

Cisco AI Strategy and Innovation

Leading in the agentic AI era

Garsen Naidu

Director, Cloud & AI Infrastructure METAC, Service Providers & Sovereign AI EMEA

Mohamed Abdallah

Leader, Cloud & AI Infrastructure METAC



October 7, 2026

Cisco Strategy

Connecting people,
places, apps, data,
and things

We securely
connect
everything to
make anything
possible

Protecting against
threats, downtime,
and risk

Executing our strategy

Build

50%+

Of R&D invested in
cloud, AI, and
cybersecurity

Buy

240+

Acquisitions and
investments over the past
10 years

Partner

Ecosystem

Strategic, technology,
regulatory, commercial, and
GTM partnerships

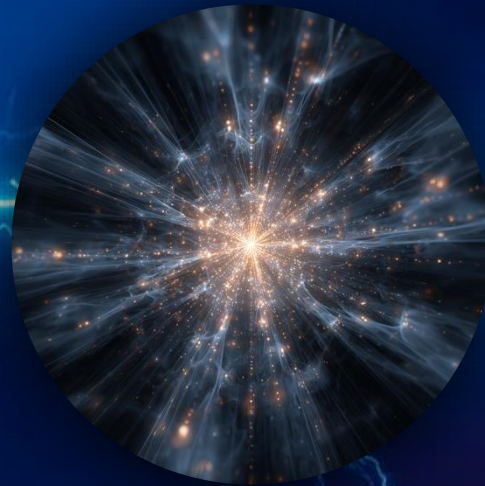
The pace of AI innovation is staggering

AI challenges you to deploy new
technology **faster than ever**
before.





Chatbots
Productivity tools



Agents
Digital coworkers



Physical
Robots and sensors

Percentage of CEOs who say
AI will significantly Impact their industry

2024 69%

Percentage of CEOs who say
AI will significantly Impact their industry

2025 94%

Data center



On premises edge



AI increasing demand for local processing and networking

25%

2027

75%

DATA PROCESSING

90%

2021

10%



AI-Ready Data Centers



Future-Proofed Workplaces

← Secure Global Connectivity →



Digital Resilience

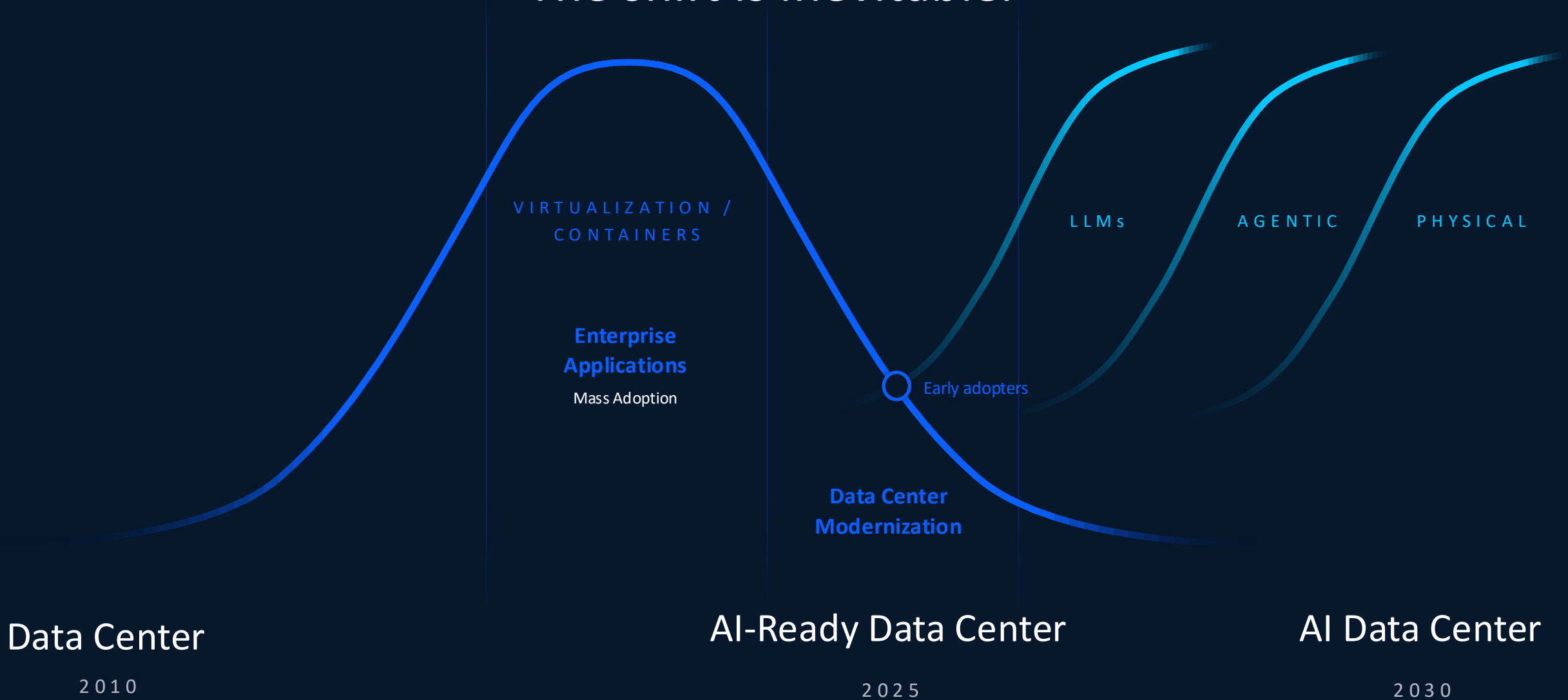


Accelerated by Cisco AI



What we know

AI is no longer optional.
The shift is inevitable.



Cisco is foundational to the world's data center buildouts

Hyperscalers

Trillion-parameter model training

Neoclouds

Cisco Confidential

Service providers

Scale-out GPU clusters for
AI cloud services

Sovereign cloud

11

Large-scale agentic AI

Enterprises

Top cloud providers will be about half of all data center capex by 2030.

The other ~half of 2030 capex—hundreds of buyers, the fastest-growing class.



Experimentation
to **production**
scale

Hyperscale to
distributed AI

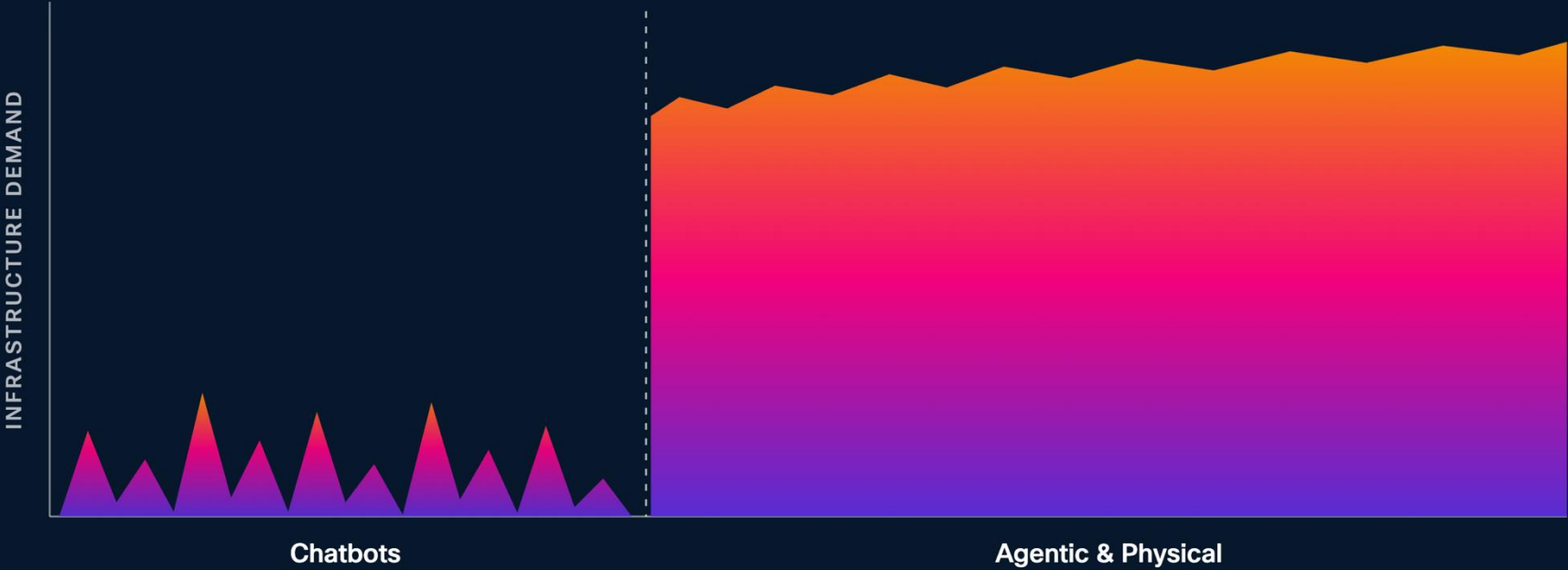
Server scale to
rack scale

Inference is the new workload

Agents are the new workforce

Tokens are the new currency

Control is the new moat



AGENTS OPERATE AT MACHINE SPEED & SCALE

Inference is the
new workload

Agents are the
new workforce

Tokens are the
new currency

Control is the
new moat

On a laptop

In a contact center

In a meeting room

In an enterprise smart switch

Agents are going to be running everywhere

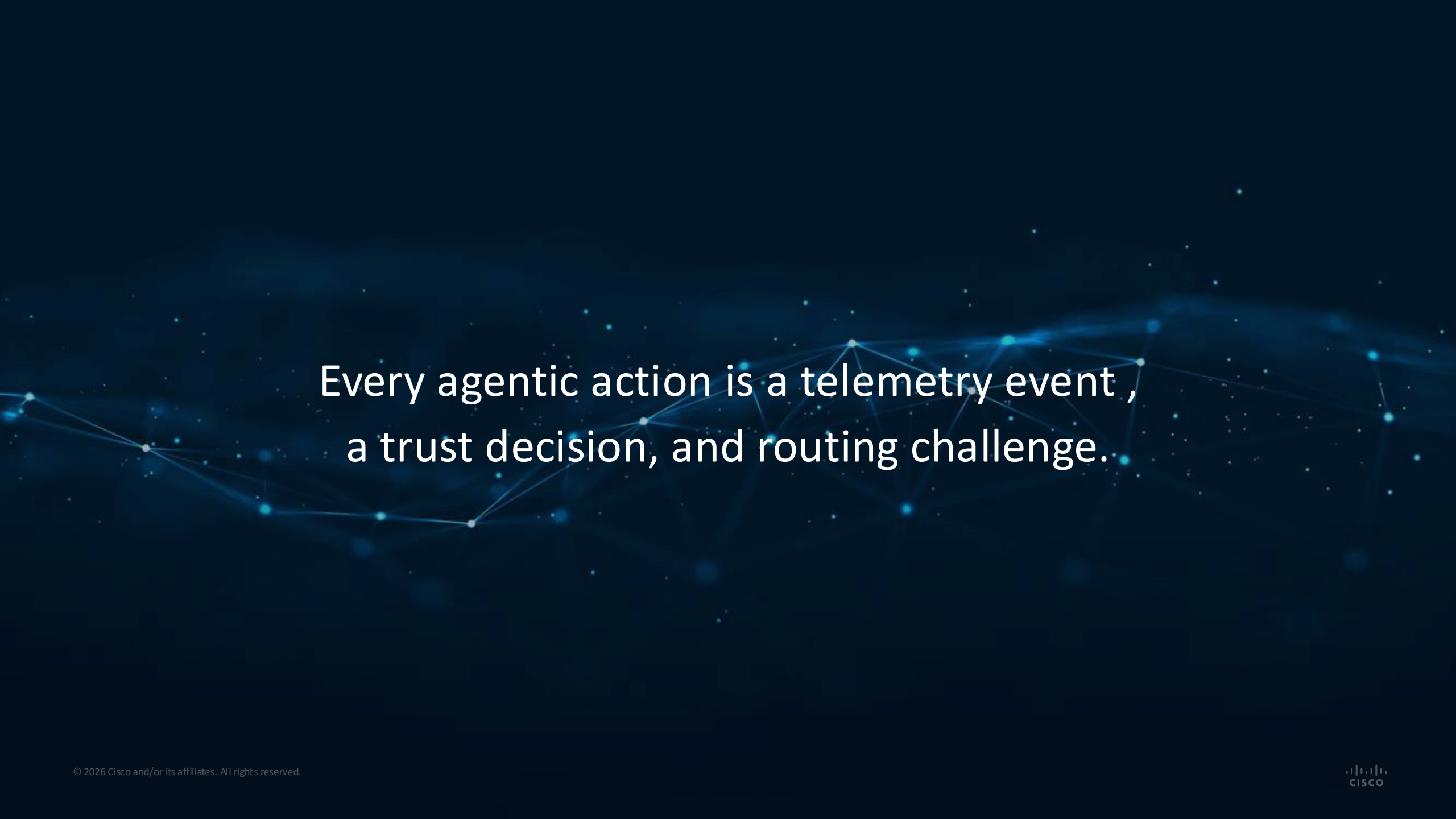
Every device, every room, every workflow — quietly, all at once.

In a branch router

In a SOC

In the data center

In a deskside device



Every agentic action is a telemetry event,
a trust decision, and routing challenge.

Inference is the
new workload

Agents are the
new workforce

Tokens are the
new currency

Control is the
new moat

AGENTS INCREASE TRAFFIC AND TOKEN USAGE

TOKENOMICS

TOKEN SPEND



20x

same models, agent volume



Agents

OpenAI

Google

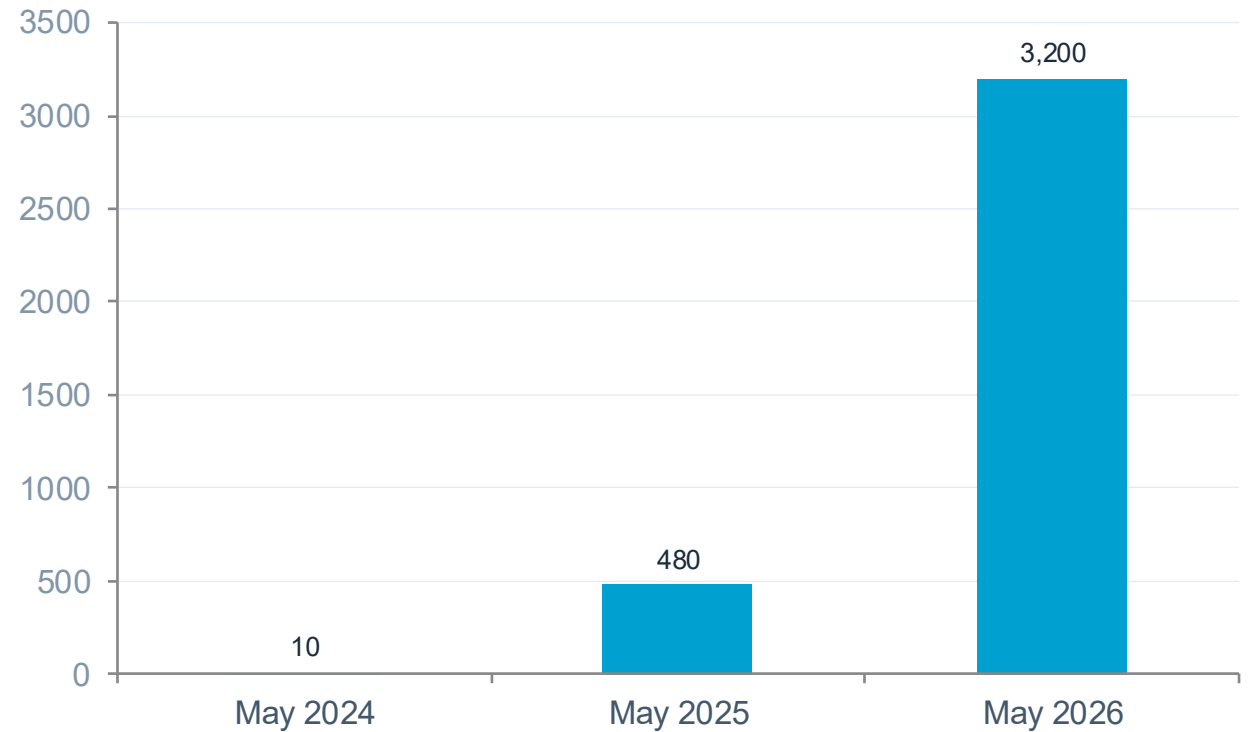
Anthropic

AI infrastructure capacity is the new revenue engine

- Neoclouds and sovereign clouds building AI infrastructure capacity
- Enterprise budgets moving from training to inferencing and agentic AI
- Reasoning and agentic models compound token demand
- Infrastructure ROI measured in tokens / second / watt

Monthly tokens processed across Google surfaces

Trillions of tokens per month, as disclosed at Google I/O keynotes · ~330x in two years



Source: Google I/O keynote disclosures (2024, 2025, 2026). Company-reported, unaudited. 3.2Q = 3,200T.

Inference is the
new workload

Agents are the
new workforce

Tokens are the
new currency

Control is the
new moat

Organizations and countries that can control their own token generation will
have a competitive advantage



Open source / open
weights models



Hybrid inferencing
architecture



Global regulatory
pressures

Our customers tell us

Standing up enterprise-grade
AI infrastructure is hard. Organizations face
complexity, fragmentation, and security concerns.



Too many tools,
not enough integration



Uncertain ROI to build
AI-ready environments



Security risks across
the AI lifecycle

Four market forces

Inference is the new workload

Agents are the new workforce

Tokens are the new currency

Control is the new moat

The AI-Ready Data Center



Powers all
workloads



Scales for
growth



Unifies
management



Secures the
entire stack



Keeps resilient

Rack-scale infrastructure

Cisco delivers AI infrastructure at any scale

Backed by CVIS

Expansion

Rack-scale AI

With Supermicro

Enterprise AI

Cisco AI PODs

Edge AI

Cisco Unified Edge

Cisco Confidential

22



Inferencing, agentic AI
at the edge



Optimization,
inferencing, agentic AI



Model training



Large-scale agentic AI
Trillion-parameter model training

AI-ready data centers | Our strategy

1

Scale
AI infrastructure

2

Fuse security into
the network

3

Deliver integrated
AI stack

Vertically integrated platform

Co-designed full-stack



Executing our strategy

Build

50%+

Of R&D invested in
cloud, AI, and
cybersecurity

Buy

240+

Acquisitions and
investments over the past
10 years

Partner

Ecosystem

Strategic, technology,
regulatory, commercial, and
GTM partnerships

AI-ready data centers | Our strategy

1

Scale
AI infrastructure

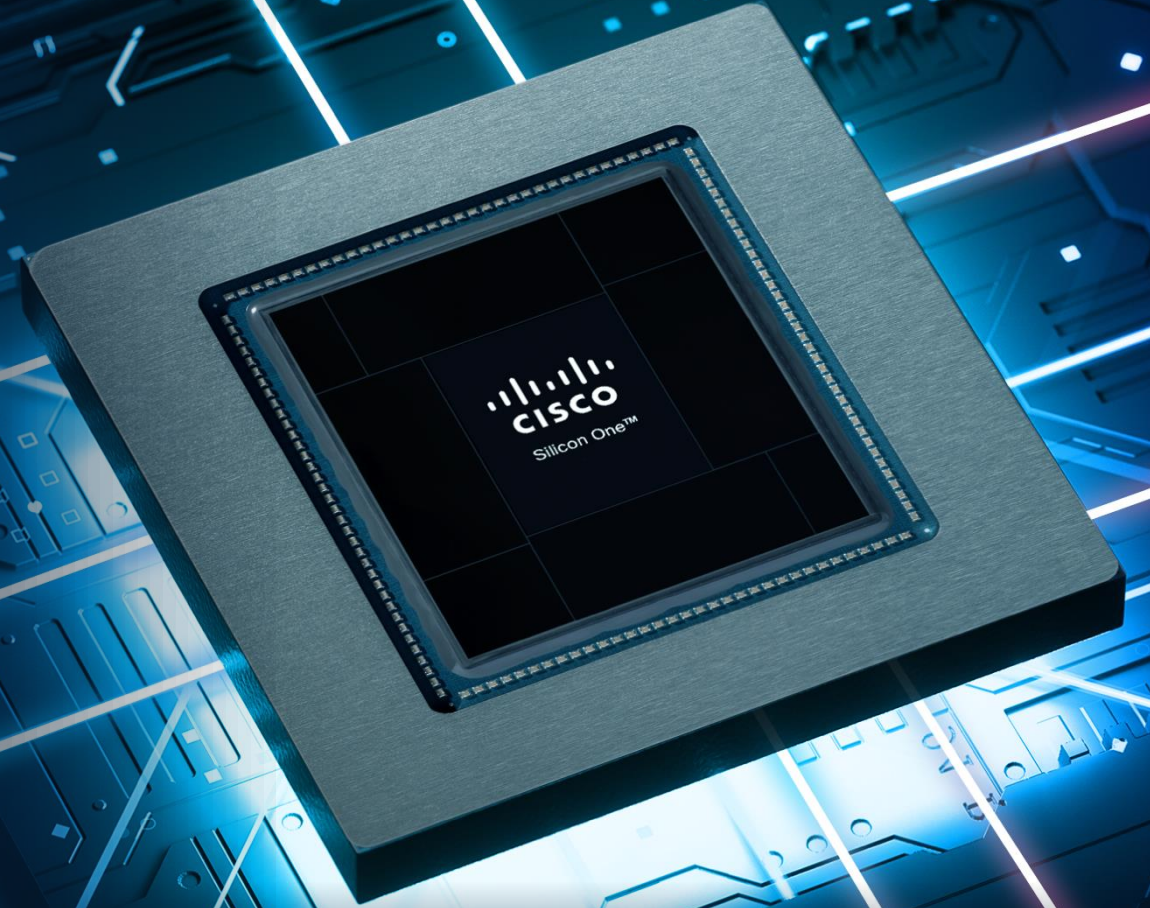
2

Fuse security into
the network

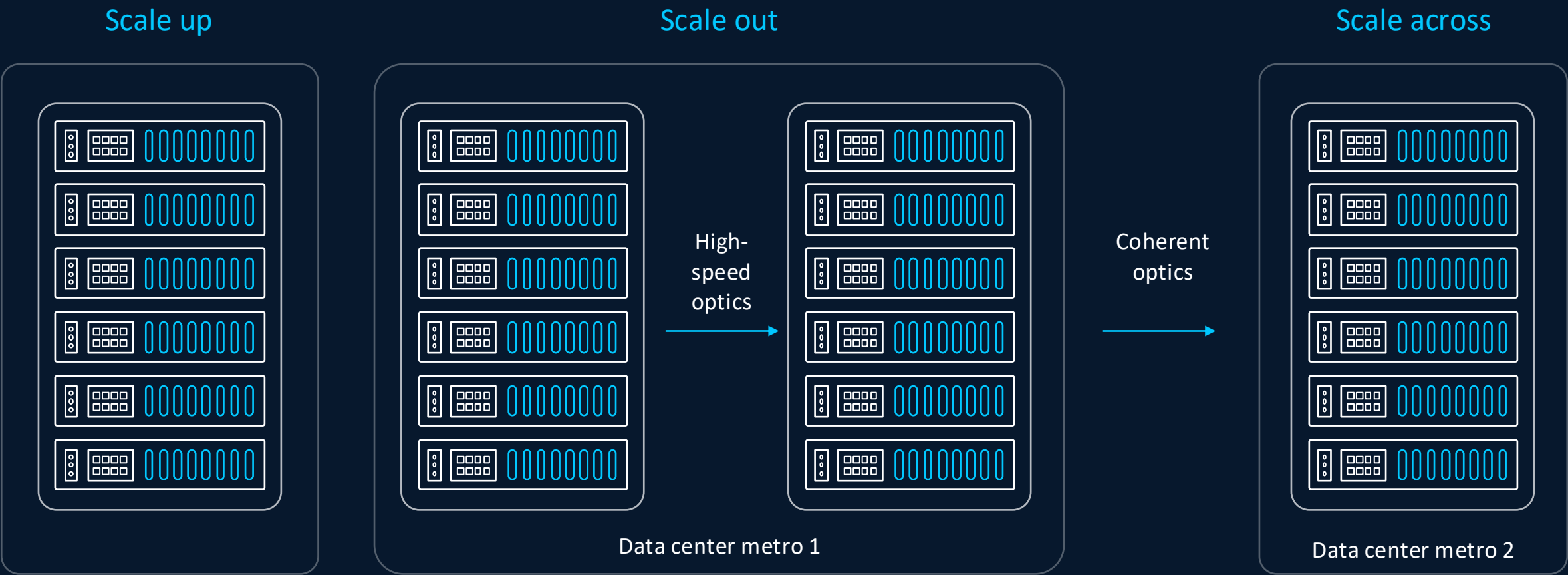
3

Deliver integrated
AI stack

Cisco Silicon One The foundation of the AI network



Agentic AI requires unprecedented scale



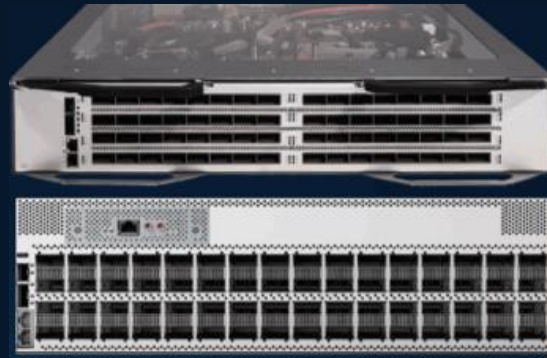
Scale-out solutions for the agentic era

SILICON



Silicon One G300

102.4T SYSTEMS



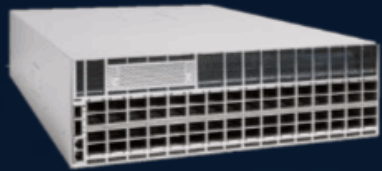
Cisco N9300 · Cisco 8100 Air- or liquid-cooled · NX-OS or SONiC

OPTICS



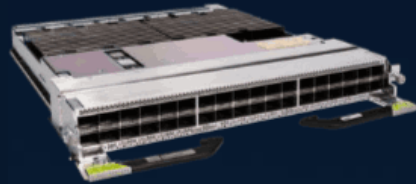
OSFP 1.6T pluggable optics 800G linear pluggable optics

Scale-across solutions for the agentic era



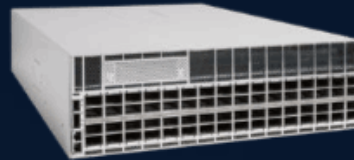
51.2T

Cisco N9364 (NX-OS)



28.8T

Cisco N9800 &
Cisco 8800 line cards



51.2T

Cisco 8223 Router
(SONiC / IOS-XR)



Pluggable optics

400G / 800G ZR & ZR+



SILICON ONE P200

We are focused on building ultra high-performance scale-out
& scale-across solutions

AI-ready data centers | Our strategy

1

Scale
AI infrastructure

2

Fuse security into
the network

3

Deliver integrated
AI stack

Aging infrastructure and processes put organizations at risk

Where attackers
look

40%

of vulnerabilities directly
impact end-of-life systems

How they
get in

31%

of all breaches now start with
vulnerability exploitation,
up from 20% in 2025

How fast they
move

8 hours

between CVE disclosure and
confirmed exploitation, down
from 21.5 days in 2025

How long to
respond

43 days

median time for organizations to
patch critical known-exploited
vulnerabilities (KEVs)

Organizations need a different approach

THE OLD APPROACH

Harden it, prove it, don't touch it

THE APPROACH IN THE AI ERA

Harden it, prove it, continuously defend it at machine speed



CADENCE

Major upgrades every 12–24 months



Continuous, CI/CD-style release readiness



MODERNIZATION

Aging gear, protocols, and scripts kept in service



Continuous removal of what can't be made safe



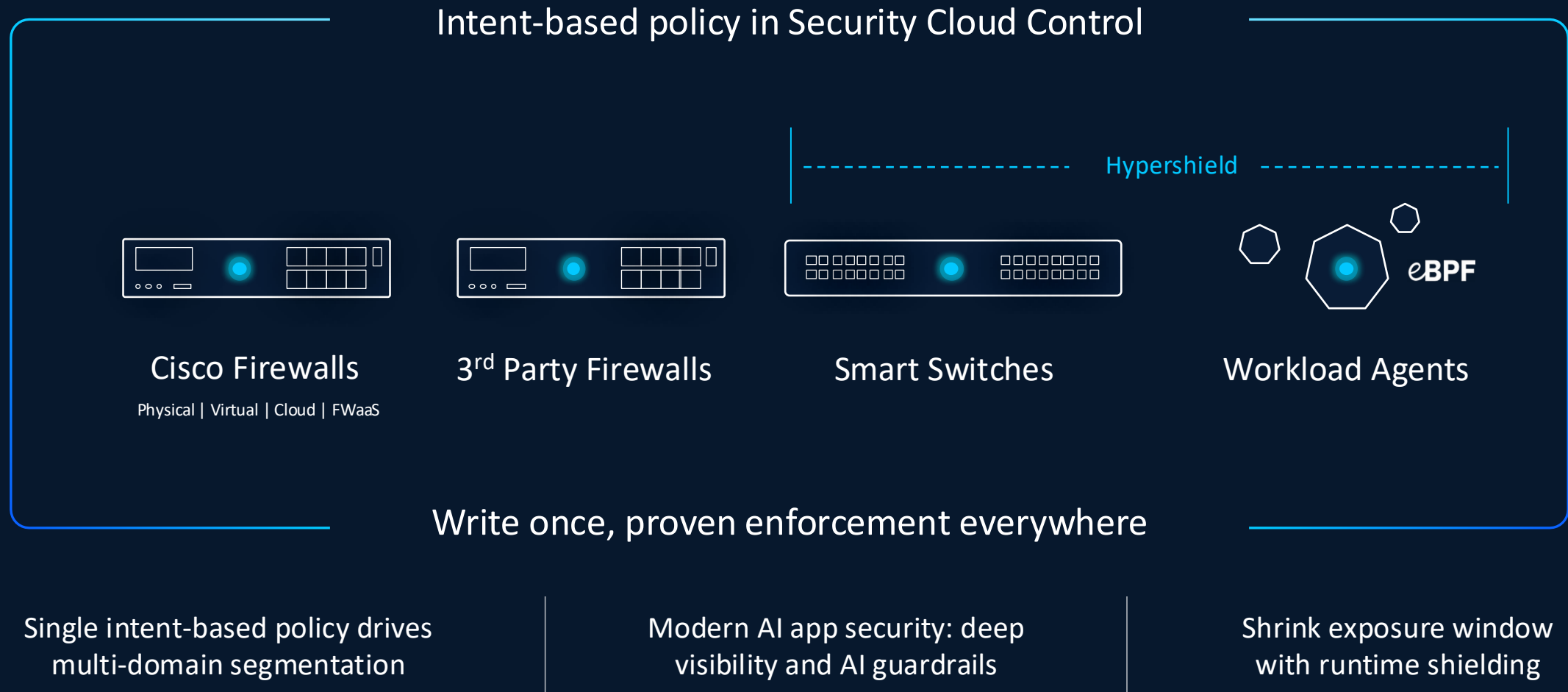
SECURITY POSTURE

Documented risk exceptions



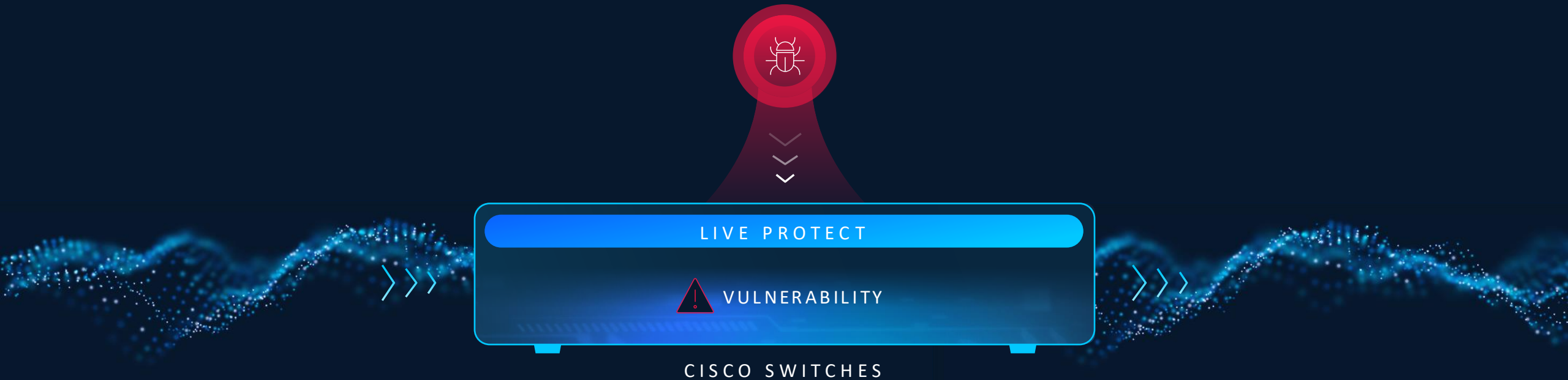
Continuous exposure reduction and containment

Hybrid Mesh Firewall | Security at the speed of change



Live Protect

Live Protect shields Cisco devices between patch windows



Apply policy
to live system

Protect now; patch in
planned window

No downtime or
switch reboot

Continuous
re-validation

AI-ready data centers | Our strategy

1

Scale
AI infrastructure

2

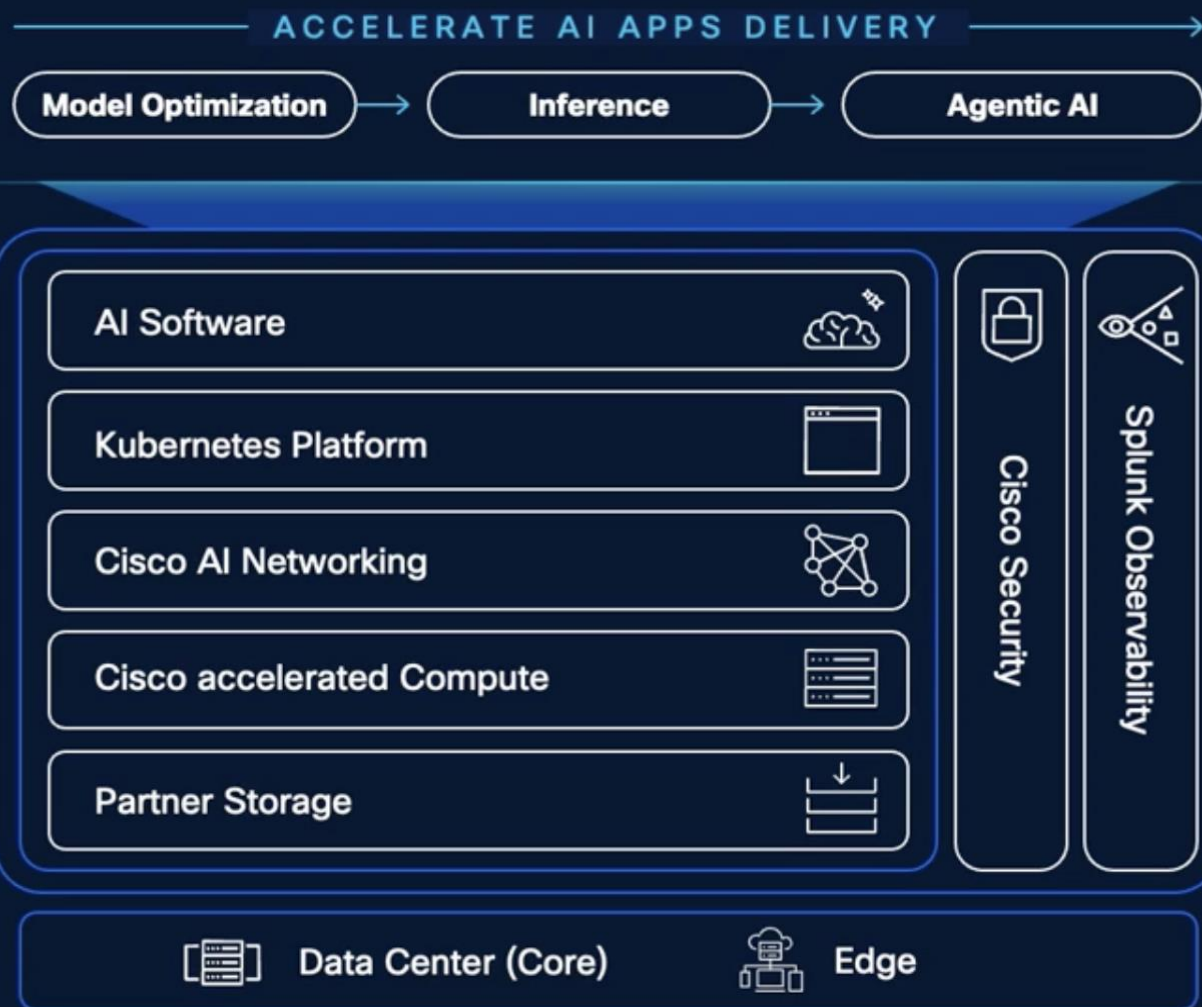
Fuse security into
the network

3

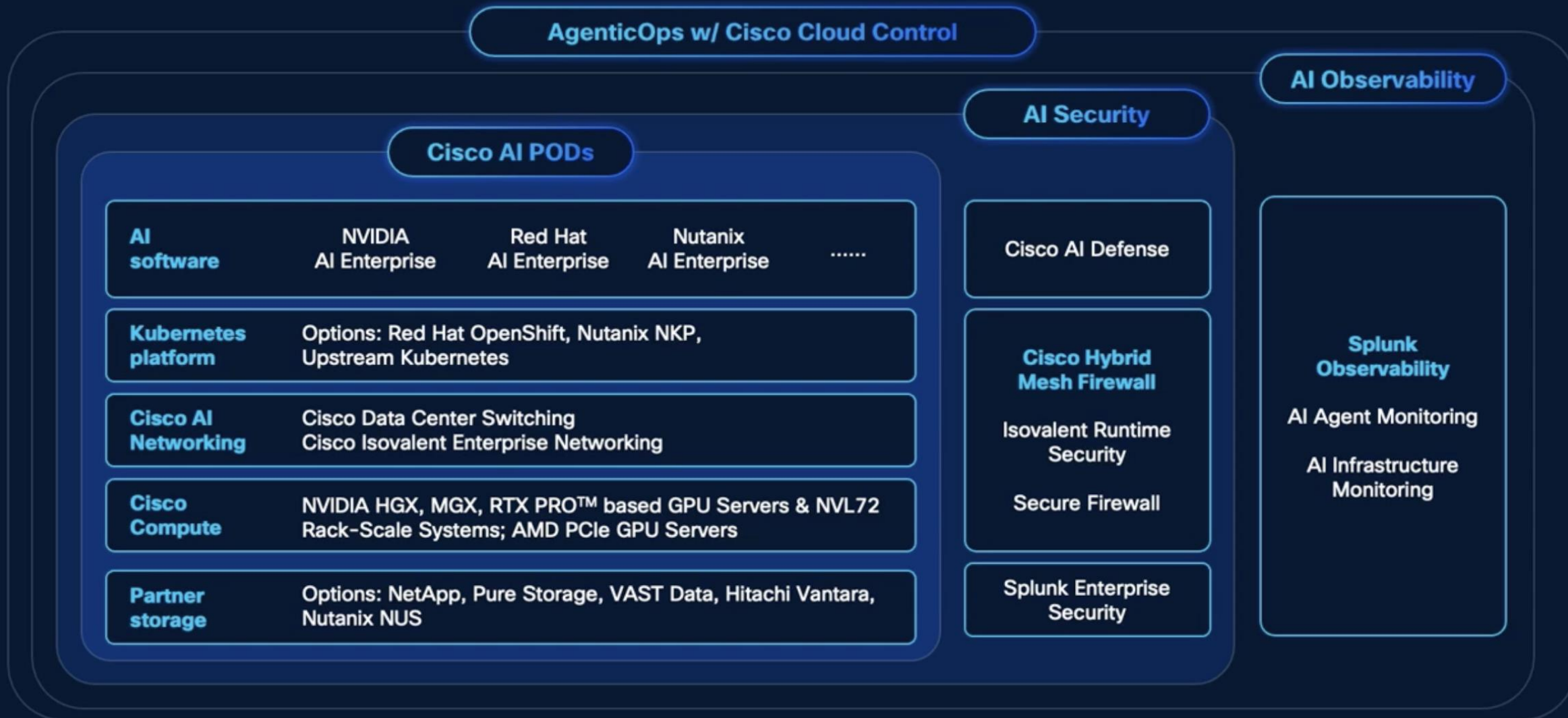
Deliver integrated
AI stack

Cisco Full-Stack AI Infrastructure

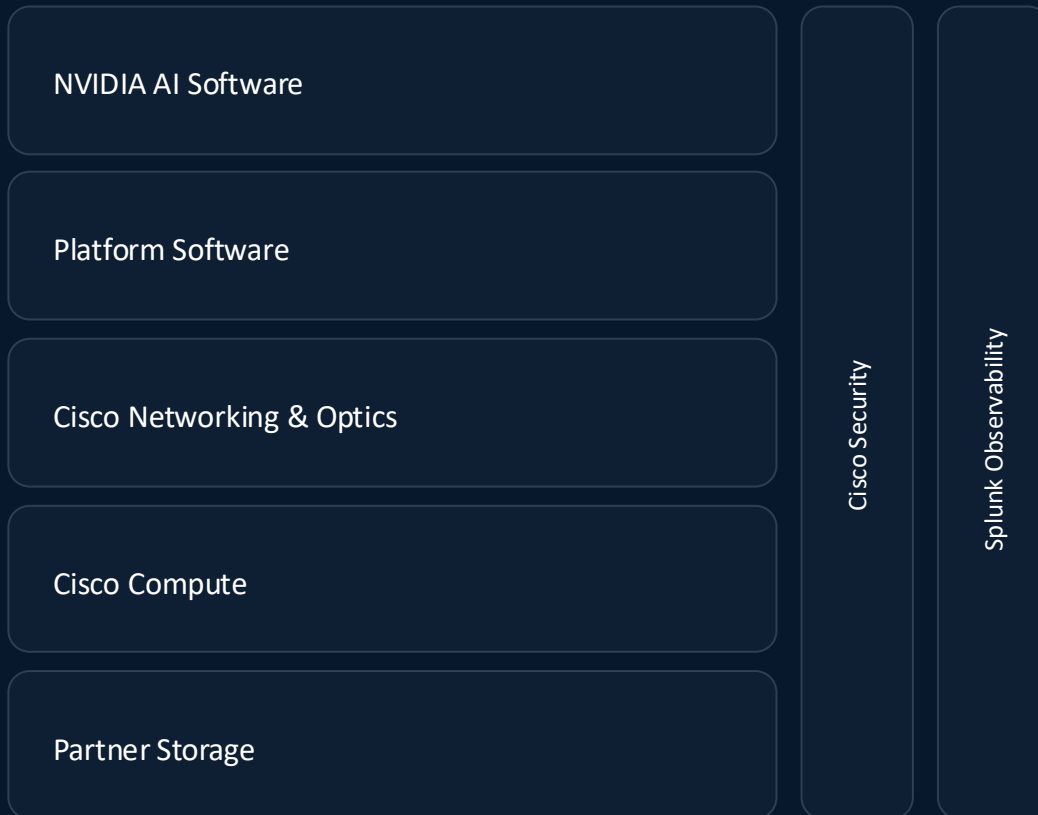
**A modular
reference design**
that combines
high-performance
infrastructure with
full-stack security
and observability



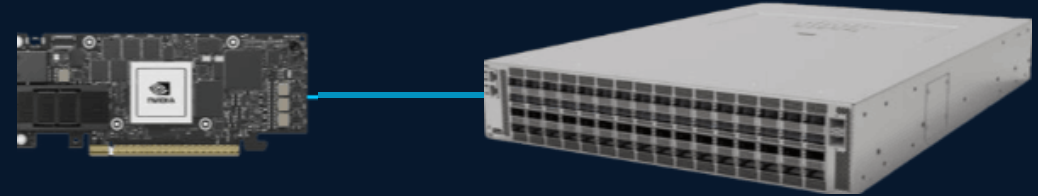
Cisco Full-Stack AI Infrastructure



Continued innovation with NVIDIA for the enterprise



Cisco Switches are fully integrated with NVIDIA Spectrum-X architecture



NVIDIA BlueField-3

Cisco N9364E (G200 2U)

Cisco's accelerated compute portfolio

Inferencing, Agentic AI
at the edge

PCIe GPU: Edge Platform

Cisco Unified Edge
NVIDIA RTX PRO



Optimization,
Inferencing, Agentic AI

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



Model Training

Dense GPU Servers

Supermicro
NVIDIA HGX
Rubin NVL8
liquid-cooled



Supermicro
NVIDIA HGX
B300 NVL8
liquid-cooled



Supermicro
NVIDIA HGX
B300 NVL8
air-cooled



Large scale Inference, Agentic AI,
Trillion Parameter Model Training

Rack-Scale Systems

Supermicro
NVIDIA
Vera Rubin
NVL72



Supermicro
NVIDIA
Grace Blackwell
300 NVL72



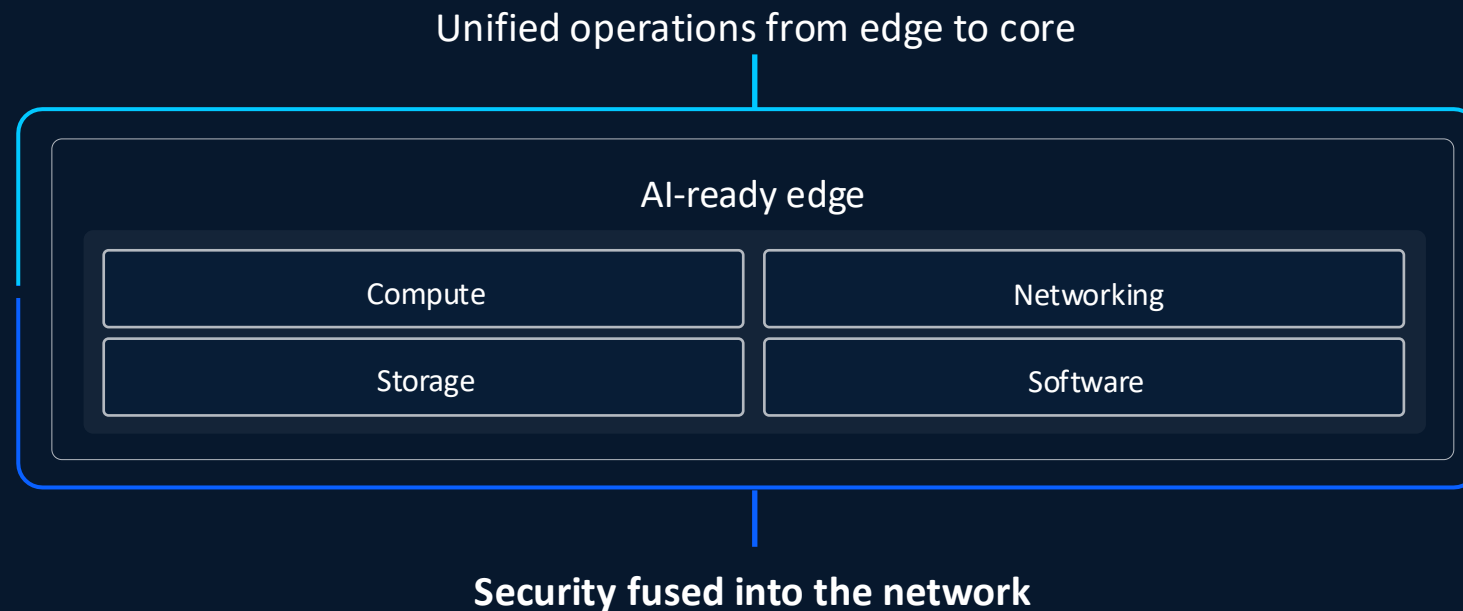
Storage Server

Cisco UCS® C225 M8N



Bringing data center power and scale to the edge

Cisco Unified Edge



Cisco Unified Edge

**Not a server.
The AI-ready platform the edge demands.**

AI-ready edge

Unified operations

Security fused into the network

Introducing Cisco Unified Edge

AI-ready edge

Compute

Storage

Networking

Software

SaaS
Management

Analytics

Security



NUTANIX



vmware
by Broadcom



Microsoft

intel



SUSE

Cisco Unified Edge: Future-Ready Performance

Integrates compute, networking, storage, and security

AI-ready edge

Compute nodes



Intel Xeon 6 SoC

CPU native AI inferencing (AMX)
Confidential compute (TDX & SGX)
Integrated Ethernet
Scalable multithreaded
cores (12, 20, 32)



UCS XE150c M8

New

UCS XE130c M8

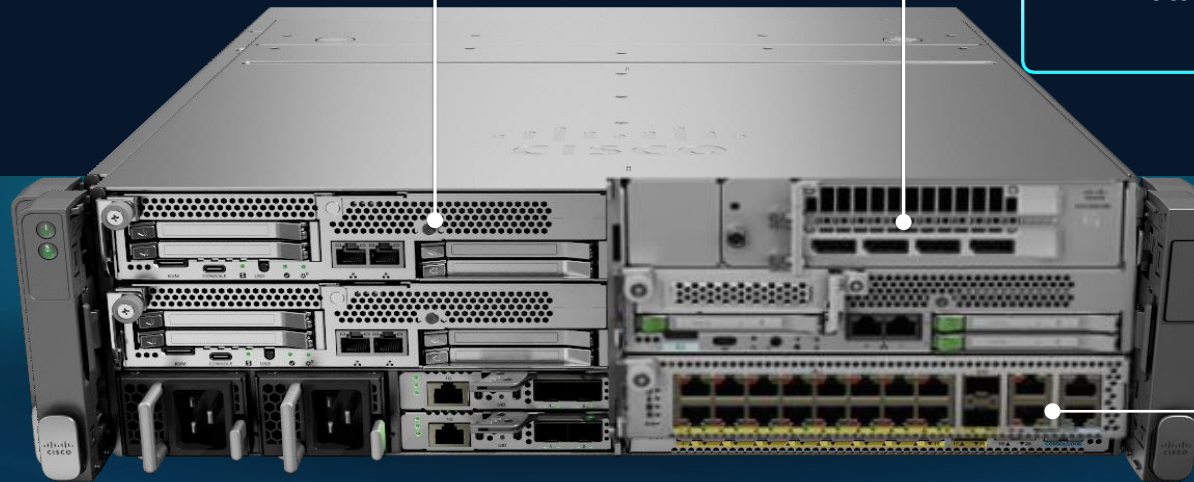
GPUs

UCS XE130c M8: NVIDIA L4
UCS XE150c M8: NVIDIA RTX Pro 4500/6000 or L40S



New

Secure Router node



NUTANIX



vmware
by Broadcom



Microsoft

intel

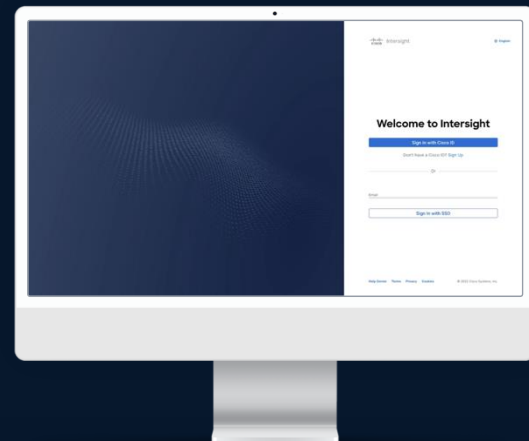


SUSE

Day 0: Zero touch provisioning

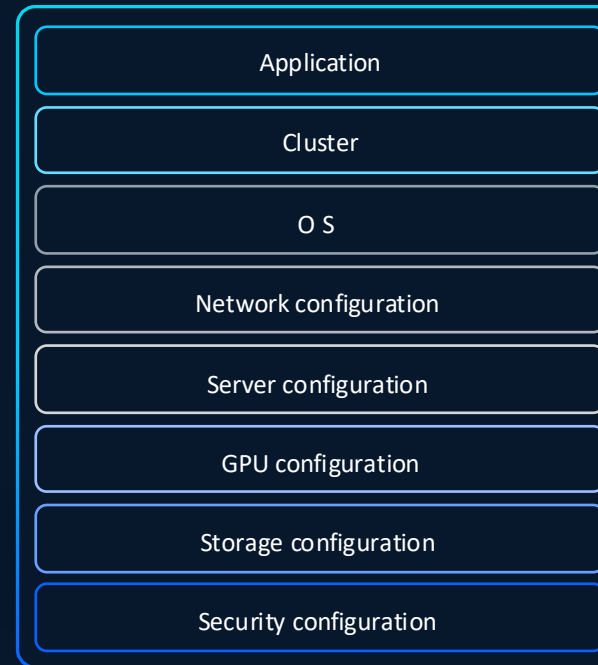
Consistent, repeatable AI-ready infrastructure deployments across multiple sites

Unified operations



Cisco Intersight

Cisco edge blueprints



Golden configurations base on Cisco validated designs

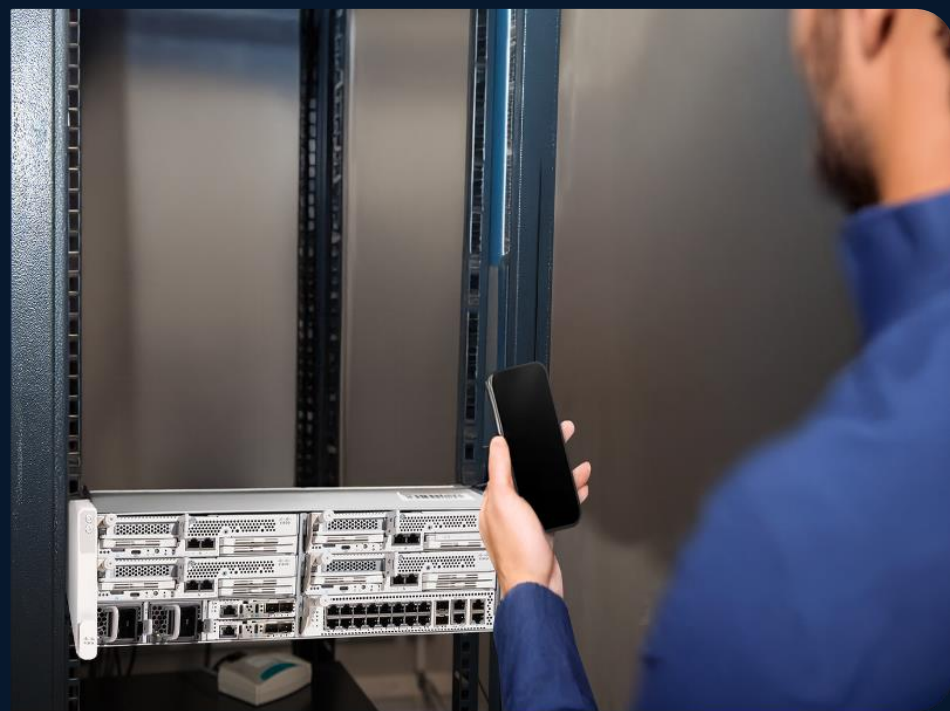
Faster deployments with fewer errors

Supported by Cisco TAC

Available in a marketplace

Day 1: NFC-based claim via mobile app

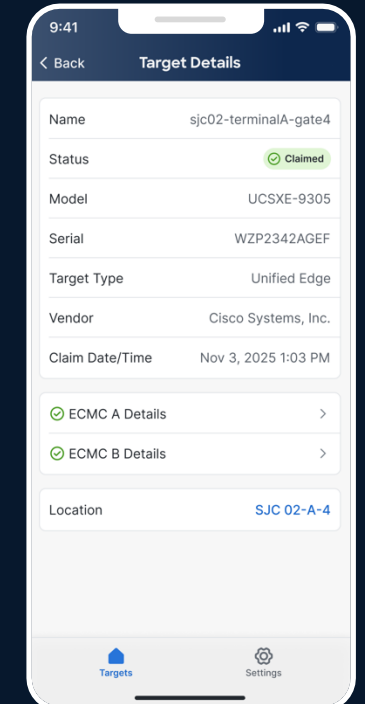
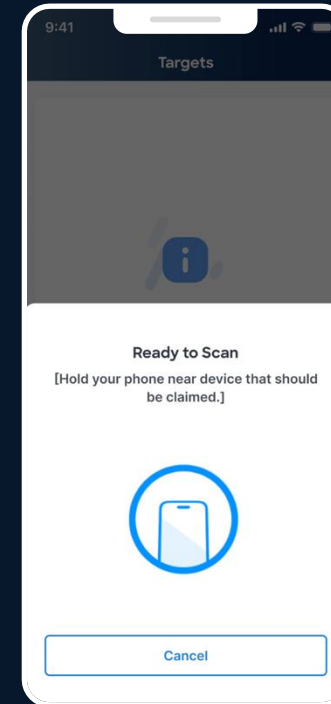
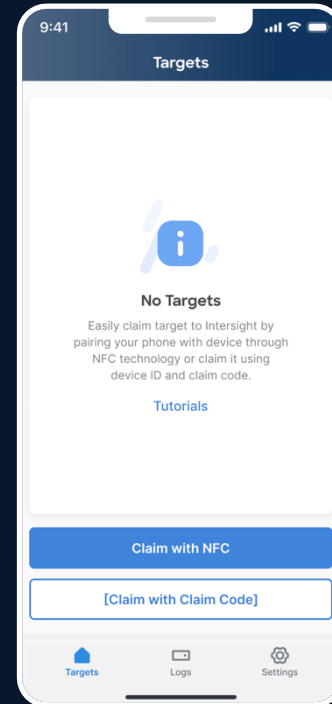
Unified operations



Fewer truck rolls – no need to onboard systems centrally

No technical expertise needed to claim device

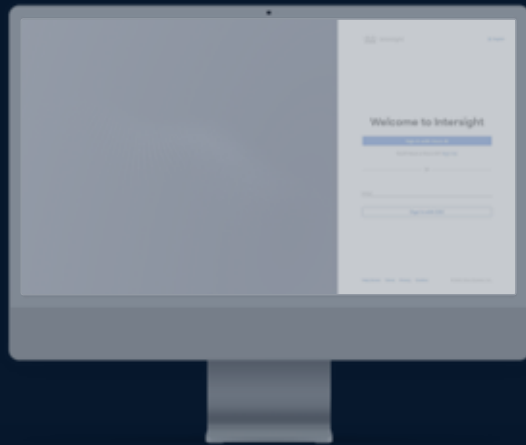
Pre-claimed capabilities accessible to authorized on-site staff



Day 2: Deployment at scale

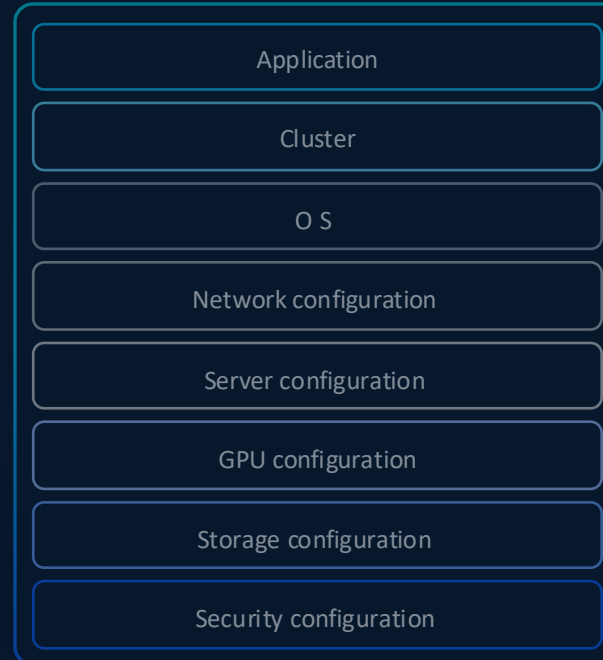
Consistent, repeatable AI-ready infrastructure deployments across multiple sites

Unified operations



Cisco Intersight

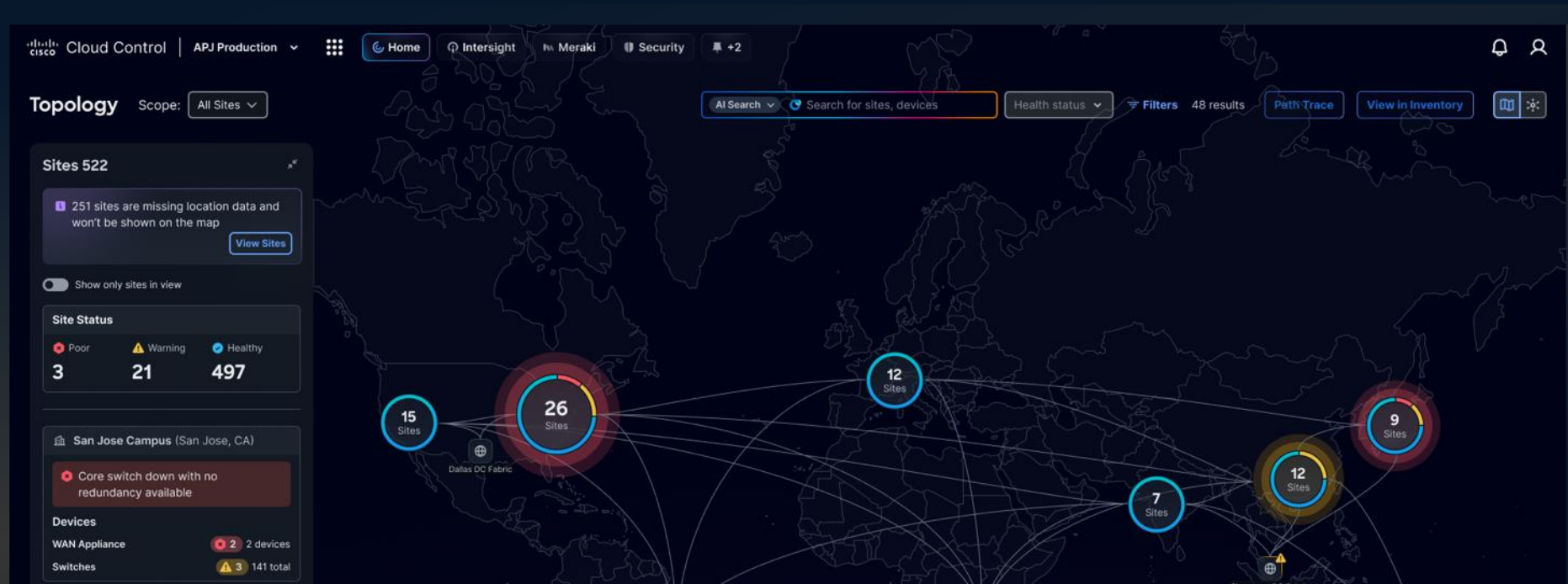
Cisco Edge blueprints (Retail, manufacturing, healthcare)



Cisco Cloud Control

One platform. Every domain. The foundation everything else is built on.

Meraki • ThousandEyes • Security Cloud Control • Splunk • Nexus Dashboard • Catalyst SD-WAN • Intersight • Webex • IQ • & more



Leverage an open ecosystem

50+ companies & growing



CUSTOMER VALIDATION



"With Cisco Secure AI Factory with NVIDIA, we no longer have to choose between performance, reliability or ease of management. NCP validation gives us the confidence that our infrastructure is optimized from day one, while the platform's rack-scale capabilities provide a seamless path to scale our AI operations as our business grows."

James Manning

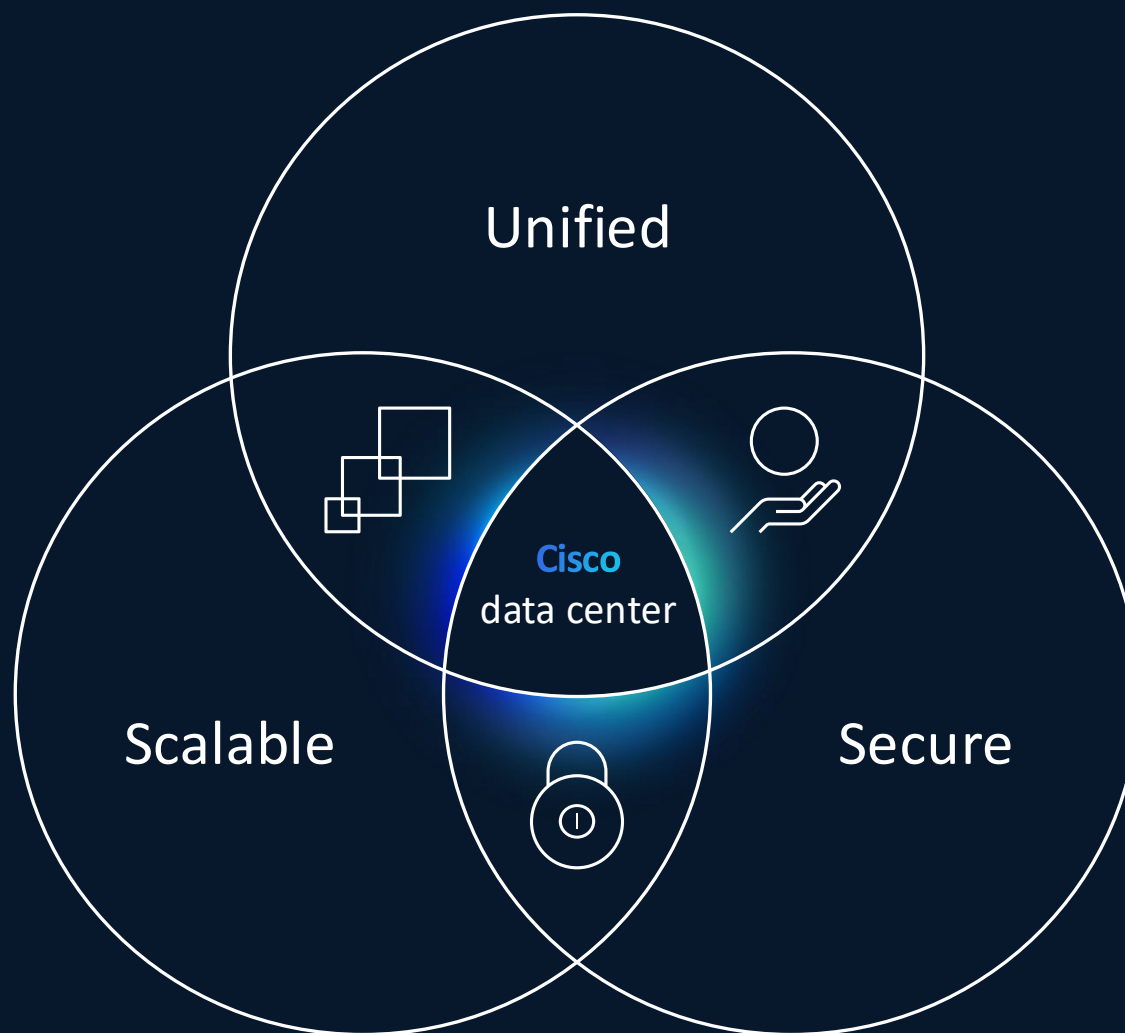
Co-Founder & CEO, Sharon AI
<https://sharonai.com/>

Cisco Confidential

Australian Sovereign AI Infrastructure Built to Solve Humanity's Greatest Challenges



Whether you're modernizing traditional workloads, or scaling for AI, the Cisco data center is...



Thank you

