CISCO Engage | Tech Day Miami

From Vision to Value

Cisco's Model for Your AI Journey

Richard Nitz
CAI Account Executive

Josh Jackson
Global AI Solutions Sales

February 19, 2026

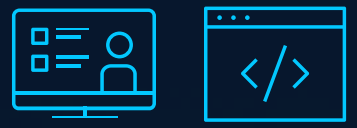
Agenda

1. AI Use Cases
2. Challenges
3. Secure AI Factory
4. Cisco’s Compute Portfolio

AI use cases across industries

Knowledgebase copilots

AI assistants



Content and code generation

Text | Images | Video | Code

Virtual agent and chatbots

Specialized domain-specific chatbots

Visual computing

Digital twins | Video analytics | Imaging and diagnostics

Language translation

Multilingual real-time communication

Detection and prediction

Forecasts | Anomalies | Insights

cisco

All organizations are defining their AI services stack

Agent ecosystem

- Chatbot applications
- AI agents
- Agent hub
- Agent platform

AI platform services

Foundational model aaS
On-prem and cloud

Custom model service
Pre-training, Fine-tuning

Data infrastructure
Vector, ingestion, embedding, retrieval

LLM tracing
Agent execution tracing

Model registry
AI artifacts, versioning, lineage, evals

Multi-tenant GPU aaS
Multi / single / fractional GPU aaS

kubernetes · PyTorch · jupyter

On-prem and cloud / SaaS data sources
- Relational databases
- NoSQL databases
- Log & metric datastores
- Documents

SaaS-based AI Services
- Data platform w/ AI / ML tools
- AI coder
- Service management
- Conversational AI for BI

cisco

Business Outcomes

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

All of this exposes key challenges for our customers' technology architectures

Infrastructure constraint

Trust deficit

Data gap

Challenges with AI projects delays time to value realization

Security vulnerabilities

AI models, frameworks, apps, and infrastructure are a new cyberattack surface with threats such as prompt injection, denial of service, and data leakage

Network performance bottlenecks

Model training and inferencing generates a lot of traffic, slowing networks and delaying time-to-value

Complex AI infrastructure deployment

Lack of high-performance infrastructure with integrated and resilient compute, network, storage, and AI software can stall projects

All of this exposes key challenges for our customers' technology architectures

Infrastructure constraint

Trust deficit

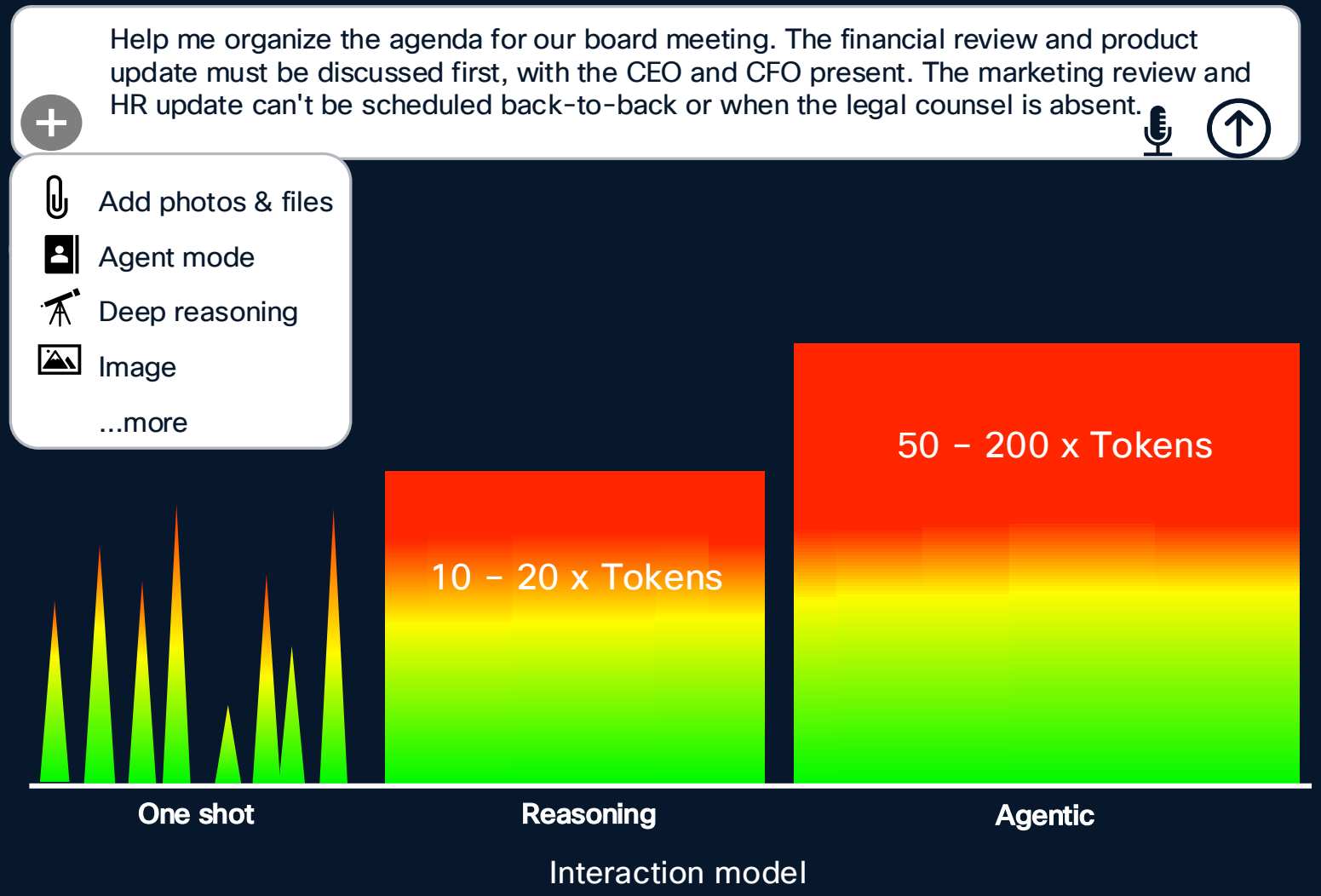
Data gap

Infrastructure is dedicated to the production of tokens

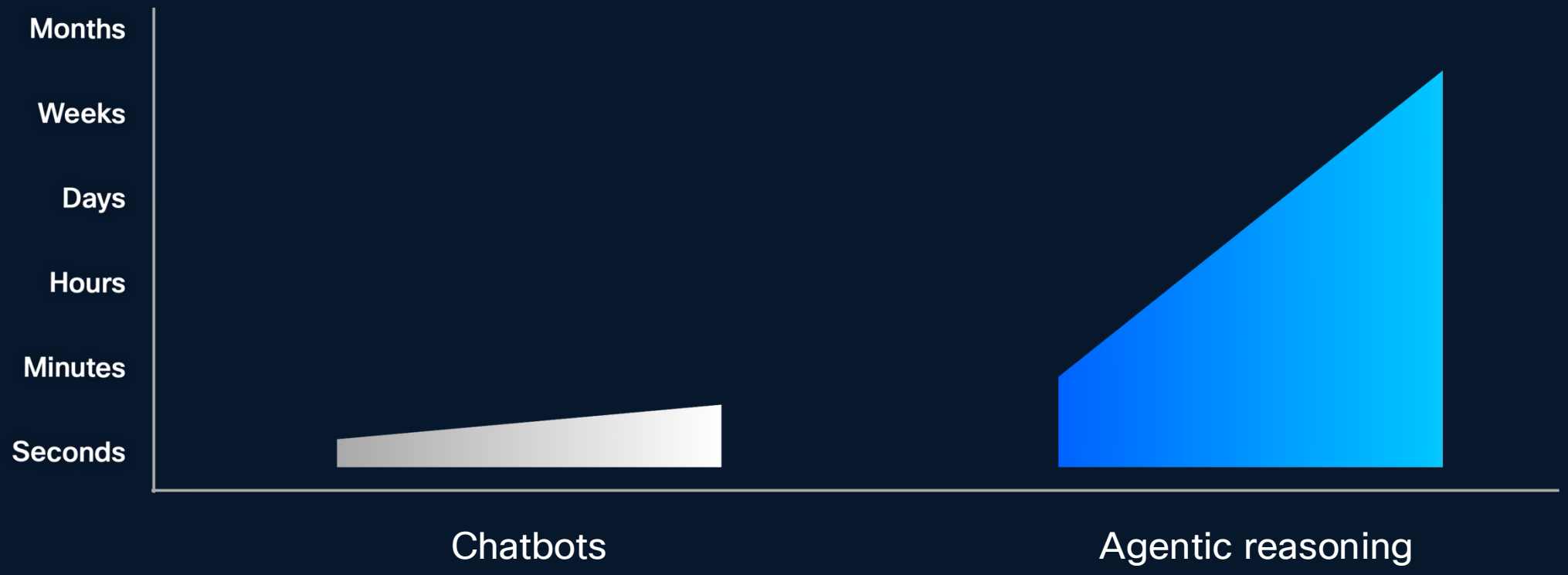
Token (noun): the atomic unit of input and output in AI systems

AI is changing: Token demand inflation

More tokens enable higher quality results and more complex tasks



Duration of autonomous execution



All of this exposes key challenges for our customers' technology architectures

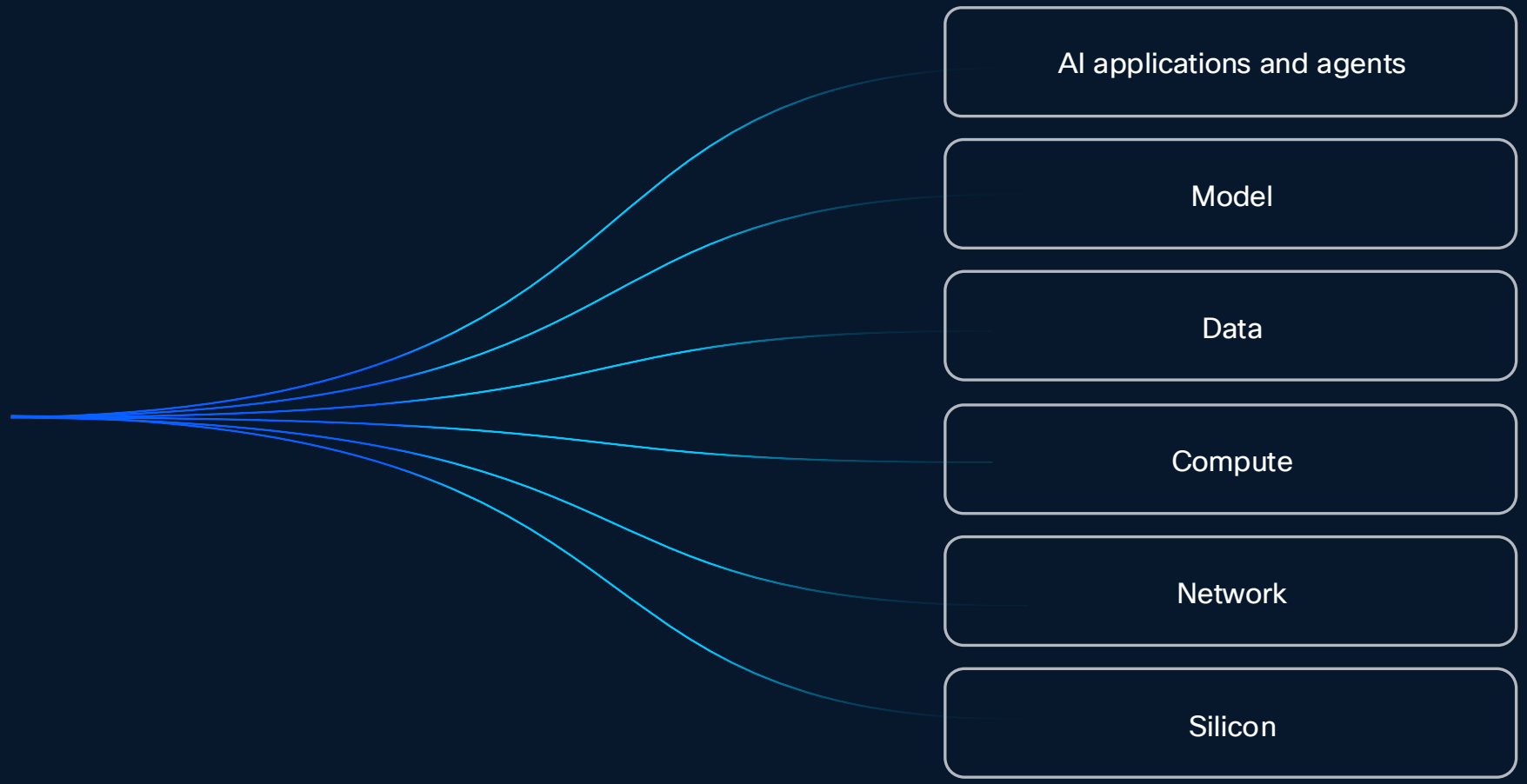
Infrastructure constraint

Trust deficit

Data gap

Time to market is hindered by insecure, untrusted AI systems

Lack of end-to-end visibility



Model threat vectors

Safety

- Profanity
- Cost harvesting / repurposing
- Harassment
- Hallucinations
- Hate speech
- Off-topic
- Toxicity
- Social division and polarization
- Self-harm
- Financial harm

Security

- Indirect prompt injection
- Infrastructure compromise
- IP theft
- Meta prompt extraction
- Prompt injection
- Model theft
- Training data poisoning
- Sensitive information disclosure
- Data exfiltration
- Model denial of service

Agent threat vectors

Identity

Access

Behavior

All of this exposes key challenges for our customers' technology architectures

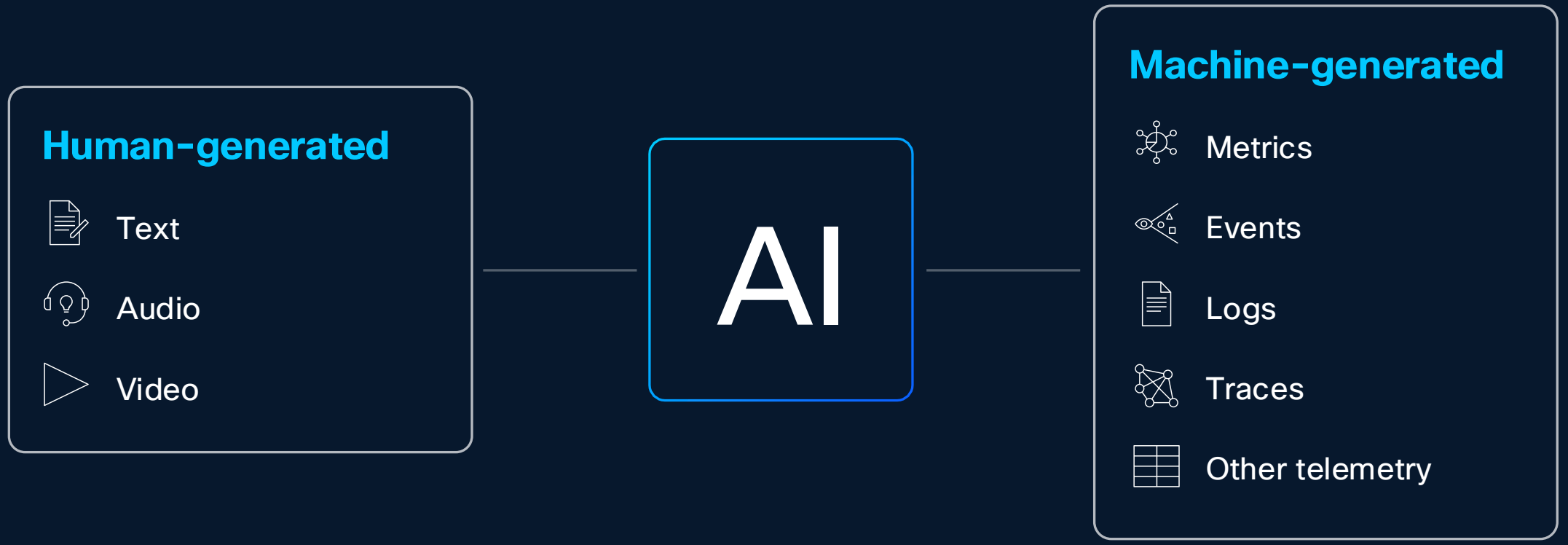
Infrastructure constraint

Trust deficit

Data gap

Data is the essential fuel for AI

More data is more context. More context means more tokens, better results, and unlocked use cases.



We’re addressing all these challenges head-on

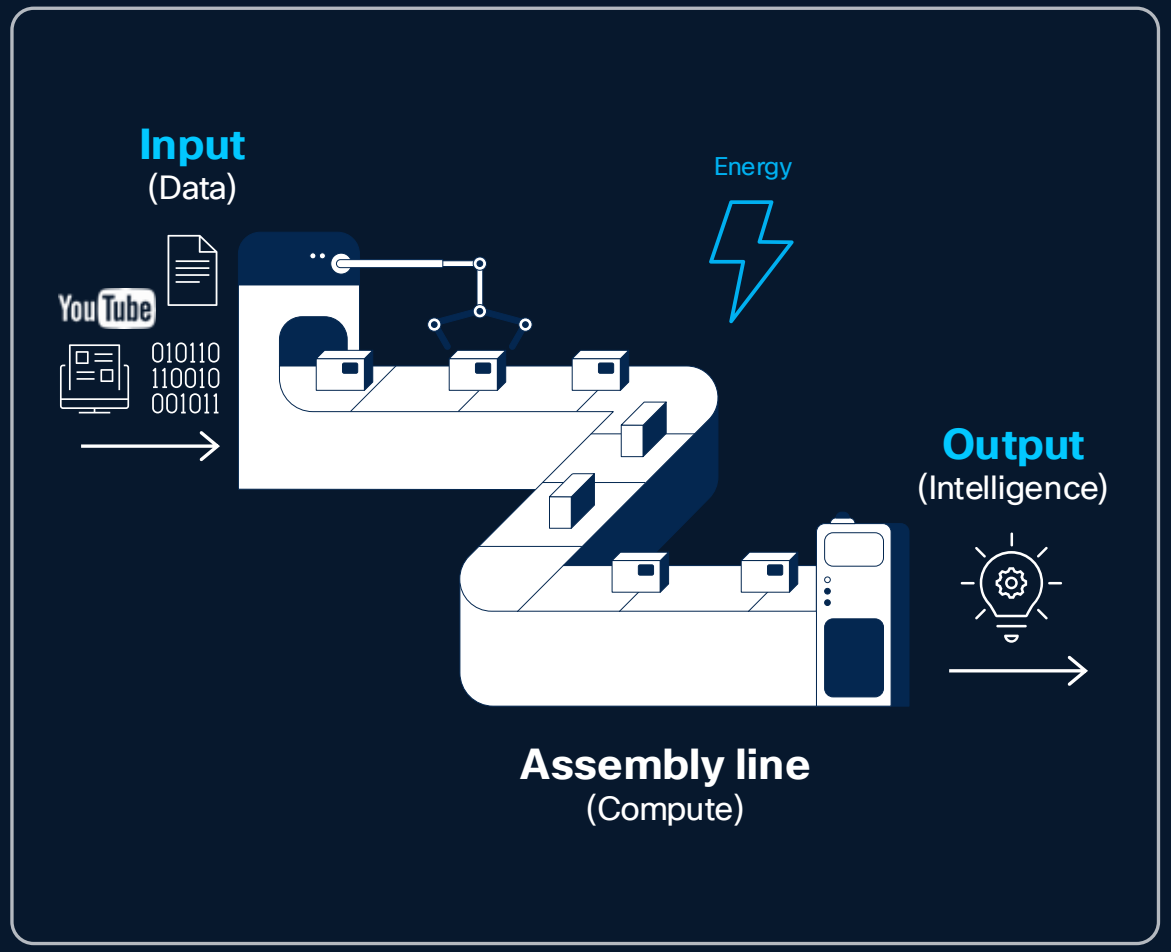
Cisco is the critical infrastructure for the AI era

Enter the Cisco Secure AI Factory with NVIDIA

What is an AI Factory?

The processing plant for tokens

Organizations everywhere are thinking about how to generate tokens as quickly, safely, and cost effectively as possible



What is needed to accelerate trusted AI outcomes?

AI factory

Secure AI factory

Executive summary

AI / GenAI benefits enterprises

- Improve customer and employee experience
- Gain competitive edge
- Increase revenue
- Save costs

Challenges delay time to value realization

- Security vulnerabilities
- Infrastructure performance
- Infrastructure deployment complexity
- People and process

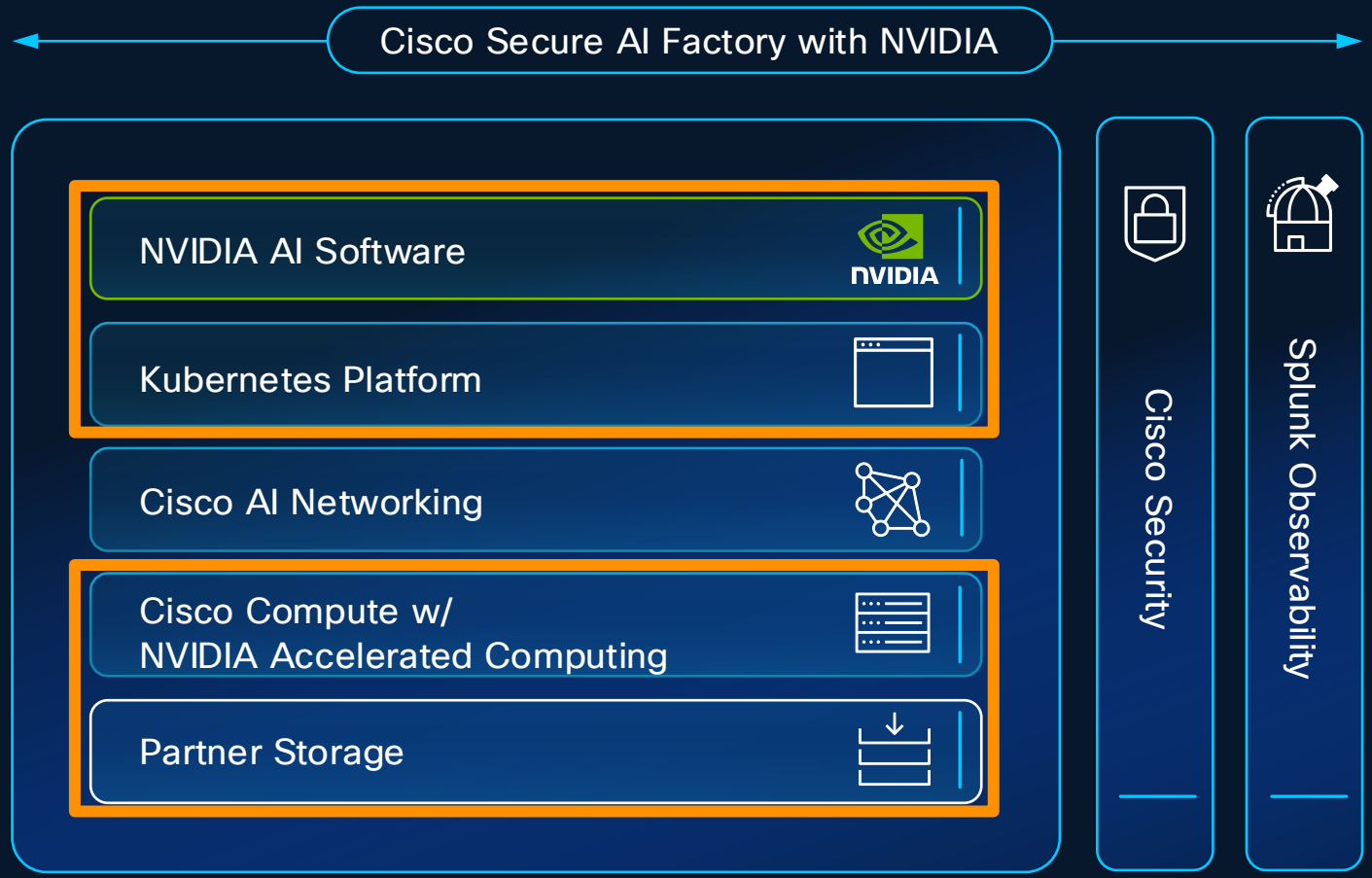
Cisco Secure AI Factory with NVIDIA accelerates AI adoption

- Security-first architecture with resiliency to protect AI models, apps, and infrastructure
- Scalable, enterprise-grade, high performance AI Infrastructure
- Modular reference design to derisk and simplify deployments

Cisco Secure AI Factory with NVIDIA

What is it?

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



IMPORTANT

SAIF is intentionally storage-agnostic but vendor-validated

Customers choose the platform that best fits their workload and regulatory environment

This partner flexibility is a major competitive advantage compared to monolithic cloud stacks

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

AI-ready data centers

Transform data centers to power AI workloads anywhere

Future-proofed workplaces

Modernize everywhere people and technology work and serve customers

Secure global connectivity

Digital resilience

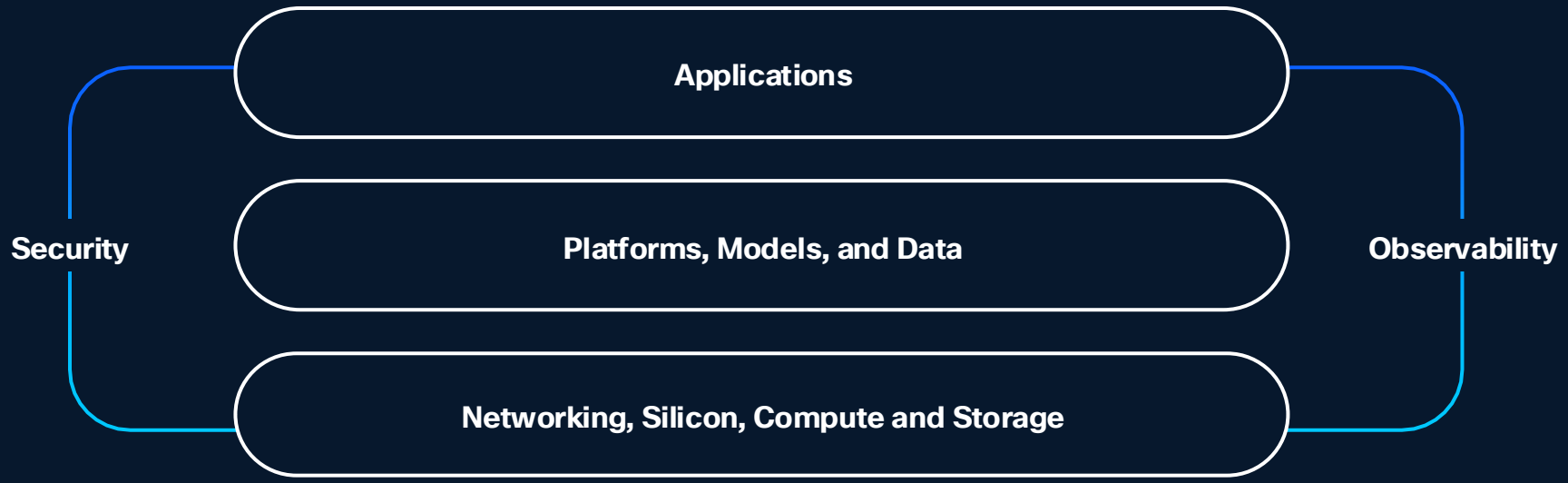
Keep the organization securely up and running in the face of any disruption

Accelerated by Cisco AI

cisco

Connecting, protecting and powering the entire stack

through Cisco technology and integrated strategic partners



AI-ready Data Center

Portfolio

Networking, Silicon, Compute and Storage

Security

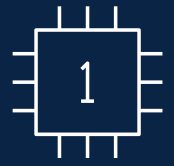
Observability

Data Centre Networking

Nexus

Hyperfabric

Silicon One



Optics

Industry leading ethernet fabrics built on custom silicon and choice of operating model

Unified Computing System

UCS-X Modular

UCS-C Rack

Unified Edge

Hyper converged

Compute that powers and optimizes any enterprise application or AI workload wherever its deployed

Full Stack Systems

AI Pod

AI Factory

Hyperfabric AI

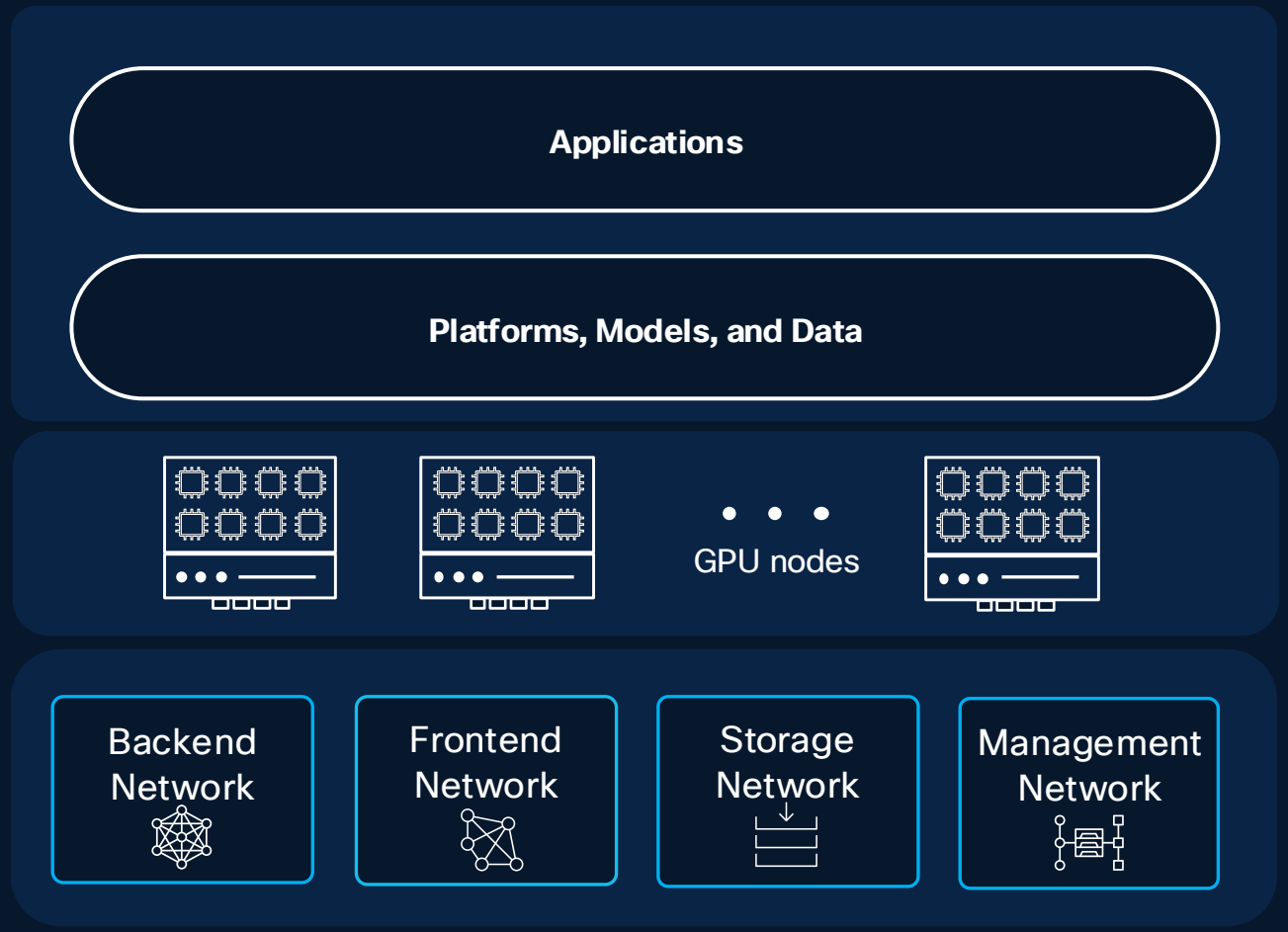
Converged Infrastructure

Validated and integrated systems that accelerate and derisk the delivery of business outcomes

cisco

AI Data Center Networking

Tightly integrated scalable network stack



AI demands a high-speed, low latency and lossless ethernet fabric

Training and distributed inference at scale demands discrete networks

Collective communication libraries directly influence network topology and routing strategies

Data Center Networking

Choice of operating model with silicon diversity

ACI with Nexus 9000

On-premise policy-based management

NXOS (ACI Mode)

Enterprise / Public Sector / Commercial

Nexus Dashboard with Nexus 9000/9100

On-premise switch management

NXOS / SONiC

Enterprise / Public Sector / Commercial / Service Provider / Tier 2 Web / Neocloud

Nexus Hyperfabric with Cisco 6000

Cloud managed Network-as-a-service

SONiC

Enterprise / Public Sector / Commercial / Neocloud

Cisco 8000

BYO management and NOS

IOS XR / SONiC

Service Provider / Tier 2 Web / Hyperscaler

Cisco Custom Silicon

NVIDIA Spectrum-X

Data Center Networking

GPU scale and Reference Architectures

ACI with Nexus 9000

On-premise policy-based management

NXOS (ACI Mode)

Nexus Dashboard with Nexus 9000/9100

On-premise switch management

NXOS / SONiC

N9300

ERA | CRA

N9100

NCP

Nexus Hyperfabric with Cisco 6000

Cloud managed Network-as-a-service

SONiC

ERA | CRA

Cisco 8000

BYO management and NOS

IOS XR / SONiC

Enterprise Reference Architecture (ERA)

Cloud Reference Architectures (CRA)

NVIDIA Cloud Partner (NCP)

cisco

Cisco Validated Designs (CVDs)

Easily deploy new systems with expert guidance

Reliable

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

Consistent

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

Comprehensive

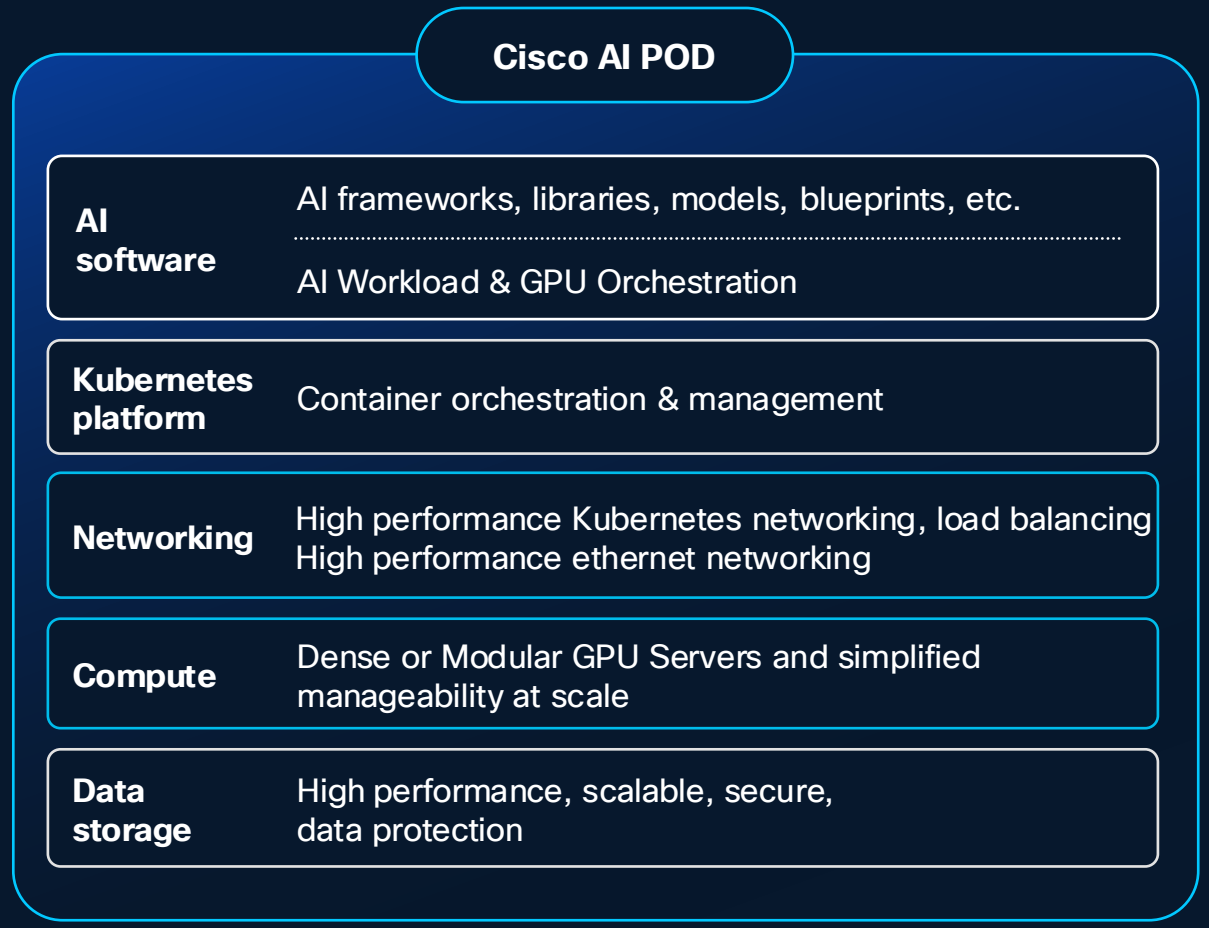
CVDs provide everything from system designs to configuration instructions to a bill of materials (BOM).

Cisco TAC support

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

Cisco AI PODs

Faster time to value with pre-configured bundles



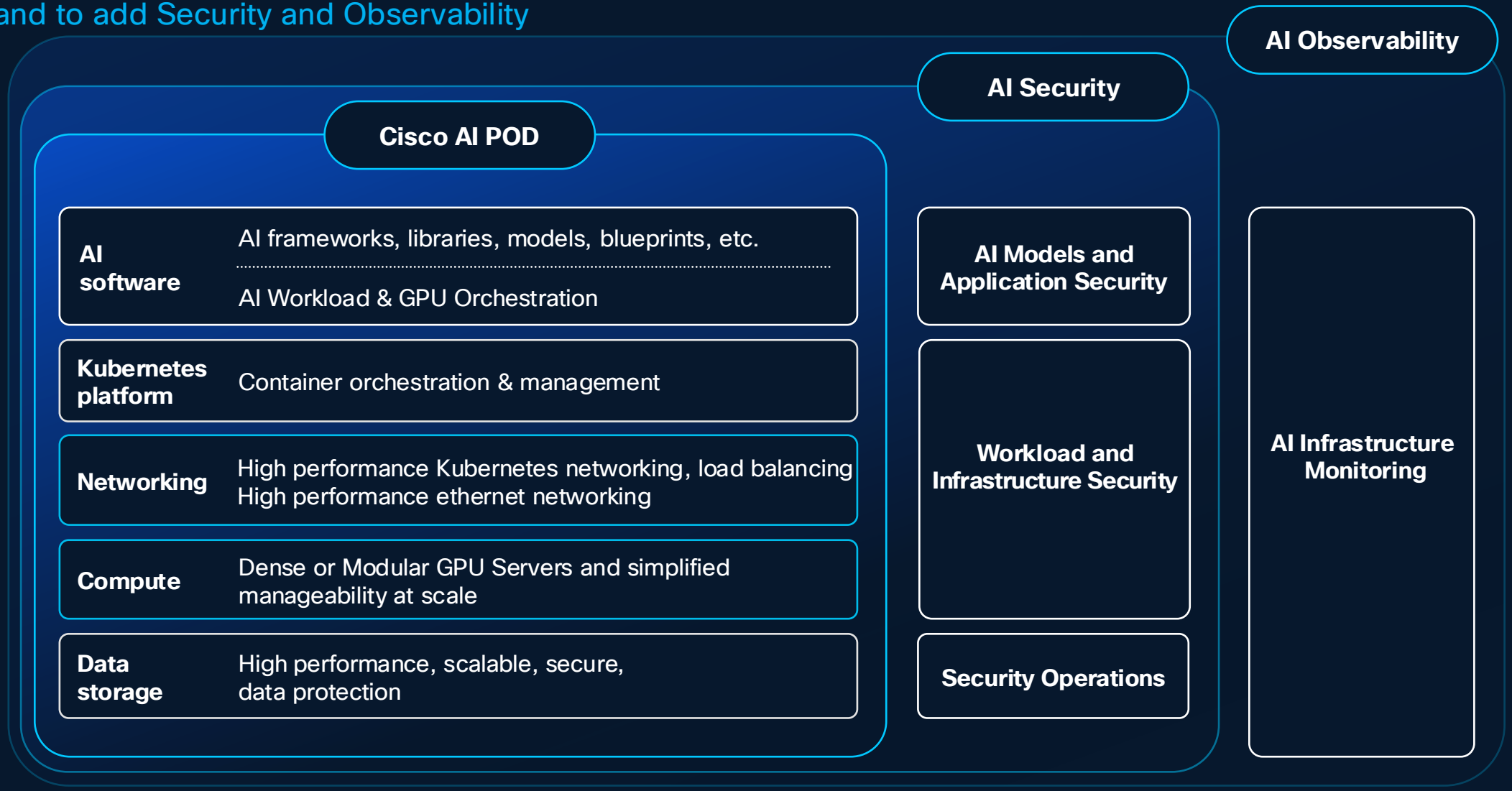
Deploy AI
with confidence

Orderable, validated AI-ready infrastructure stacks

Fully supported stack including Cisco and
3rd party components

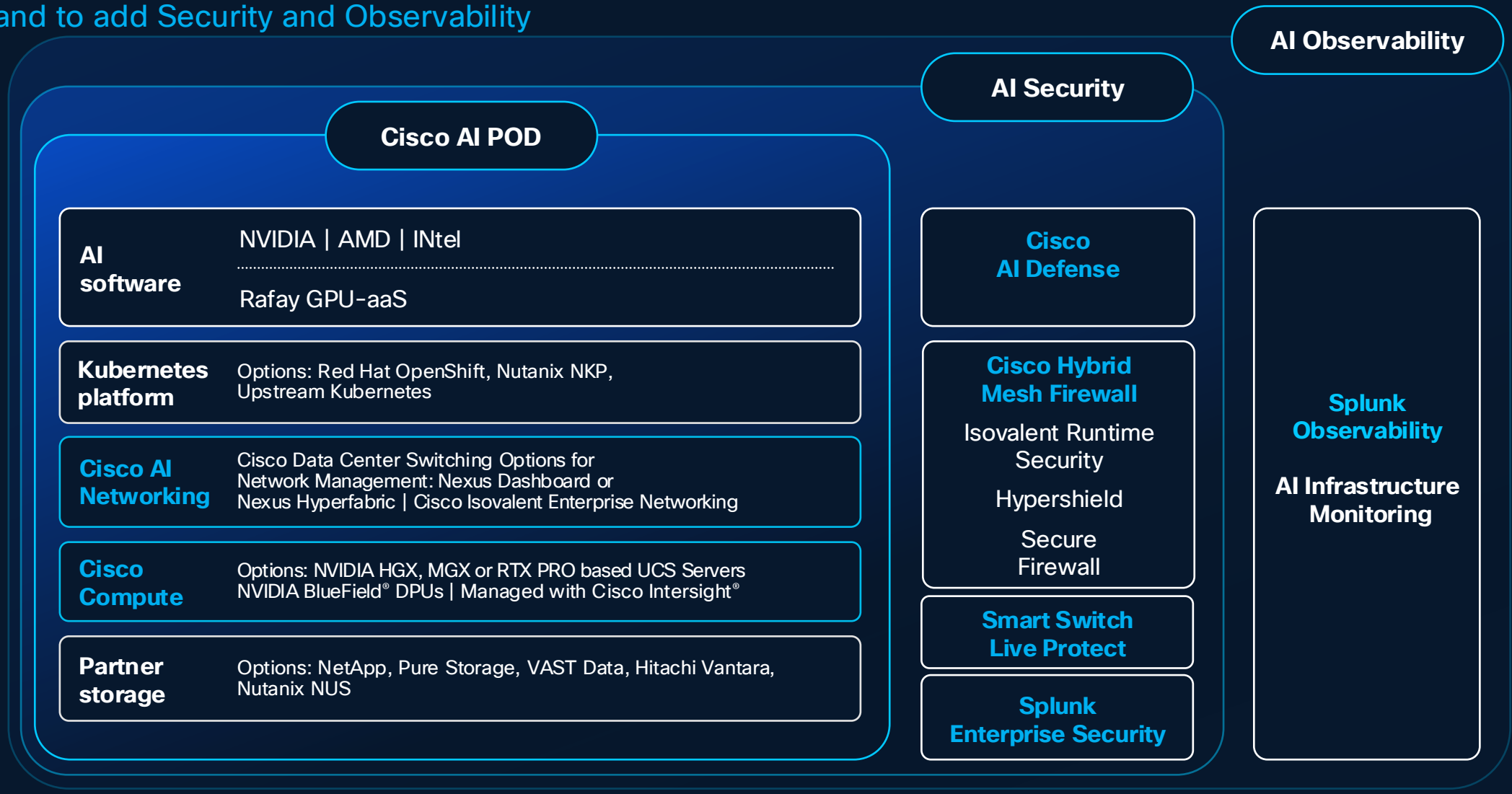
Cisco AI PODs

Expand to add Security and Observability



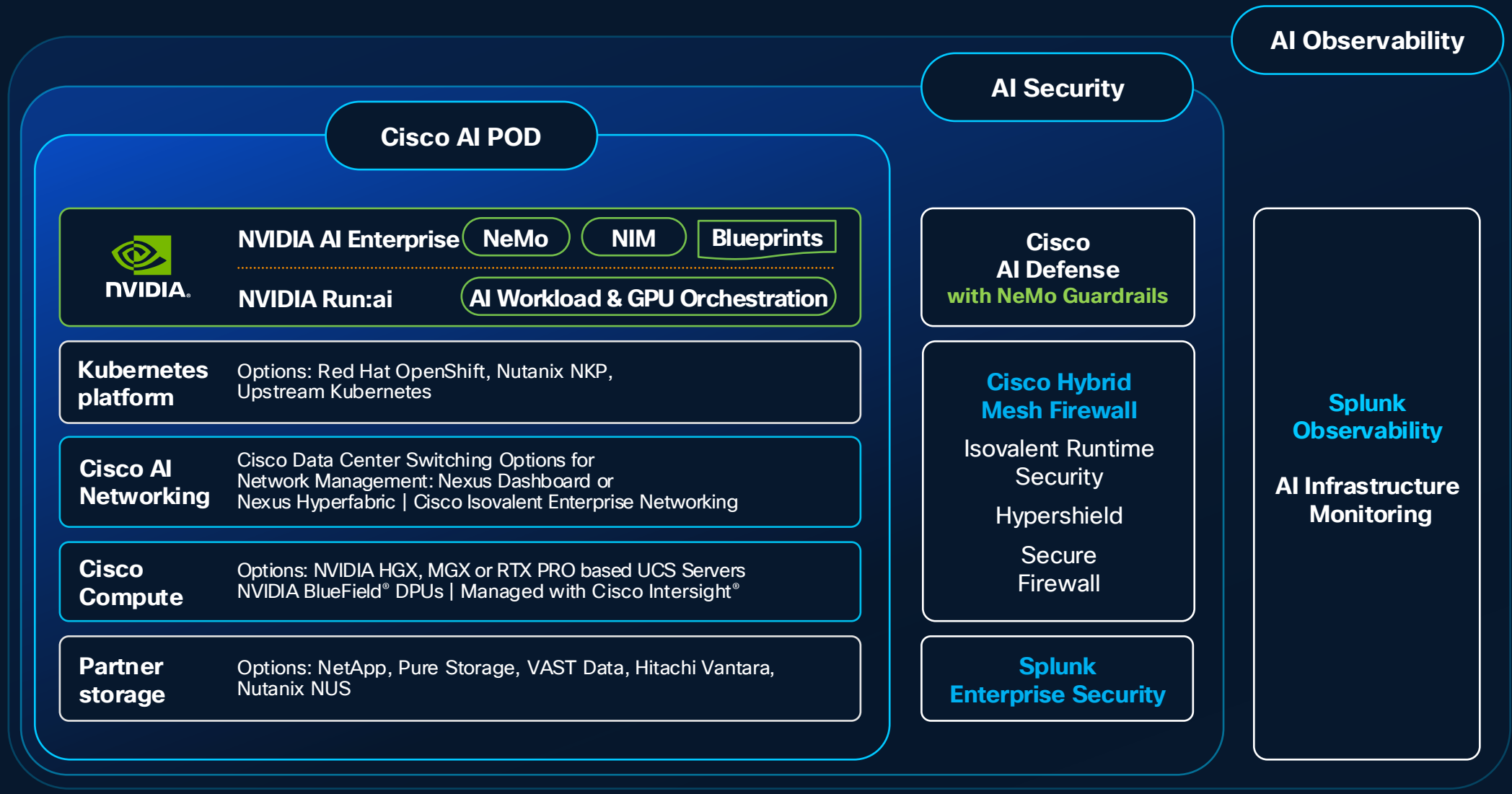
Cisco AI PODs

Expand to add Security and Observability

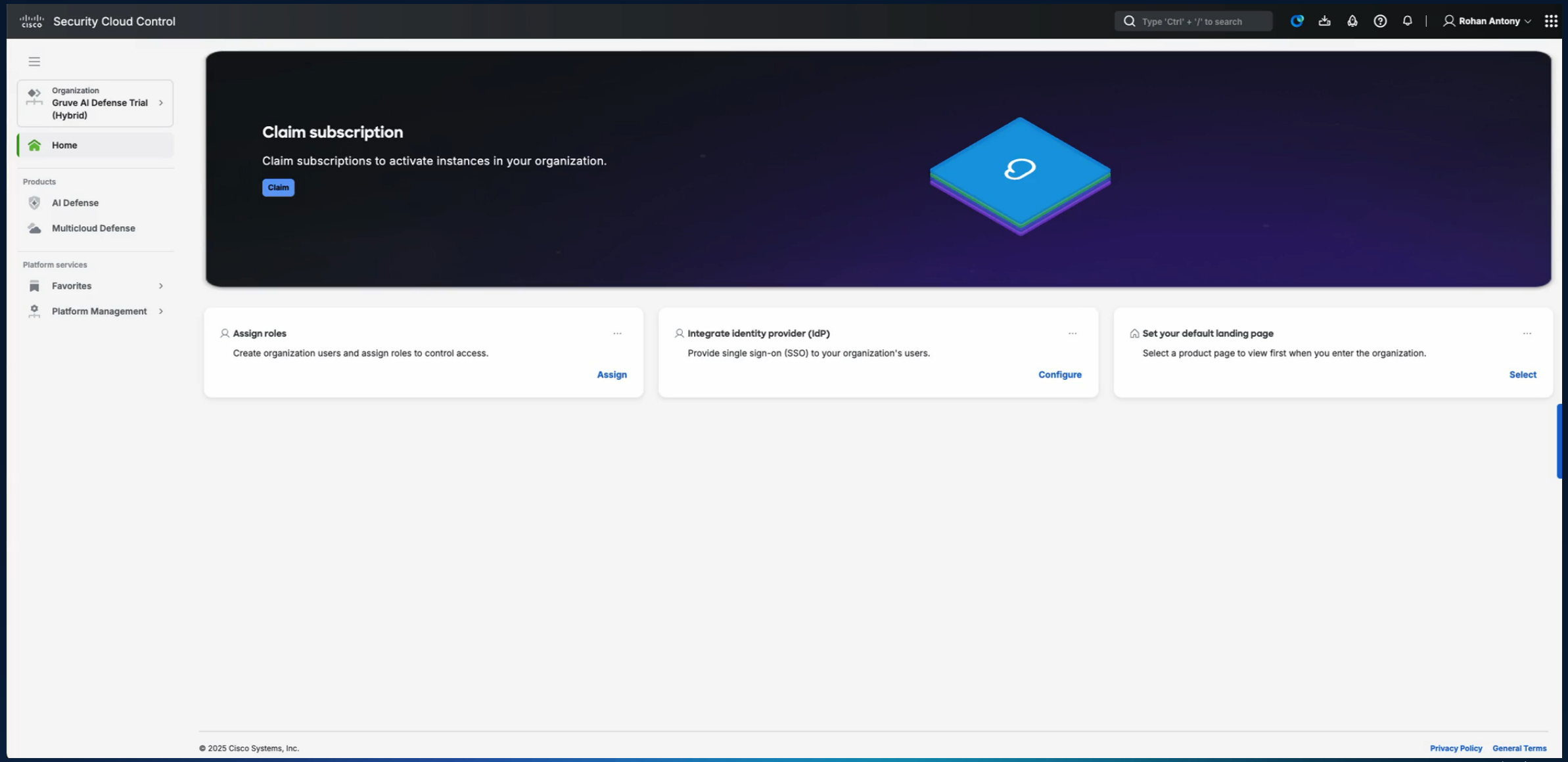


Cisco Secure AI Factory with NVIDIA

Expanded partnership to accelerate AI adoption in the enterprise



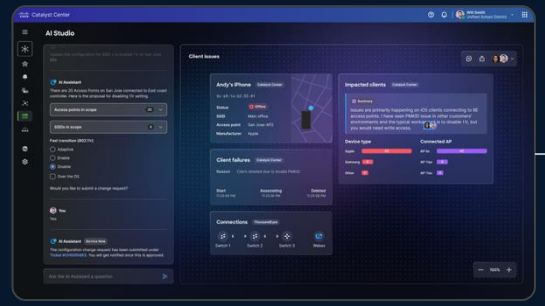
Cisco AI Defense



Cisco + Vast Data + Nvidia

AgenticOps Across IT

AI Canvas



Campus and Branch

Catalyst Center
Topology, client details, location, etc.

webex
Voice and video experience

Cisco Meraki
Topology, client details, location, etc.

ThousandEyes
WAN, Internet, App Insights

SD-WAN
WAN Details

Identity Intelligence
User trust level, identity checks & reasons

Data Center

Nexus Dashboard
Data center network management

Nexus Hyperfabric
Data center network management

Intersight
Unified management, automation, security

Security and Observability

splunk>
Cisco and third-party insights

Firewall
Security & connection events

ISE
Authentication Insights

Duo
Authentication & compliance

Secure Access
Private & SAAS Resource Access

XDR
Related Threat Incidents

Cisco Compute Portfolio

UCS – AI use case focused servers

CISCO INTERSIGHT®

Validated solutions for AI

Build the model
Training

Optimize the model
Fine-tuning and RAG

Use the model
Inferencing

Dense GPU

Modular (w/GPU Expansion) and Rack

Unified Edge

Demanding AI

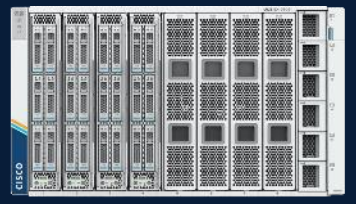
Mainstream and Edge AI

cisco

Infrastructure

Cisco UCS Compute Portfolio

Blade



UCS X-Series
X9508 Chassis

IFM Module

New

UCS X-Series Direct

UCS X210c M7

New

UCS X210c M8

UCS X410c M7

UCS B200 M6

New

UCS X215c M8

New

UCS X580p
PCIe Gen5 node
PCIe Gen5 switch module

Rack

New

UCS C240 M8E3S
36 EDSFF E3.S1T

New

UCS C240 M8SX
28 HDD/SDD/NVMe

New

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

New

UCS C220 M8E3S
16 EDSFF E3.S1T

New

UCS C220 M8S
10 HDD/SSD/NVMe

UCS C220 M7N
10 NVMe

New

UCS C245 M8SX
28 HDD/SDD

New

UCS C225 M8S
10 HDD/SSD

New

UCS C225 M8N
10 NVMe

AI Servers

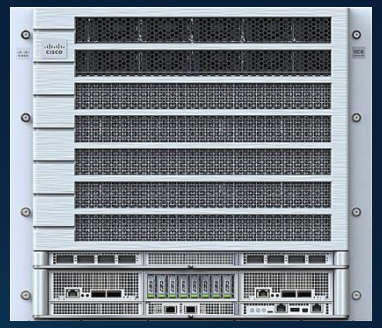
New

UCS C885A M8
8RU Dense GPU Server

New

UCS C845A M8
4RU MGX Server

New



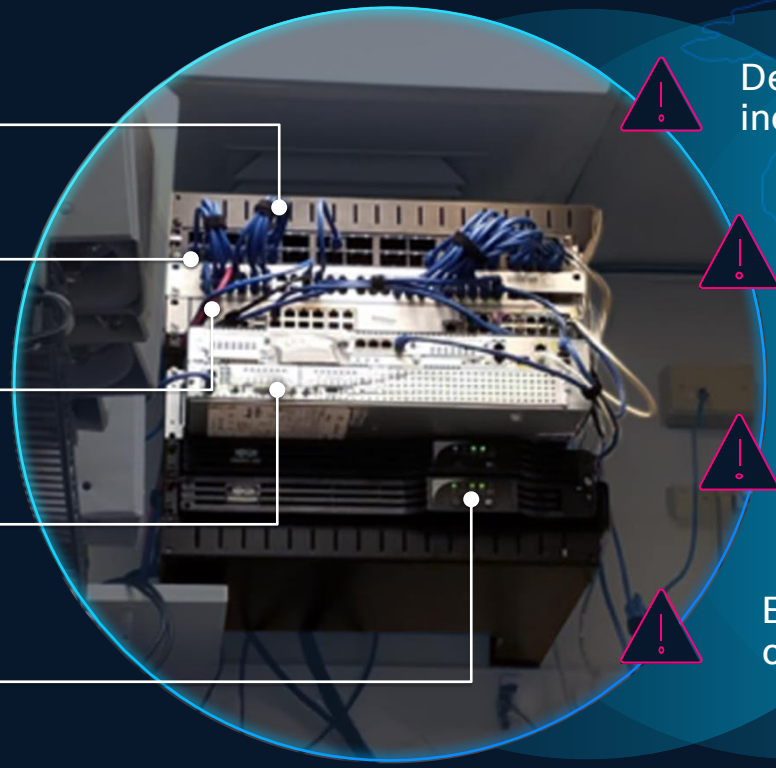
UCS C880A M8
10RU Dense GPU Server

Edge

New

Unified Edge
UCS XE9305 Chassis
UCS XE130c M8
Compute Nodes

Legacy edge infrastructure



Wireless

Switch

Router

Server

Battery backup

Deployment inconsistency

Product incompatibility

Limited technical expertise

Environmental constraints

Deployed **everywhere**

Operational complexity

Security risks

Cisco Unified Edge: Future-Ready Performance

Integrates compute, networking, storage, and security

AI-ready edge

Compute node

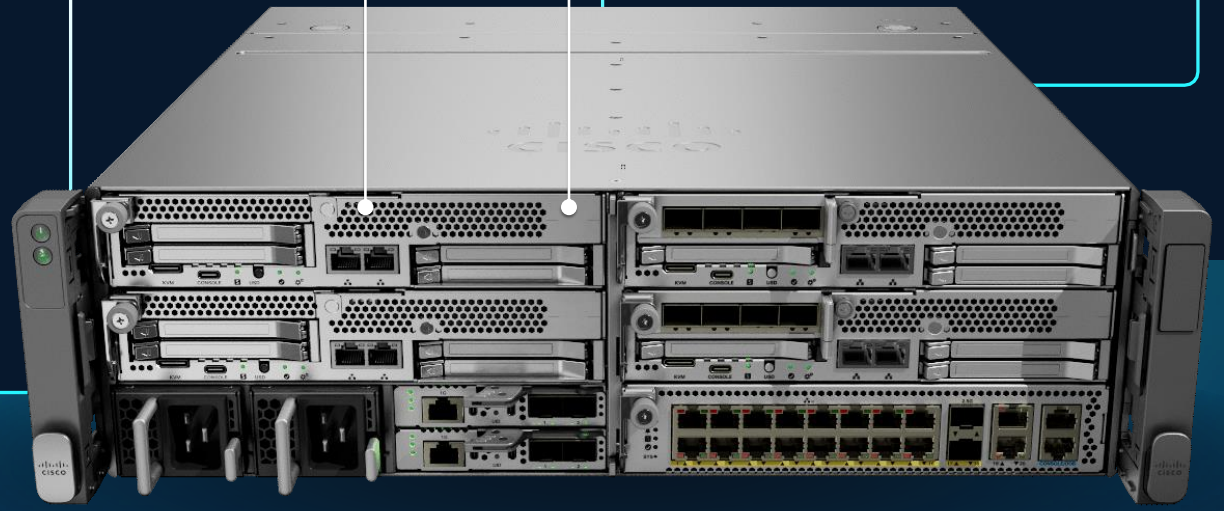
- Compute
- Storage
- Software

GPU

Half-height/half-length GPU

NVIDIA L4 first

Additional GPUs on roadmap



Intel Xeon 6 SoC

CPU native AI inferencing (AMX)

Confidential compute (TDX & SGX)

Integrated Ethernet

Scalable multithreaded cores (12, 20, 32)

NUTANIX · Red Hat · vmware by Broadcom · Canonical Ubuntu · Microsoft · intel · cisco · SUSE

Intersight – fleet deploy, operations and support

Centralized Management

Global Policies

Intuitive Experience

Enhanced Support

Proactive Guidance

Secure and Extensible

SaaS Delivered

Comprehensive Automation

Single Pane of Glass

SaaS Consumption Model

Free customers from care and feeding of management tools and eliminate upgrade dependencies

Seamless Extensibility

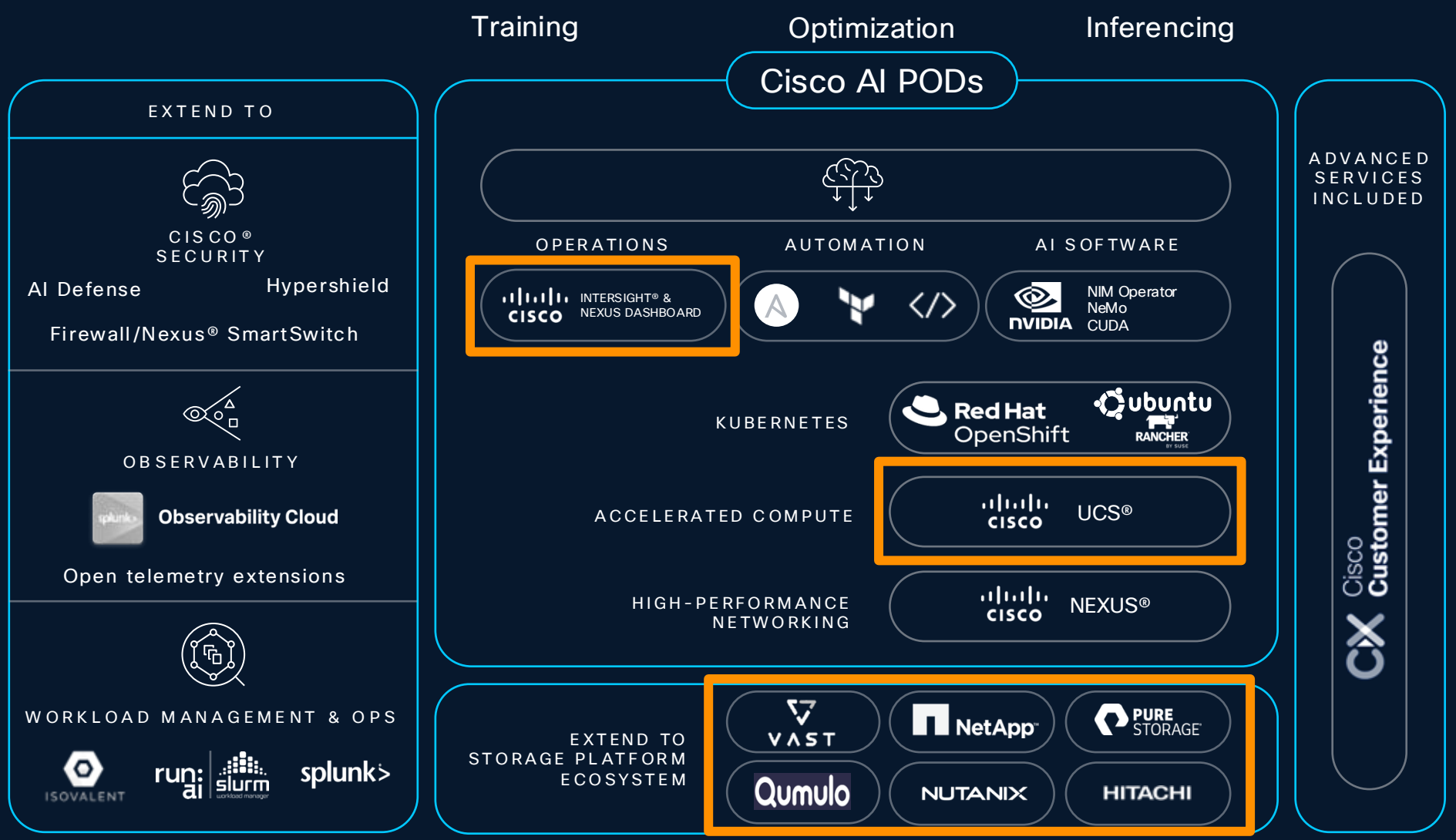
Simplify management across technologies and geography

Continuous Feature Integration

Rapid development, delivery and customer feedback

Cisco AI PODs

Introducing AI POD “Integrated Offerings”



BYO AI tools:

RAFAY
Kubeflow
jupyter
Apache Airflow
Weights & Biases
mlflow
neptune.ai
kedro
comet
ZenML
CLEARML
PREFECT
Flyte
mongoDB

cisco

Infrastructure to power AI

Services to accelerate the value of AI

Security for AI, AI for security

Data to drive insights and context

Software to unlock productivity

Cisco is bringing these together to make your enterprise AI journey easier

Resources to learn more

Cisco Compute

View on cisco.com/go/ucs

AI-Ready Infrastructure

View on cisco.com

Isovalent Enterprise Platform

View on Isovalent.com (now part of Cisco)

Cisco Compute YouTube channel

Visit youtube.com

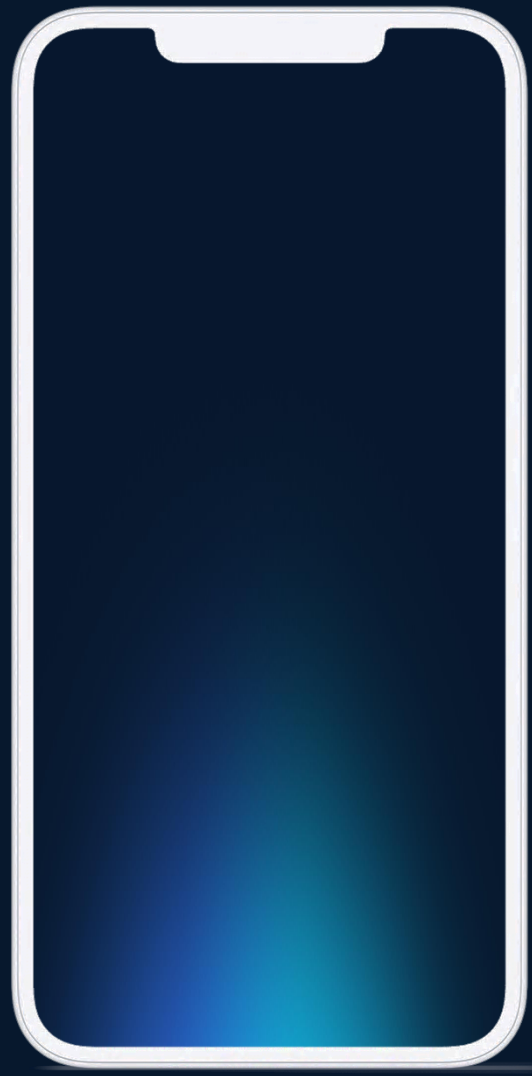
Blogs

Visit blogs.cisco.com/datacenter

Online community

Visit Data Center and Cloud online community

Interested in a Unified Edge Test Drive?



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

Thank you