CISCO Engage | Tech Day Twin Cities

From Model to Mission

Safeguarding Data in AI-Driven Environments

Matt Camacho
Leader Solutions Engineer - Security

Brian Jacklin
Solutions Engineer – Cloud and AI Infrastructure

4/16/2026

Agenda

1. AI Security – Level Set
2. What is an AI Factory
3. What's a Secure AI Factory
4. Secure AI Factory – into Zero Trust Architecture

All organizations are defining their AI Services Stack

Agent Ecosystem

Chatbot Applications

AI Agents

Agent Hub

Agent Platform

SaaS based AI Services

Foundational Model aaS
On-prem & 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 & Cloud /SaaS Data Sources

Relational Databases

NoSQL Databases

Log & Metric Datastores

Documents

AI Platform Services

Data Platform w/ AI/ML Tools

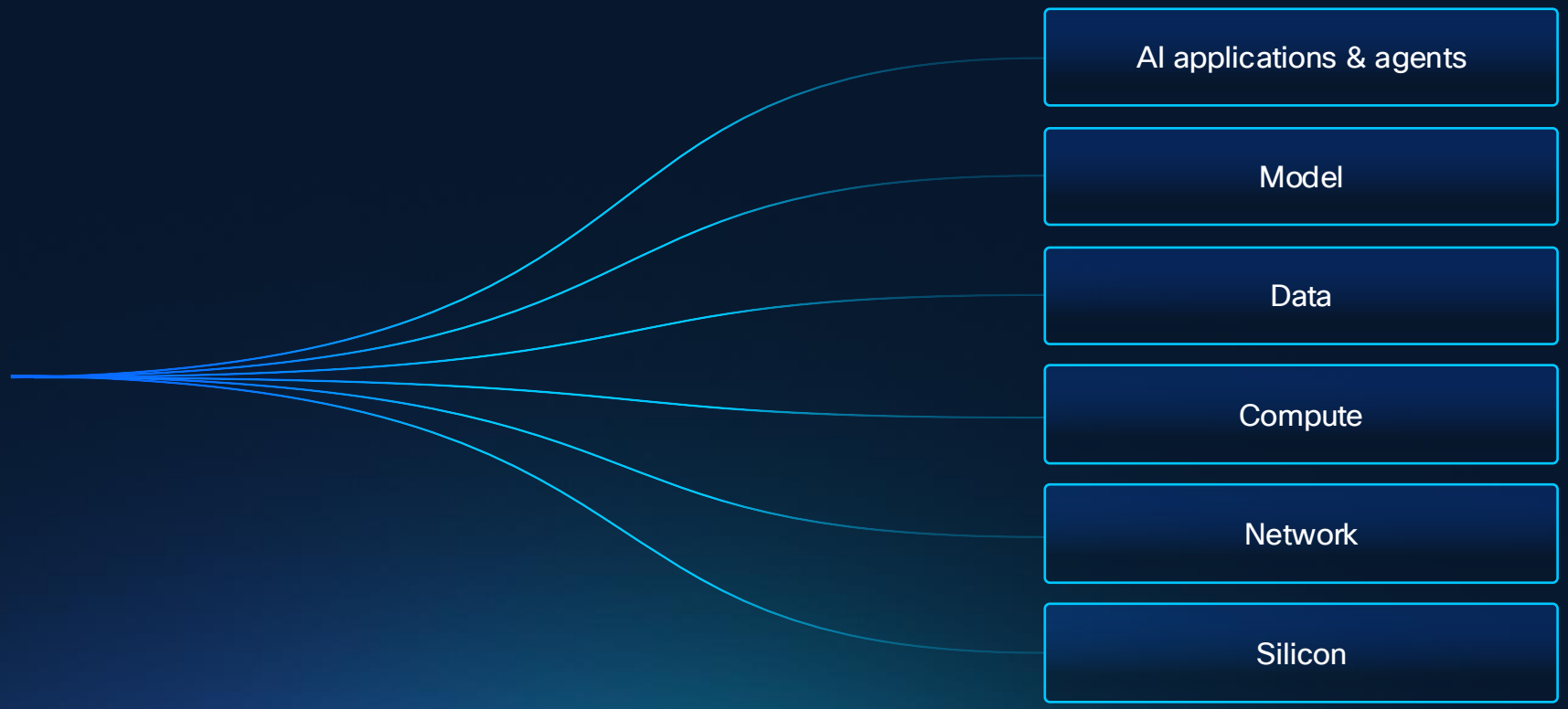
AI Coder

Service Management

Conversational AI for BI

cisco

Lack of **end-to-end** visibility



Model threat vectors

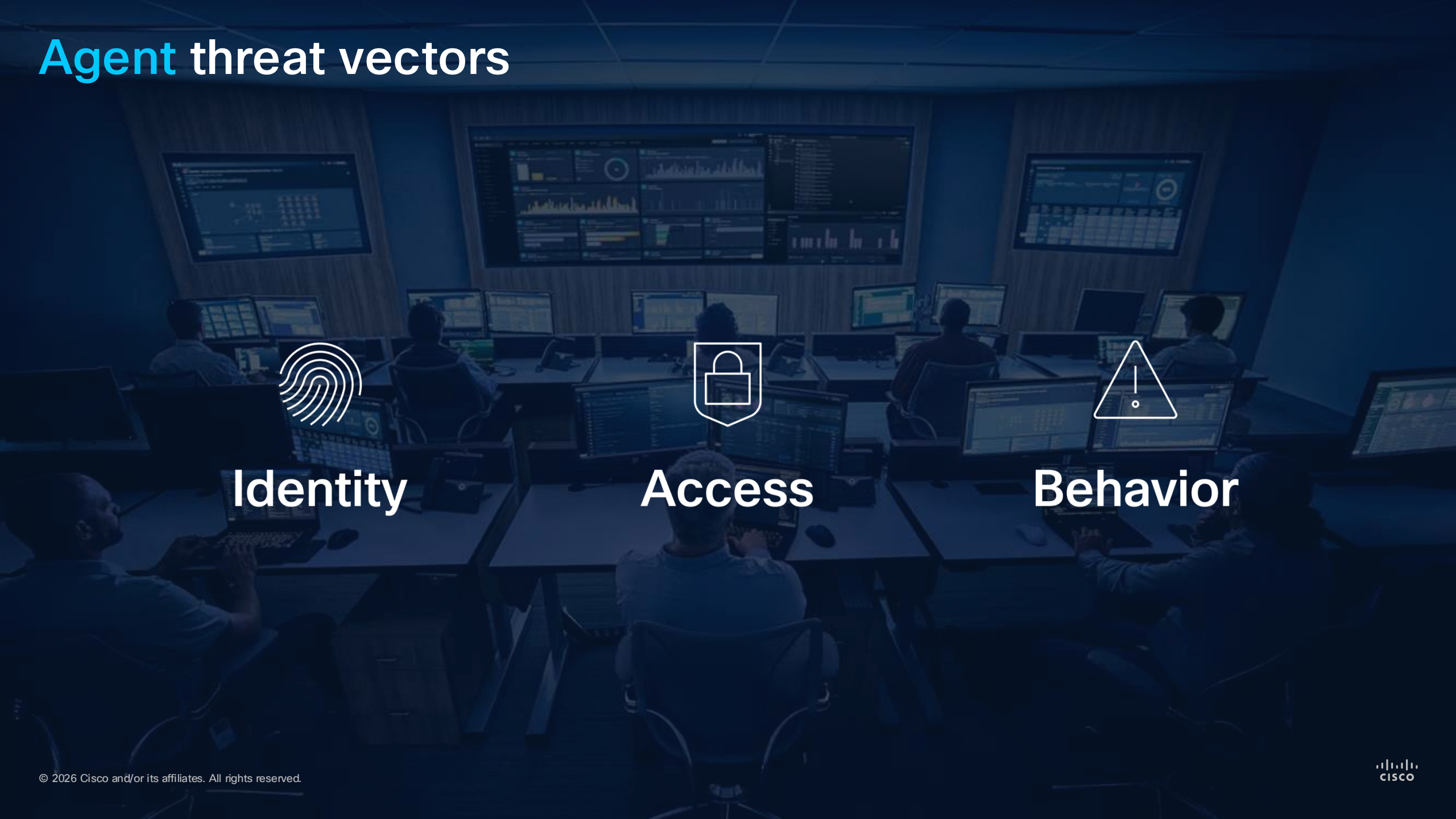
Safety

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

Agent threat vectors



Identity

Access

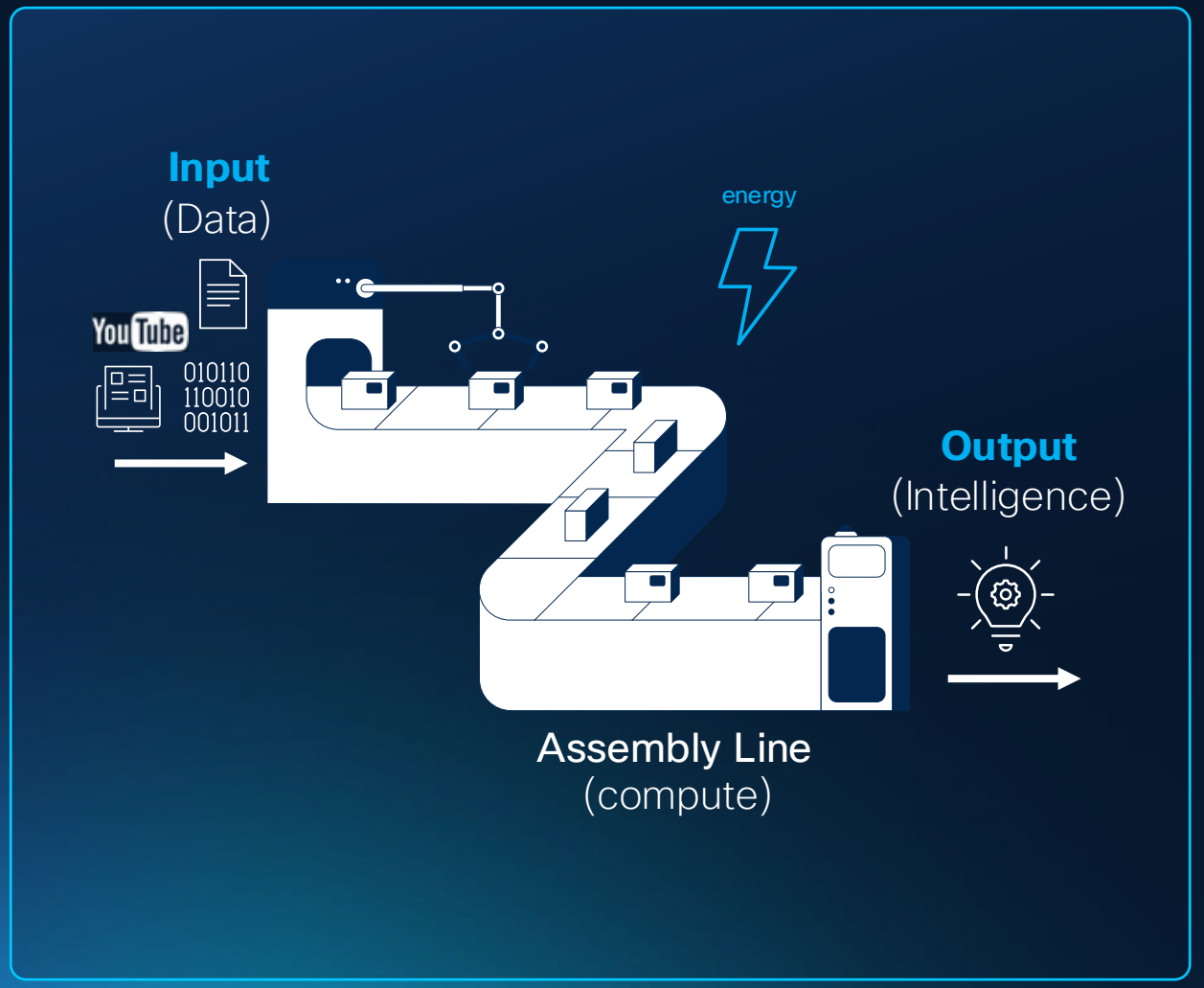
Behavior

We’re addressing all of these challenges head-on
Cisco is the critical infrastructure for the AI era

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.



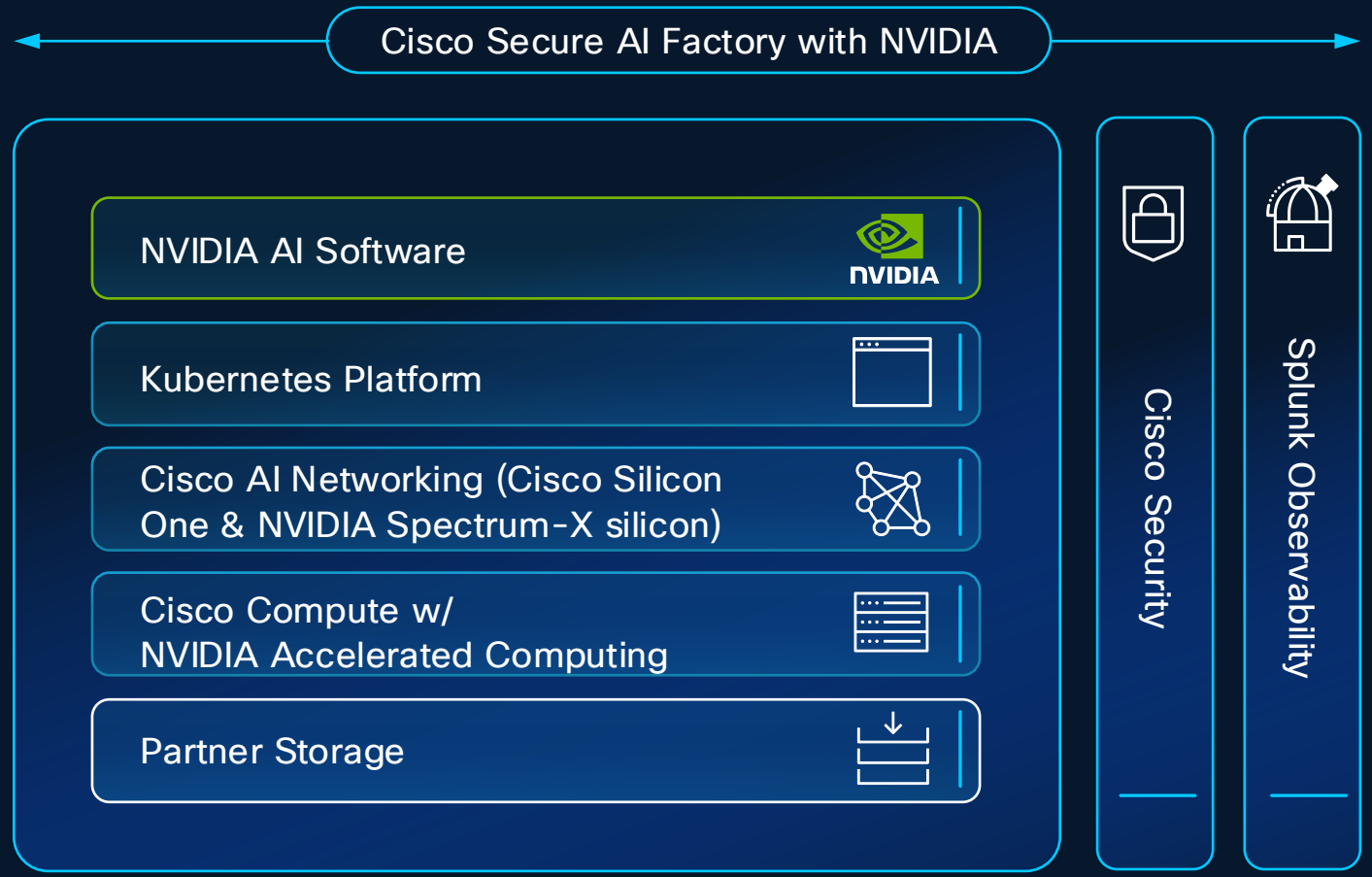
What is needed to accelerate trusted AI outcomes?

☐ AI factory

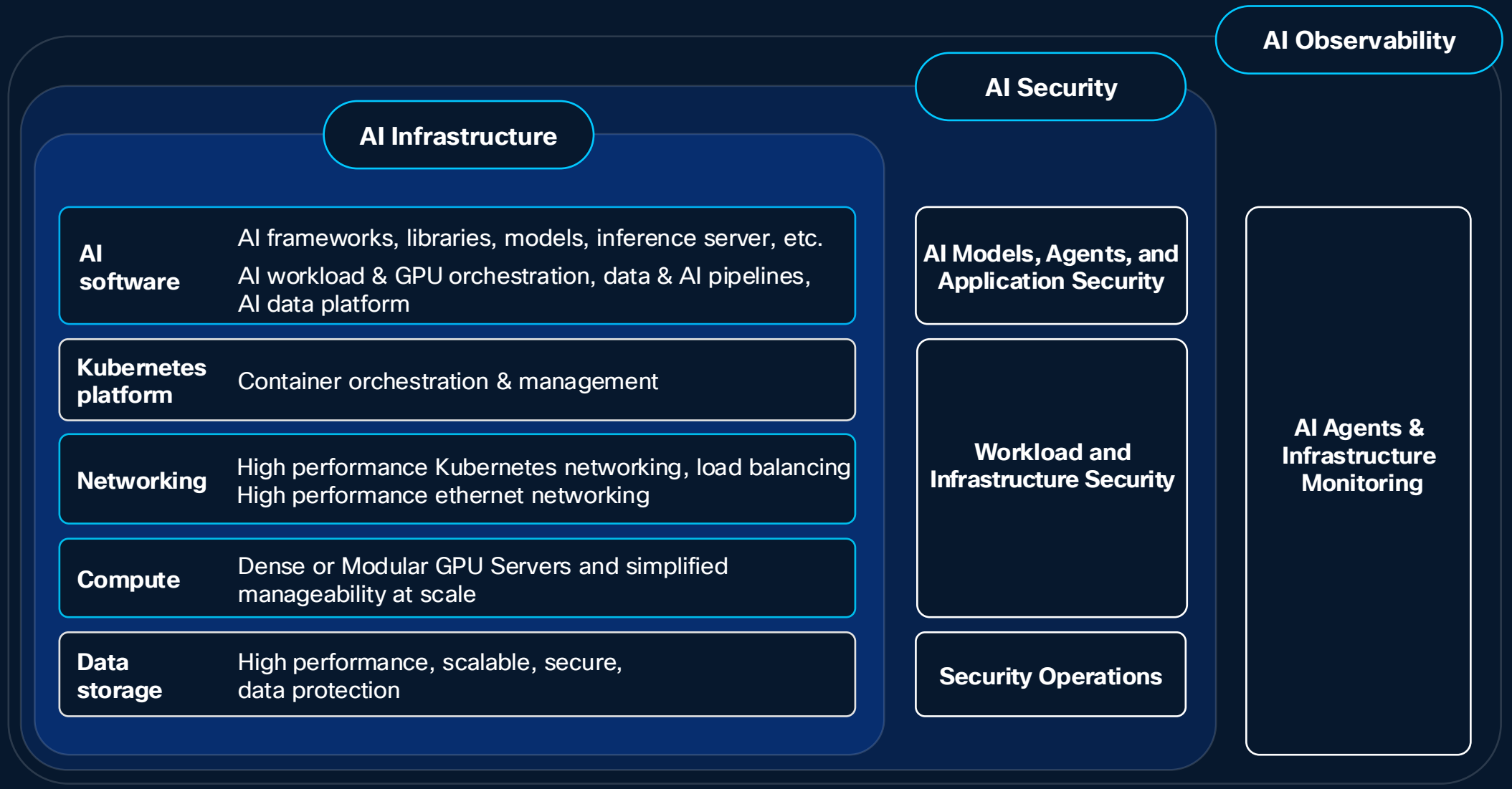
☑ **Secure** AI factory

Cisco & NVIDIA: Accelerating AI Adoption

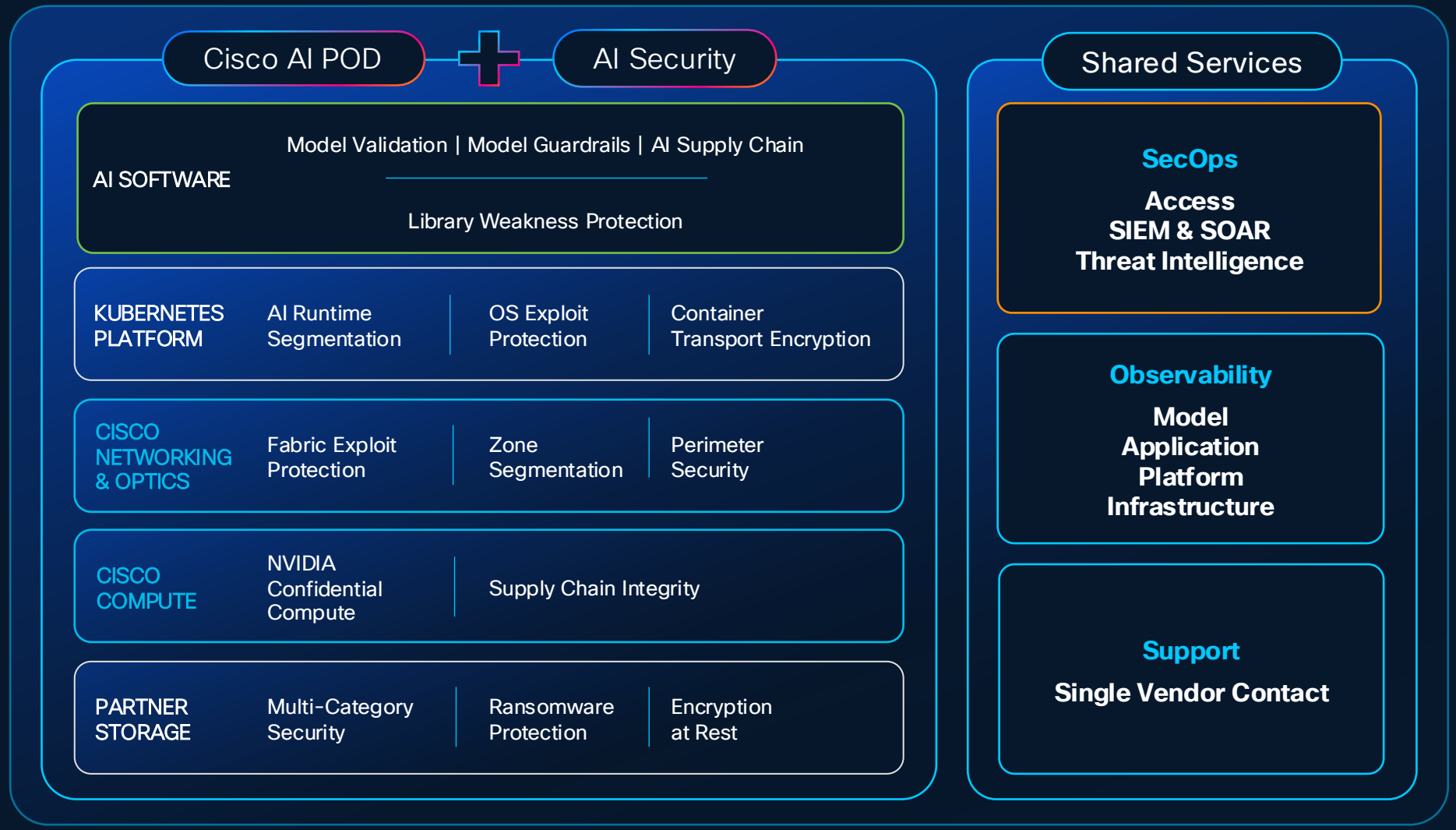
- A modular reference design that combines high-performance infrastructure with full-stack security and observability
- Extend NVIDIA's Spectrum-X architecture to include Cisco Silicon and Cisco Optics



Key capabilities of Cisco Secure AI Factory with NVIDIA



Key Security Capabilities At Every Layer



AI Model Security

Core AI Defense Capabilities within a Secure AI Factory with NVIDIA

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

Detection: AI Model & Application Validation

Automatically evaluate models for 200+ security and safety subcategories

45+ Prompt Injection Attack Techniques

- Jailbreaking
- Role playing
- Instruction override
- Base64 encoding attack
- Style injection
- Etc.

30+ Data Privacy Categories

- PII
- PHI
- PCI
- Branded content
- Privacy infringement
- Etc.

20+ Information Security Categories

- Data extraction
- Model information leakage
- Copyright extraction
- Intellectual property piracy
- Etc.

50+ Safety Categories

- Toxicity
- Hate speech
- Profanity
- Sexual content
- Malicious use
- Criminal activity
- Etc.

Guardrail Categories

Security

- Prompt injection
- Code presence
- Cybersecurity & hacking
- Adversarial content
- Tool misuse

Privacy

- Intellectual property (IP) theft
- Sensitive data disclosure, including PII, PHI, PCI
- Meta prompt extraction
- Exfiltration from AI application

Safety

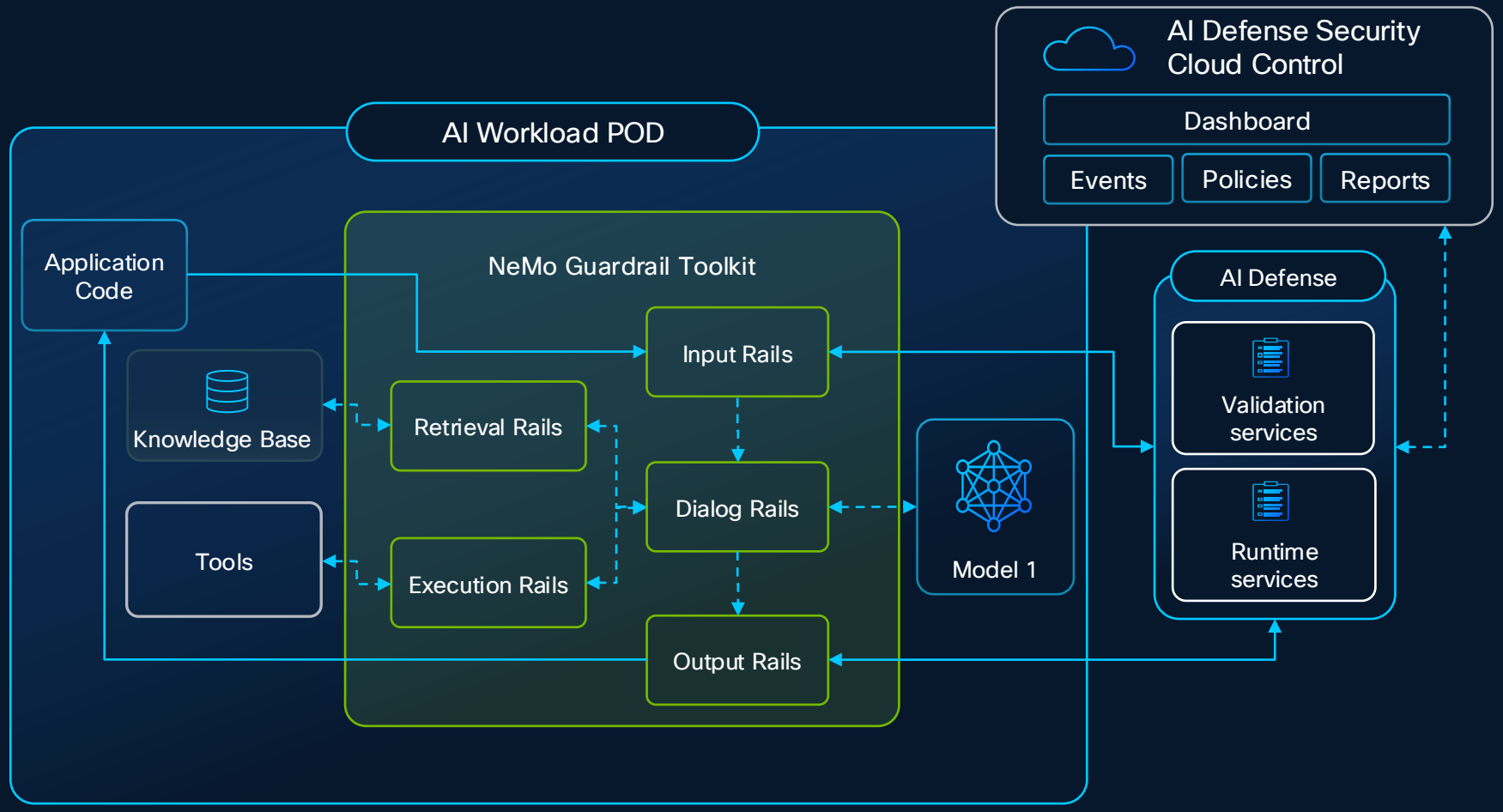
- Hate speech & profanity
- Sexual content
- Harassment
- Violence & public safety threats
- Rogue agents

Guardrails map directly to AI security standards from OWASP, NIST & MITRE

Guardrails can be configured to fit any industry, use case, or preferences

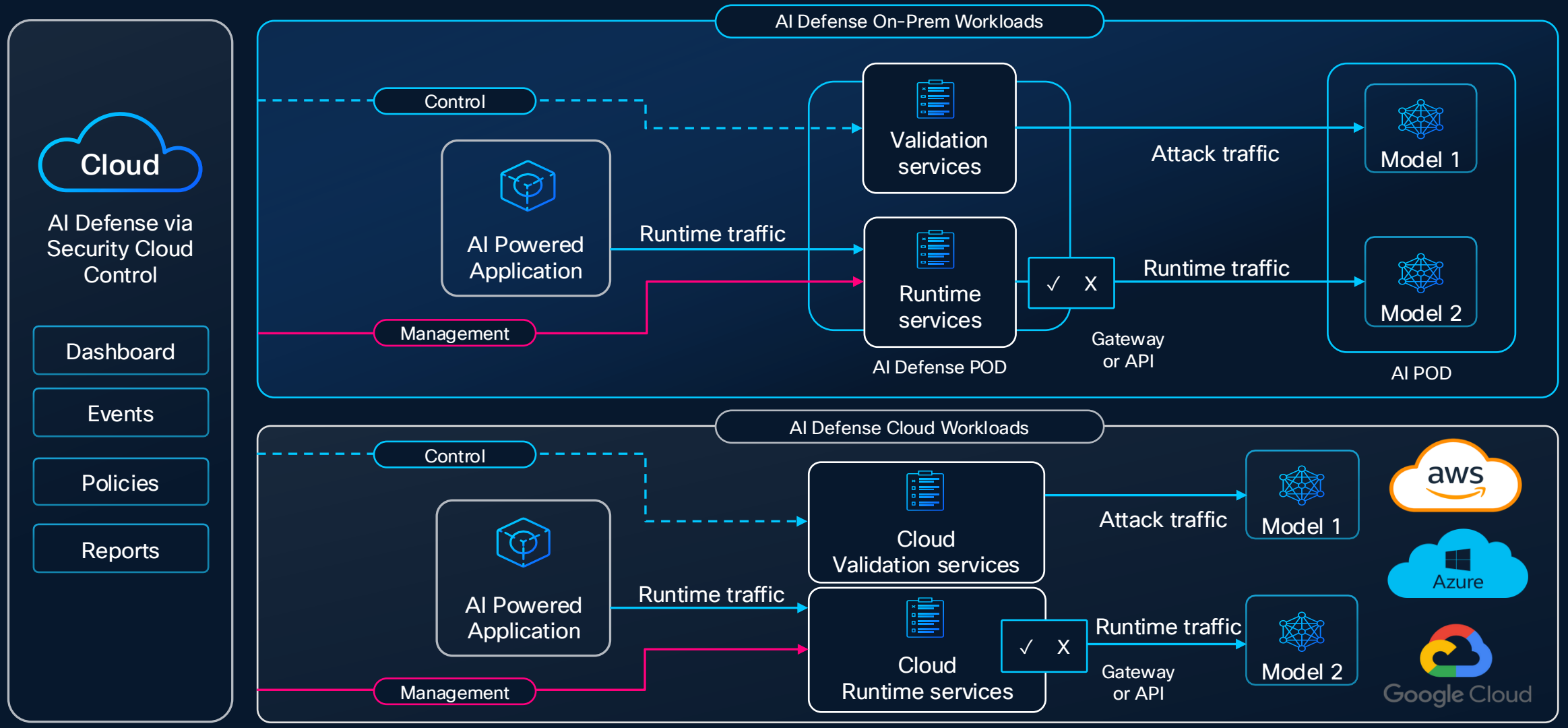
AI Defense & Nvidia – NeMo Guardrails Integration

- AI Defense provides input & output guardrails via API disposition
- Common guardrail policy for on-premises and cloud deployments.
- Supports additional guardrail types included in the NeMo Guardrail Toolkit
- Nvidia documentation link



Cisco AI Defense

Cloud Managed – Hybrid Enforcement

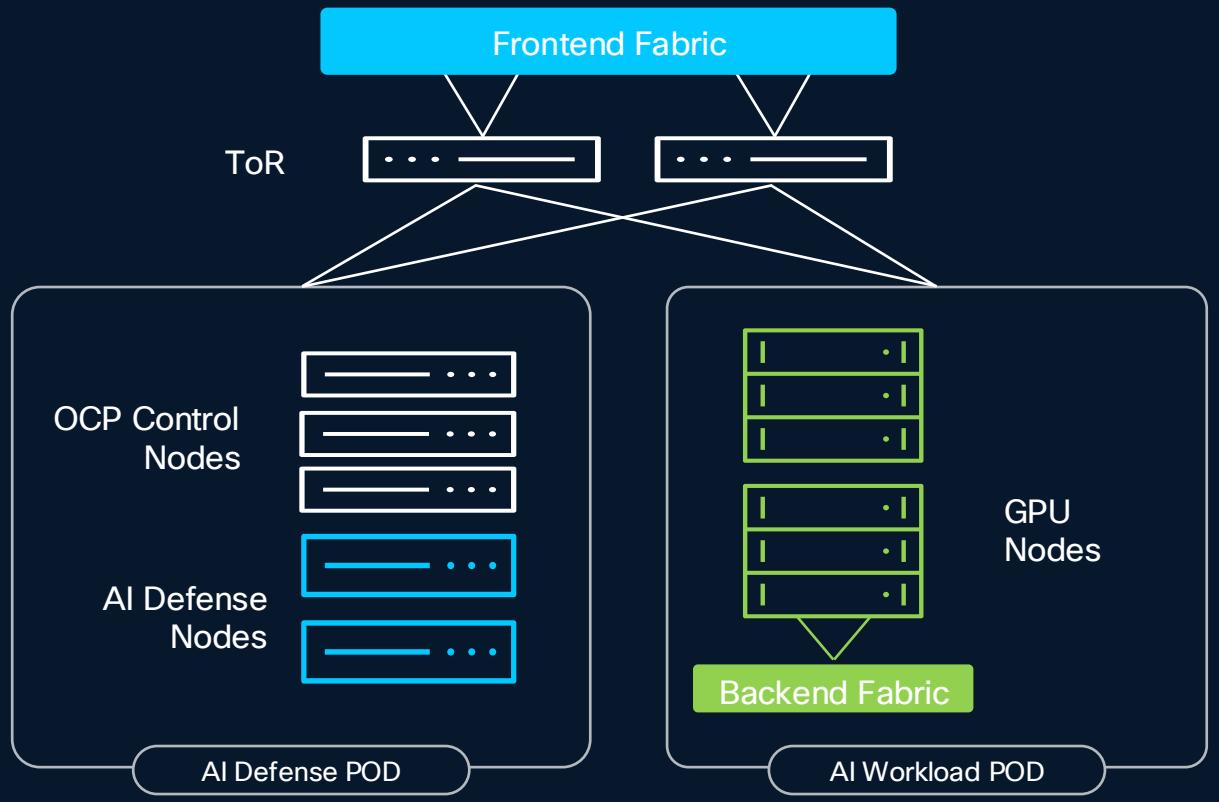


AI Defense for On-Prem Workloads

Supports Validation and Runtime Protection Capabilities

Supported AI Defense Node Configurations			
Size	Small	Medium	Large
Hardware Model	UCS C845A	UCS C845A	UCS C845A
Hardware Quantity	2	2	3
GPUs Included	4 L40S per C845A	8 L40S per C845A	8 L40S per C845A
Networking Supported	1/10Gb, 25/50 Gb 100/200 Gb	1/10Gb, 25/50 Gb 100/200 Gb	1/10Gb, 25/50 Gb 100/200 Gb
Load Supported	100 Req/s 20 Apps	200 Req/s 40 Apps	300 Req/s 60 Apps

AI Defense POD Reference Architecture



Segmentation for AI

Cisco Hybrid Mesh Firewall

Define policy once, enforce everywhere

Cisco Firewalls

Physical | Virtual | Cloud | FWaaS

3rd Party Firewalls

NVIDIA Bluefield DPU Firewall on AI Servers

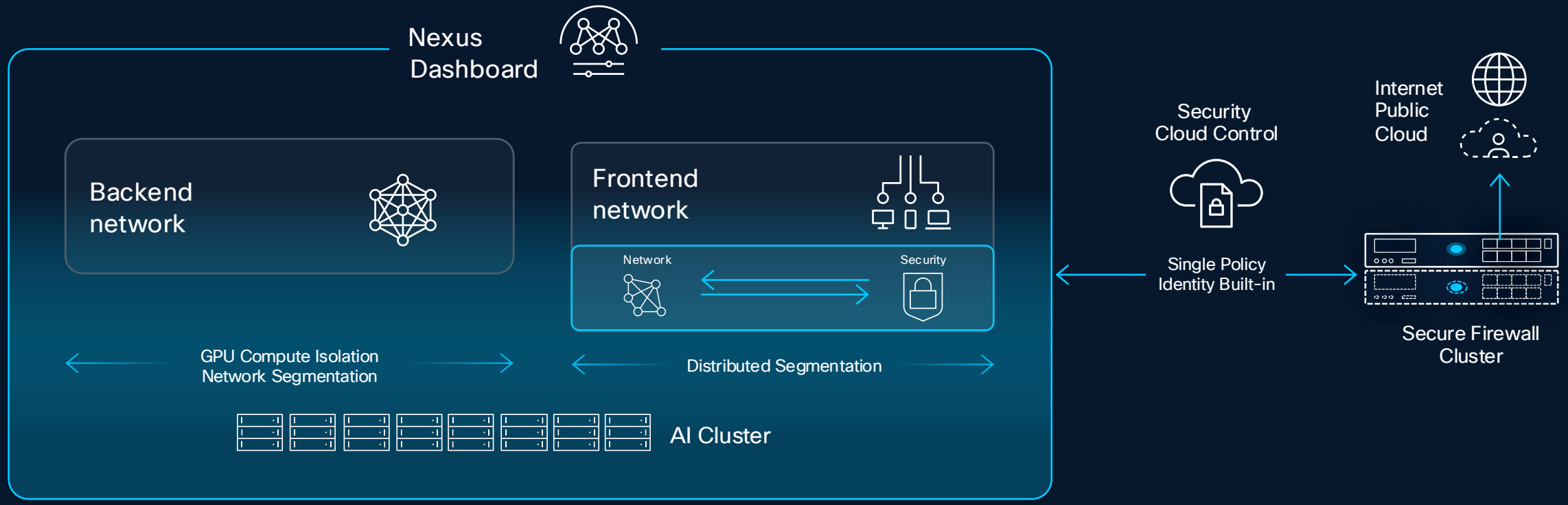
Smart Switches

eBPF

Workload Agents

Perimeter Security for an AI Factory

with Cisco's Hybrid Mesh Firewall

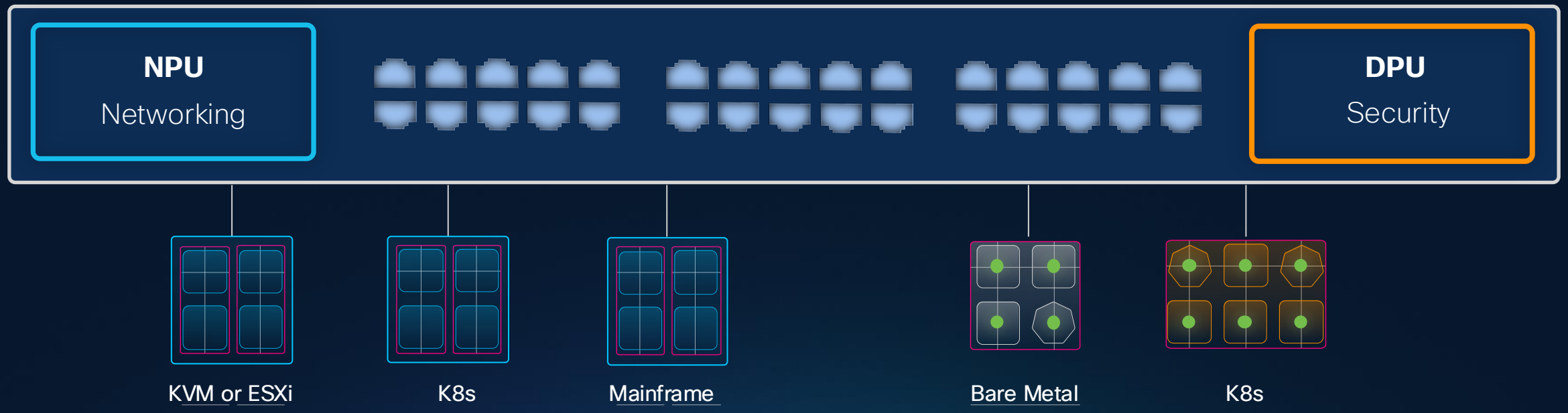


Write policy once, enforce across the mesh

Cisco's Hybrid Mesh Firewall solution allows for the creation of advanced firewall capabilities implemented at the perimeter network (e.g L7 AppID, IDS/IPS, URL Filtering, SSL Decryption) along with L3 & L4 policies at frontend top of rack with Smart Switches.

Segmentation with Smart Switches

Turn every switch port into an enforcement point



Nexus Networking

VLAN / VRF / VXLAN
Network context

Hypershield Security

800Gbps of stateful performance
Instant protection for new workloads

Nexus Smart Switch

Unmatched Flexibility, Performance, and Efficiency

Cisco Nexus 9300
Series Smart Switches

Networking

cisco
Silicon One

Services

DPU

Rich NX-OS Features and Services

- Routing Switching
- EVPN/MPLS/ VXLAN/SR
- Rich Telemetry
- Line-rate Encryption
- Power Efficiency

Software-defined Stateful Services

- Distributed Security

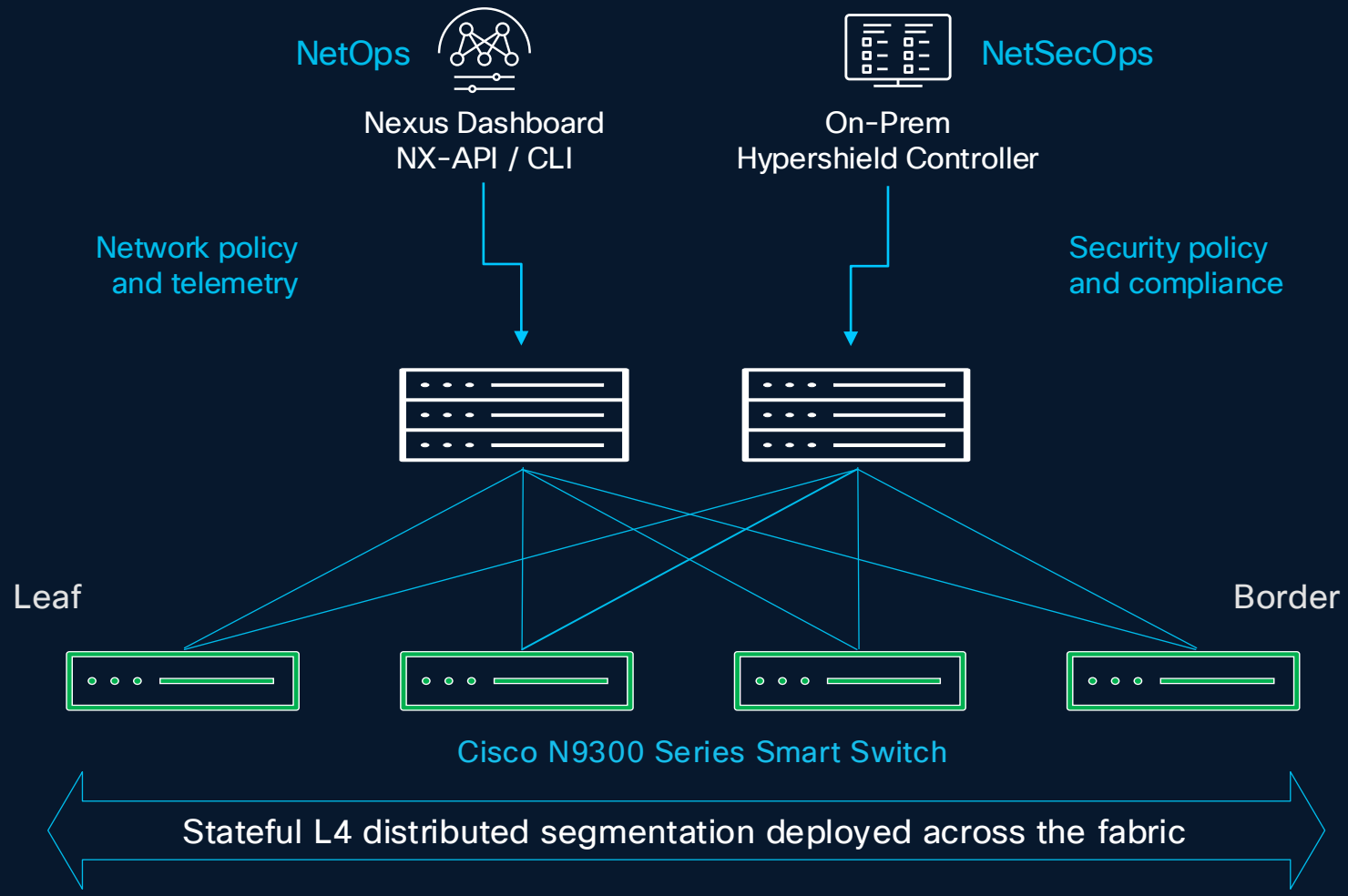
Future Use Cases

- Load Balancer
- Large-Scale NAT
- Event-Based Telemetry
- DoS Protection

cisco

Smart Switch Use Case

Top of Rack L4 Segmentation

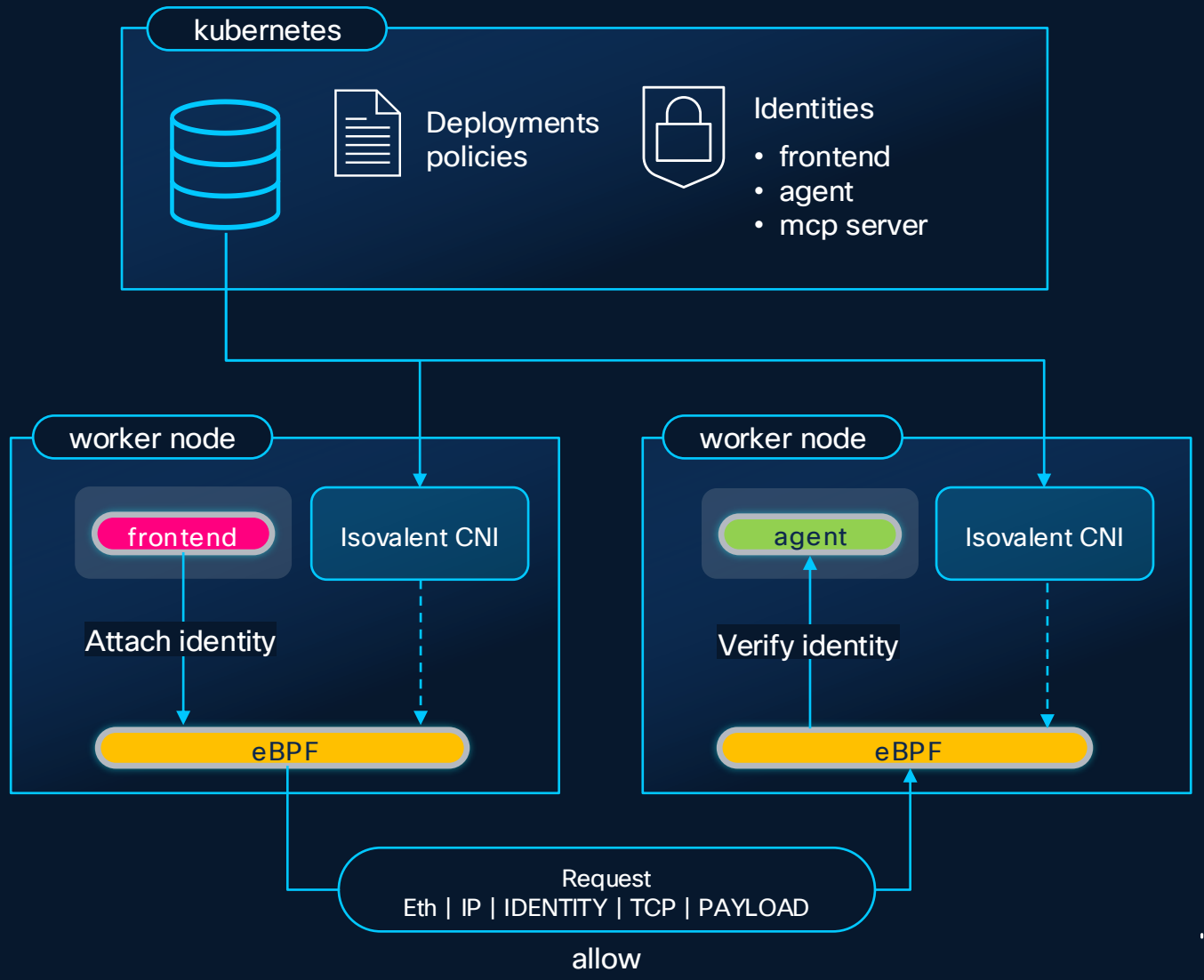


Identity-based Workload Security

Network filtering

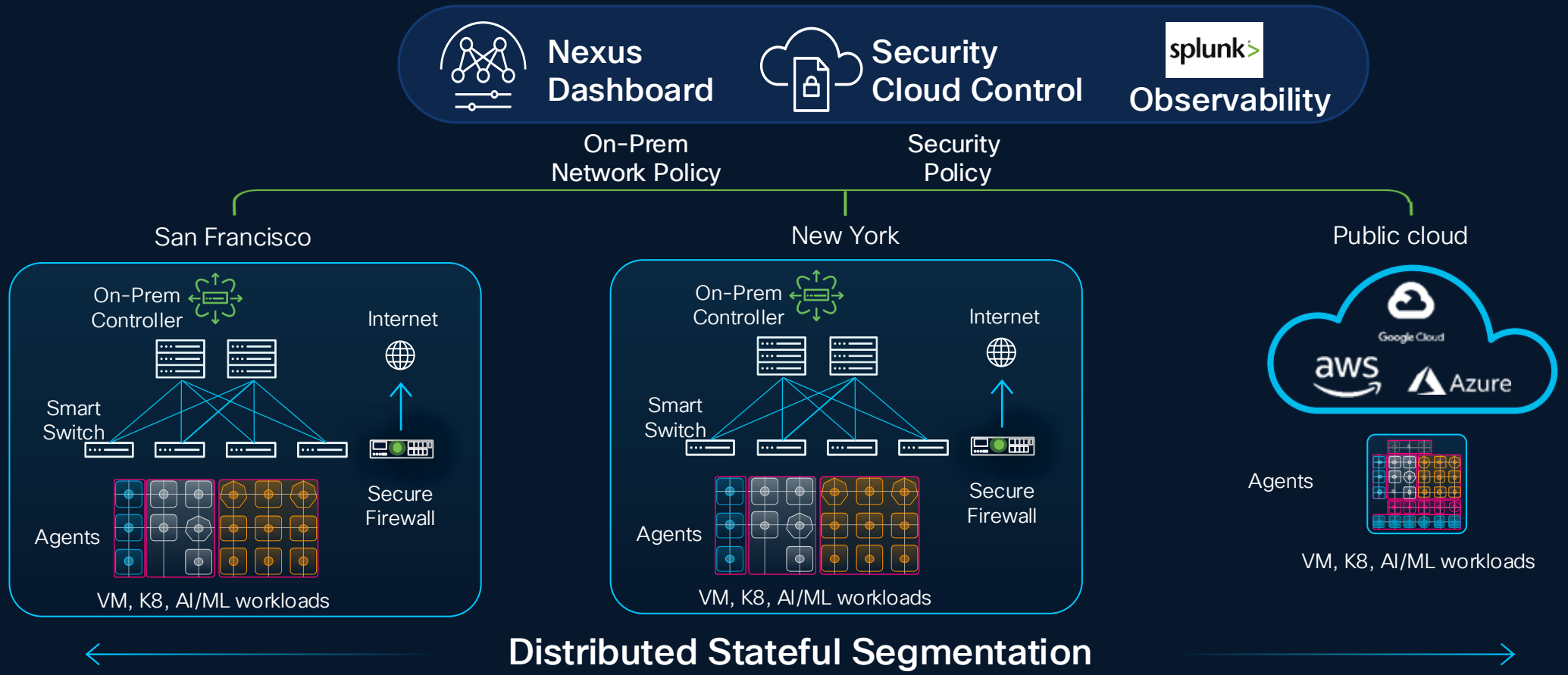
- Kubernetes networking
- Load balancing
- Kubernetes services
- Identity-based security
- L7 policies

```
apiVersion: "cilium.io/v2"
kind: CiliumNetworkPolicy
metadata:
 name: "agent-rule"
spec:
 endpointSelector:
  matchLabels:
   role: agent
 ingress:
 - fromEndpoints:
  - matchLabels:
    role: frontend
```



Security infused in Data Center and Cloud

Unified Policy across Smart Switch, Workload & Perimeter

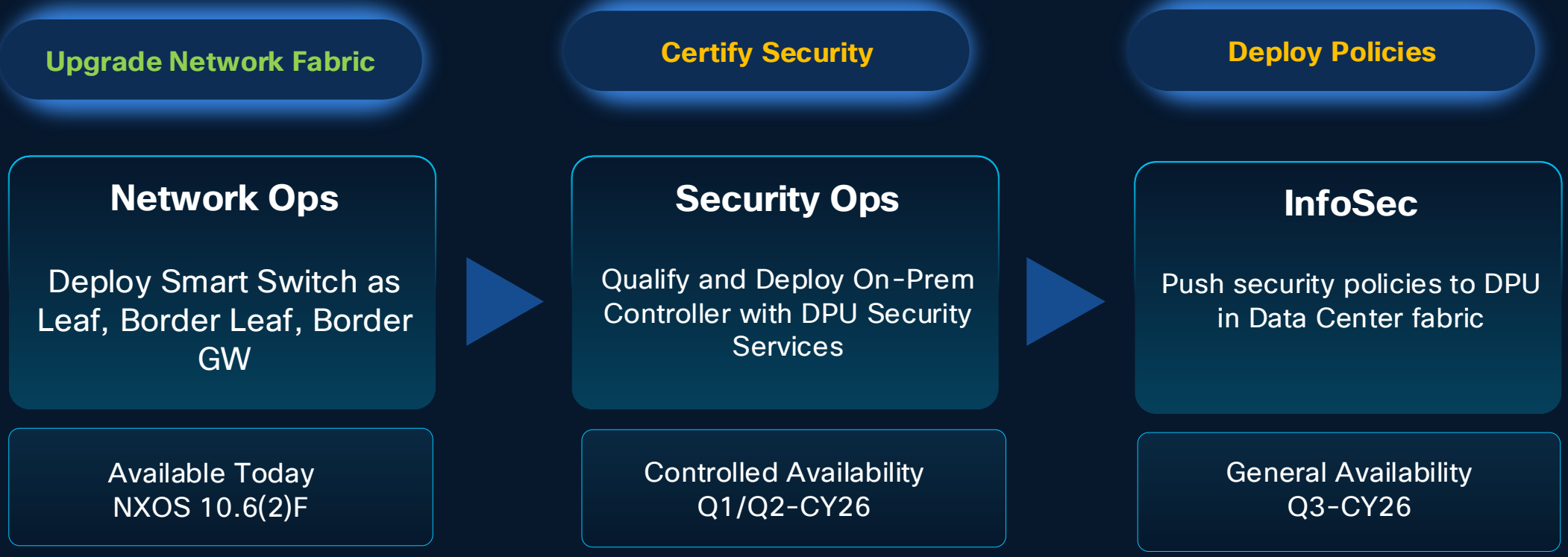


- Global or air-gapped management
- Unified Policy Enforcement
  Smart switch, workload, perimeter
- 800G security with large scale
- Always On high availability
- Full visibility from data center to cloud

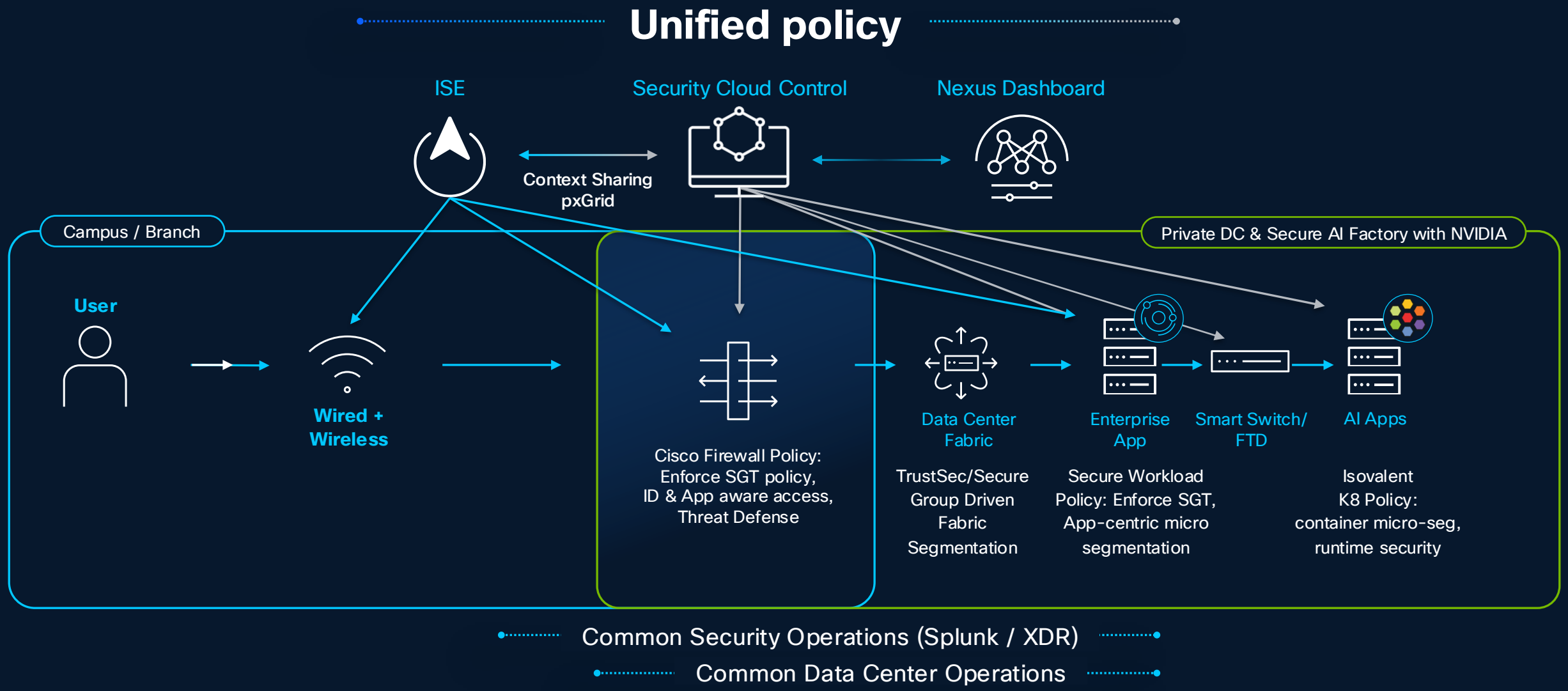
Smart Switch & Hypershield Roadmap

Shipping	Committed (Q3 CY26)	Future
Smart Switch Network Mode VXLAN-EVPN Fabric Multi-Site VXLAN-EVPN Fabric BGP routed Fabric Classic LAN	**Distributed Segmentation with Hypershield** Top of Rack Segmentation Zone Segmentation – ACI Fabric AI Frontend Fabric	**Nextgen (High density 100G) Smart Switch** 48p 100G, 8p 400G Smart Switch 1.6T+ DPU Services SFP-DD / DSFP ports
NXOS 10.6(2)F Nexus Dashboard 4.2	NXOS 10.7(1)F Nexus Dashboard 4.3 Hypershield 1.0	Future

Customer Rollout Journey



Secure AI Factory with NVIDIA's Place In A Zero Trust Architecture



Thank you


CISCO