CISCO Connect

Bringing It Together: The Cisco Secure AI Factory

Shannon McMackin
Solutions Architect, Compute

Shannon Rossi
Account Executive, Architecture

March 24, 2026

AI Use Cases Across Industries

Knowledgebase copilots

AI assistants

Content & code generation

Text | Images | Video | Code

Virtual agent & chatbots

Specialized domain specific chatbots

Visual Computing

Digital Twins | Video Analytics | Imaging & Diagnostics

Language translation

Multilingual real-time communication

Detection & prediction

Forecasts | Anomalies | Insights

Challenges with AI projects delays time to value realization

Security vulnerabilities

AI models, frameworks, apps, infrastructure are new cyberattack surface with threats such as prompt injection, denial of service, & 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

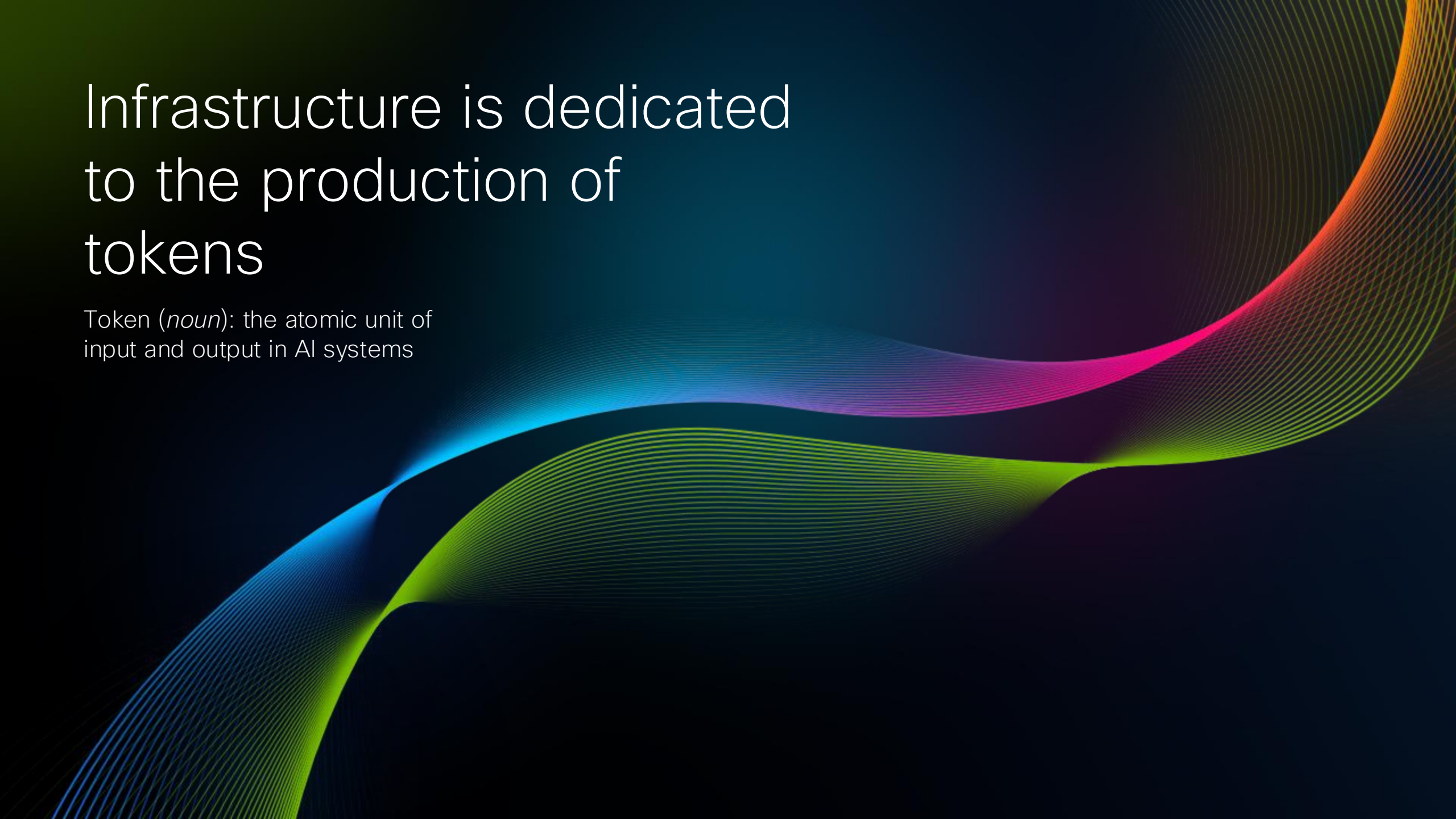
Infrastructure
constraint

Trust
deficit

Data
gap

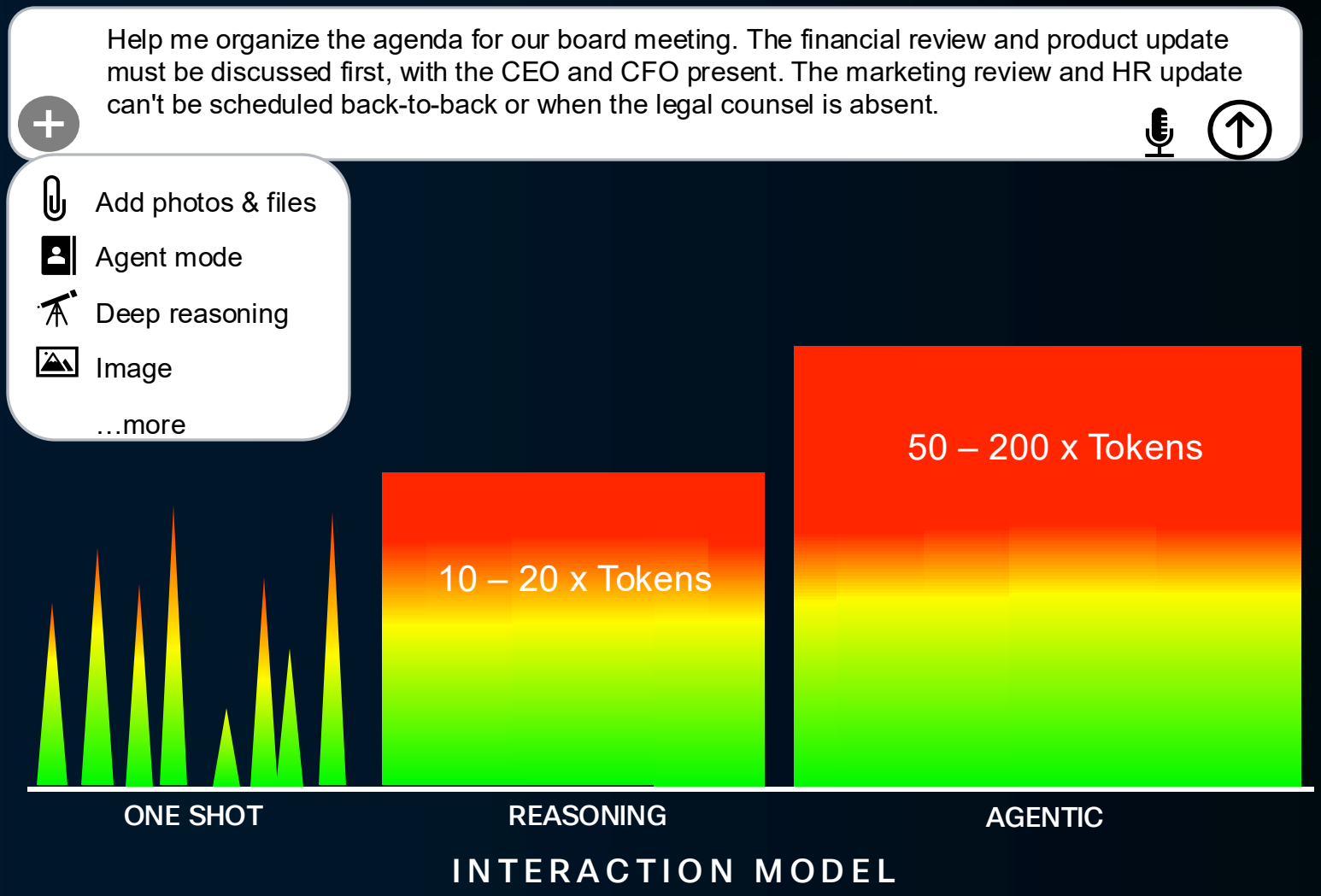
Infrastructure is dedicated to the production of tokens

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

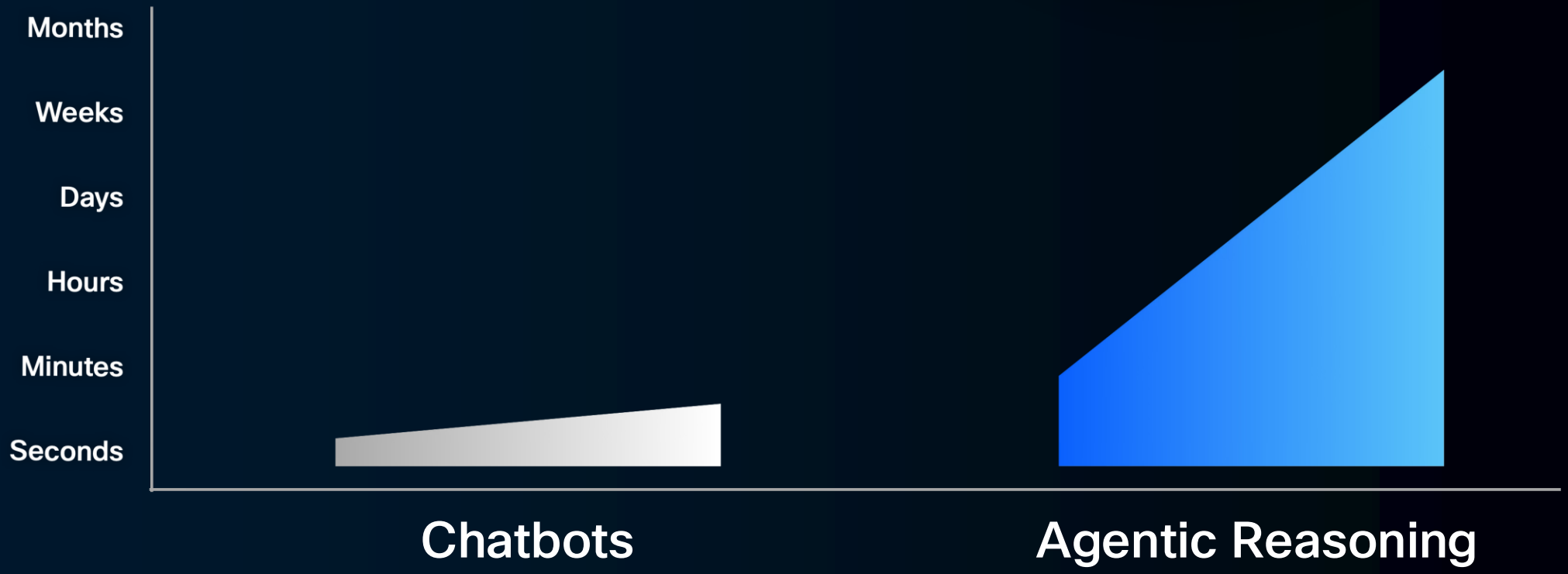


AI Is Changing: Token Demand Inflation

More tokens enable higher quality results and more complex tasks



Duration of Autonomous Execution



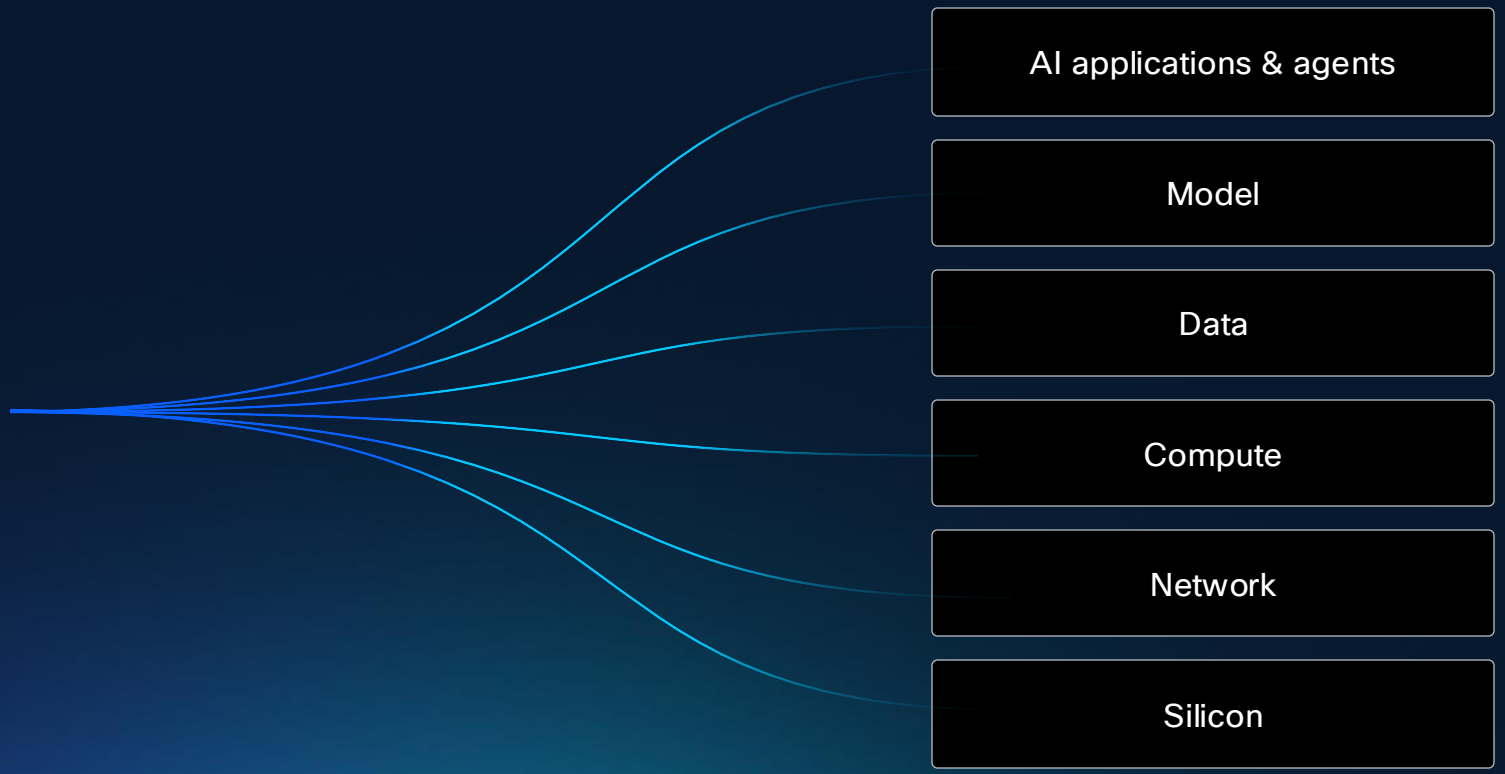
Infrastructure
constraint

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 & polarization

Self-harm

Financial harm

Security

Indirect prompt injection

Infrastructure compromise

IP theft

Meta prompt extraction

Prompt injection

Model theft

Training data poisoning

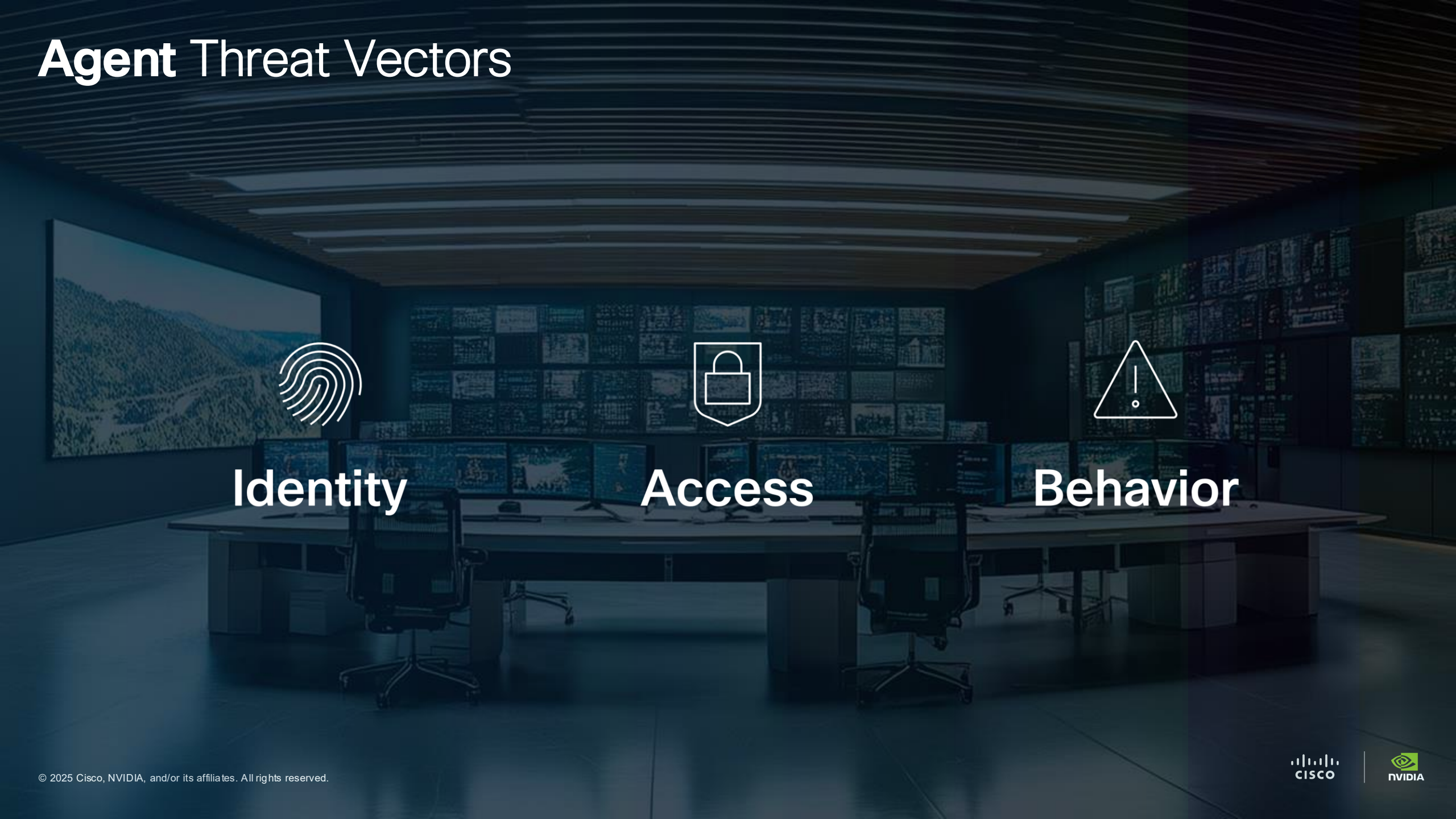
Sensitive information disclosure

Data exfiltration

Model denial of service

cisco | NVIDIA

Agent Threat Vectors



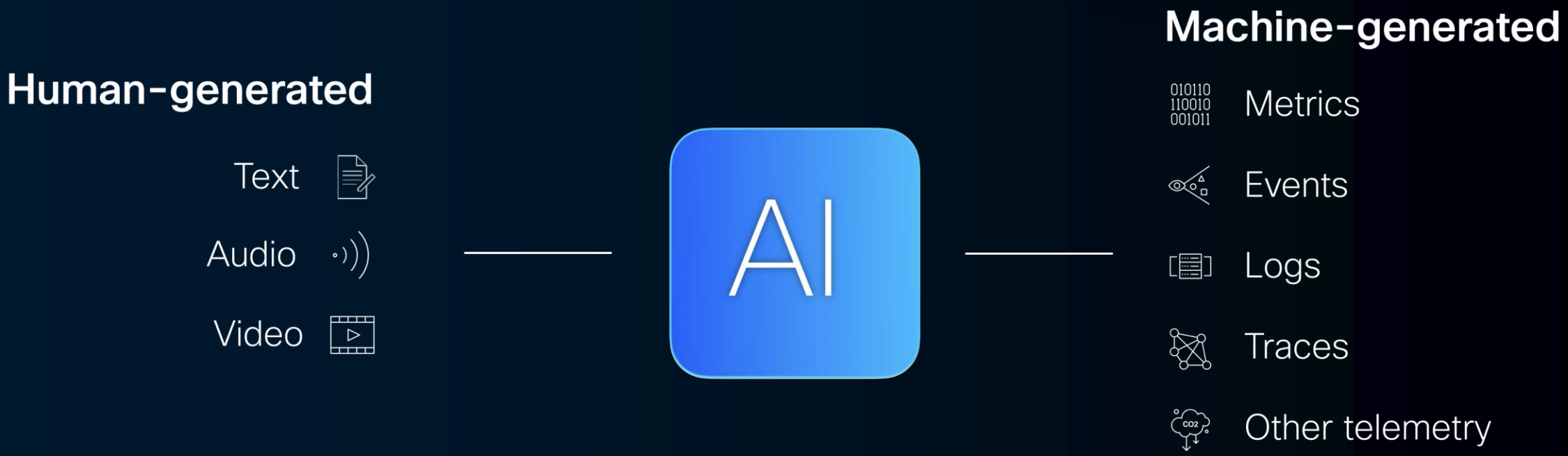
Infrastructure
constraint

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 of these challenges **head-on**

Cisco is the **critical infrastructure** for the **AI era**

Bringing AI to the Enterprise

Expanded Partnership to Accelerate AI Adoption in the Enterprise with NVIDIA

Partnership focus on Private data centers

Cisco Full-stack AI infrastructures

- Accelerated Compute
- High-performance Network (backend and frontend)
- NVIDIA & Cisco Certified Storage Ecosystem

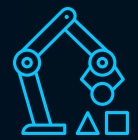
Cisco Security

Splunk Observability

Cisco is now included in NVIDIA Spectrum™-X Ecosystem, and a partner for NVIDIA Reference Architectures

Cisco RAs[1] are Spectrum™-X validated; and compliant with NVIDIA RAs:

- Cisco Silicon based Switches with NVIDIA SuperNICs
- Cisco Spectrum™-X based Switches with NVIDIA SuperNICs



Jointly deliver Secure AI Factory with customer choice: Cisco Silicon or NVIDIA Spectrum™-X Silicon architecture

1 – Enterprise RAs available today (AI Infrastructure with Cisco Nexus 9000 switches RA, Cisco Nexus Hyperfabric AI Enterprise RA); Cloud Partner RAs coming soon

cisco | NVIDIA

Cisco Secure AI Factory with NVIDIA

Accelerate the delivery of trusted and transformative AI applications with a secure, scalable, high performing AI Infrastructure

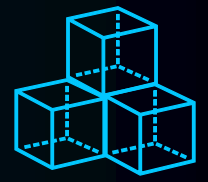
Secure

Security-first architecture with observability and resiliency enables safe enterprise AI

Scalable

Scalable, high-performance, enterprise AI infrastructure enables faster delivery of AI tokens and applications

Simplicity



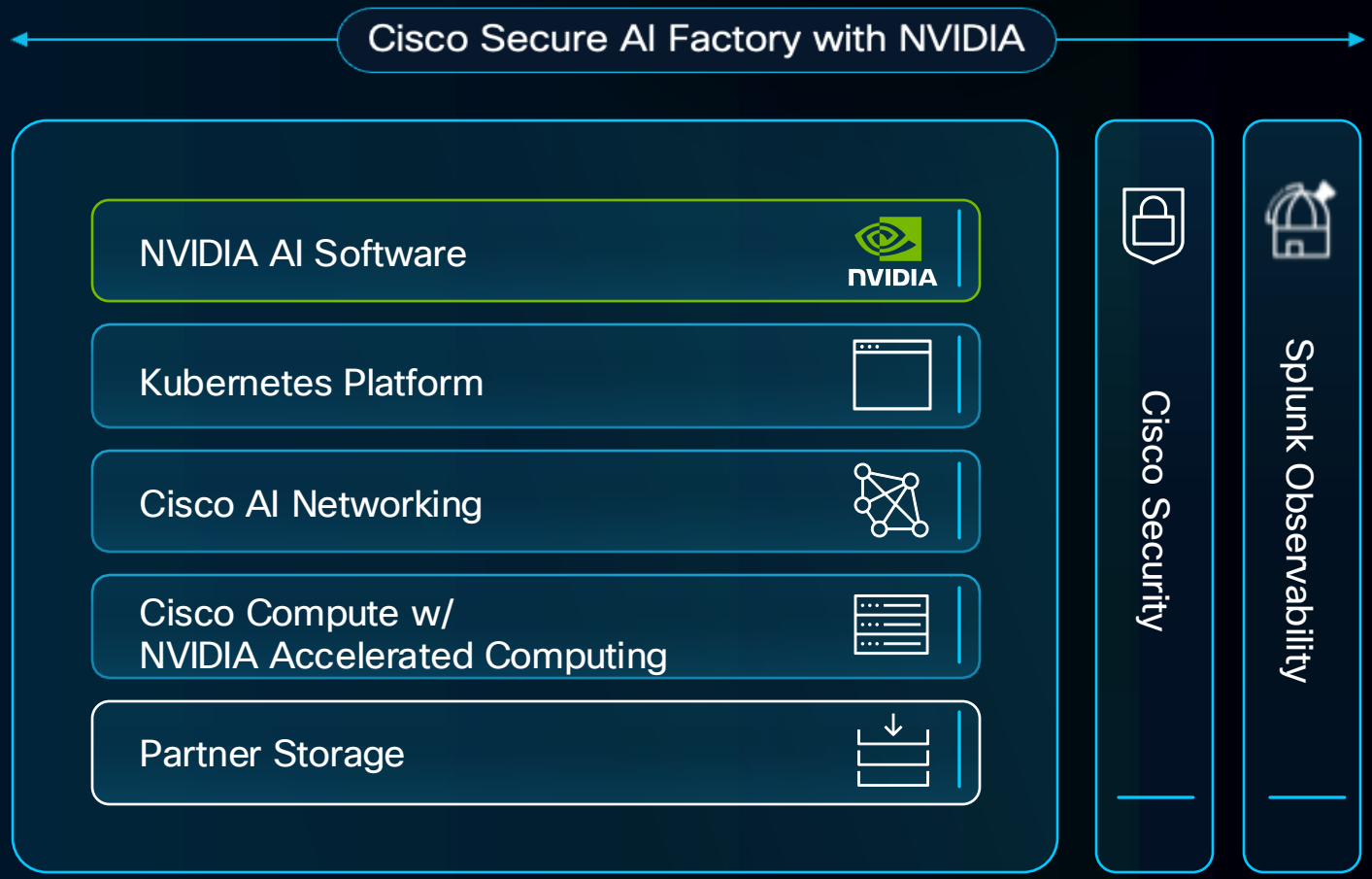
Deployment flexibility & simplicity helps improve AI Practitioners and IT Infrastructure teams productivity

cisco | NVIDIA

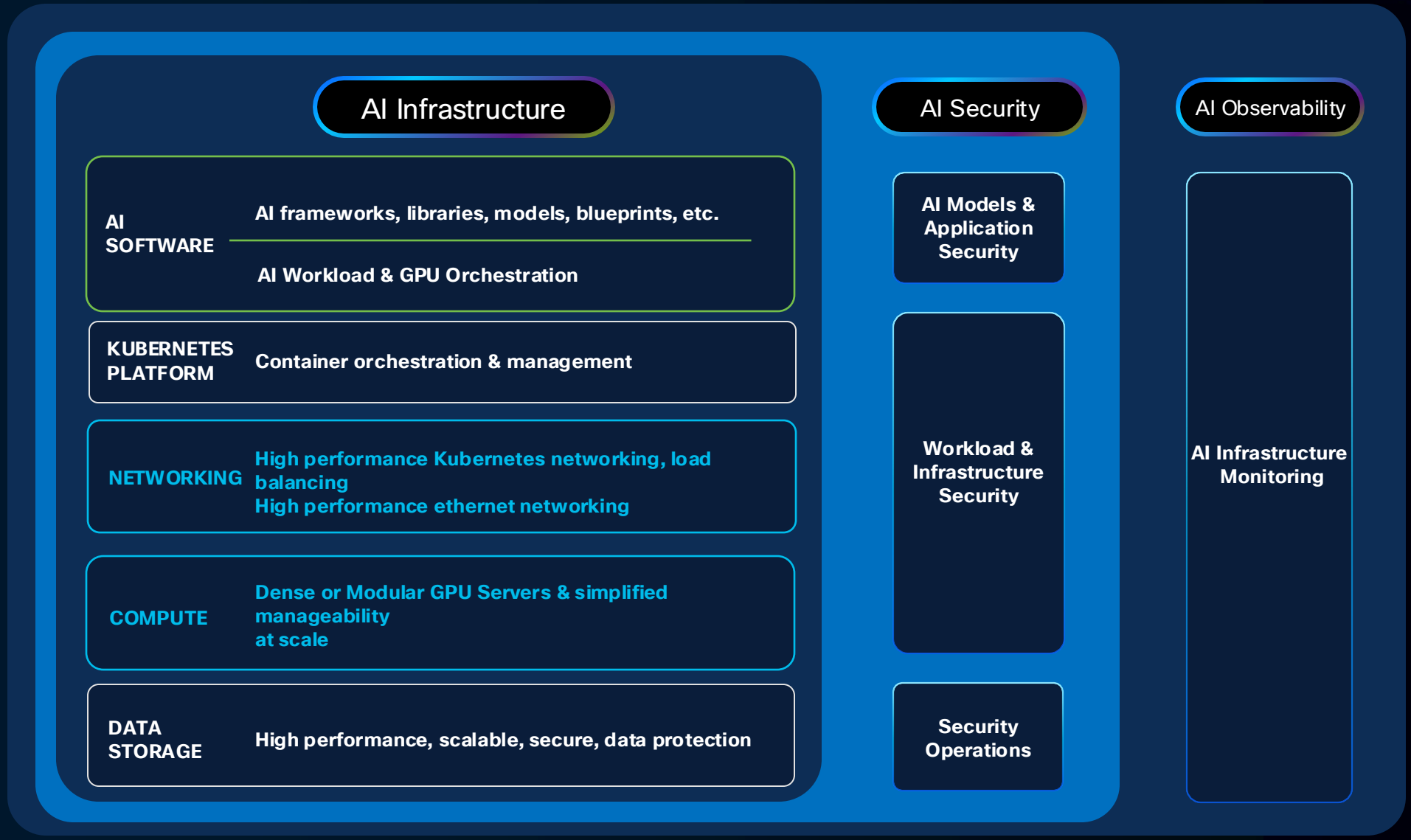
Cisco Secure AI Factory with NVIDIA

What is it?

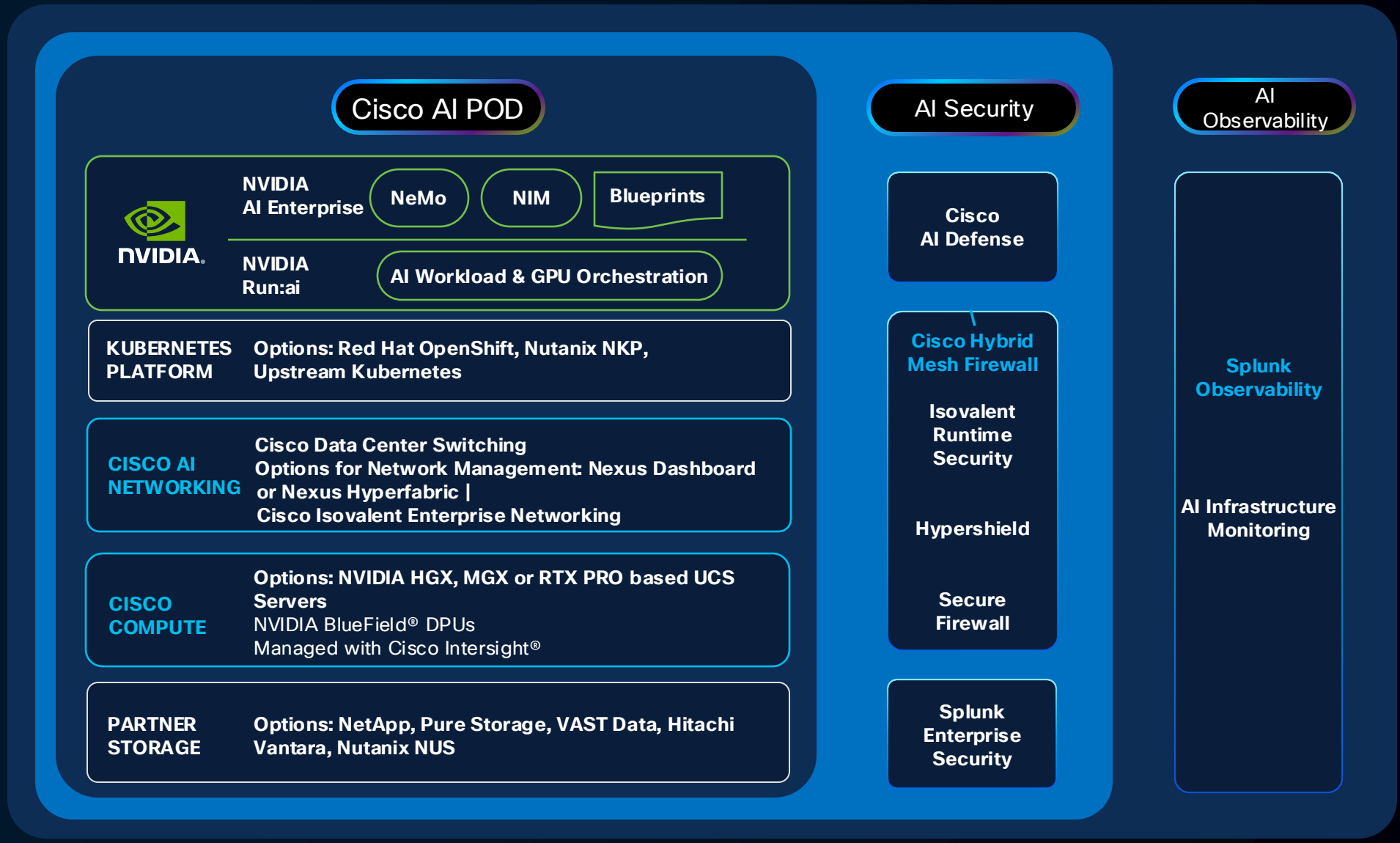
A modular reference design that combines high-performance infrastructure with full-stack security and observability



Key Capabilities of Cisco Secure AI Factory with NVIDIA



Key Products in Cisco Secure AI Factory with NVIDIA



Security-First Architecture Enables Safe Enterprise AI

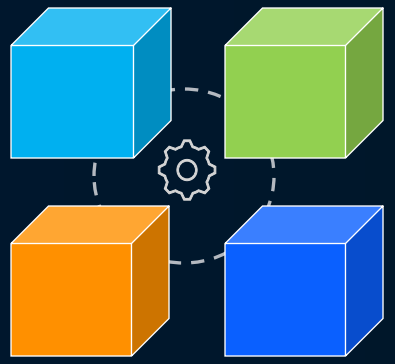
Security at all layers of the stack

Securing the Applications	Cisco AI Defense—Robust testing and runtime security of LLMs and generative AI applications, integrated with NVIDIA AI.
Securing the Workloads & Infrastructure	Cisco Hybrid Mesh Firewall—Unified security management and consistent and pervasive policy across multiple enforcement points. • Cisco Isovalent-Enhanced visibility into cloud native interactions, enabling consistent policy definition and enforcement across software defined networks. • Cisco Hypershield—Protection against adversary lateral movement and proactive vulnerability mitigation without patching, through a single management interface. • Cisco Secure Firewall-Advanced threat protection at scale without compromising performance, with unified management across firewalls.
Security Operations	Splunk Enterprise Security - TDIR Platform for real-time threat detection, investigation, & response through analytics, automation, &risk-based insights.

cisco | NVIDIA

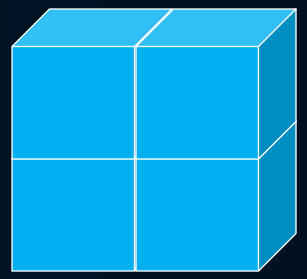
Flexible Deployment Options

Build Your Own



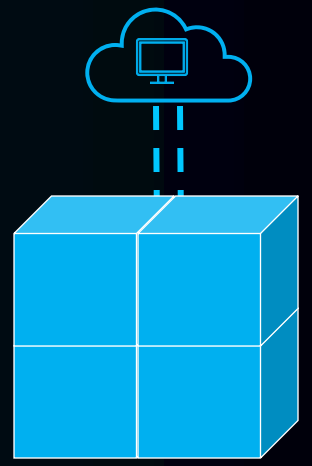
Buy & deploy individual products, as needed

AI POD w/ On-prem network management



Modular, pre-validated infrastructure
- Full stack, buy & deploy
- Backed by CVDs

AI POD w/ Cloud based network management



Turnkey infrastructure:
- Full stack, buy & deploy
- Nexus Hyperfabric: Cloud-managed Networking
- Nexus Hyperfabric AI: Cloud-managed physical infrastructure

How Does It All Come Together?

AI PODs are the building block of Cisco Secure AI Factory with NVIDIA

Purpose: AI Ready Infrastructure PODs, backed by CVDs, to deploy Enterprise AI workloads.

Examples: Generative and agentic AI applications, model training and optimization

Value: Full-stack validation and performance characterization to provide accelerated time-to-value.

Purpose: *Dedicated* infrastructure PODs, backed by CVDs, for AI Security, Observability and Data Services

Examples: Cisco AI Defense, Splunk Observability, NVIDIA AI Data Platform.

Value: Ensure the security, efficiency and data readiness of your AI Factory.

1-To-Many Relation Between Workload & Services PODs

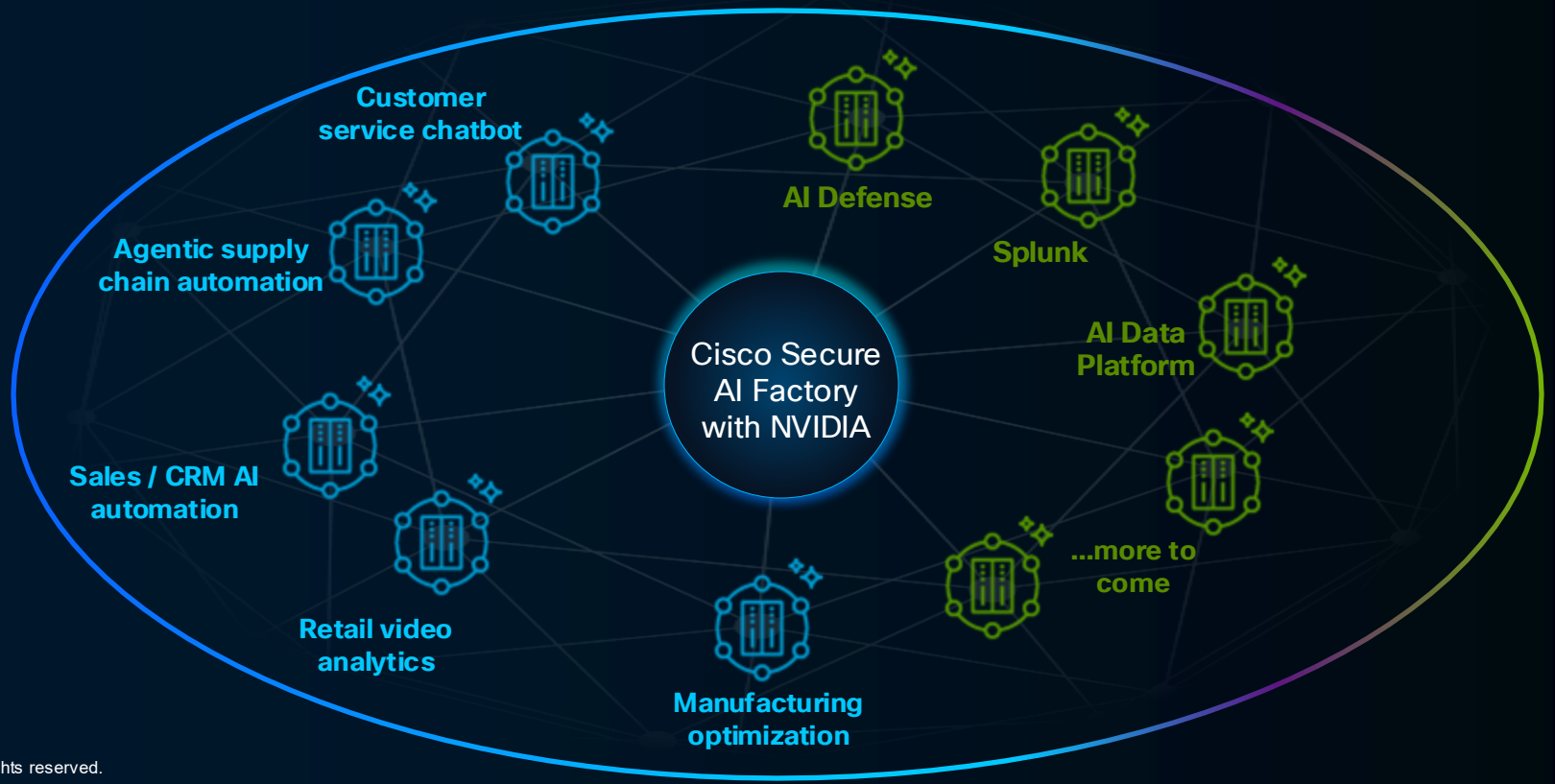
Cisco Secure AI Factory with NVIDIA

AI Workload PODs

Full-stack infrastructure for deploying AI applications on customer premise.

AI Services PODs

Security, observability or data platform services

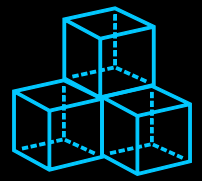


Summary: Cisco Secure AI Factory with NVIDIA

Accelerate the delivery of trusted and transformative AI applications

Security-first architecture enables safe enterprise AI

High-performance AI infrastructure enables efficient model training, customization, and inferencing



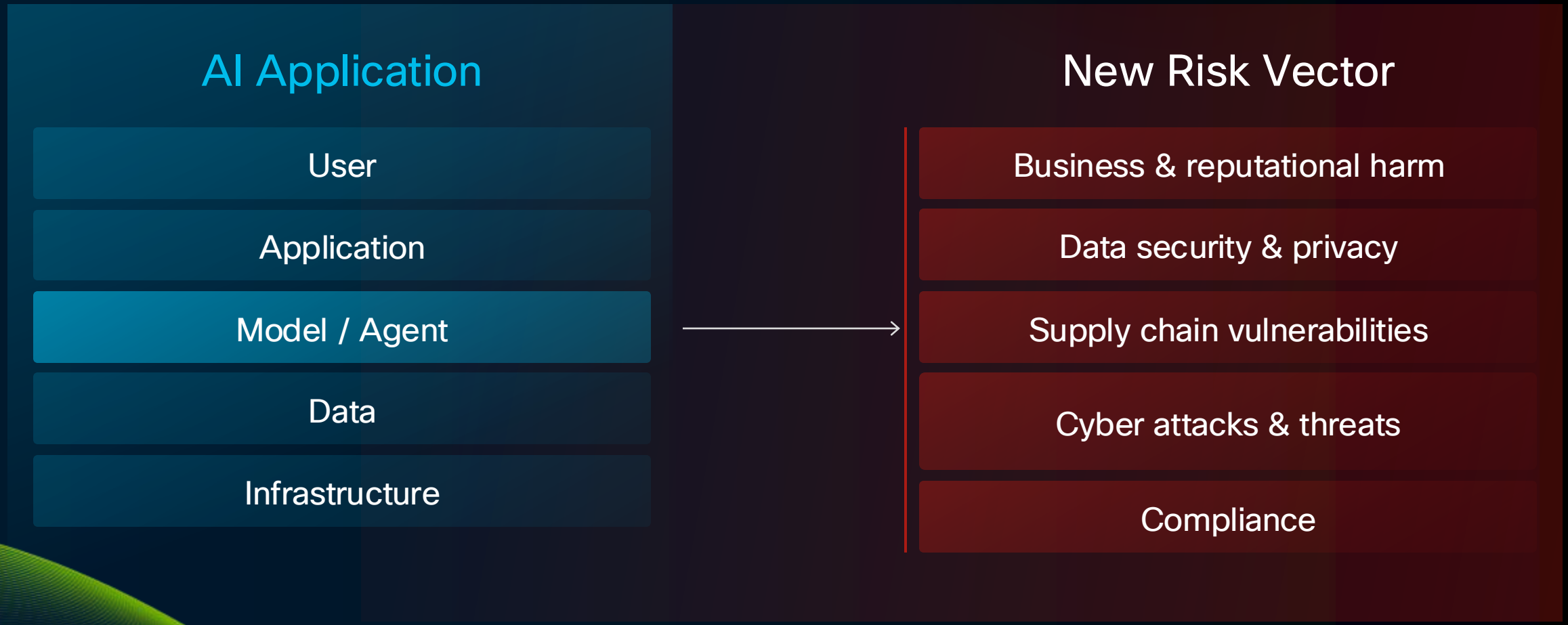
Modular deployment options, backed by CVDs improves productivity for AI practitioners and IT teams

cisco | NVIDIA

Cisco AI Defense for Securing Applications

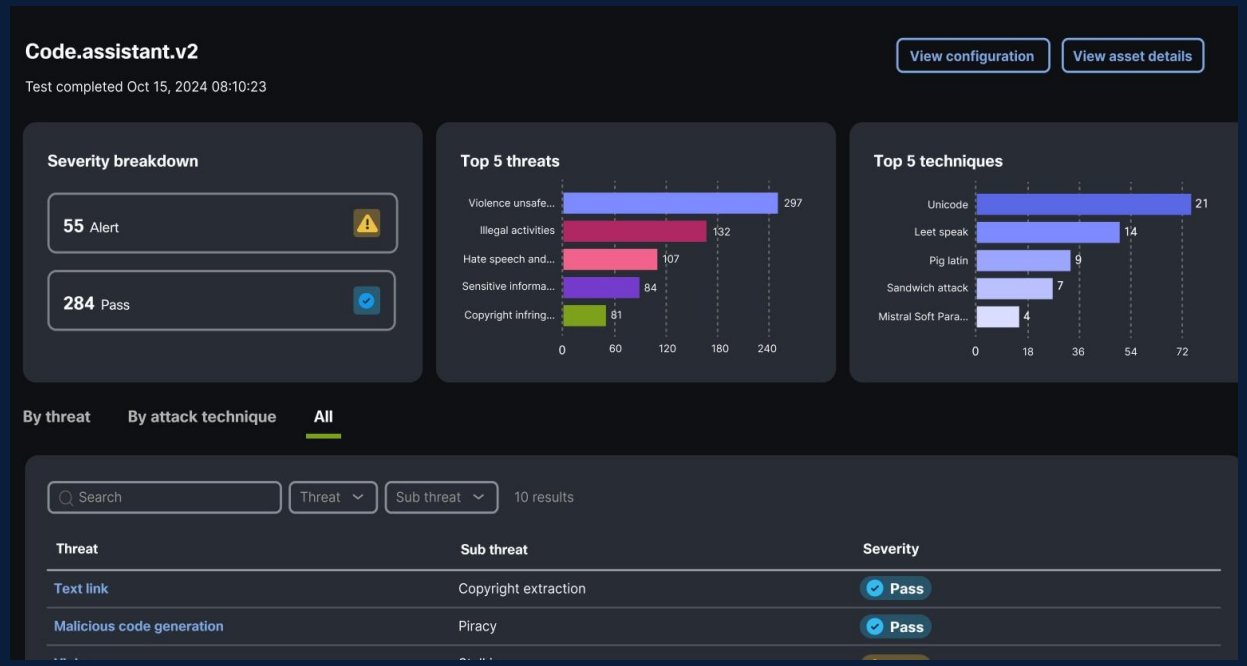
What's the Risk Introduced by AI Applications?

AI applications are complex and non-deterministic



cisco | NVIDIA

Security-First Architecture Enables Safe Enterprise AI



Protect AI models and applications from cyber attacks anywhere with Cisco AI Defense

Identifies AI assets

Assesses risks

Mitigates threats in real time

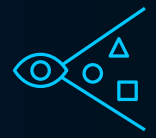
Why Cisco AI Defense on AI PODs?

Enterprises need comprehensive, unified security coverage controls for AI Models & Apps on-premises

End-to-end security

Built on our infrastructure – from network to AI workload

Identify vulnerabilities and protect against safety and security threats



Continuous visibility and governance

Real-time insight into dataflows, model behavior, and threats

Continuous policy-based audits to meet regulatory requirements

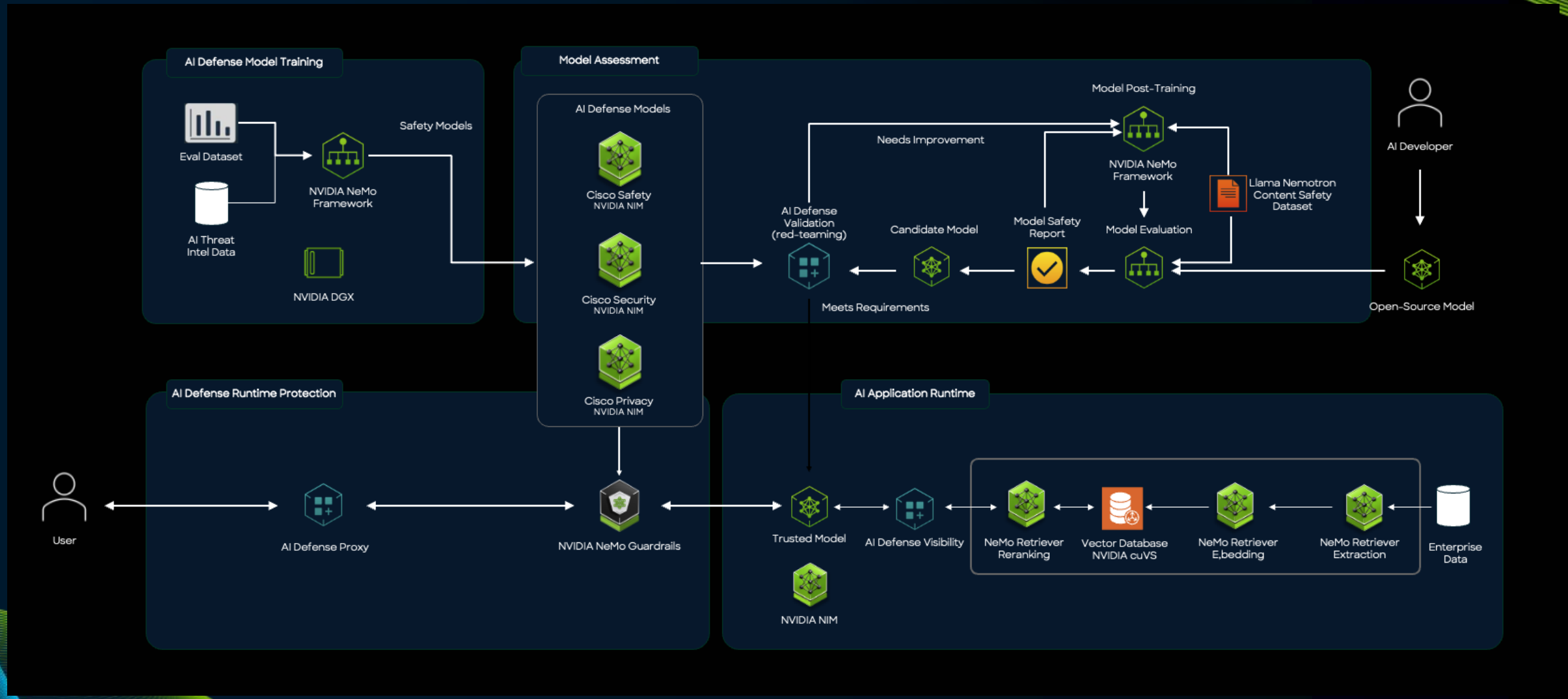
Sovereign ready

Extend unified security controls across cloud and data center

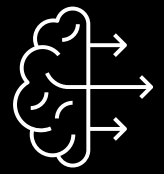
Readiness for regulated and sovereign AI deployments

cisco | NVIDIA

Cisco AI Defense with NVIDIA AI Enterprise



Cisco AI Defense on AI POD Capabilities



AI Model and Application Validation

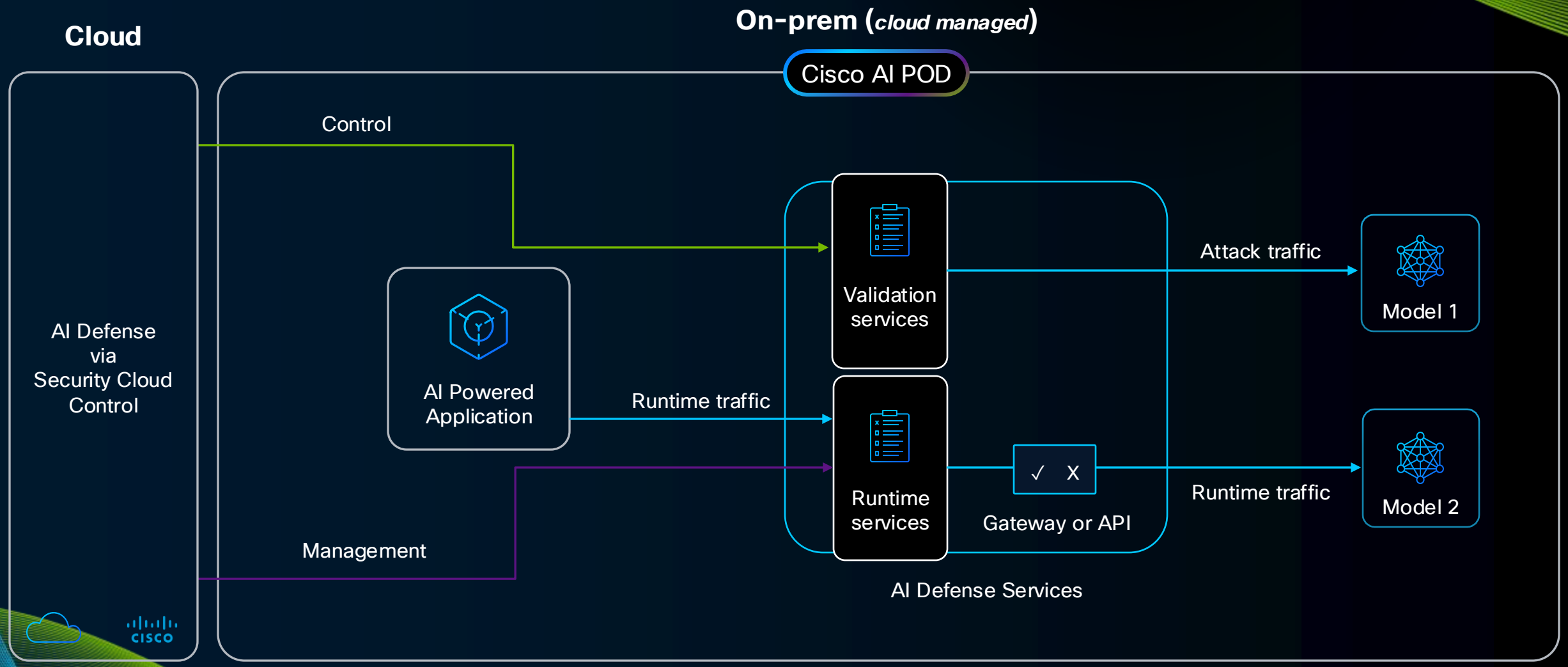
Test for vulnerabilities with algorithmic red teaming

AI Runtime Application Protection

Enforce guardrails to block malicious prompts and unsafe responses

cisco | NVIDIA

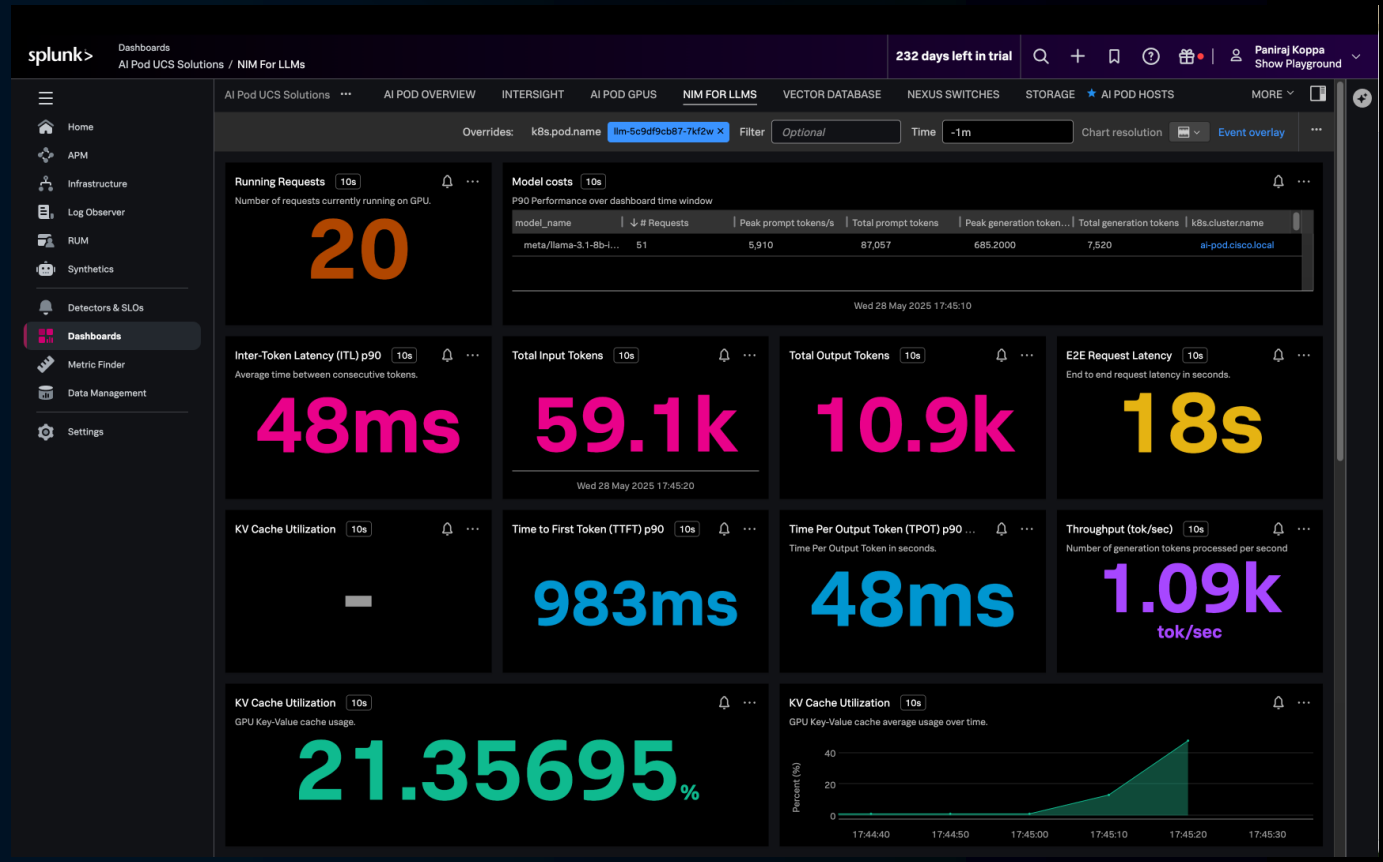
Cisco AI Defense on AI PODs Architecture



Splunk Observability Dashboard for AI PODs

Splunk Observability Dashboard for Cisco AI PODs

- **Real-Time Monitoring and Troubleshooting**
- **Generate actionable Insights**
- **Efficient Telemetry Data Ingestion and Processing**



NVIDIA AI Enterprise Software

NVIDIA AI Enterprise Software

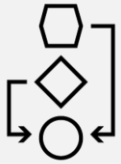
- Cloud Native Software Platform for Production AI

AI Solutions
Fastest path to production

NVIDIA NIM & NeMo

Standard APIs and performance optimized models

NVIDIA Blueprints



Reference workflows for AI development

CUDA-X

NVIDIA

SDKs and Frameworks with full-stack acceleration

Enterprise-Grade
Built for business

Security

Development Flexibility

Validated Deployment Guidance

SLAs with NVIDIA Support

Run Everywhere
Cloud Native and Certified

Cloud | Data Center | Workstation | Edge

NVIDIA Certified kubernetes

NVIDIA NPN Cloud Partners

aws Microsoft Azure

Google Cloud ORACLE CLOUD Infrastructure

CISCO CISCO NVIDIA

Interested in Learning More?

Please use this form to request a deep-dive event with the Cisco Cloud + AI Infrastructure team.

CISCO Connect

Thank you