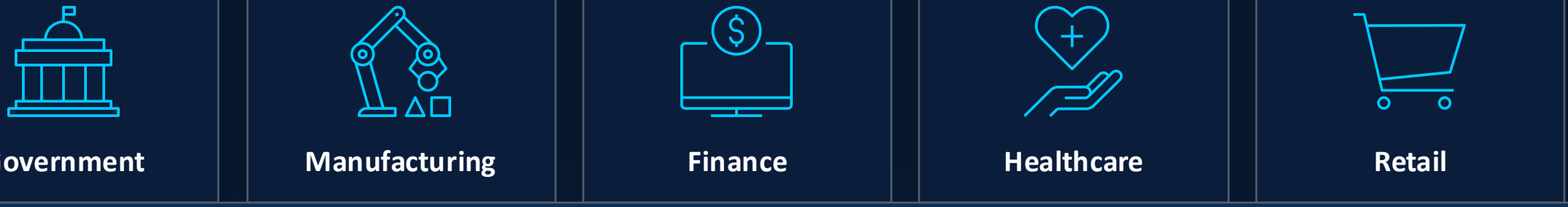
Umělá inteligence

Nejde jen o infrastrukturu v DC

Cisco Tech Club

Artificial intelligence outcomes span every industry



Knowledgebase copilots

AI assistants

Content & code generation

Text | Images | Video | Code

Reporting & data analytics

Summarize texts | Generate visualization

Language translation

Multilingual real-time communication

Virtual agent & chatbots

Specialized domain specific chatbots

Detection & prediction

Forecasts | Anomalies | Insights

Build the Model | Train

Optimize the Model | Fine-tune & RAG

Use the Model | Do Inferencing

Common AI Challenges

Unclear business objectives & priorities

Unclear direction hinders cross team collaboration, creates confusion, and hampers acquisition of necessary skills

Security vulnerabilities

AI models, frameworks, apps, and supporting infrastructure represent a new cyberattack surface

Performance bottlenecks

Model training and inferencing generates a lot of traffic, slow networks and delays time-to-value

Complex AI infrastructure deployment

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

Cisco AI PODs

A scalable architecture, built to support any AI workload simply & efficiently

Deploy AI with confidence

Cisco CVD, NVIDIA ERA

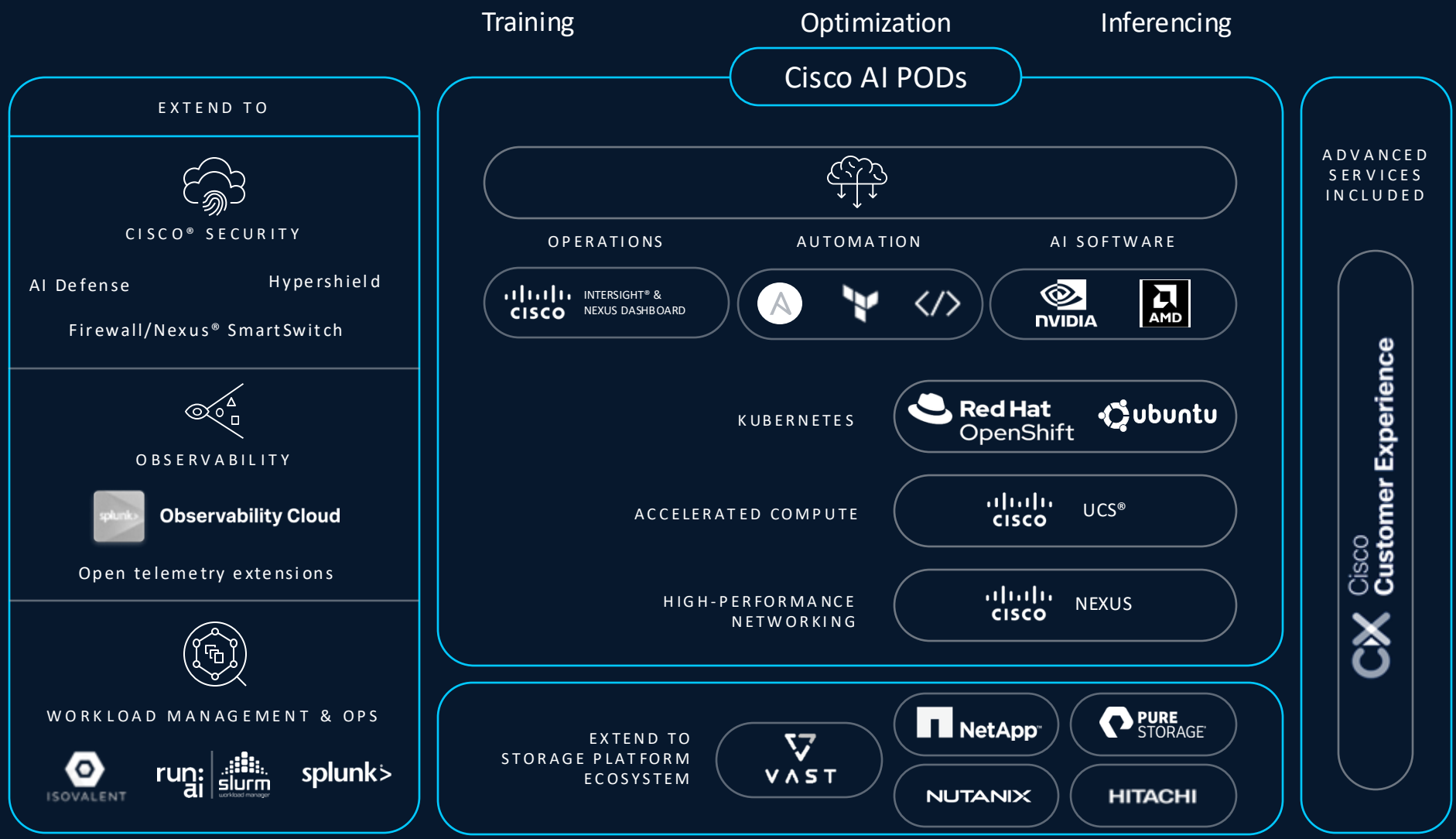
Fully supported stack including Cisco and 3rd party components

Cisco CX Success Track

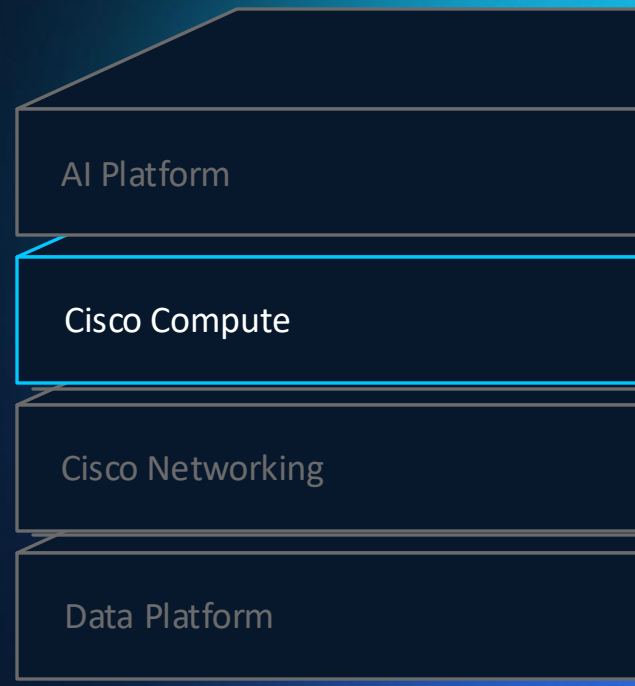
Orderable, use case driven AI-Ready infrastructure stacks

Inferencing. Optimization. Training.

Incremental, atomic-level – or- fabric-based cluster scale



AI Compute for Cisco AI PODs



Compute AI portfolio

Address AI workloads with visibility, consistency, and control

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

Build the model
Training

Optimize the model
Fine-tuning and RAG

Use the model
Inferencing

RTX PRO SERVER
Supporting RTX PRO 6000 Blackwell Server Edition GPUs

Cisco UCS®
GPU-dense servers
PCIe and NVLink Servers

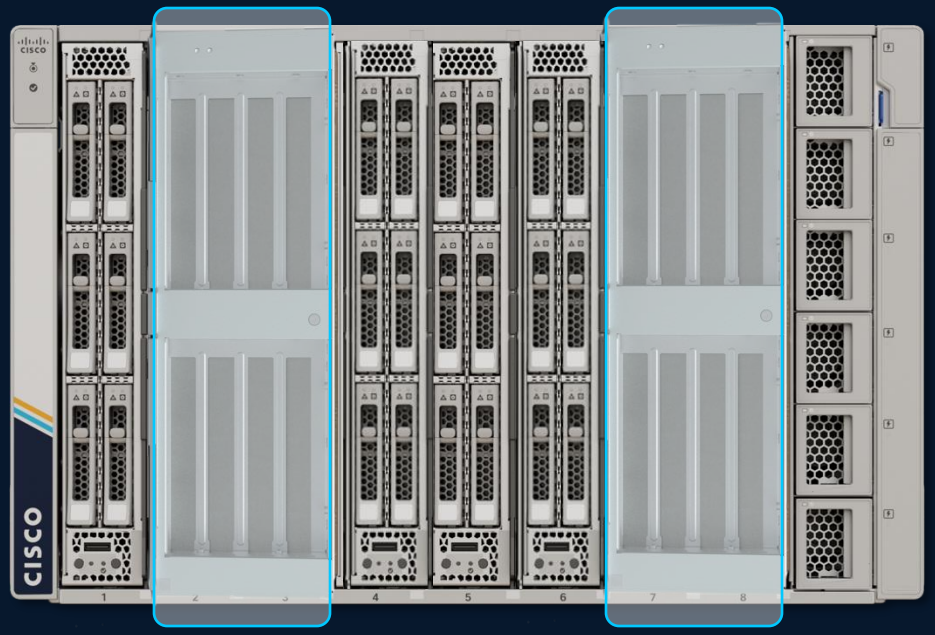
Cisco UCS blade (with GPU extensions) and
rack servers

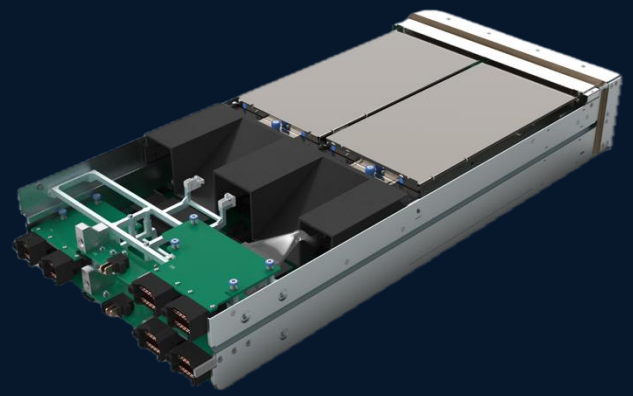
Enterprise AI edge

Dense compute for demanding AI

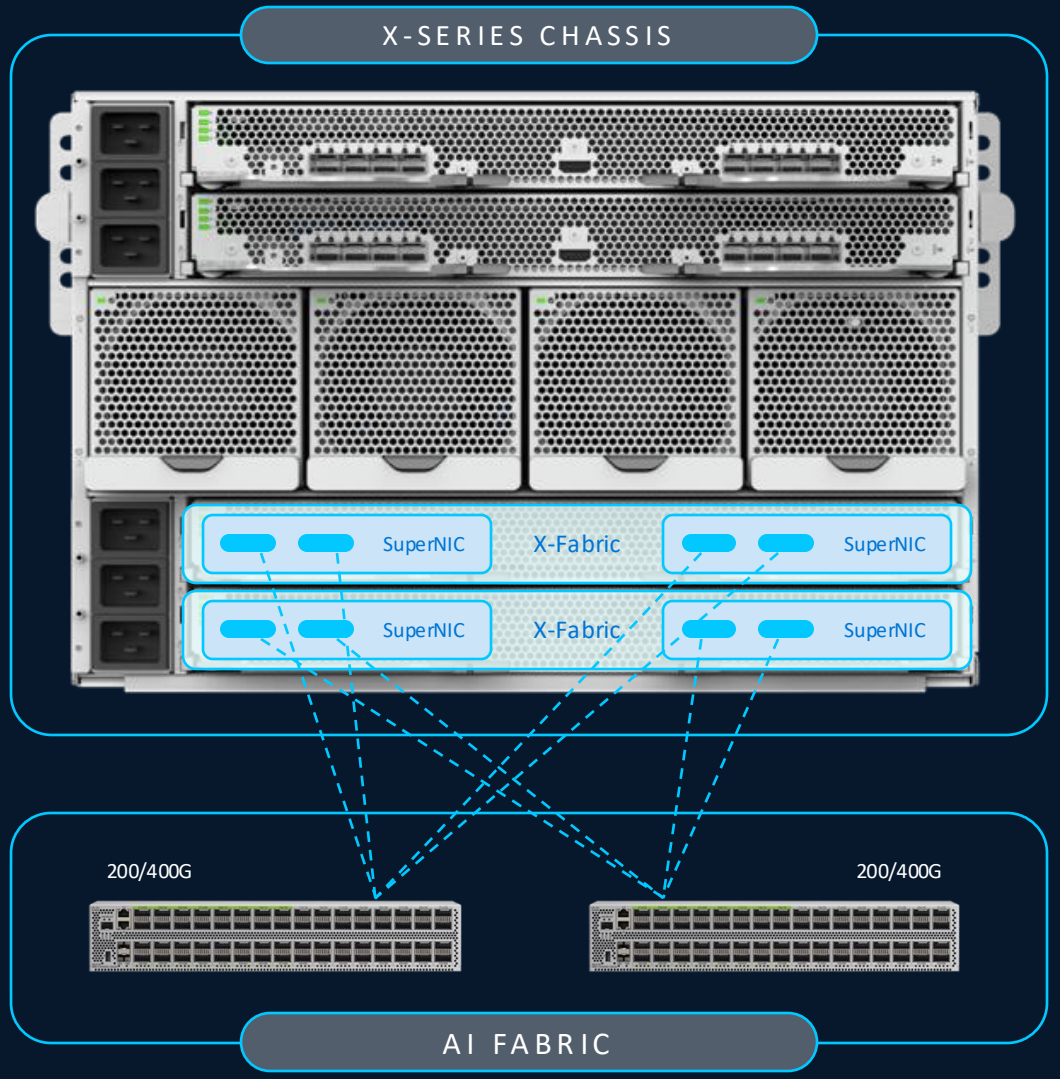
Full-stack AI with compute and networking

UCS X – build for AI as well

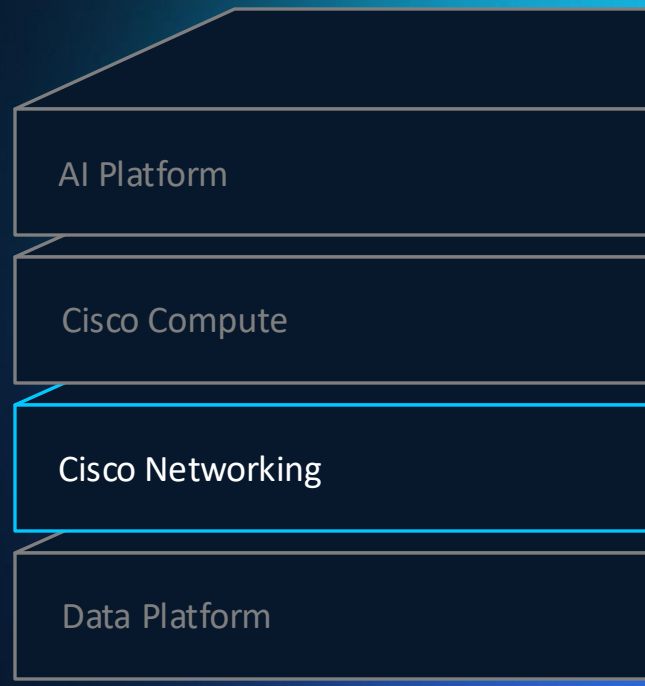




- 4x FHFL GPU and PCIe G5 GPU support (Nvidia, AMD, Intel)
- NVLink bridge support
- Support up to 600W FHFL GPU
- Policy based GPU management
- Ability to share GPUs across two Compute nodes
- GPU Direct Support over RDMA
- GPU Backend(East-West Traffic) network support



Cisco AI PODs Network Fabric Options

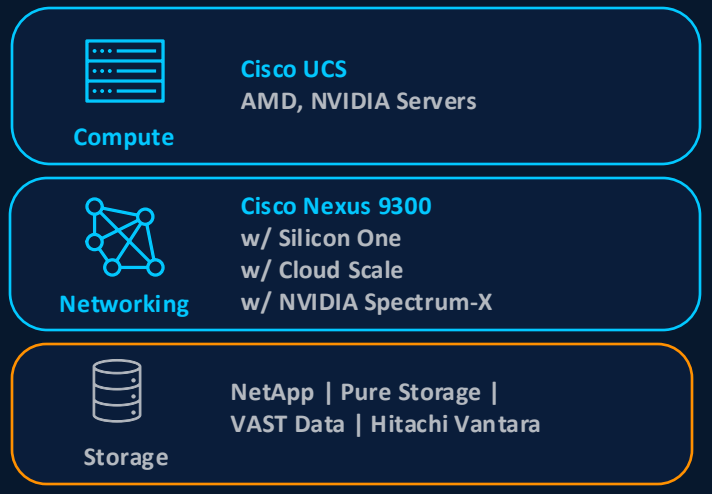


Cisco AI PODs Solution Options

Choose Based on AI Use-Case and Operational Model

Nexus Dashboard

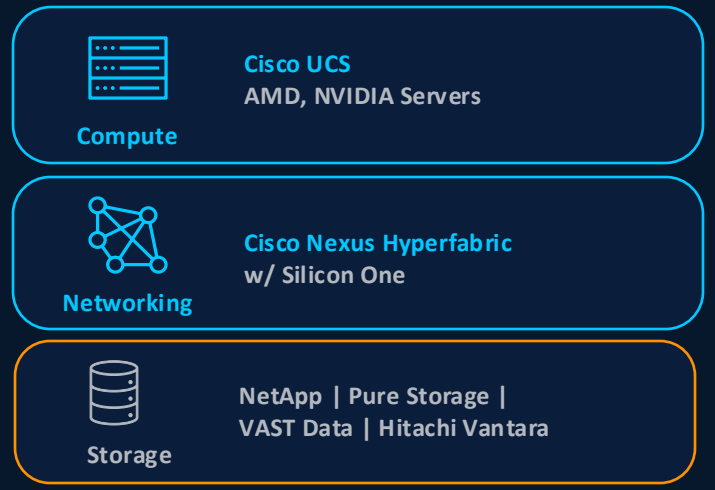
On-premises managed network



Nexus 9300 managed with Nexus Dashboard
NVIDIA Enterprise Reference Architecture
Flexible design consistent performance at any scale
Dedicated AI or multi-purpose DC

Nexus Hyperfabric

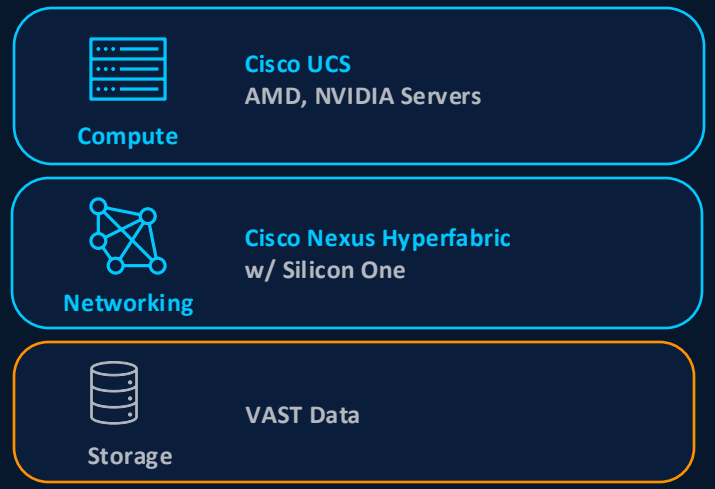
Cisco-managed network cloud controller



Cloud-managed fabric-as-a-service
Deploy anywhere, easy to scale
BYO servers, GPUs, storage, software stack
Multi-purpose DC with AI & non-AI services

Nexus Hyperfabric AI

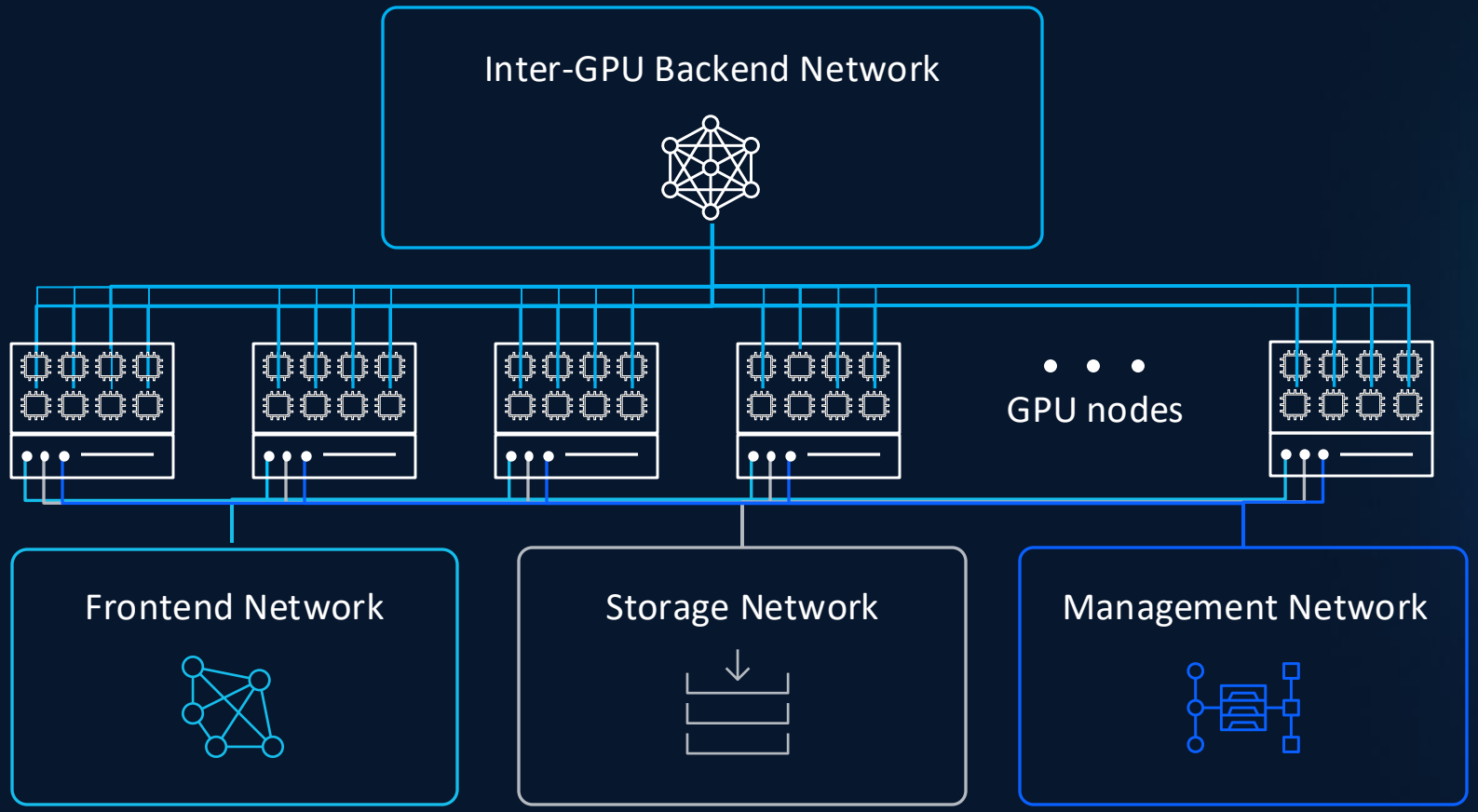
Cisco-managed network cloud controller



Full stack: network, servers, GPU, storage
NVIDIA Enterprise Reference Architecture compliant
Blueprints encompass the entire cluster
Monitoring spans network to compute NIC

Customizable ⟷ *Prescriptive*

AI Networking



Cisco Nexus high density 800G & 400G fixed switches

Cisco AI PODs embedding the latest Nexus Switches

Nexus 9300 64-port 800G Switch

512-wide radix	Fully shared packet buffer	Advanced load balancing	Low Latency

N9364E-SG2-Q or N9364E-SG2-O

Compact 2RU 51.2T Switch

G200 ASIC (5nm) | 100G SerDes | 256MB packet buffer

64 800G ports | Up to 128 line-rate 400G ports (2x400G breakout)

Choice of QSFP-DD800 or OSFP ports

Cisco NXOS spine and AI/ML spine/leaf capable

Ultra Ethernet Consortium MEMBER

Nexus 9332D-GX2B
32p 400G
8p MACsec/CloudSec

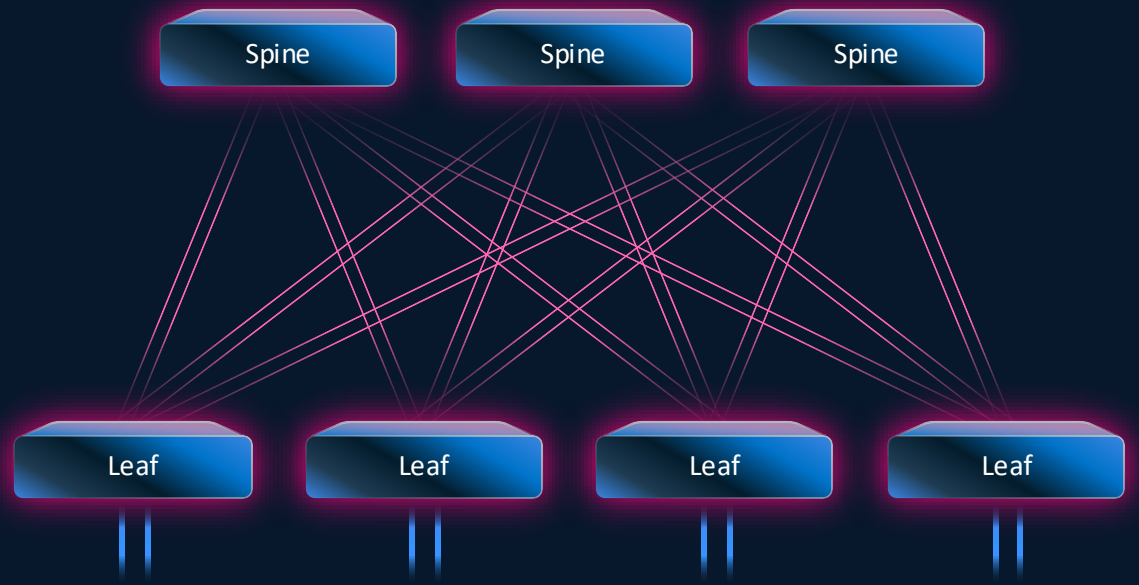
Nexus 9364D-GX2A
64p 400G
16p MACsec/CloudSec

ACI Leaf, ACI Spine, and NX-OS
25.6T, 19.2T, and 12.8T 400G switches
120MB smart buffer

Security
MACsec and CloudSec

Telemetry
FT, FTE, SSX, INT-XD

Powering AI Fabrics with Cisco Intelligent Packet Flow



Ultra Ethernet READY

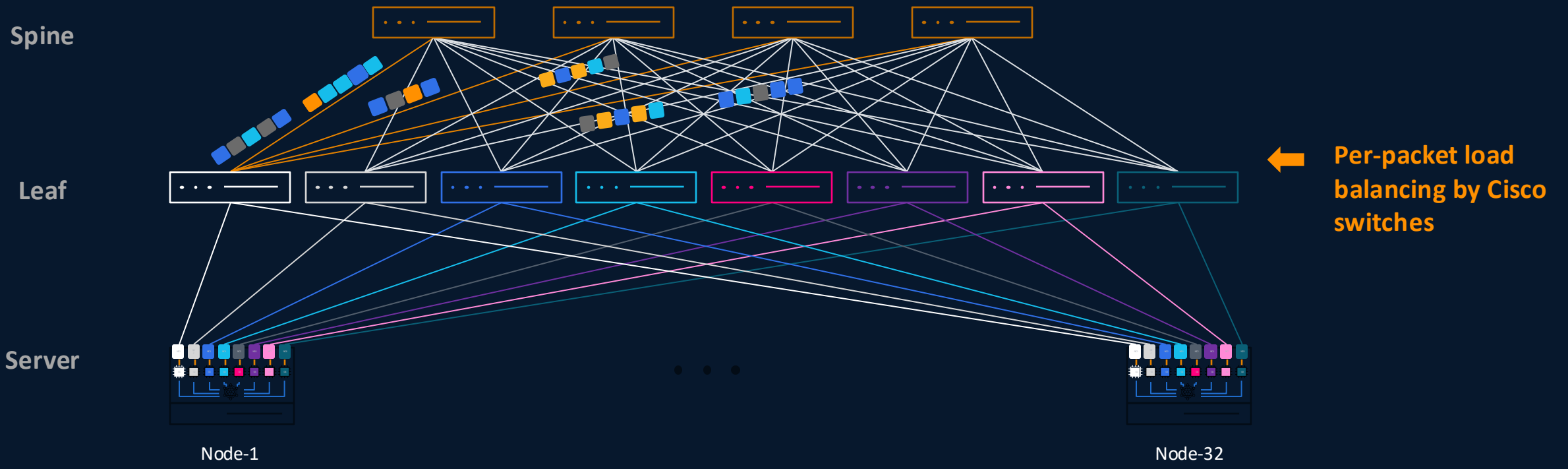
Advanced load balancing

FEATURES

Dynamic Load Balancing (DLB) with Load and Congestion Awareness	Per Packet and Selective Packet Spray (Ex: RDMA vs. non-RDMA) + NVIDIA Spectrum-X integration*	ECMP Static Pinning
Weighted Cost Multi Path Load Balancing (h-WCMP) with Dynamic Load Balancing	Policy Based Flowlet Load Balancing (DSCP, ACL…) + RoCEv2 Header Filter*	High Entropy ECMP

*Introduced with NX-OS 10.6.1F

Per-Packet Load Balancing



Cisco Per-packet load balancing

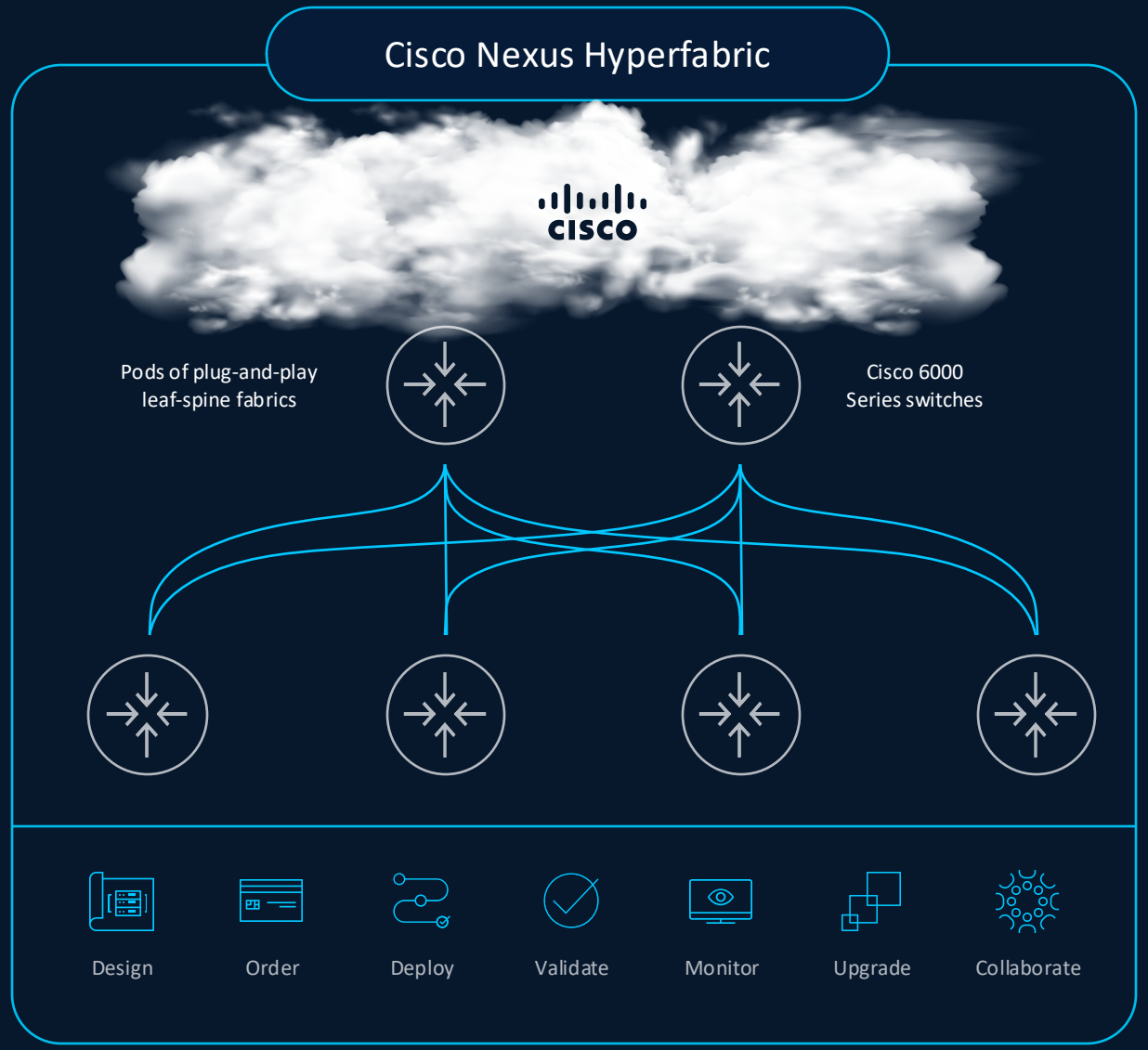
- Traffic is forwarded on a **per packet basis**, every packet will be hashed on different output port
- Expected that receiver will put packet in order (**server NICs do the packet reordering**)

Cisco Nexus Hyperfabric

Cloud Managed fabric option for Cisco AI PODs

AVAILBLE NOW

- Design, deploy and operate on-premises fabrics located anywhere
- Easy enough for IT generalists, application and DevOps teams
- Outcome driven by a purpose-built vertical stack



Cisco AI PODs

A scalable architecture, built to support any AI workload simply & efficiently

Deploy AI with confidence

Cisco CVD, NVIDIA ERA

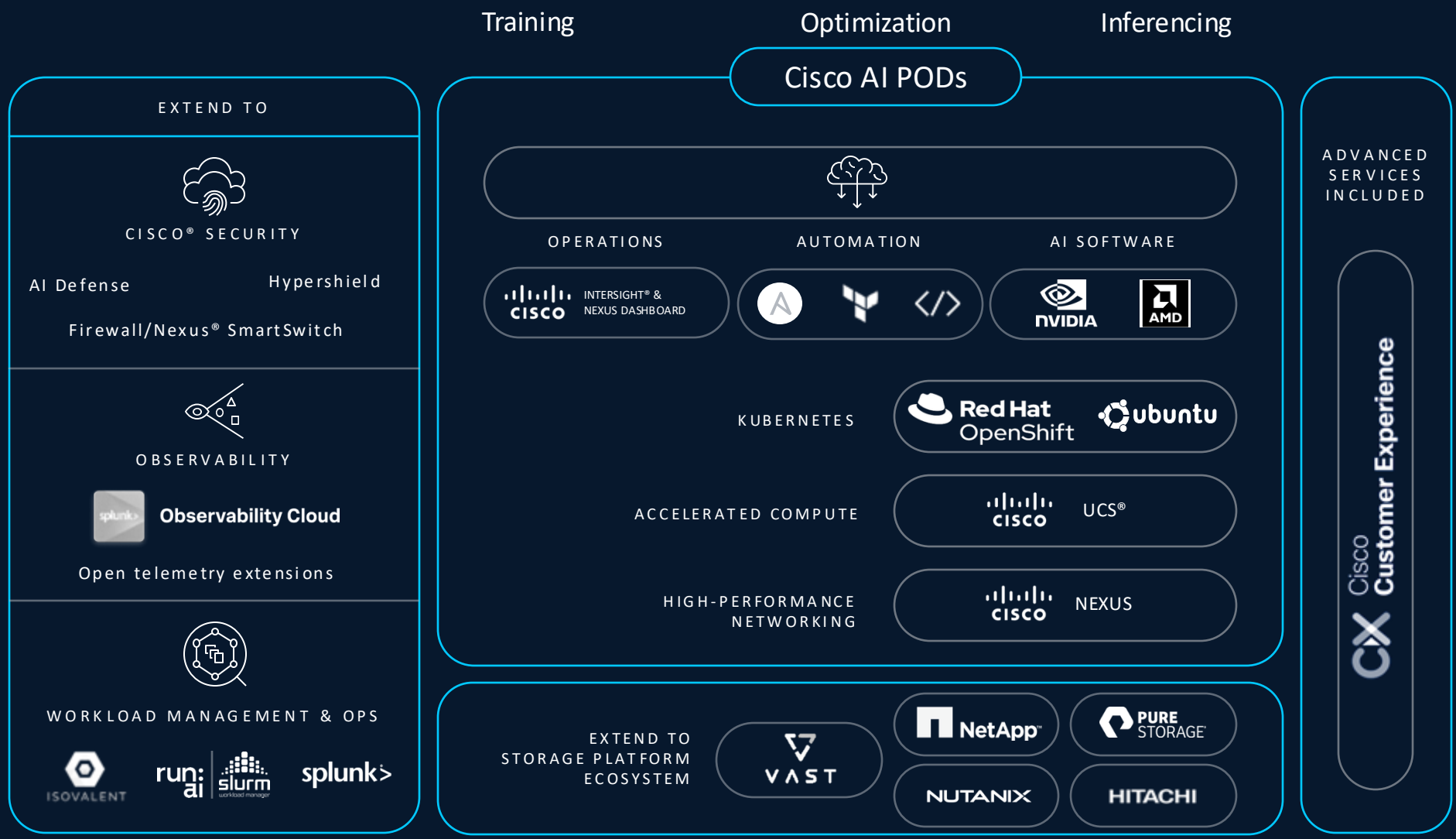
Fully supported stack including Cisco and 3rd party components

Cisco CX Success Track

Orderable, use case driven AI-Ready infrastructure stacks

Inferencing. Optimization. Training.

Incremental, atomic-level – or- fabric-based cluster scale



Security in Cisco AI PODs

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
Securing the Workloads	Cisco Hypershield—Protection against adversary lateral movement and proactive vulnerability mitigation without the need for patching, all from a single management interface, integrated with NVIDIA AI.
Securing the 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 smooth policy definition and enforcement across software defined networks.

What’s the risk?

AI applications are complex and non-deterministic

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.

AI Runtime Protection

Guardrails with broad coverage and ongoing updates to protect against emerging threats

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

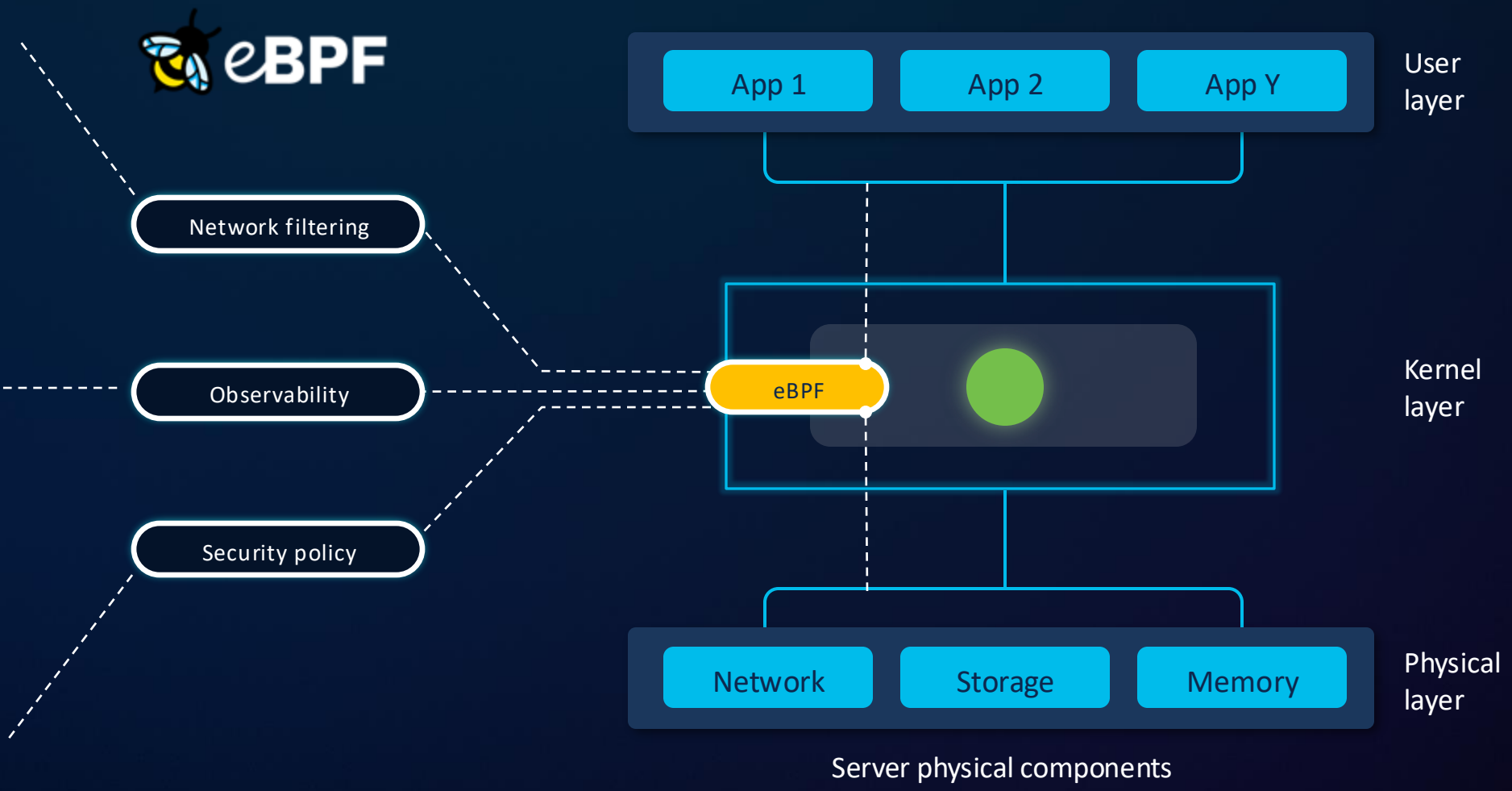
Cisco Hypershield

eBPF – Foundation of Hypershield

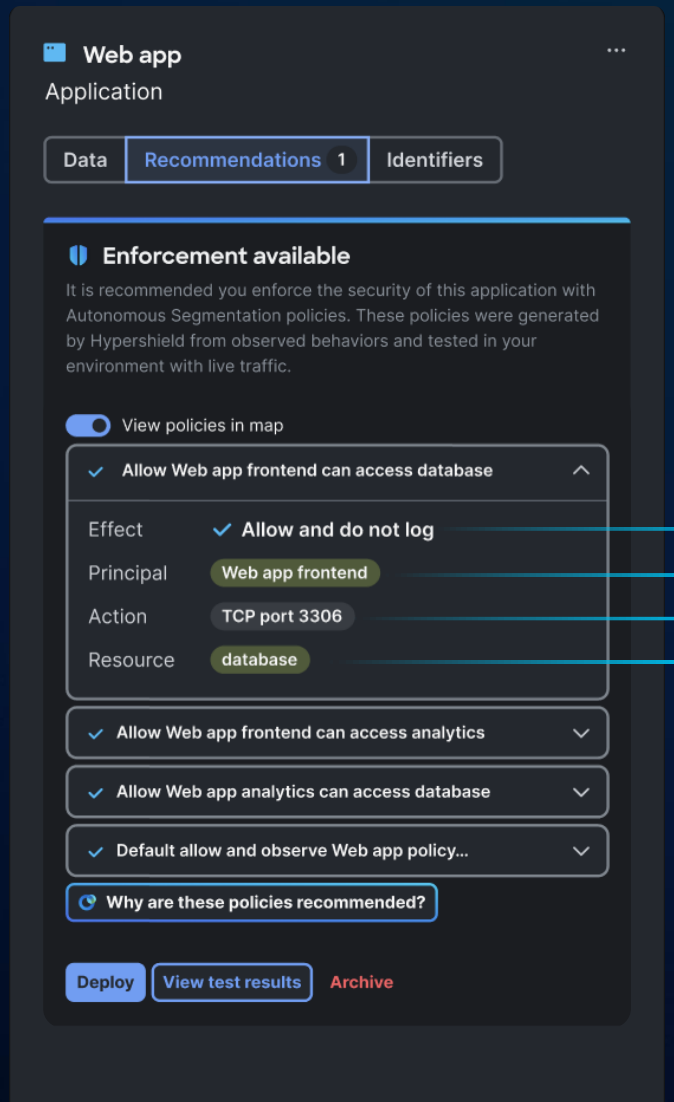
- Kubernetes networking
- Load balancing
- Kubernetes services
- Identity-based security
- L7 policies

- Dependencies map (service and flows)
- Monitoring and alerting
- App monitoring

- Monitor process execution
- Runtime security policies
- Real-time enforcement



Hypershield AI recommends policy

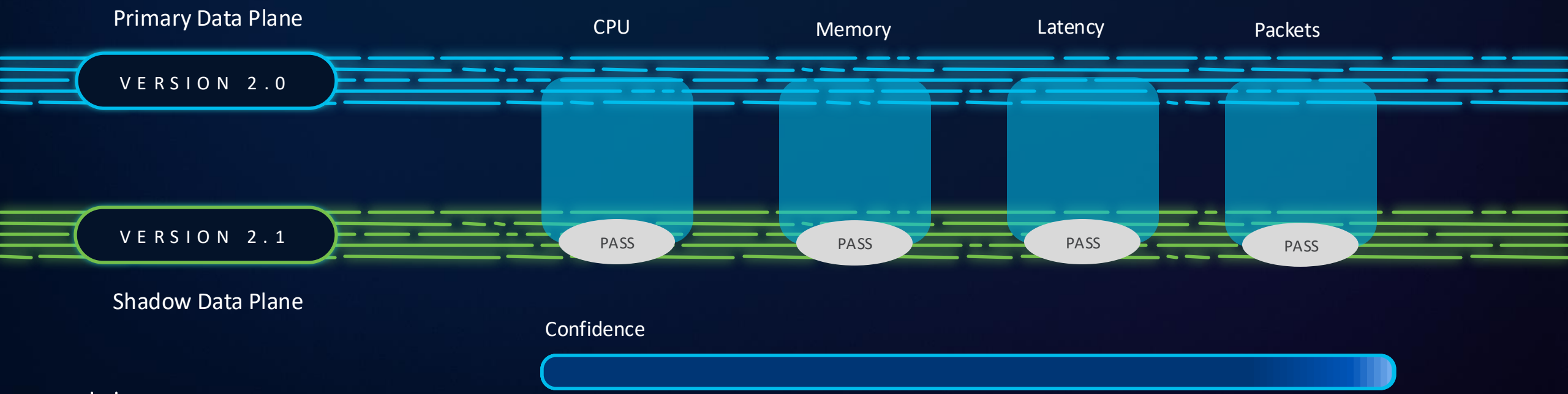


AI-generated rules to enforce app security (flexible vs restrictive)

Effect	Allow (flexible policy)
Principals	‘Web app frontend’ object
Actions	Communication port
Resources	‘Database’ object

Hypershield Digital Twin

Self-qualifying updates



Nexus Smart Switch

Unmatched Flexibility, Performance, and Efficiency

Cisco
Smart Switches

Cisco Nexus 9300 Services Accelerated Switch

Networking

Cisco Silicon One

- 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

Hypershield

AMD PENSANDO

- Software-defined Stateful Services
- Programmable at all layers: add new services without HW change
- Scale-out services with wire-rate performance
- Power down DPU complex when not used

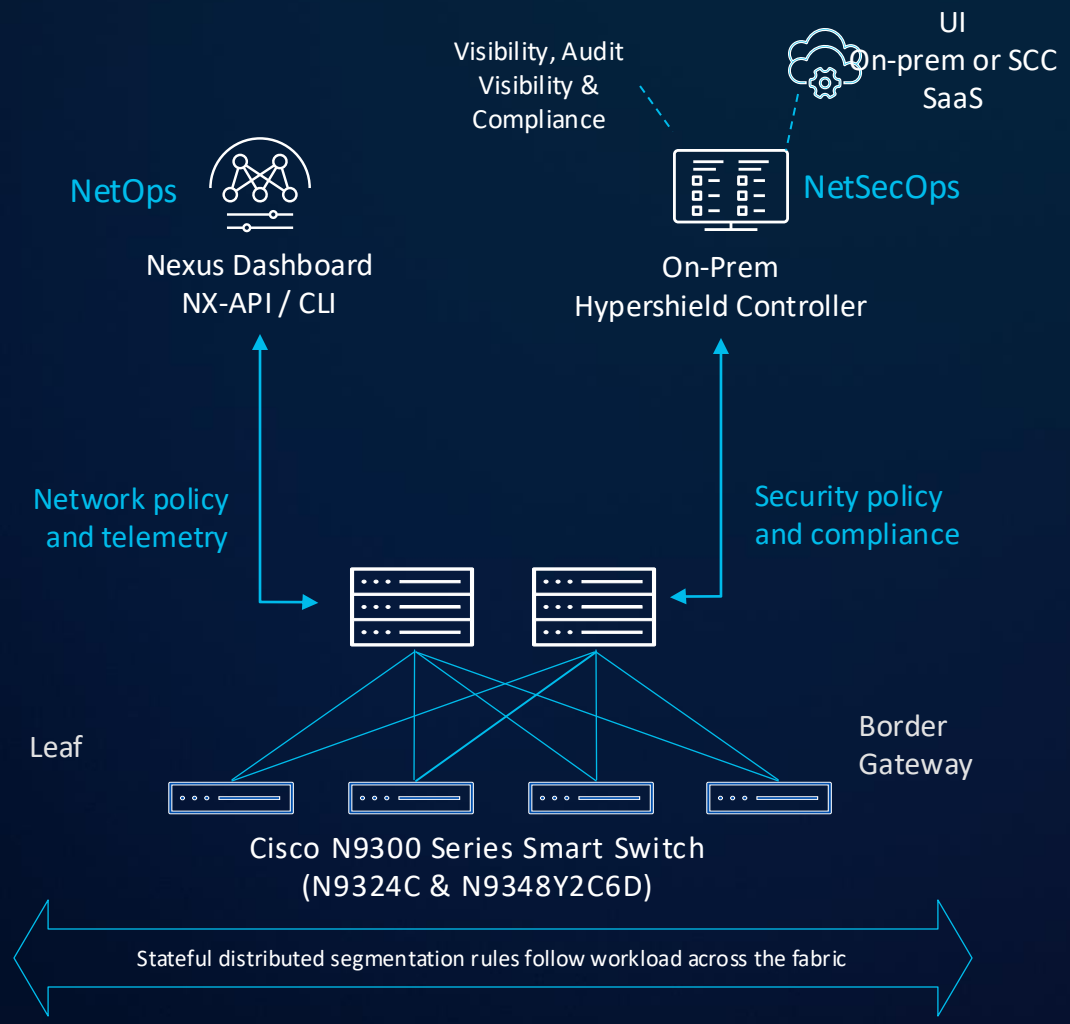
Distributed Security

IPSEC Encryption | Large-Scale NAT | Event-Based Telemetry | DoS Protection

Future Use Cases

Smart Switch “Networking & Security” Use Case

Top of Rack L4 Segmentation – November GA



Security Infused in Data Center Fabric

Version: NXOS 10.6(2), Hypershield 1.2

Smart Switches: N9348Y2C6D-SE1U, N9324C-SE1U

Fabric: VXLAN-EVPN, VXLAN-multi-site, BGP fabric, brownfield

Segmentation: VRF/VLAN + CIDR rules, stateful/stateless, 100K rules, 800G throughput (final scale based on benchmarking)

Policy: CRD schema, policy validation and canary rollout/rollback

Hypershield: Air-gap ready on-prem controller* and optional Security Cloud Control SaaS

Upgrade: NXOS CLI for DPU load, SMU for Hypershield agent

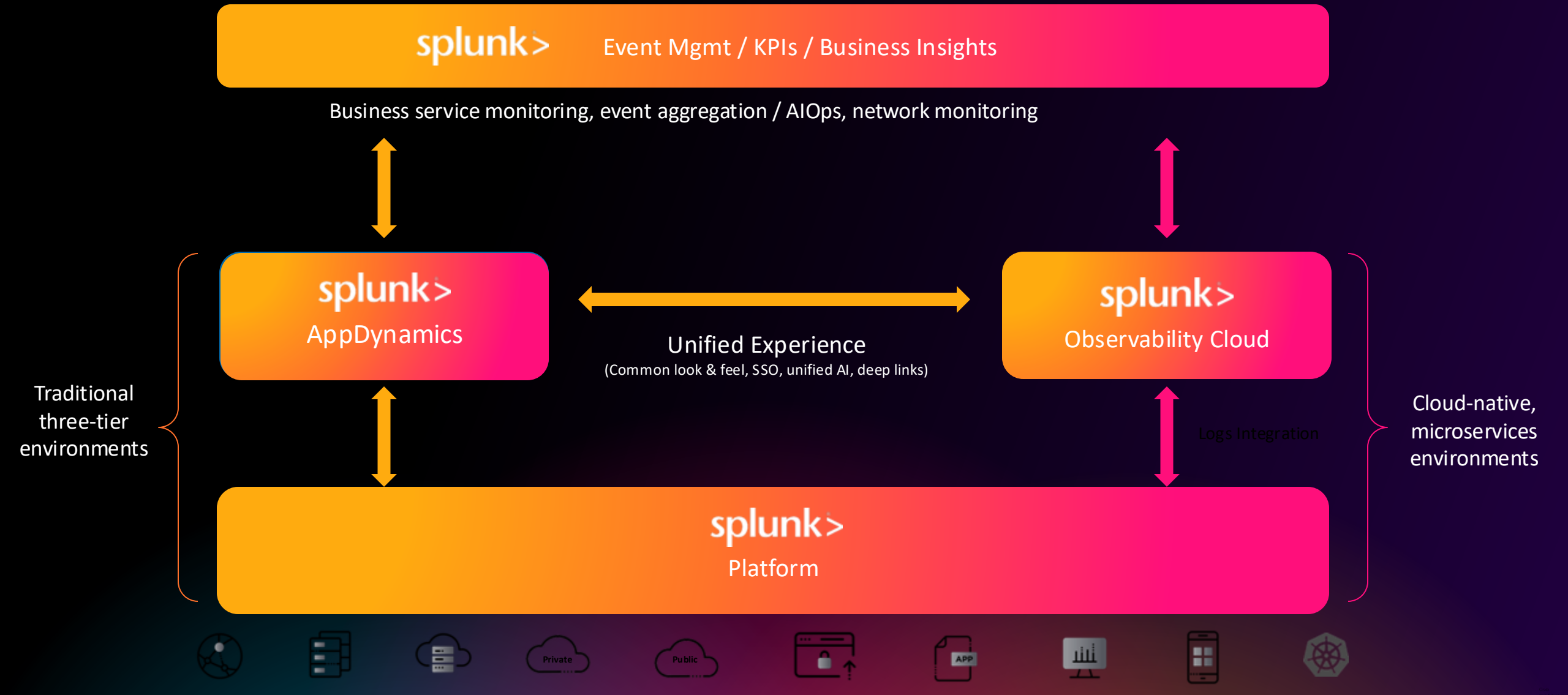
Observability: Nexus Dashboard, Splunk, Prometheus/Grafana

*More On-prem controller details coming soon

Observability in Cisco AI PODs

Integrated Full-Stack Observability

A view of the combined portfolio



Workloads

Security | Observability | Network Ops | Custom Applications

Cisco Data Fabric

AI-powered data management

Federated Search and Analytics

AI-Native Experiences and Platform for AI

Machine Data Lake

Sources

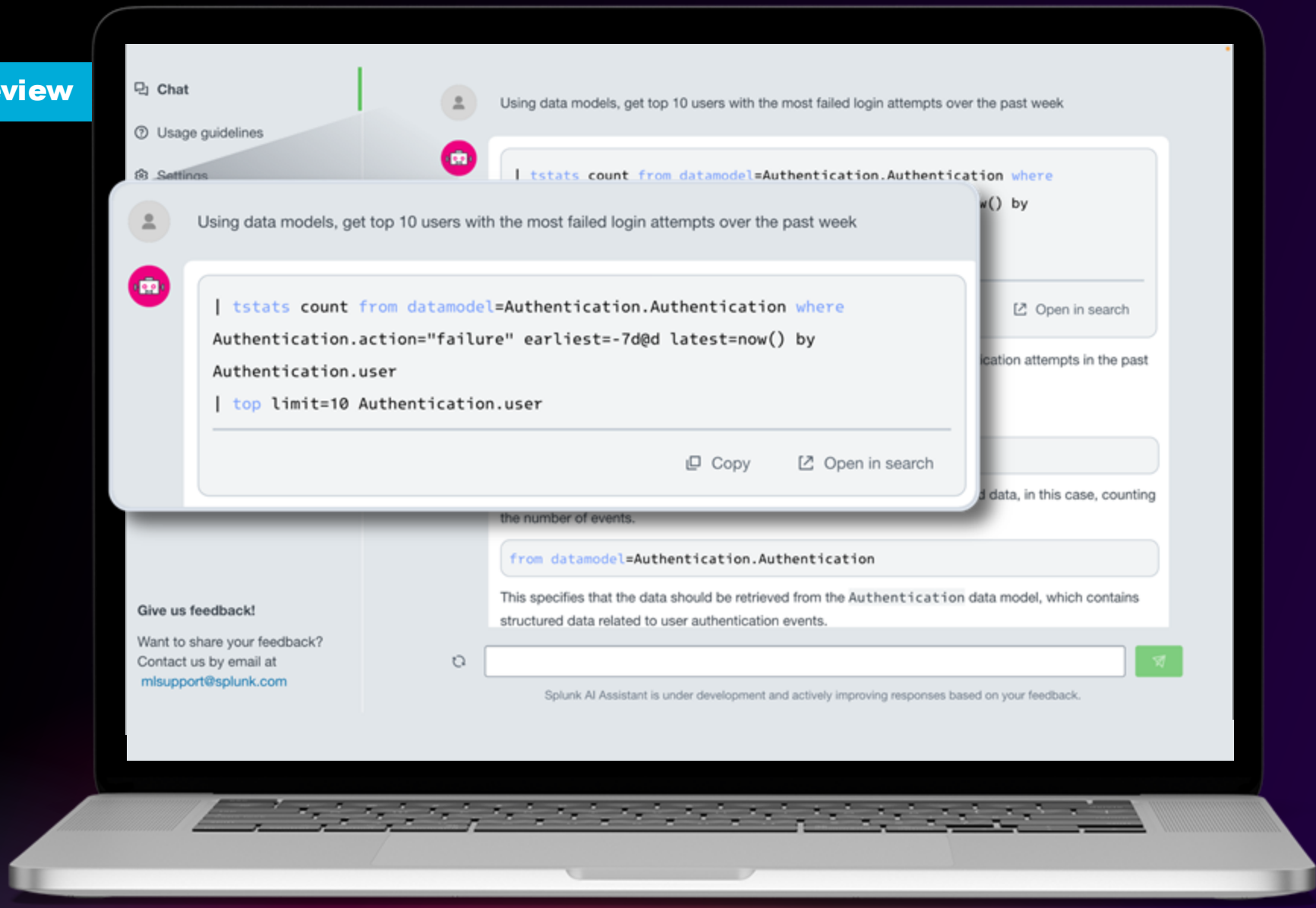
Infrastructure | Applications | Security | Users & Devices

PLATFORM

Preview

Splunk AI Assistant

Empower more people to search in Splunk using natural language



Cisco AI PODs

A scalable architecture, built to support any AI workload simply & efficiently

Deploy AI with confidence

Cisco CVD, NVIDIA ERA

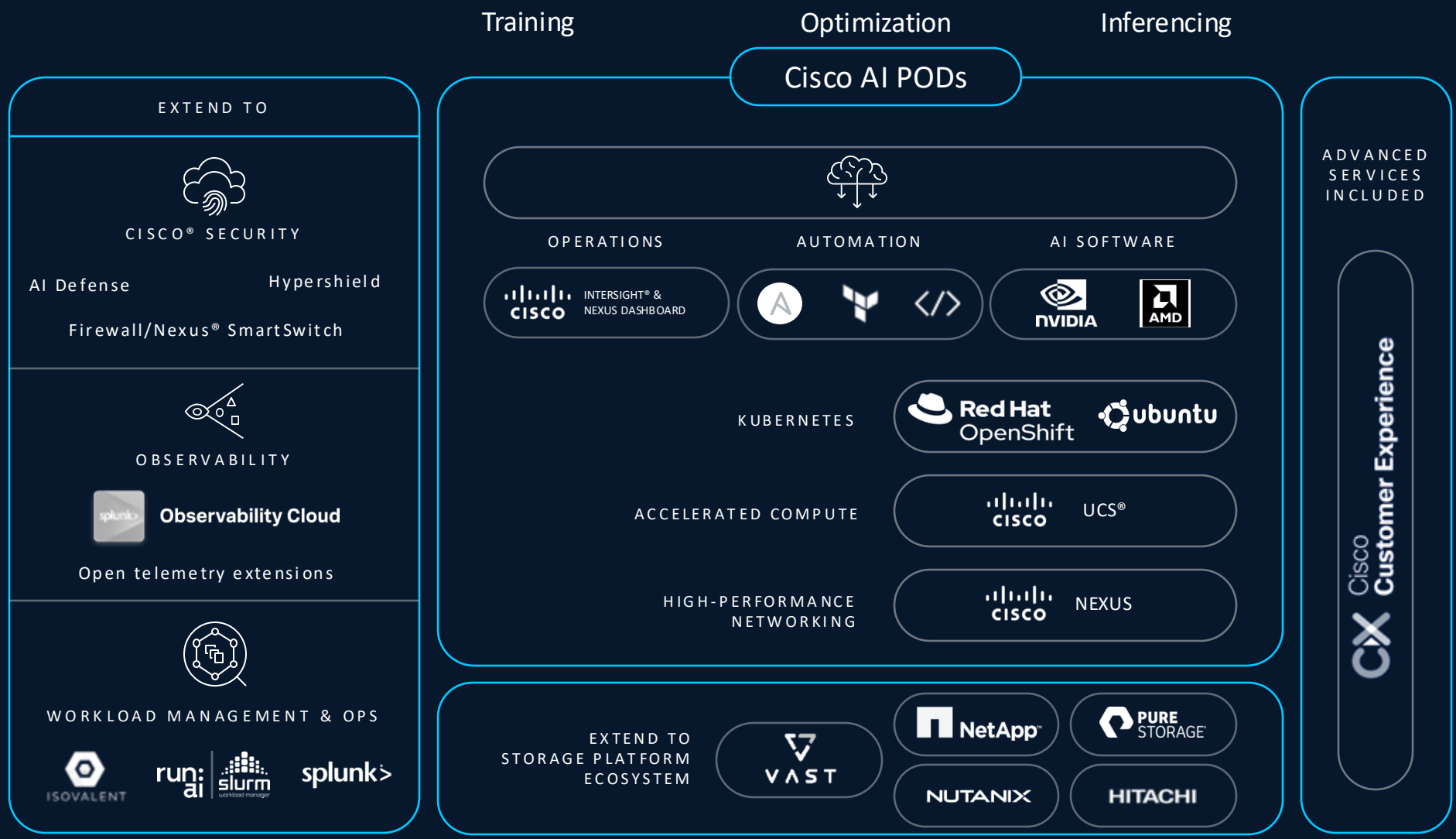
Fully supported stack including Cisco and 3rd party components

Cisco CX Success Track

Orderable, use case driven AI-Ready infrastructure stacks

Inferencing. Optimization. Training.

Incremental, atomic-level – or- fabric-based cluster scale



Děkujeme za Vaši pozornost

Následující Tech Club webinář:

02.12. Nová generace přepínačů Cisco a pozicovaní fabric řešení pro podnikové sítě

Přednášející: Jaromír Pilař

Registrovat se můžete na oficiálním webu **Cisco Tech Club webináře**

Následující webinář

Tech Club portál