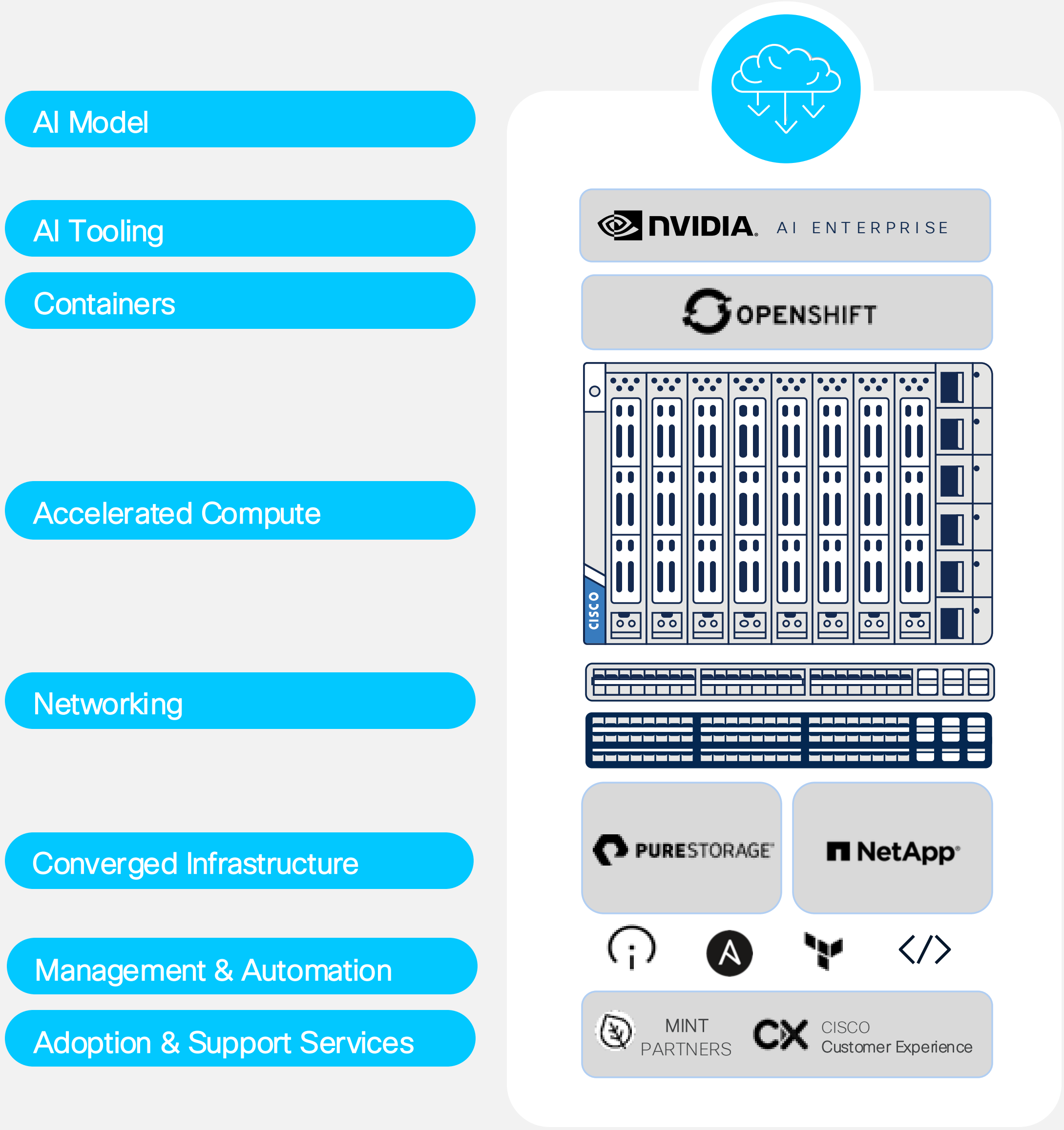
AI PODs

Deploying AI with confidence

Confidently deploy AI-ready infrastructure with pre-designed full stack architecture bundles for targeted AI use cases.

Leverage automation frameworks for rapid deployment and adoption of infrastructure.

Operate with best-in-class single-support model for your AI deployment architecture, include enterprise support for select Operations Support System (OSS) tools and libraries



AI Model

AI Tooling

Containers

Accelerated Compute

Networking

Converged Infrastructure

Management & Automation

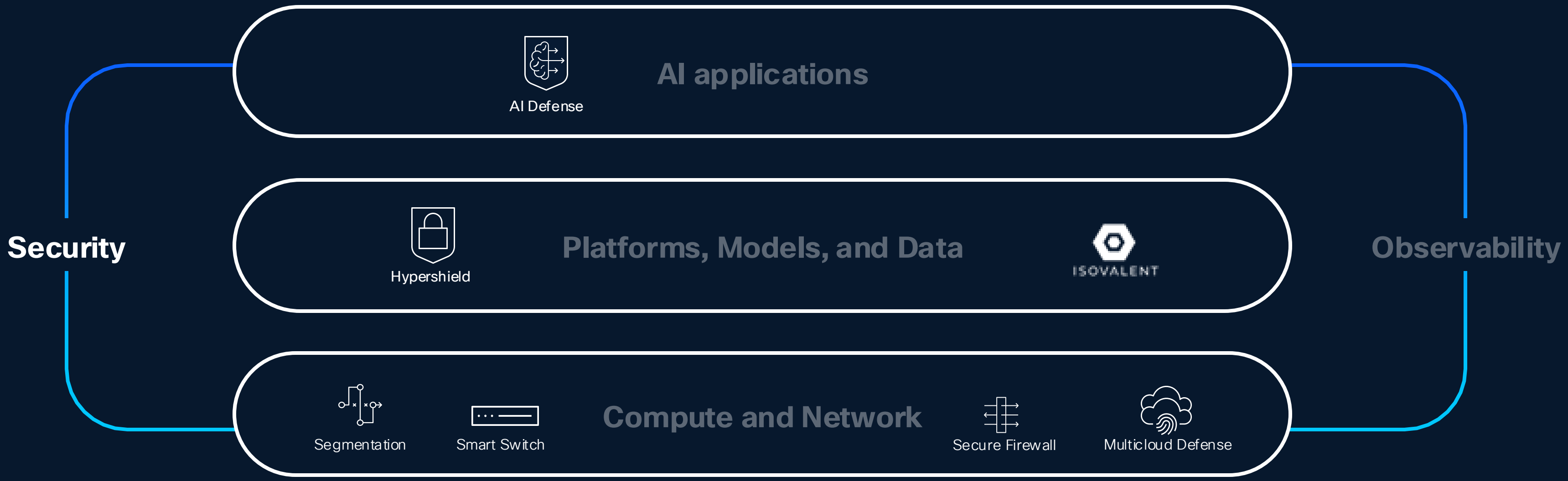
Adoption & Support Services

Cisco AI PODs for Inferencing

Typical use case	Edge Inferencing (7B-13B Parameter)	RAG Augmented Inferencing (13B-40B+ Parameter)	Large-Scale RAG Augmented Inferencing	Scale-Out Inferencing Cluster (Inferencing Multiple Models)
Hardware Specification (Required)	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 • X-Series FI9108 100G	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 2x Fabric Interconnect • 6536 100G	Large 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 2x Fabric Interconnect • 6536 100G	Scale-Out 4x X210C compute nodes • 8x Intel 5th Gen 6548Y+ • 4 TB System Memory • 20x 1.6 TB NVMe drives 4x X440p PCIe • 8x NVIDIA L40S 2x Fabric Interconnect • 6536 100G
Software specification (Required)	Cisco Intersight • Essentials Nvidia AI Enterprise • Essentials	Cisco Intersight • Essentials Nvidia AI Enterprise • Essentials	Cisco Intersight • Essentials Nvidia AI Enterprise • Essentials	Cisco Intersight • Essentials Nvidia AI Enterprise • Essentials
Default Components (Optional)	OpenShift • OpenShift Container Platform • Single-Node Controller’	OpenShift • OpenShift Container Platform • X210c Control Plane Cluster Networking • 2x Nexus switches (93600CD-GX or 9332D-GX2B) • Nexus Dashboard appliance	OpenShift • OpenShift Container Platform • X210c Control Plane Cluster Networking • 2x Nexus switches (93600CD-GX or 9332D-GX2B) • Nexus Dashboard appliance	OpenShift • OpenShift Container Platform • X210c Control Plane Cluster Networking • 2x Nexus switches (93600CD-GX or 9332D-GX2B) • Nexus Dashboard appliance
Add-On	CI Storage FlashStack FlexPod	CI Storage FlashStack FlexPod	CI Storage FlashStack FlexPod	CI Storage FlashStack FlexPod

Full Stack Protection

Security from ground to cloud



Full Stack Protection

Hyper Distributed Security

Reduce attack surface and ensure compliance with consistent security policies

AI-native Management

Real-time visibility, streamlined workflows with centralized control and AI-driven insights

AI Model Protection

Discover shadow AI, deploy AI guardrails and protect models and apps during runtime

Nexus Smart Switch

Redefining Network Security

Programmability Performance Flexibility Efficiency

Cisco Nexus 9300 Services Accelerated Switch

- Rich NX-OS Features and Services
- High-speed connectivity and scalable performance
- Optimized for latency and power efficiency

Routing/ Switching

EVPN/MPLS/ VXLAN/SR

Rich Telemetry

Line-rate Encryption

Power Efficiency

- Software-defined Stateful Services
- Programmable at all layers: add new services without HW change
- Scale-out services with wire-rate performance

Distributed Firewall

Future Use Cases

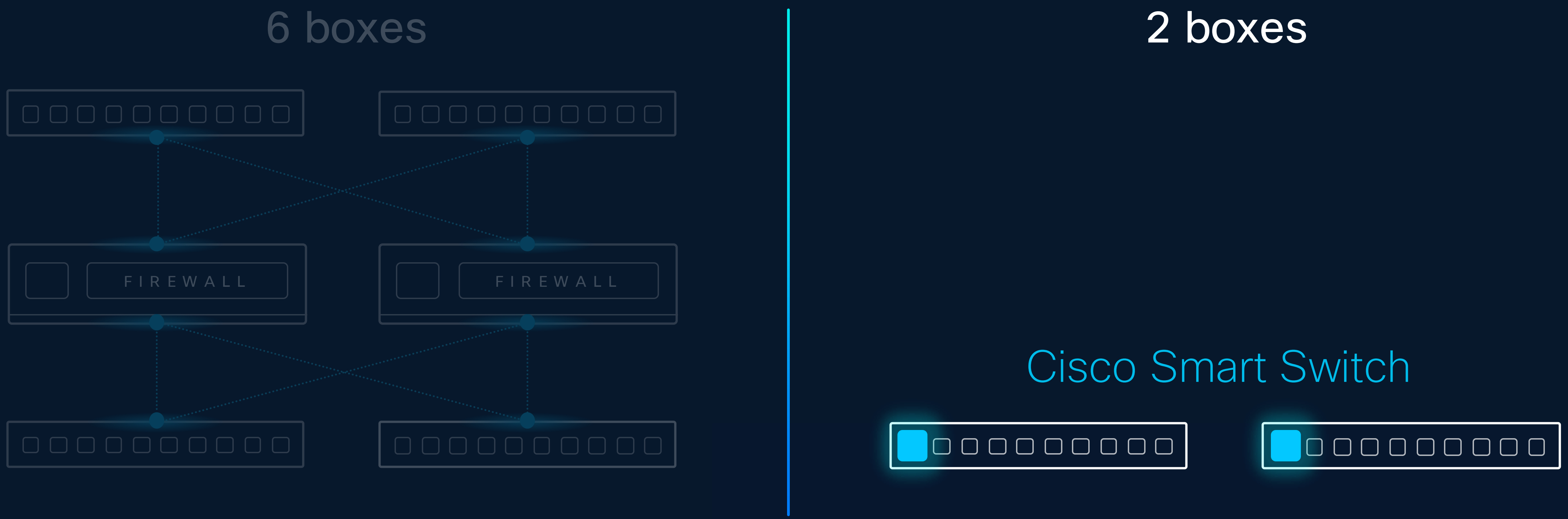
IPSEC Encryption

Large Scale NAT

Event-Based Telemetry

DoS Protection

Unprecedented ROI



Power ↓

Software licenses ↓

Optics ↓

Support contracts ↓

Cables ↓

• • •

Cisco Smart Switches integrated with Hypershield security

Cisco N9300 Series Smart Switches

N9324C-SE1U

24-port 100G

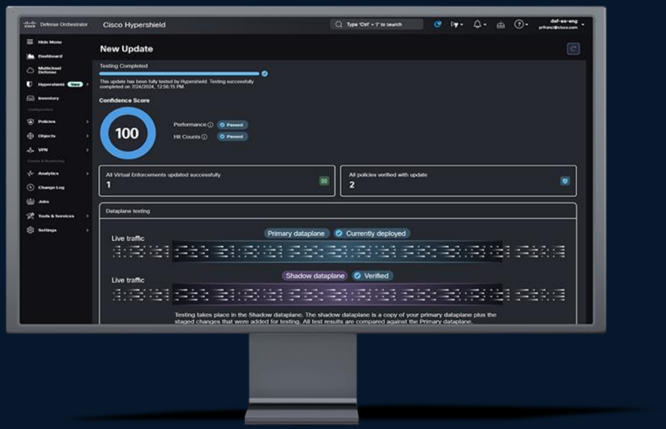
800G Services Throughput

N9348Y2C6D-SE1U

48-port 25G, 6-port 400G, 2-port 100G

800G Services Throughput

Cisco Hypershield



New Update

Testing Completed

This update has been fully tested by Hypershield. Testing successfully completed on 7/24/2024, 12:56:15 PM.

Confidence Score

100

Performance: Passed

Hit Counts: Passed

All Virtual Enforcements updated successfully: 1

All policies verified with update: 2

Dataplane testing

Live traffic — Primary dataplane — Currently deployed

Live traffic — Shadow dataplane — Verified

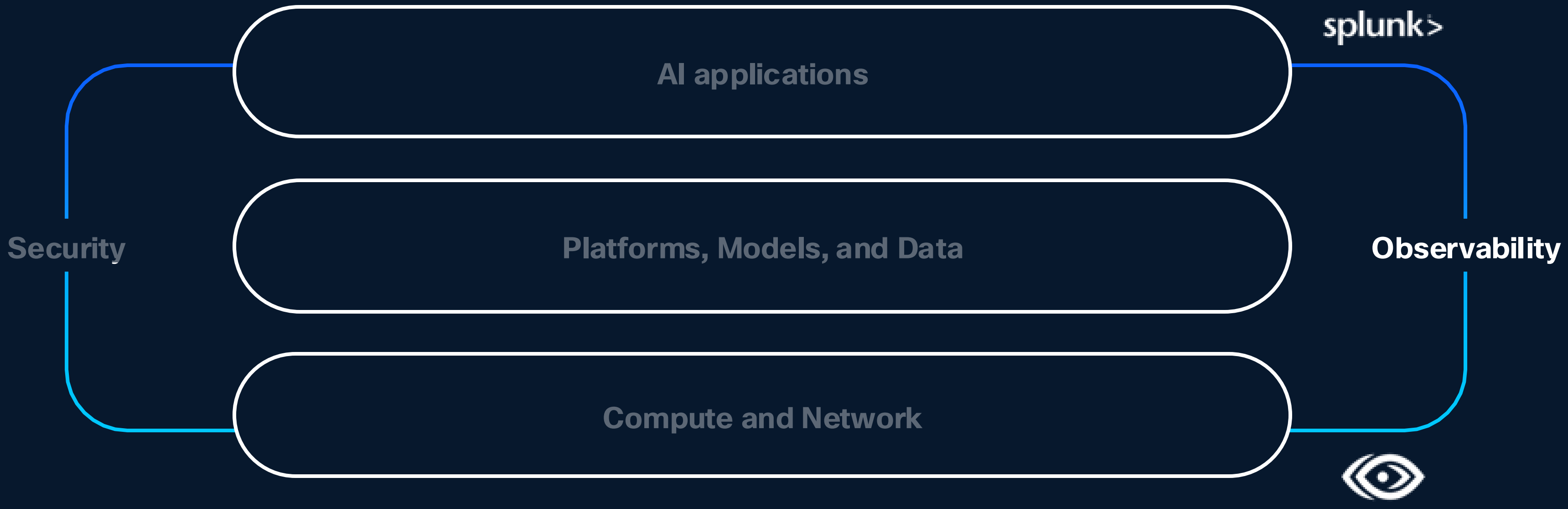
Testing takes place in the Shadow dataplane. The shadow dataplane is a copy of your primary dataplane plus the staged changes that were added for testing. All test results are compared against the Primary dataplane.

Use cases

- Cloud Edge
- Zone-based segmentation
- Data Center Interconnect (DCI)
- Top of Rack segmentation and enforcement

Observability

End to end visibility and insights to stay secure, compliant and resilient



See Everything

Complete Visibility

Surface insights and correlate across the full stack, every location and each experience

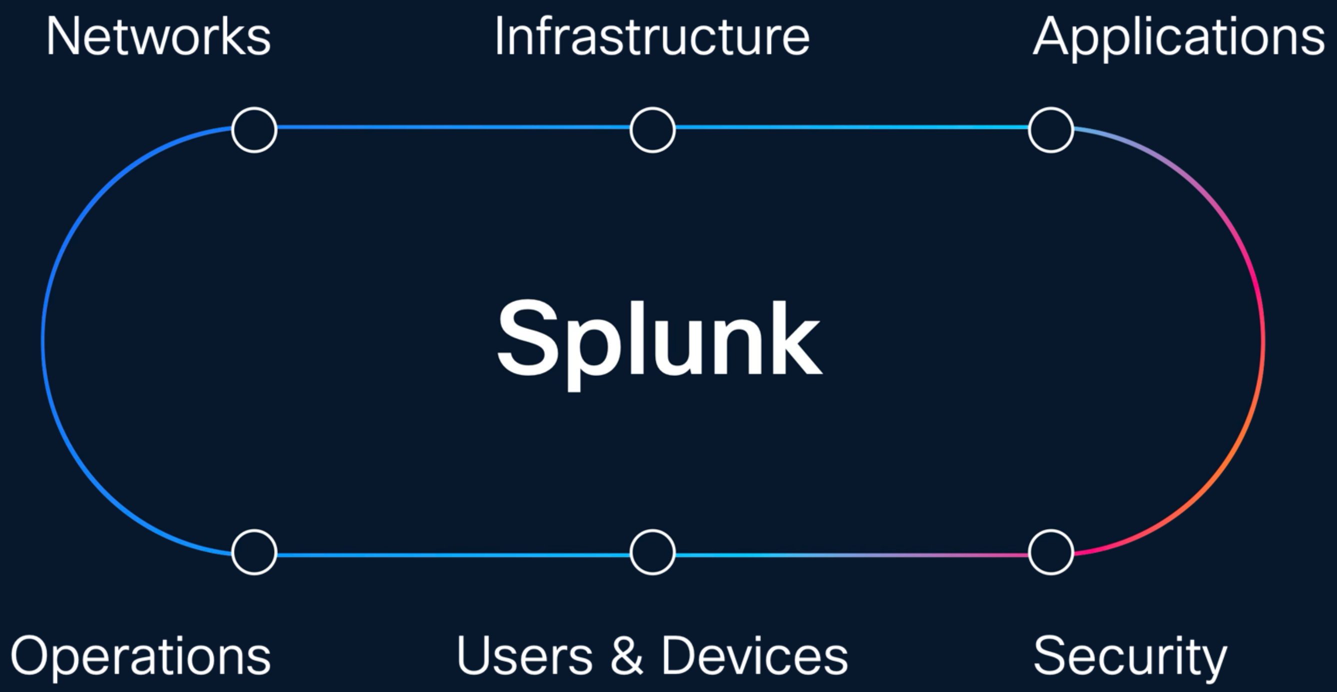
Service Intelligence

Ai driven incident prediction, detection and resolutions pre-integrated with the Cisco portfolio

Integrated AI

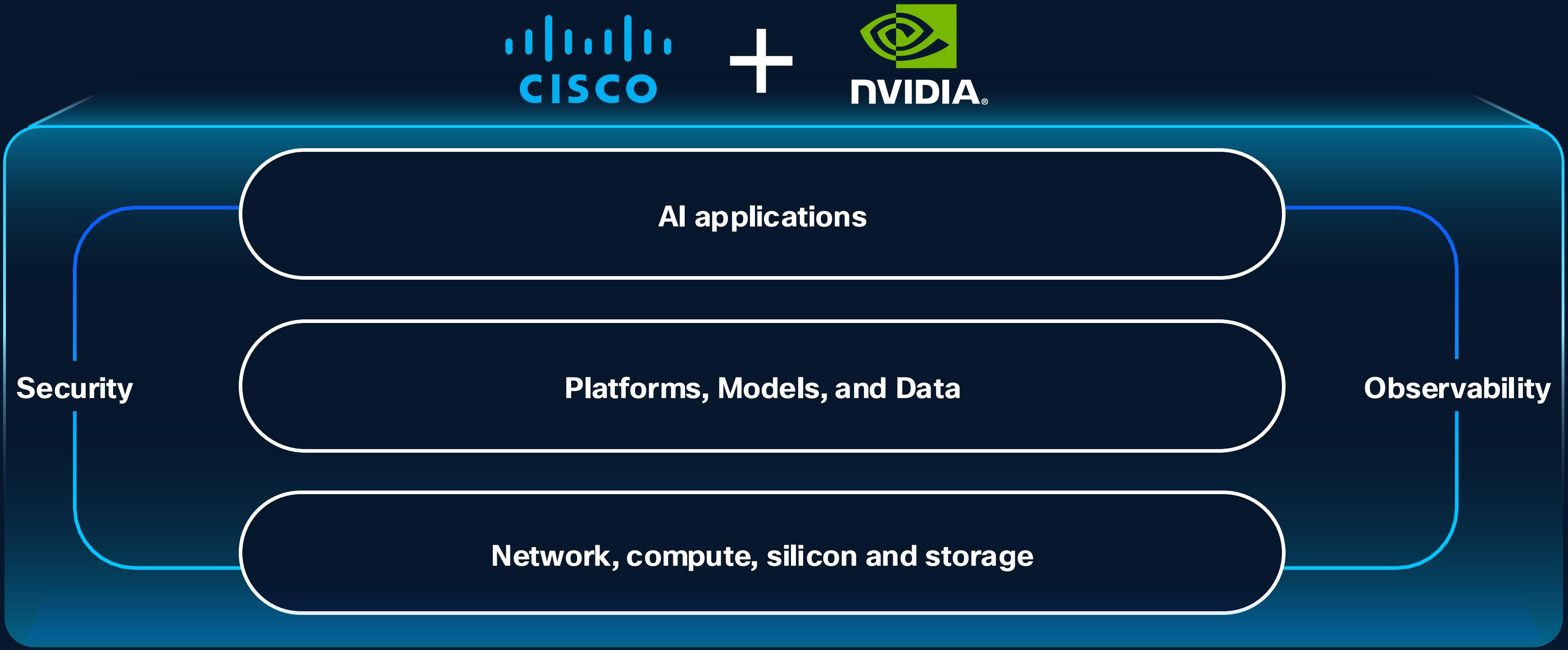
Use AI assistants to develop complex analysis or bring you own models via out-of-the-box or ecosystem tooling

We have **unmatched ability** to solve this problem



Cisco Secure AI Factory

Reference architecture for Cisco and NVIDIA infrastructure working together


CISCO
+
NVIDIA
AI applications
Security
Platforms, Models, and Data
Observability
Network, compute, silicon and storage

Secure AI

Security First AI

Embedded security at every layer ensures the models you build, or use are compliant and protected

High Performance

High performance networking, compute, storage and security delivered as vertically integrated or modular stacks

Pre-validated

Reduce risk and accelerate deployment with certified Nvidia Enterprise reference architecture (ERA) and Cisco validated designs (CVD)

Only Cisco unifies **networking, compute, security,** and **observability** to deliver AI-ready data centers.

Thank You

cisco