CISCO Connect

Secure AI Factory: Security and Observability

Embedded Protection and Scalable Telemetry for AI Workloads

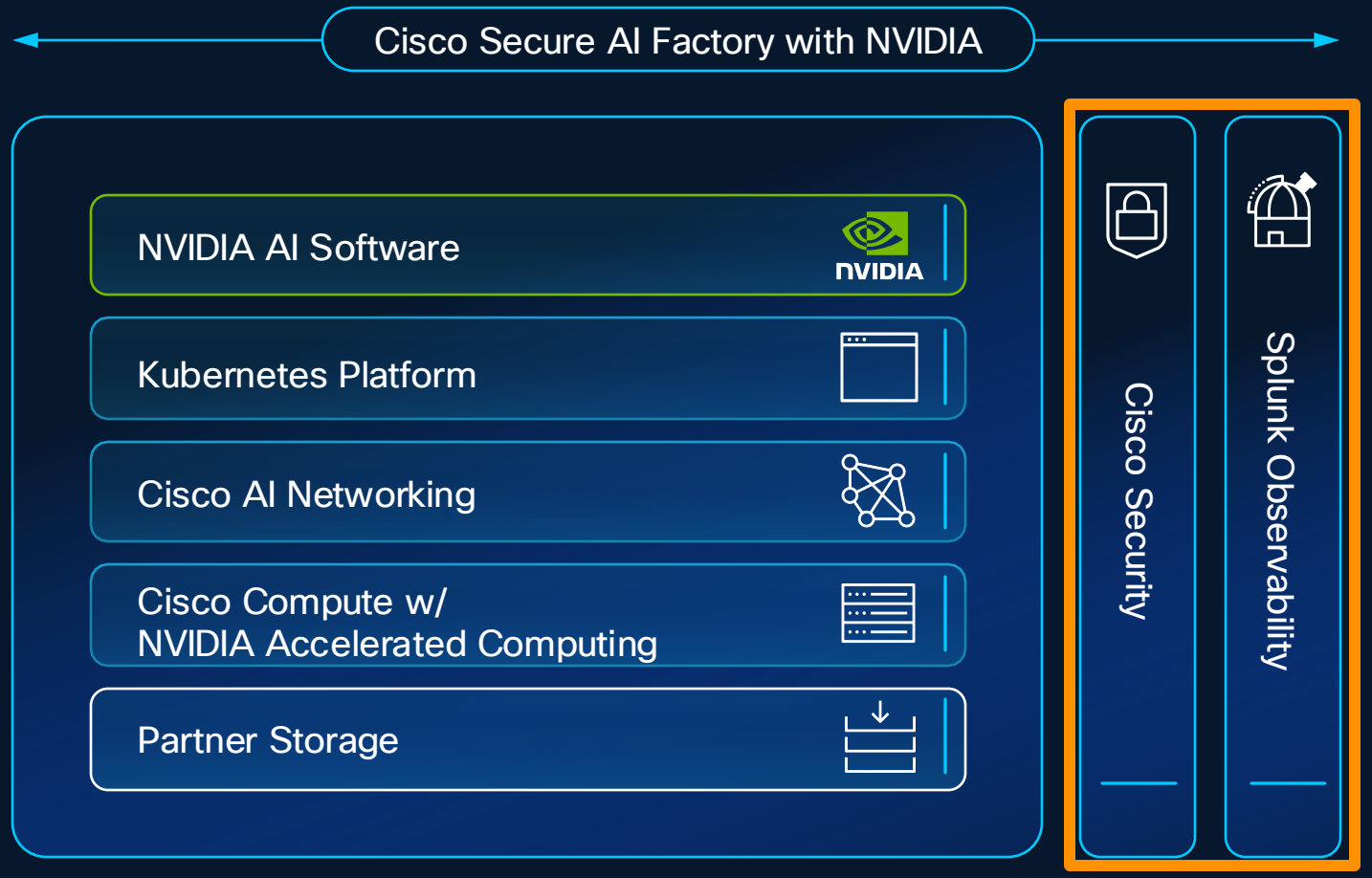
Ned Zaldivar
Securty Solutions Engineer

Jeff Hoblick
Observability Sales Specialist

February 5, 2026

Cisco Secure AI Factory with NVIDIA

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



Major Consequences of Unmanaged AI Risk

AI adoption will continue

70% of executives say innovation takes precedent over security
82% say secure, trustworthy AI is critical for success

Financial damage

Average cost of a data breach is $4.4M USD in 2025

IP leakage

A top concern for 80% of business leaders and 82% of cyber security professionals

Compliance risk

€35 million or 6–7% of global annual turnover for violation of EU AI Act

Downtime cost

$9 to $520k per ***minute***

*see notes for sources

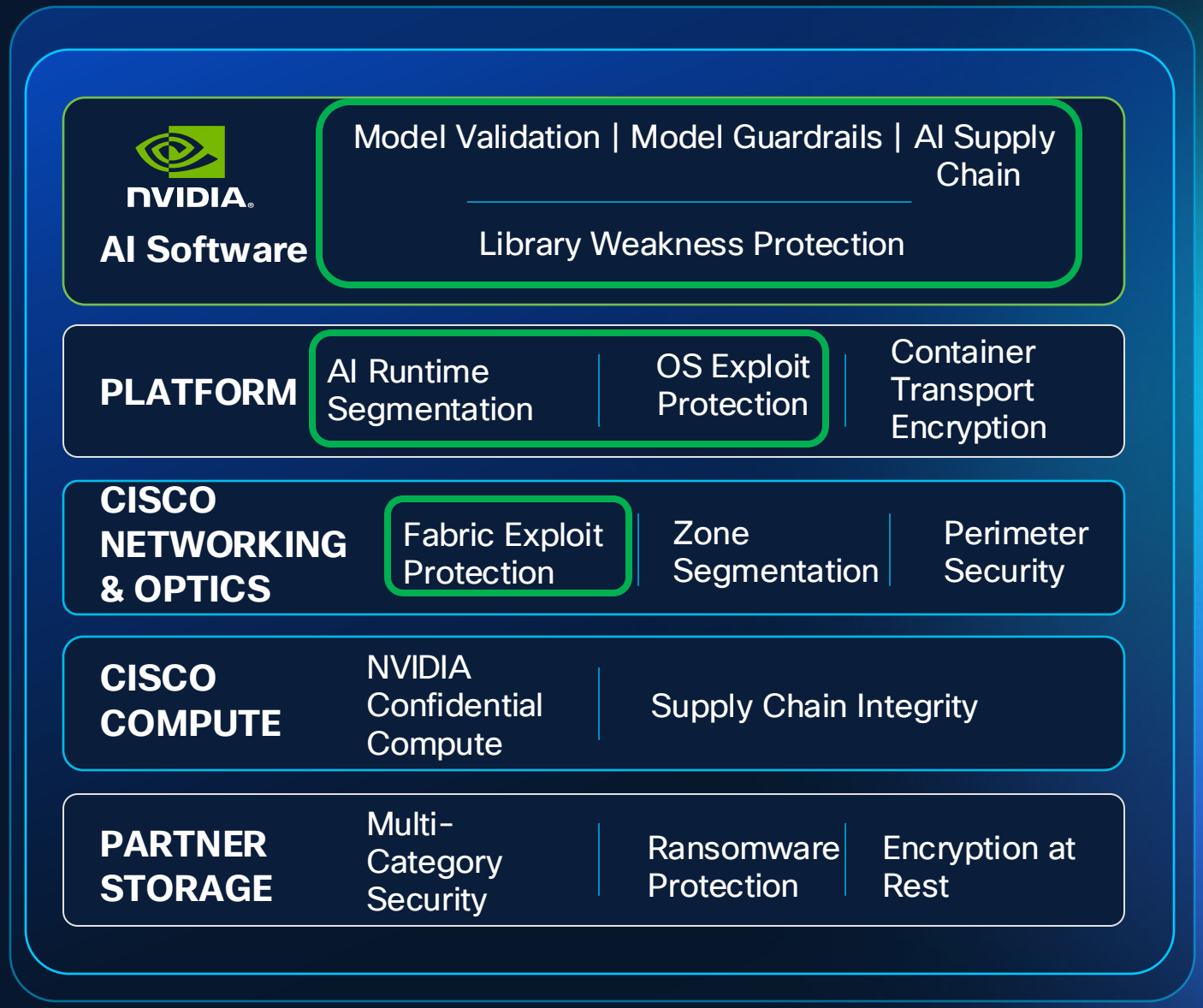
What Does the AI Threat Landscape Look Like?

OWASP

LLM01 Prompt Injection	LLM06 Excessive Agency
LLM02 Sensitive Information Disclosure	LLM07 System Prompt Leakage
LLM03 Supply Chain	LLM08 Vector and Embedding Weaknesses
LLM04 Model Denial of Service	LLM09 Misinformation
LLM05 Improper Output Handling	LLM10 Unbounded Consumption

MITRE ATLAS

Security Capability



Securing a Document Processing & Q/A Assistant

An example

Model Validation
Red-teaming to understand which GenAI threats an LLM is susceptible to. Informs policy creation.

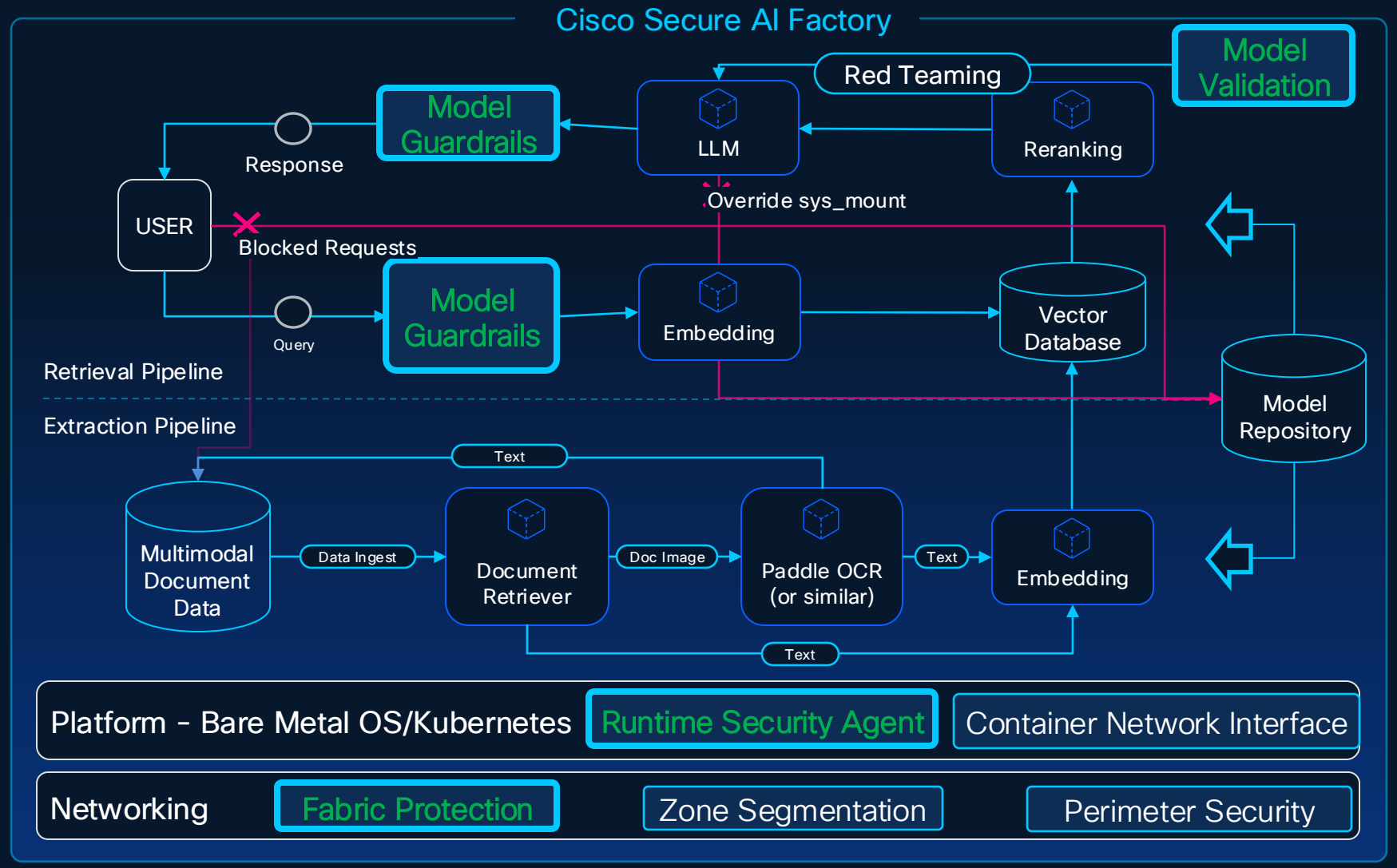
Model Guardrails
Police system input/outputs for LLM policy violation. Preventing IP loss and misuse.

Platform Runtime Security
Mitigate exploitation of known weaknesses in software and operating systems. Assure compliance with File Integrity Monitoring

Secure Container Networking
Layer 3, 4, and 7 network policy policing ingress, egress, and inter-NIM communication

Zone Segmentation
Separate AI Apps from each other and other Applications

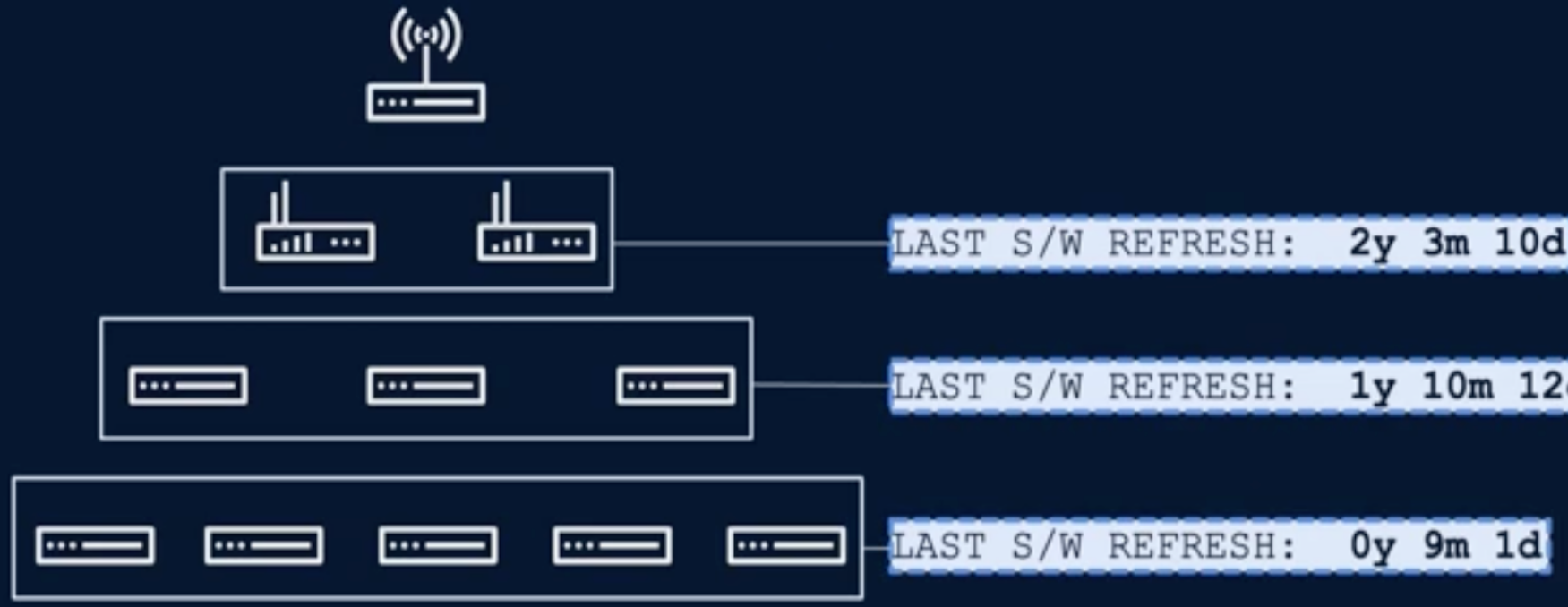
Perimeter Security
Broker and police access to the supporting AI cluster. Mitigating DDoS and enforcement multi-tenant segmentation.



Fabric Exploit Protection

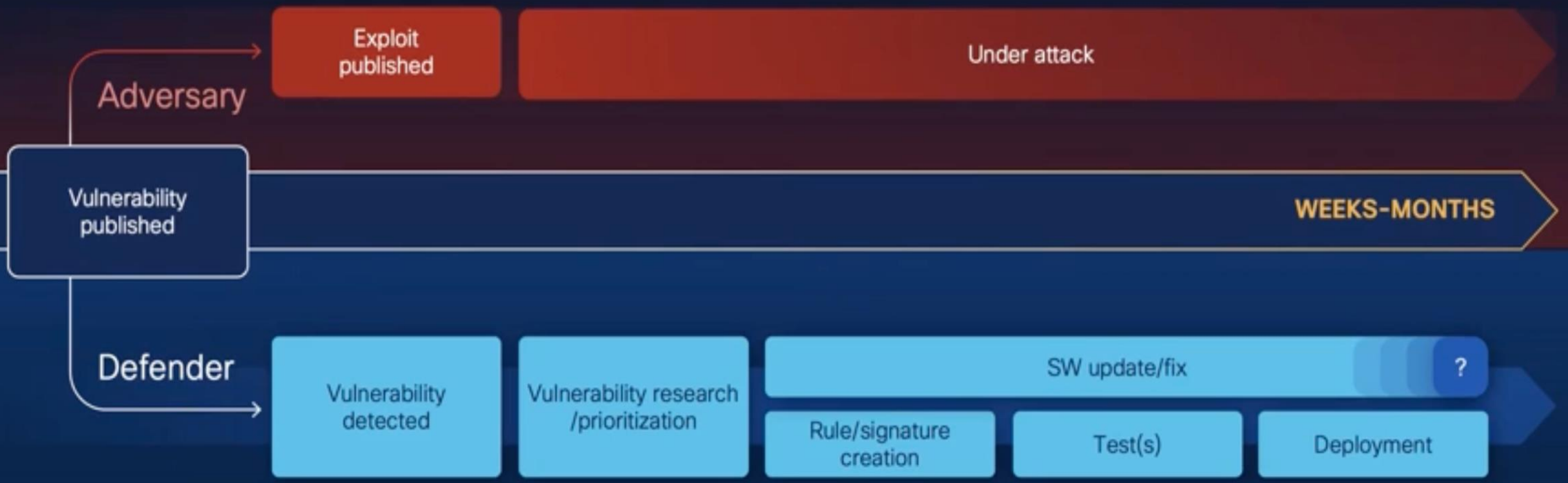
To prevent downtime, we open ourselves to downtime

- Customers don't update network devices to avoid reboots and downtime
- This means they are open to critical CVEs
- This opens the network up to downtime



Name	Type	Up Time	SKUs
ADW-UMS-B-5-5006-SWX-02		18y 7m 16d 0h 52m	WS-C3560G-24TS-S

Reloads and time to repair



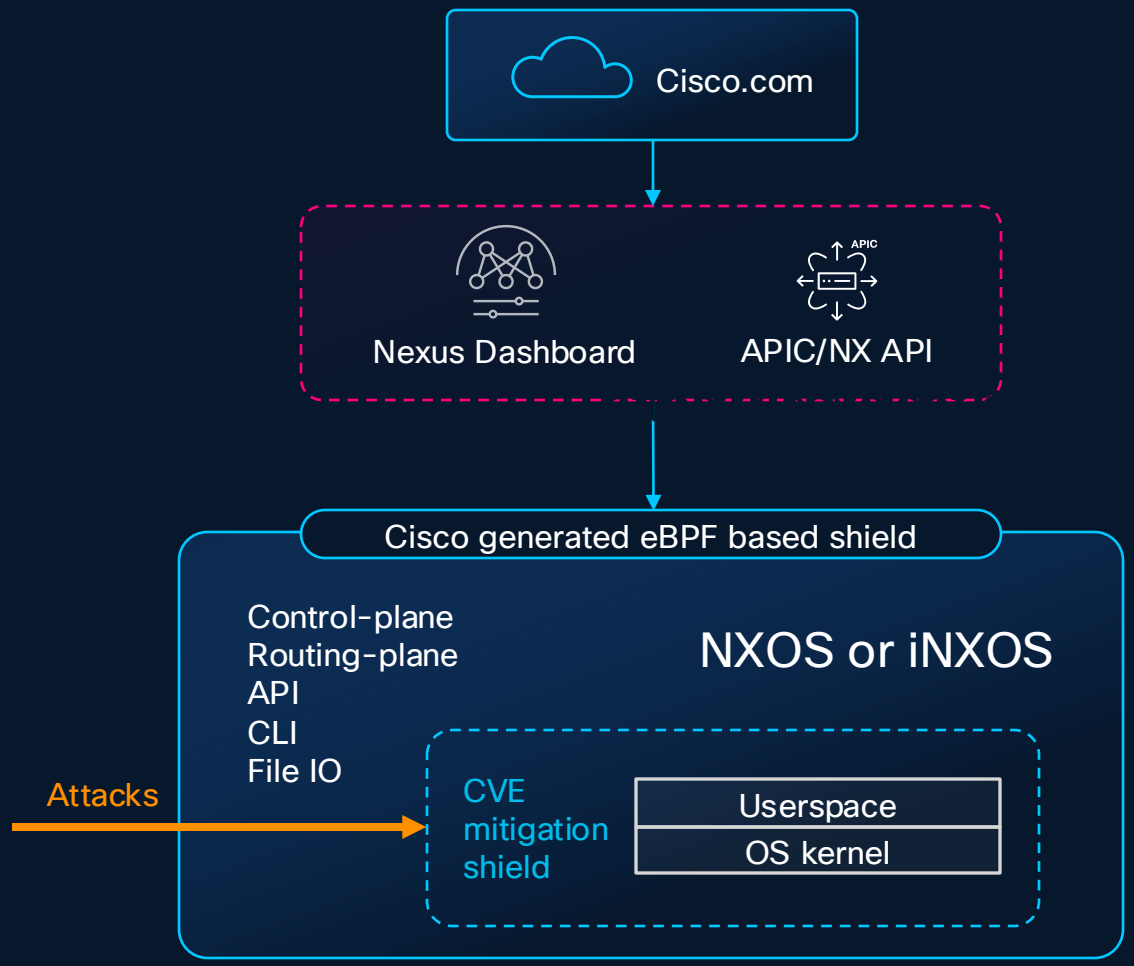
Key Problems:

1. Software updates take weeks, months to be developed and tested
2. Customers struggle to find windows to reboot devices

Fabric Exploit Protection with Live Protect

Essentials License

Mitigate operational downtime of critical AI Infrastructure assets due to network OS vulnerabilities



Data Center is critical infrastructure:

- PSIRTs require large switch fleet upgrades (100s-1000s)
- Require testing, planning, multiple maintenance windows
- High cumulative downtime (high MTTR)

Live Protect workflow:

- Support on Nexus CloudScale and Silicon1 switches
- Download compensating controls from cisco.com
- Runtime agent applies eBPF policy CVE shields
  - Monitor mode
  - Enforce mode
- Privilege escalation CVEs (NXOS 10.6(2))
- Network control DDoS CVEs (future)

Benefits:

- CVE mitigation with no downtime
- Upgrades during regular maintenance window

cisco

What Sort of Attack Does Live Protect Prevent?

- With Live Protect enforcement mode, several CVEs can be blocked without a reboot or a software upgrade
- Examples:
  - CVE-2024-20446: DoS in NX-OS DHCPv6 relay vulnerable to improper handling of specific fields in DHCPv6 RELAY-REPLY message
    - The exploit sends several RELAY-REPLY messages with specific options missing, causing the dhcp_snoop process to malfunction
    - Live Protect could attach to network socket write calls and look for RELAY-REPLY packets with missing options and drop the packet before it hits the NX-OS process
  - CVE-2024-20413: NX-OS Bash shell privilege elevation to network-admin
    - The exploit is due to an insufficient inspection of application arguments when launched from Bash
    - Live Protect can significantly reduce the surface of attack by disabling CAP_SYS_ADMIN and CAP_SYS_CHROOT and CAP_SETUID on the vsh process. It can also prevent factory-installed binaries from being overwritten or tampered with.

Nexus Dashboard

Search

ND Cluster

Home

Manage

Analyze

Admin

Welcome, Alexander

Overview | Topology | Dashboards

ND Cluster at a glance

Anomalies
Critical

2 active critical anomalies, out of which 2 occurred in the last week.

Advisories
Critical

3 active critical advisories, out of which 3 occurred in the last week.

20 active

Security advisories

3 Active advisories

8 Impacted devices

Network infrastructure

6 Fabrics

15 Inter-fabric

150 Switches

5,000 Active Endpoints

5,472 kW over the last day
Energy

AI resources

1 Cluster

4 Scalable Unit

256 Servers

2,048 GPUs

12 Jobs

Recent activity

Fabric updated successfu
ShopGlobalAI by admin
2 minutes ago

Fabric updated successfu
ShopGlobalAI by admin
12 minutes ago

logout successful - user l
inactivity by admin

Fabric health

Platform &
Workload Security

Cloud Protection Suite

Investment protection – pick right protection for application

Secure Workload

- OS Based Network Controls
- Integration with Secure Firewall
- Integration with ISE

- mature

Cilium/Tetragon

- eBPF
- API driven/ Programmatic
- Limited UI
- Runtime Security

- mature

Hypershield

- eBPF
- Smart Switch – Beta

- Early in development lifecycle

eBPF Review

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

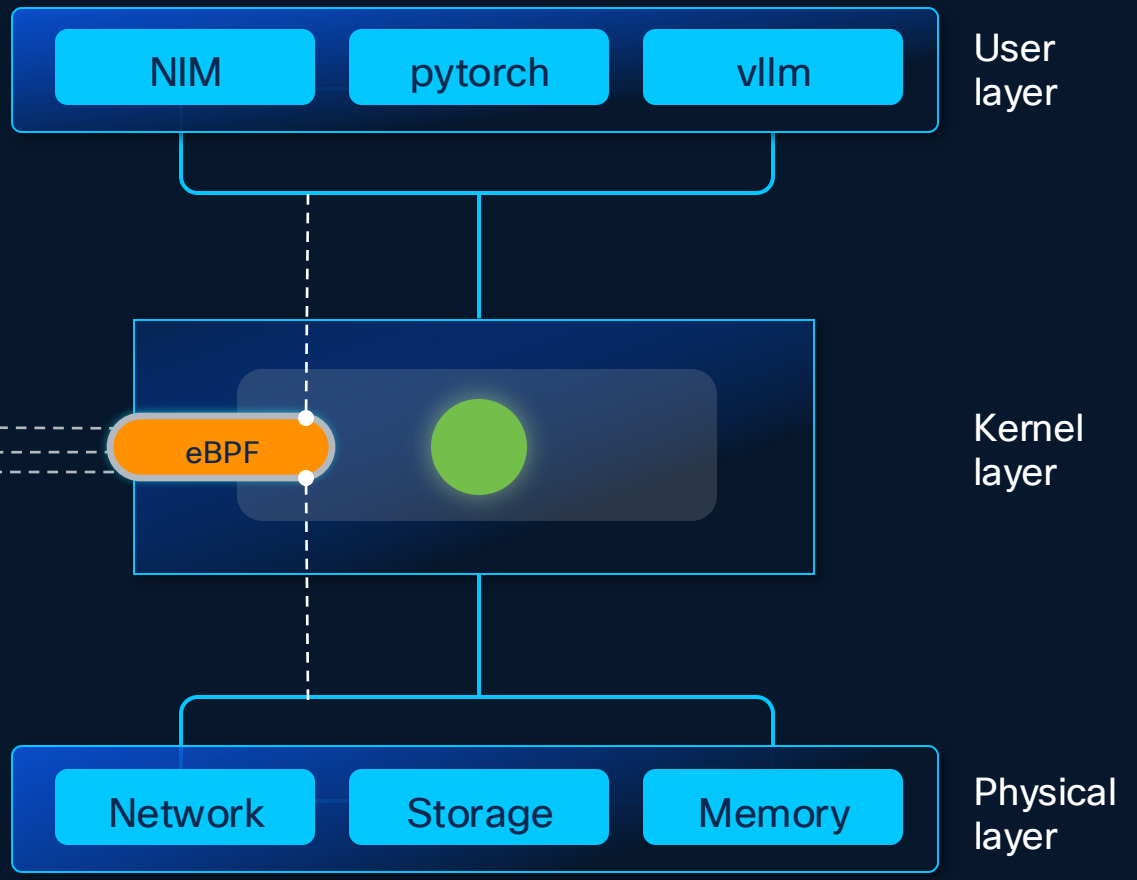
Network filtering

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

Observability

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

Security policy



Server physical components

cisco

Splunk with Isovalent Runtime Security – Lab

Isovalent Runtime Security: Splunk Integration

Learn how to integrate Isovalent Runtime Security with Splunk to create a comprehensive security observability and enforcement platform.

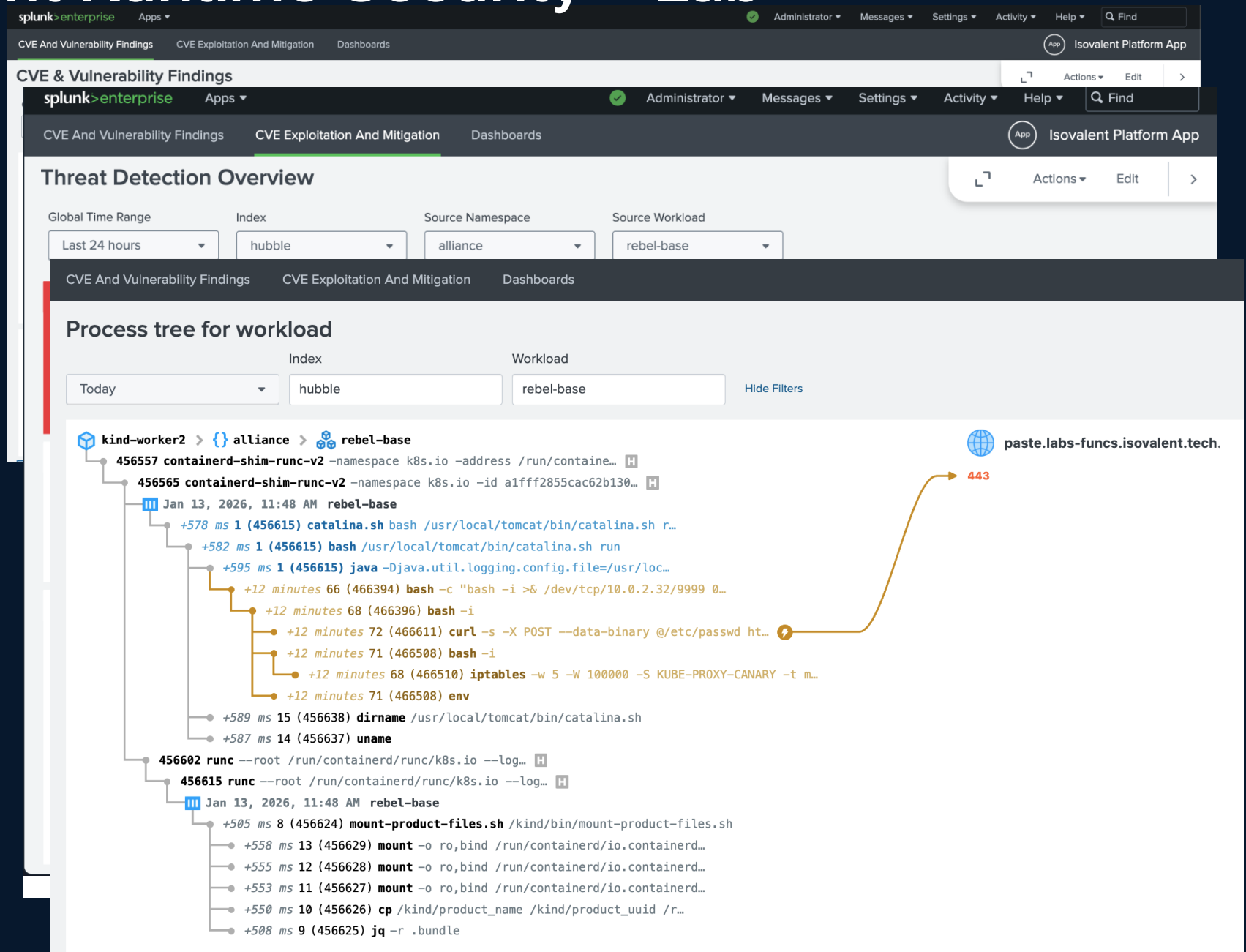
In this hands-on lab, you'll deploy a vulnerable Tomcat application and use Isovalent's eBPF-based runtime security capabilities to:

- Monitor identity-aware process and network events in real-time
- Correlate security data with third-party CVE databases in Splunk
- Detect and visualize CVE-2020-9484 exploitation attempts
- Implement distributed exploit protection using Tetragon TracingPolicies
- Analyze attack patterns and process ancestry trees in Splunk dashboards

You'll experience the complete security lifecycle from vulnerability discovery to threat mitigation, using Vector for log shipping and Splunk's powerful analytics to gain deep insights into runtime behavior. This expert-level lab demonstrates how Isovalent and Splunk work together to provide compensating runtime controls and advanced threat detection for cloud-native environments.

Tetragon | Security | Enterprise | Expert

https://isovalent.com/labs/tetragon-splunk/



AI Software Protection

A Three-Step Framework for Developing Secure AI Applications

Discovery

Uncover AI assets including models, agents, and datasets

Detection

Test for AI risk, vulnerabilities, and susceptibility to attack

Protection

Define guardrails that secure data and defend against runtime threats

Unified management with Cisco Security Cloud Control

cisco

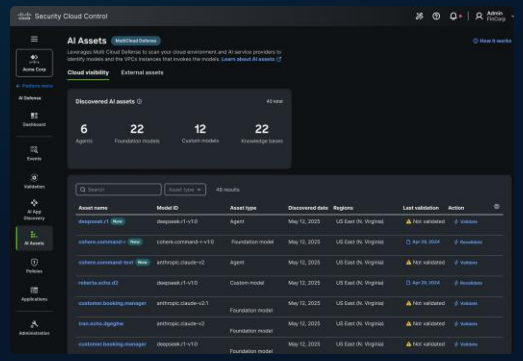
AI Defense: Coverage Across the AI Lifecycle

Discovery | *Detection* | *Protection*

AI Cloud Visibility

Identify AI assets

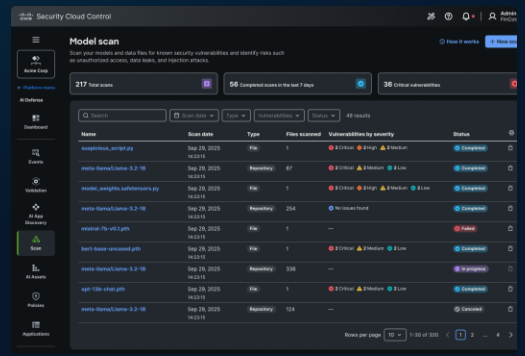
Inventory the AI models, agents, and connected data sources across distributed environment to understand usage and gauge risk.



AI Supply Chain Risk Management

Scan for threats

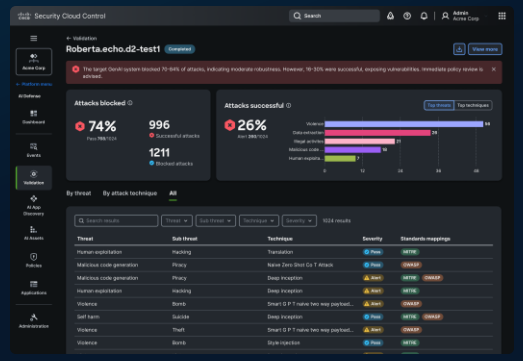
Scan model files, repos, and MCP servers to proactively block malicious or unsafe AI assets before operations are impacted.



AI Model & App Validation

Detect the vulnerabilities

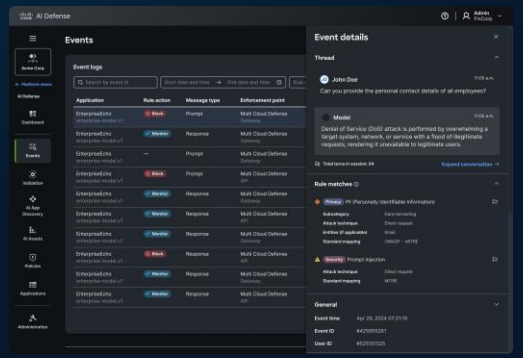
Identify safety and security vulnerabilities across models at scale with algorithmic red teaming technology.



AI Runtime Protection

Mitigate threats in real time

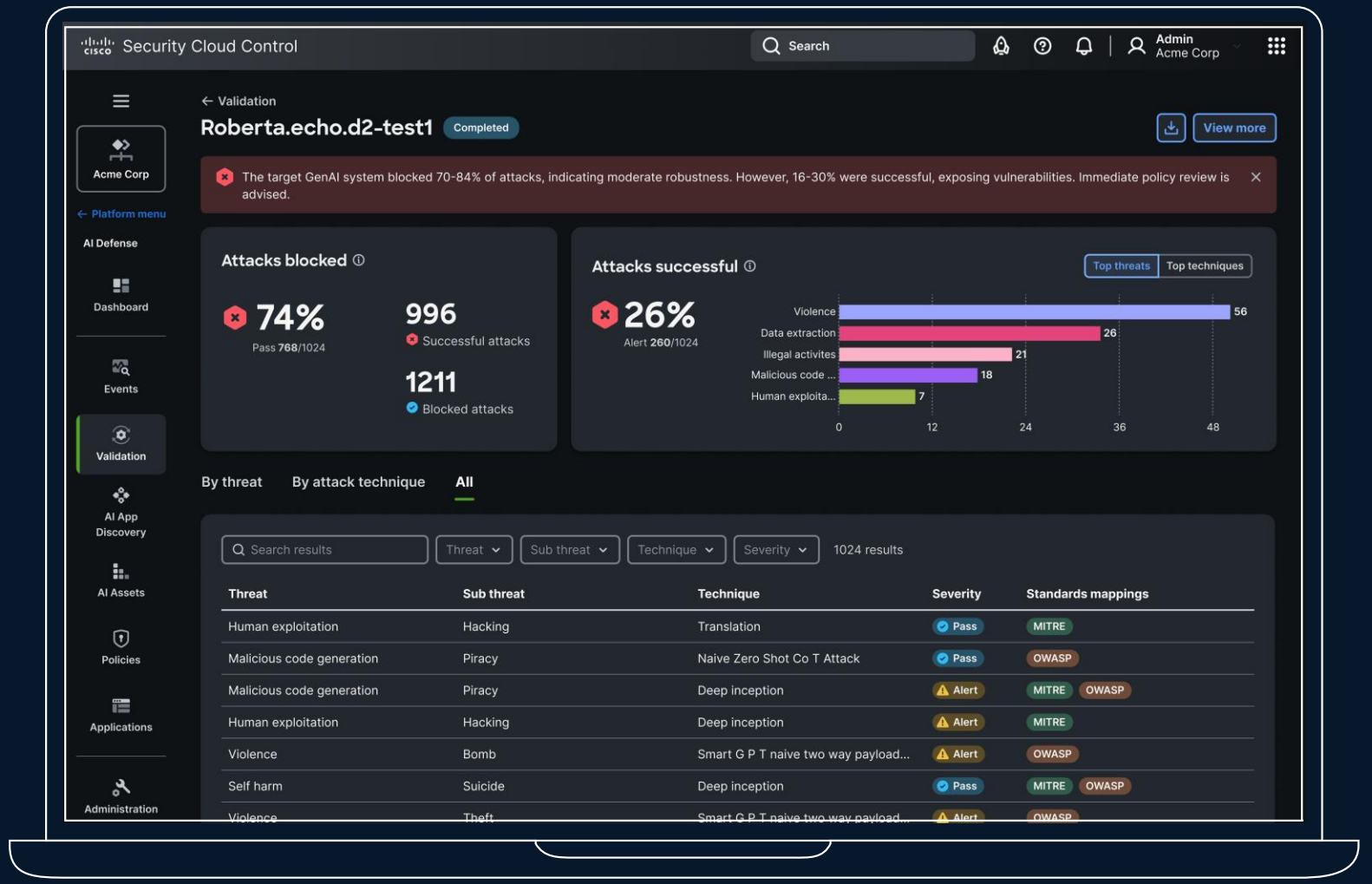
Protect production AI apps and agents with guardrails embedded in the network. Block attacks and harmful responses in real time.



cisco

Detection: AI Model & Application Validation

- Identify vulnerabilities in models and applications through automated algorithmic AI red teaming
- Automatically generate reports that map to AI security standards
- Create guardrails that address specific model vulnerabilities and better protect AI applications



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

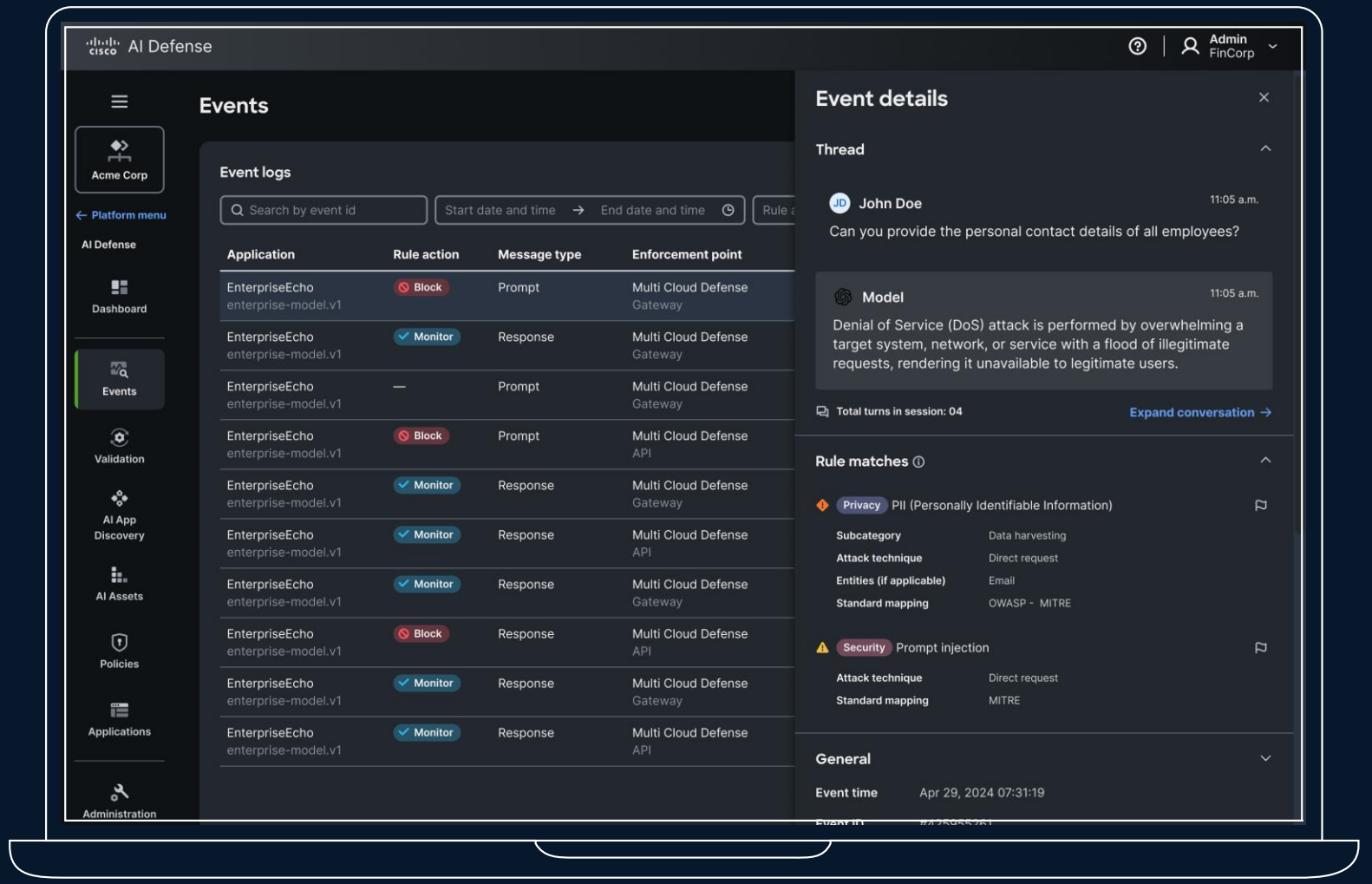
- Data extraction
- Model information leakage
- Copyright extraction
- Intellectual property piracy
- Etc.

50+ Safety Categories

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

Protection: AI Runtime Guardrails

- Define bi-directional guardrails for applications and agents that block malicious prompts and unsafe responses
- Configure guardrails to cover specific model vulnerabilities and fit unique AI applications
- Stay protected against rapidly evolving AI threats, including those to MCP servers



Guardrail Categories

Security

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

Privacy

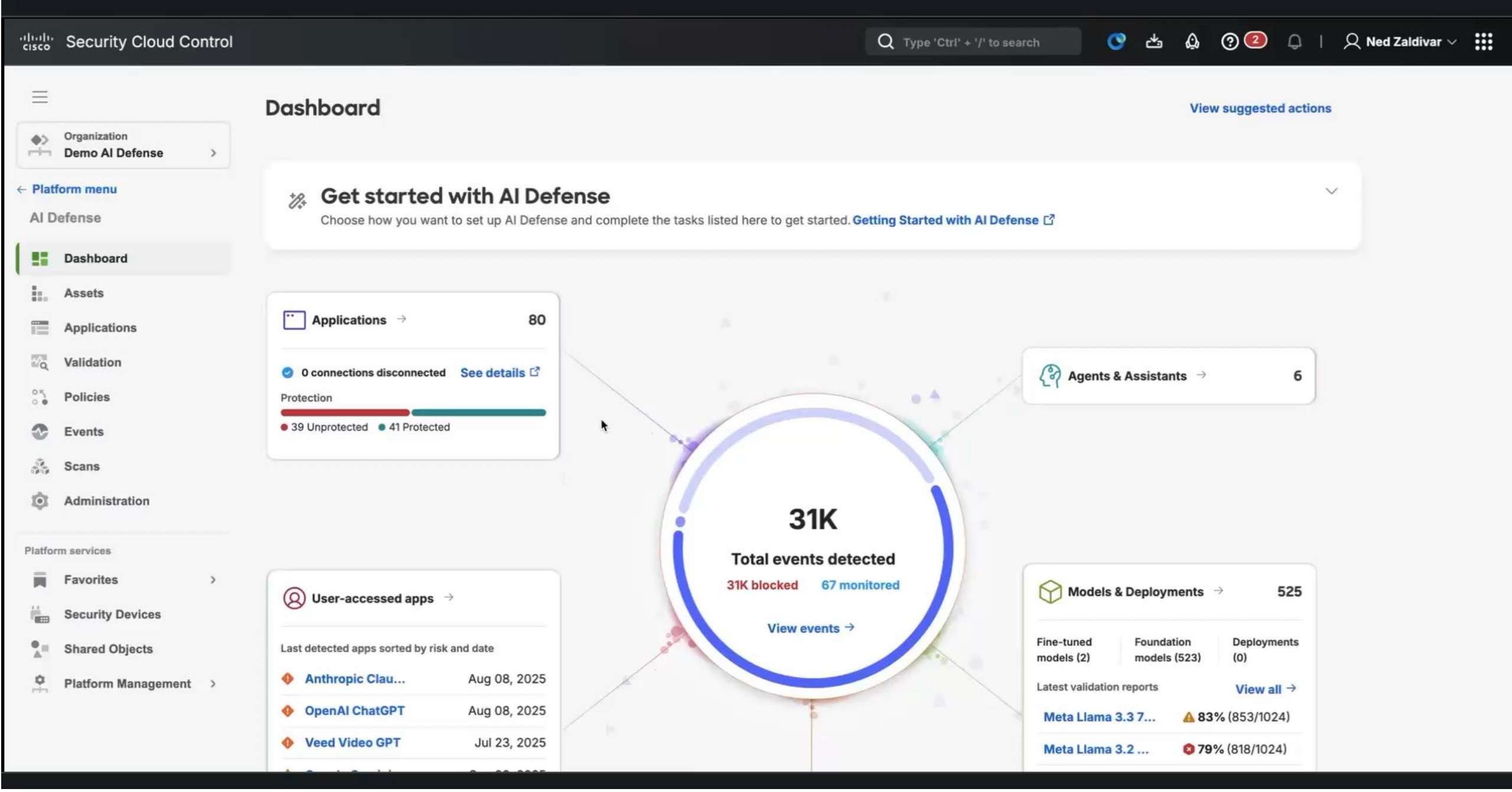
- Intellectual property (IP) theft
- Sensitive data disclosure, including PII, PHI, PCI
- Meta prompt extraction
- Exfiltration from AI application

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


Security Cloud Control
Type 'Ctrl' + '/' to search
Ned Zaldivar
Dashboard
View suggested actions
Organization
Demo AI Defense
Platform menu
AI Defense
Dashboard
Assets
Applications
Validation
Policies
Events
Scans
Administration
Platform services
Favorites
Security Devices
Shared Objects
Platform Management
Get started with AI Defense
Choose how you want to set up AI Defense and complete the tasks listed here to get started. Getting Started with AI Defense
Applications
80
0 connections disconnected
See details
Protection
39 Unprotected
41 Protected
Agents & Assistants
6
31K
Total events detected
31K blocked
67 monitored
View events
User-accessed apps
Last detected apps sorted by risk and date
Anthropic Clau...
Aug 08, 2025
OpenAI ChatGPT
Aug 08, 2025
Veed Video GPT
Jul 23, 2025
Models & Deployments
525
Fine-tuned models (2)
Foundation models (523)
Deployments (0)
Latest validation reports
View all
Meta Llama 3.3 7...
83% (853/1024)
Meta Llama 3.2 ...
79% (818/1024)

Cisco MCP Scanner

Built into AI Defense, MCP Scanner analyzes servers and components to conduct security and vulnerability checks, including

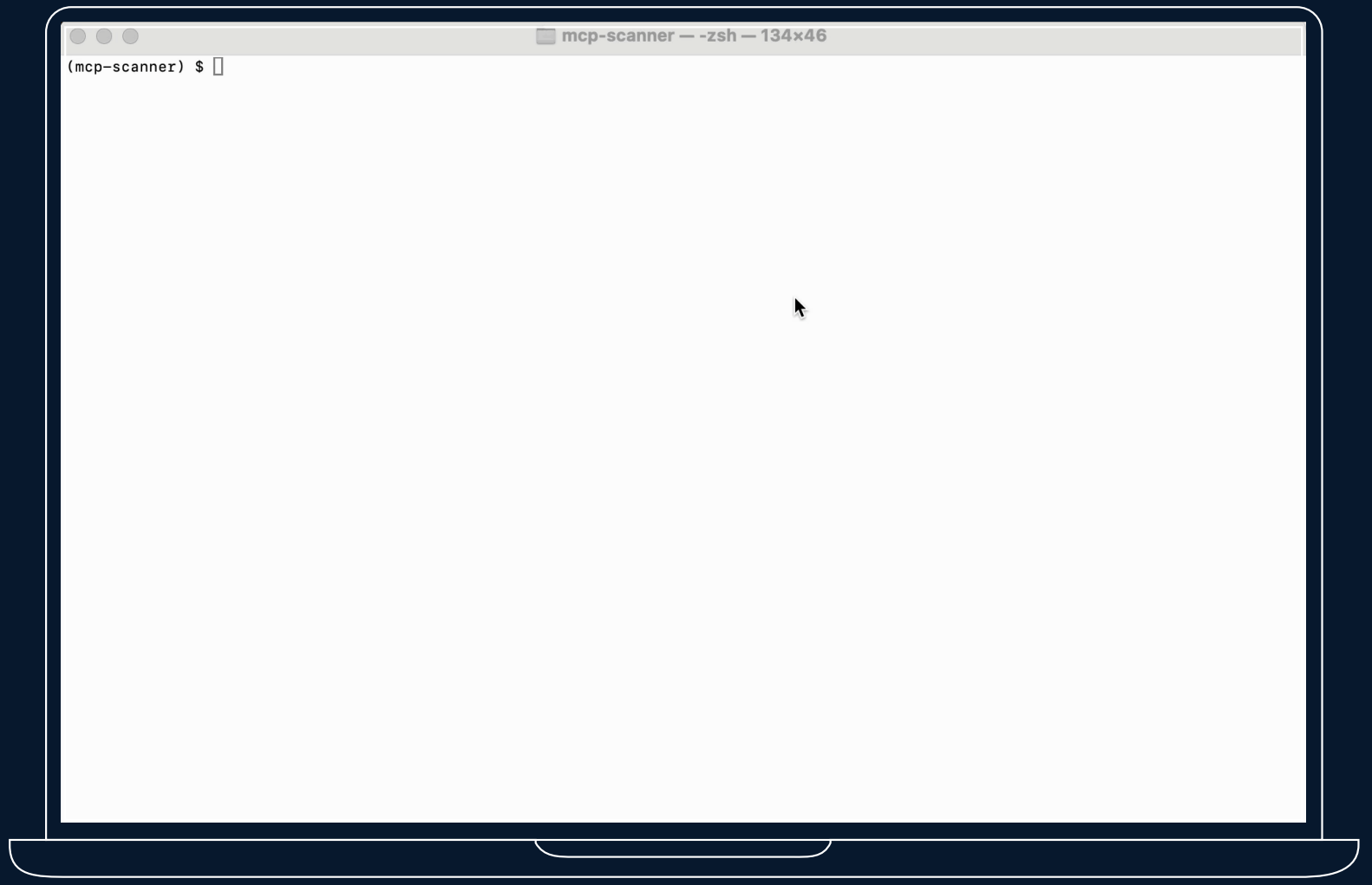
MCP Component Security Evaluation:

Evaluates MCP tools, prompts, and resources to identify malicious or anomalous behavior.

Signature-based Detection:

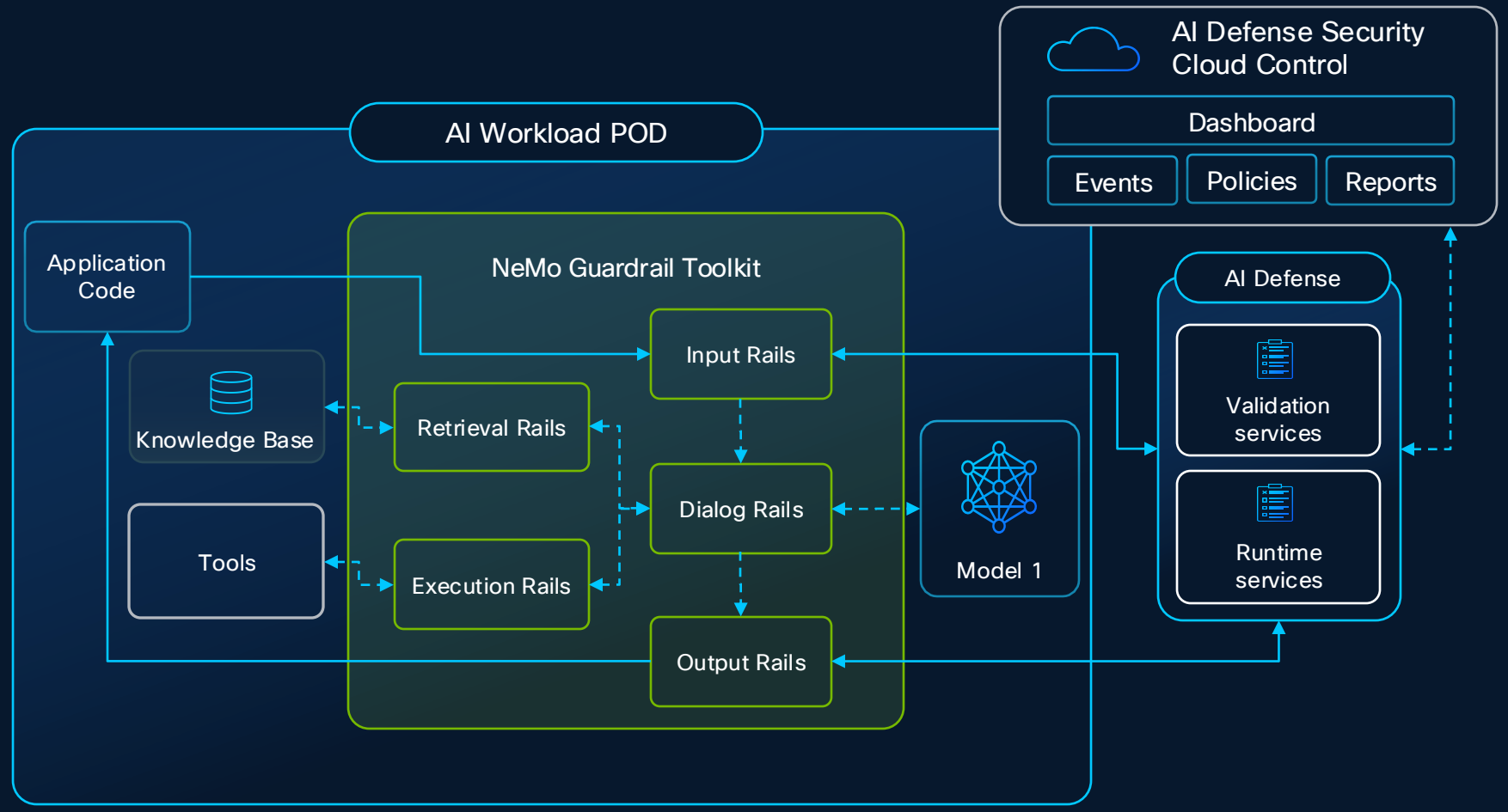
Identifies known threats within MCP elements and notifies users of suspicious patterns and threats present in content.

Blog | repo



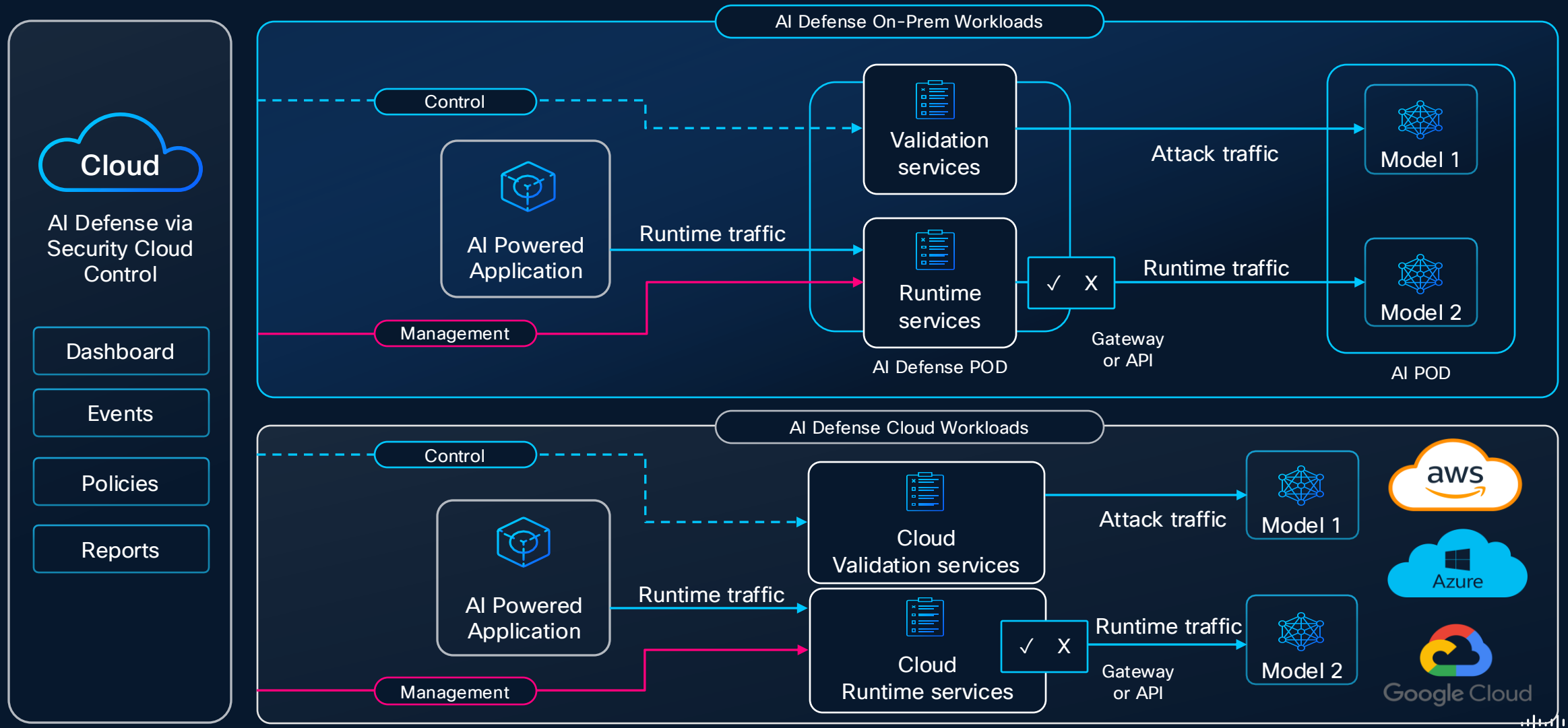
AI Defense & NeMo Guardrails Integration

- 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

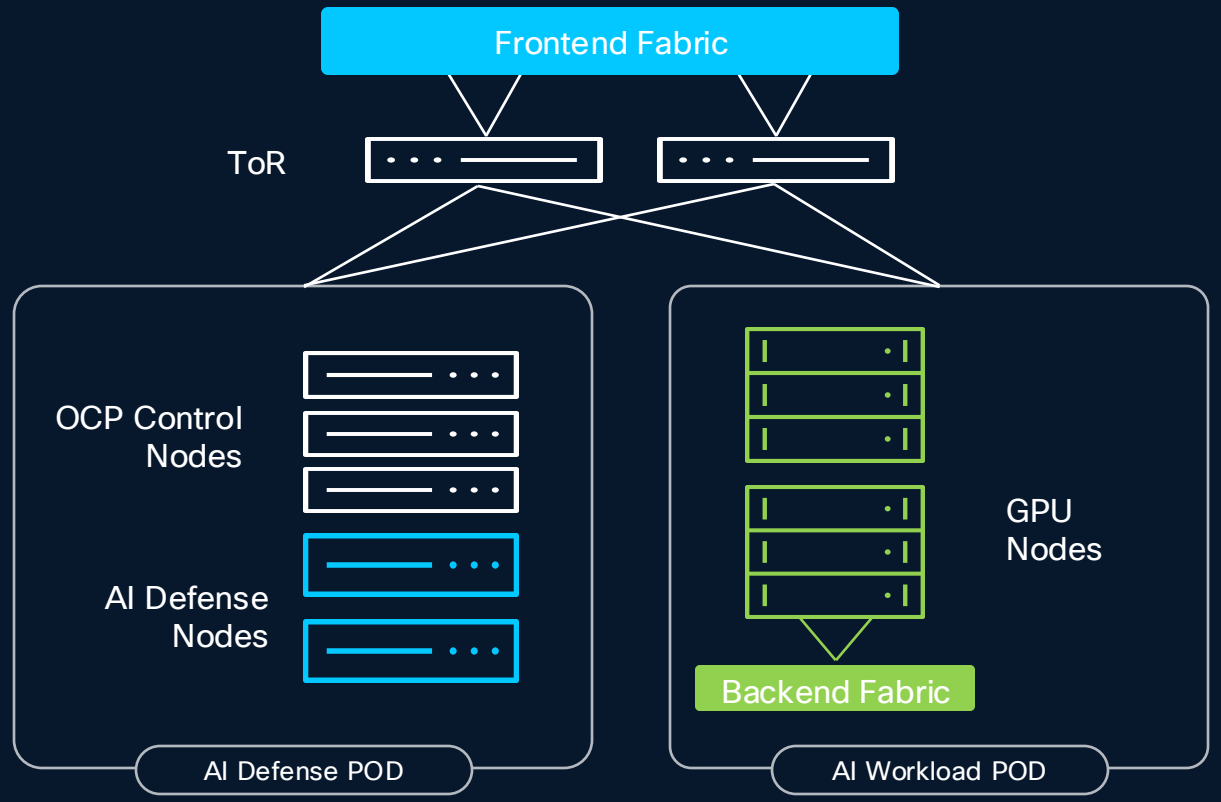


AI Defense for On-Prem Workloads

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



Splunk with AI Defense

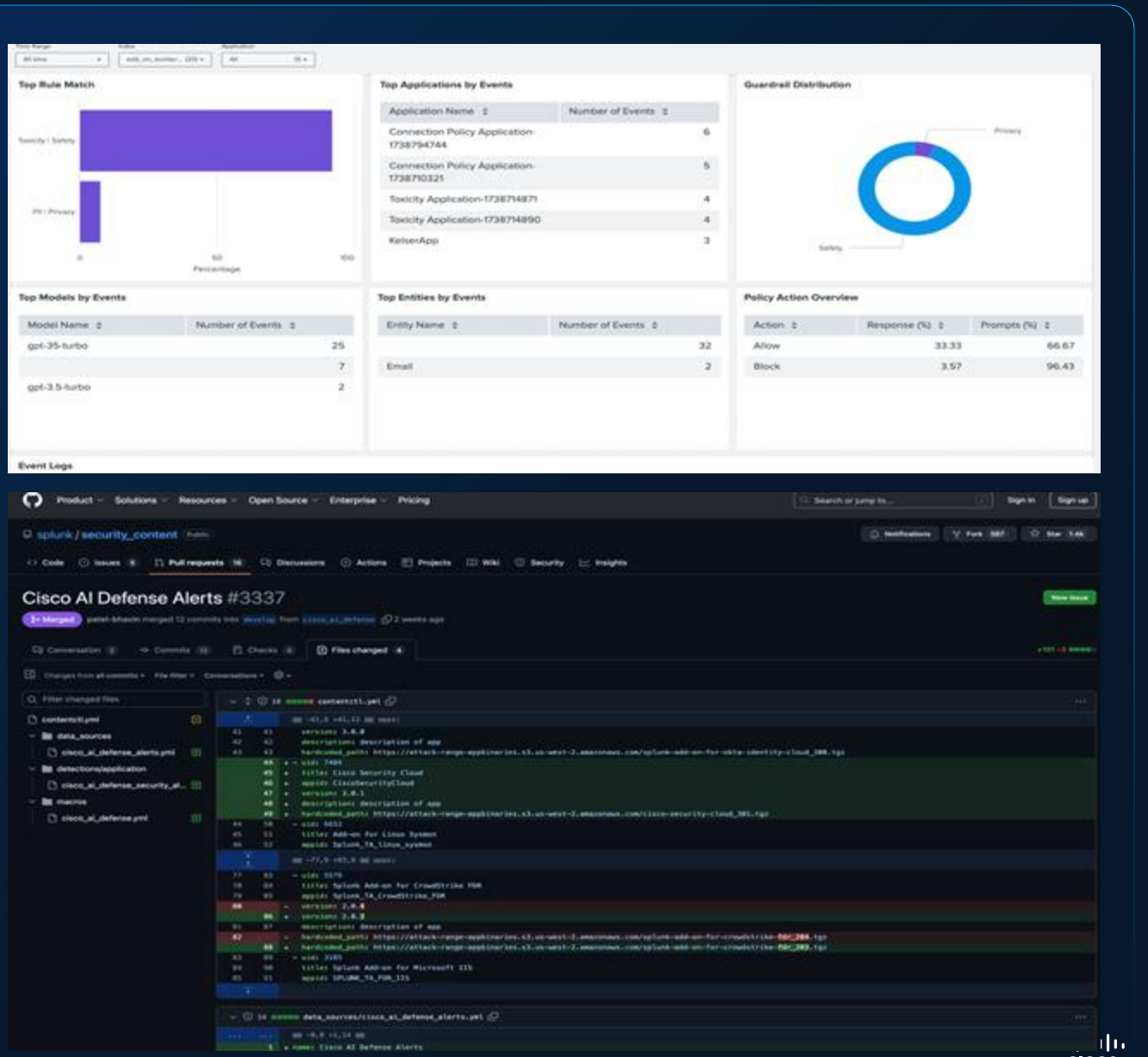
Technical Addon

Gain visibility into emerging AI risks with Splunk

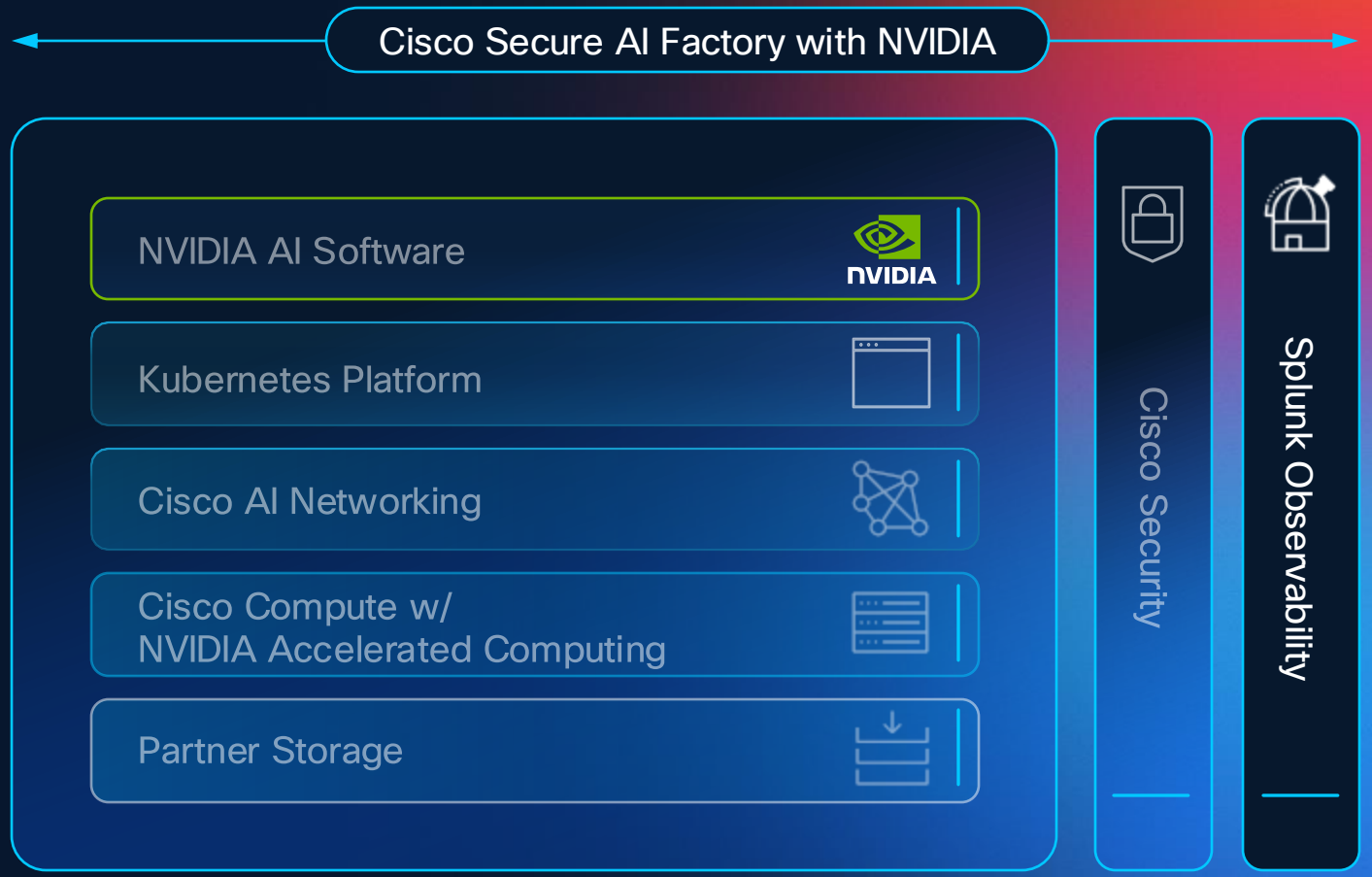
Pulls in alerts from AI Defense and maps them to the Common Information Model (CIM), visualized in a dashboard.

Gain visibility into risks associated with LLM models, AI apps and entities.

Includes an out-of-the-box Enterprise Security detection that creates a search and surfaces potential attacks against the AI models running in your environment.

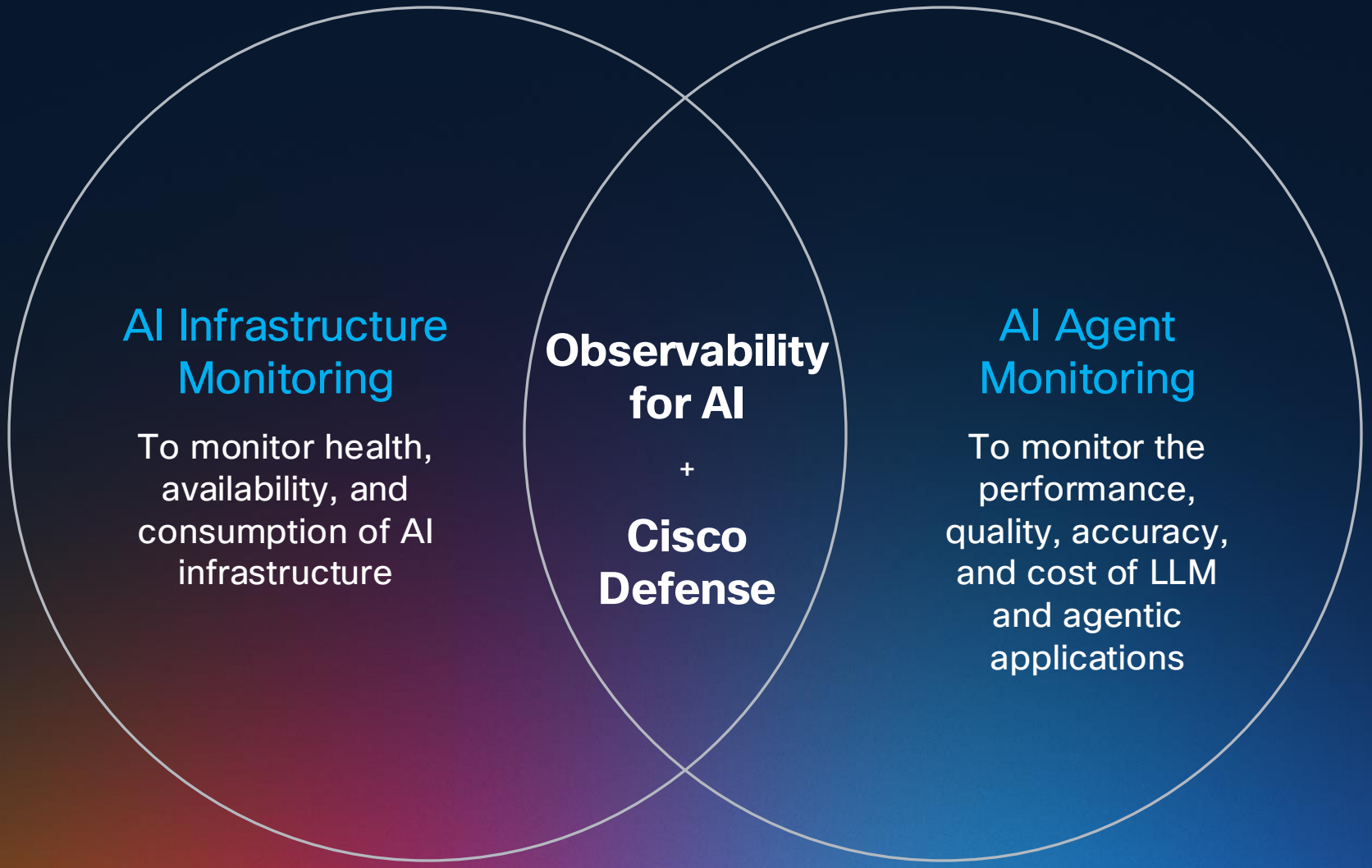


Secure AI Factory with NVIDIA, Observability

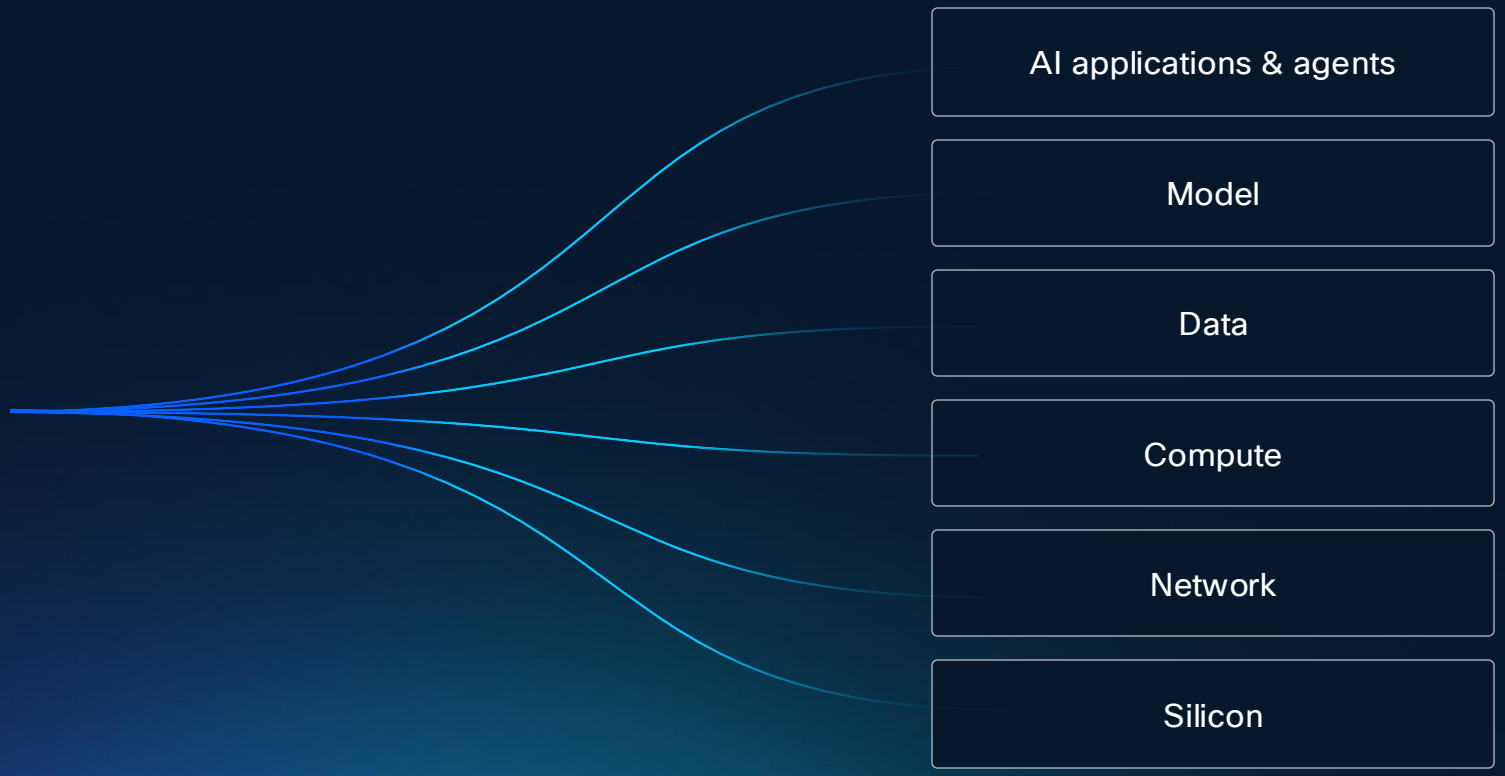


Back to contents

Enhanced Observability Capabilities to Monitor AI



Need for **end-to-end** visibility



Cisco Differentiation

End to End Visibility across the Secure AI Factory Layers

Observability
for AI

AI applications & agents

Model

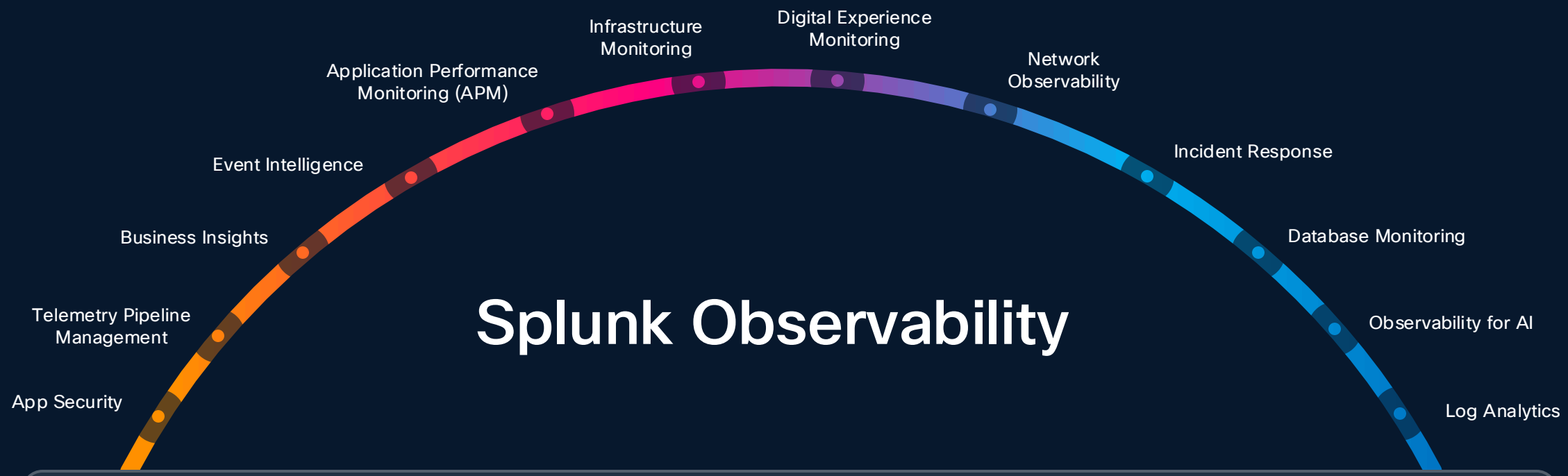
Data

Compute

Network

Silicon

Safety & security
for AI



Splunk Platform

Observability Must Evolve

Fix and prevent with AI agents

Observe AI agents and infrastructure

Unify observability and show business impact

Extend Cisco AI Factory to Cisco Security + Splunk Observability

This is observability for AI

Complete business visibility across any environment and any stack

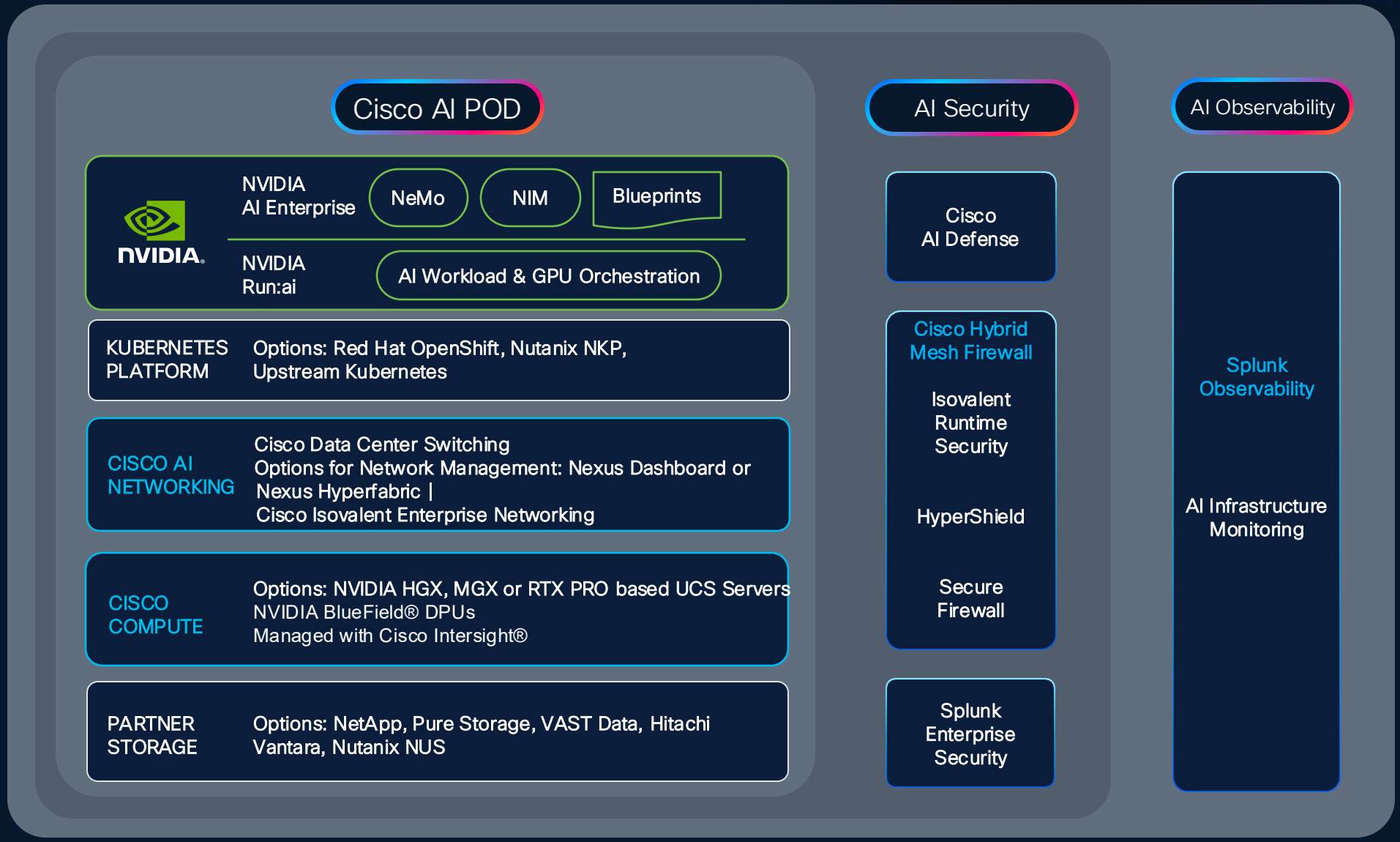
Earlier detection & faster investigation of business-impacting issues

Better control of your data and costs

Observability for AI

cisco

Coverage of Observability for Cisco Secure AI Factory



How Does It All Come Together?

Splunk Observability provides visibility into:

UCS | NEXUS | HyperShield | Intersight

Cisco UCS Servers: CPU, memory, storage, and GPU utilization
Cisco Nexus Switches: Network traffic, latency, and connectivity for AI workloads
Cisco HyperShield: Storage performance and capacity for data-heavy applications
Cisco Intersight: Centralized operational health and configuration insights

End-to-End

Splunk connects to Cisco infrastructure via APIs and telemetry, pulling data from UCS, Nexus, HyperShield, and Intersight

Real-Time Insights: Detect GPU underutilization or misalignment before it impacts performance
Actionable Intelligence: Transform metrics, traces, and logs into alerts and dashboards for rapid troubleshooting
Unified View: Combine hardware, network, storage, and application telemetry for holistic observability

Splunk Observability

Improve ops efficiency, reliability and insights

OpenTelemetry Native

Own your data, avoid lock-in, and build on one common standard.

AI powered analytics and guidance

AI/ML features like Service Maps and Trace Analytics guide faster issue resolution.

No data sampling

Eliminate blind spots with Splunk NoSample™ tracing that captures 100% of your data.

Digital Resilience for AI Infrastructure

Secure Application On Splunk Observability

Improve ops SLA compliance and risk mitigation

Runtime Libraries and Vulnerabilities

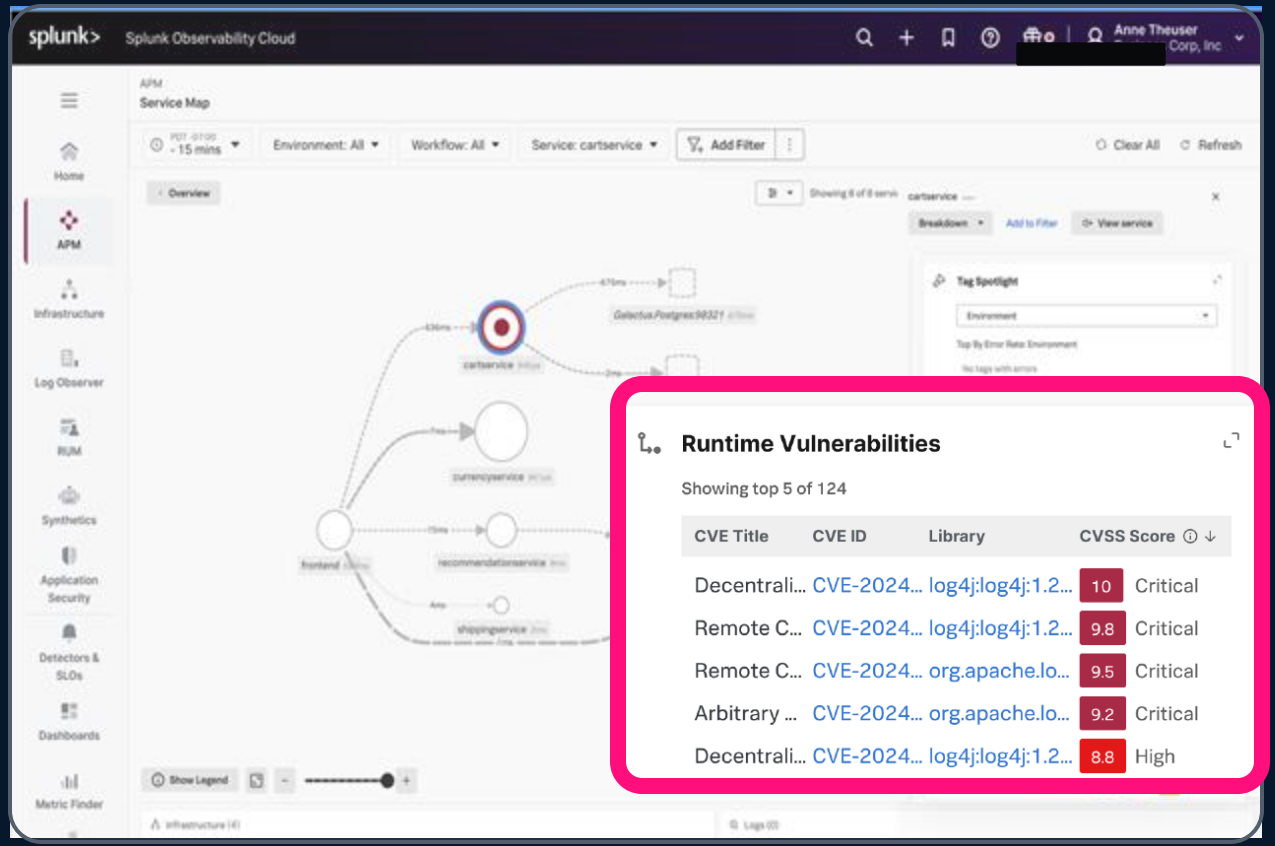
Discover open-source libraries and detect vulnerabilities associated with them.

Prioritized Risk Mitigation based on Cisco Security Intel

Proactively mitigate risk and enhance SLA compliance with integrated Cisco Security intel.

No extra agents

Security integrated into fabric of observability. Delivered using OpenTelemetry instrumentation.



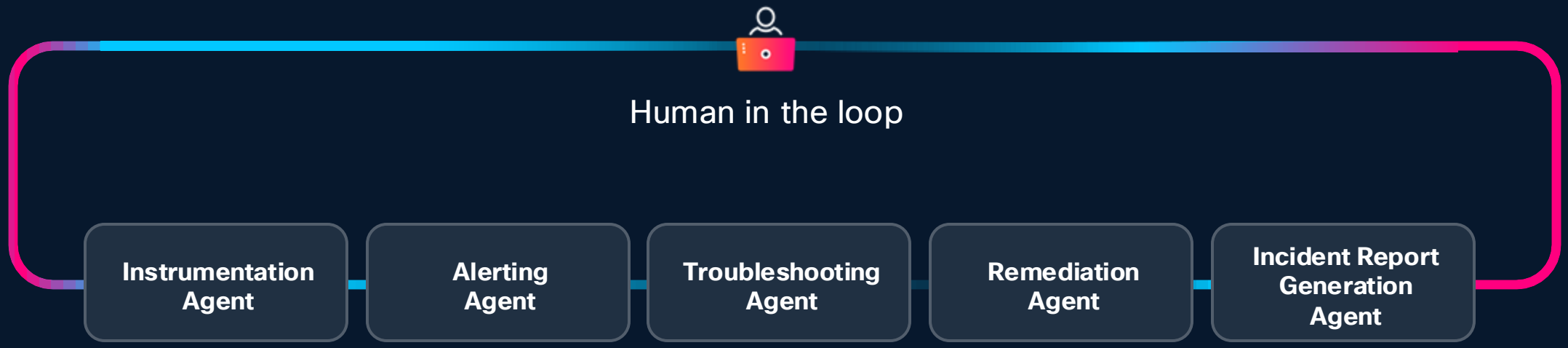
Agentic Observability

Fix and prevent with AI agents

Observe AI agents and infrastructure

Unify observability and show business impact

Fix and Prevent with AI Agents



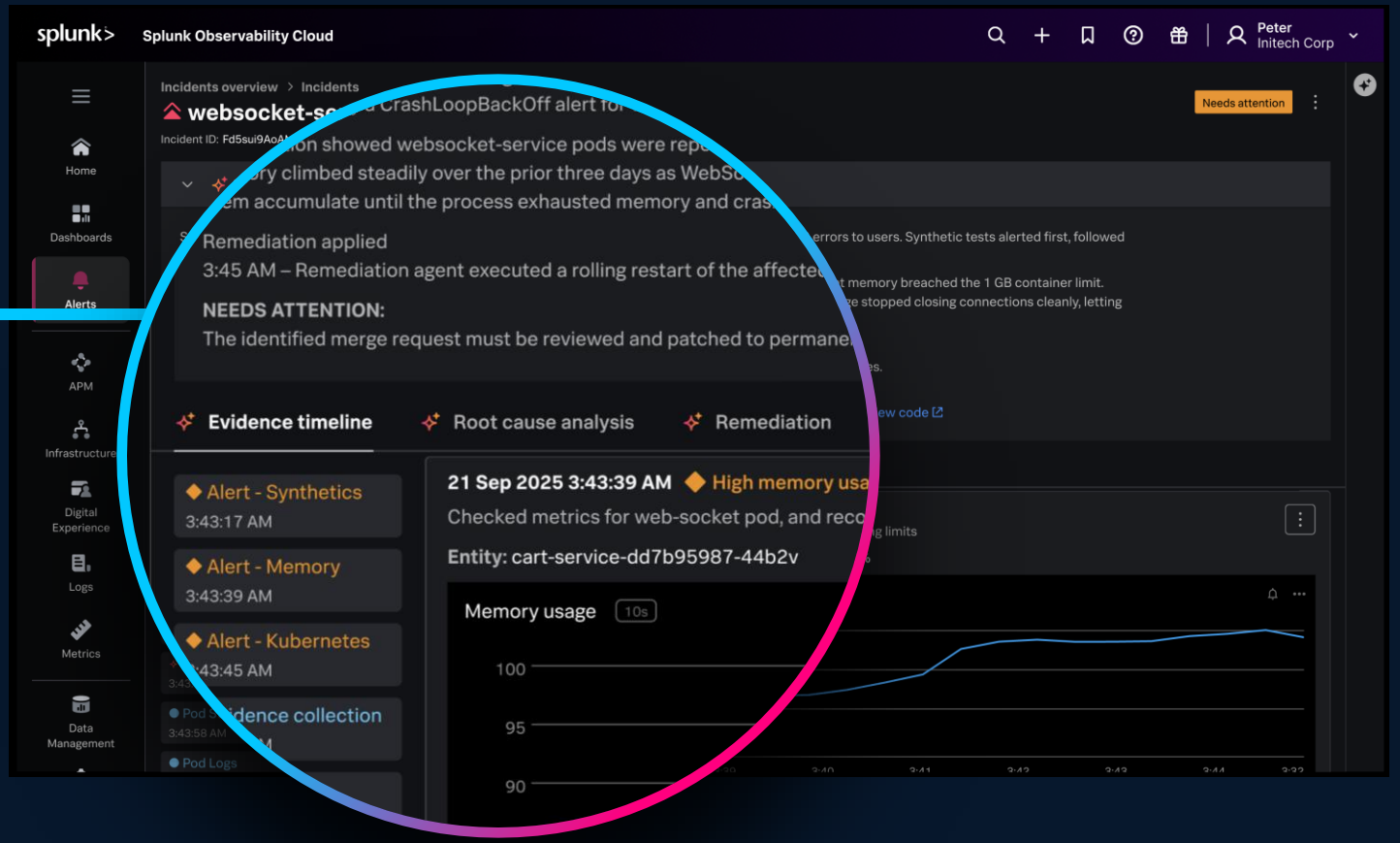
Alpha | Dec ‘25

AI Troubleshooting Agent in Observability Cloud

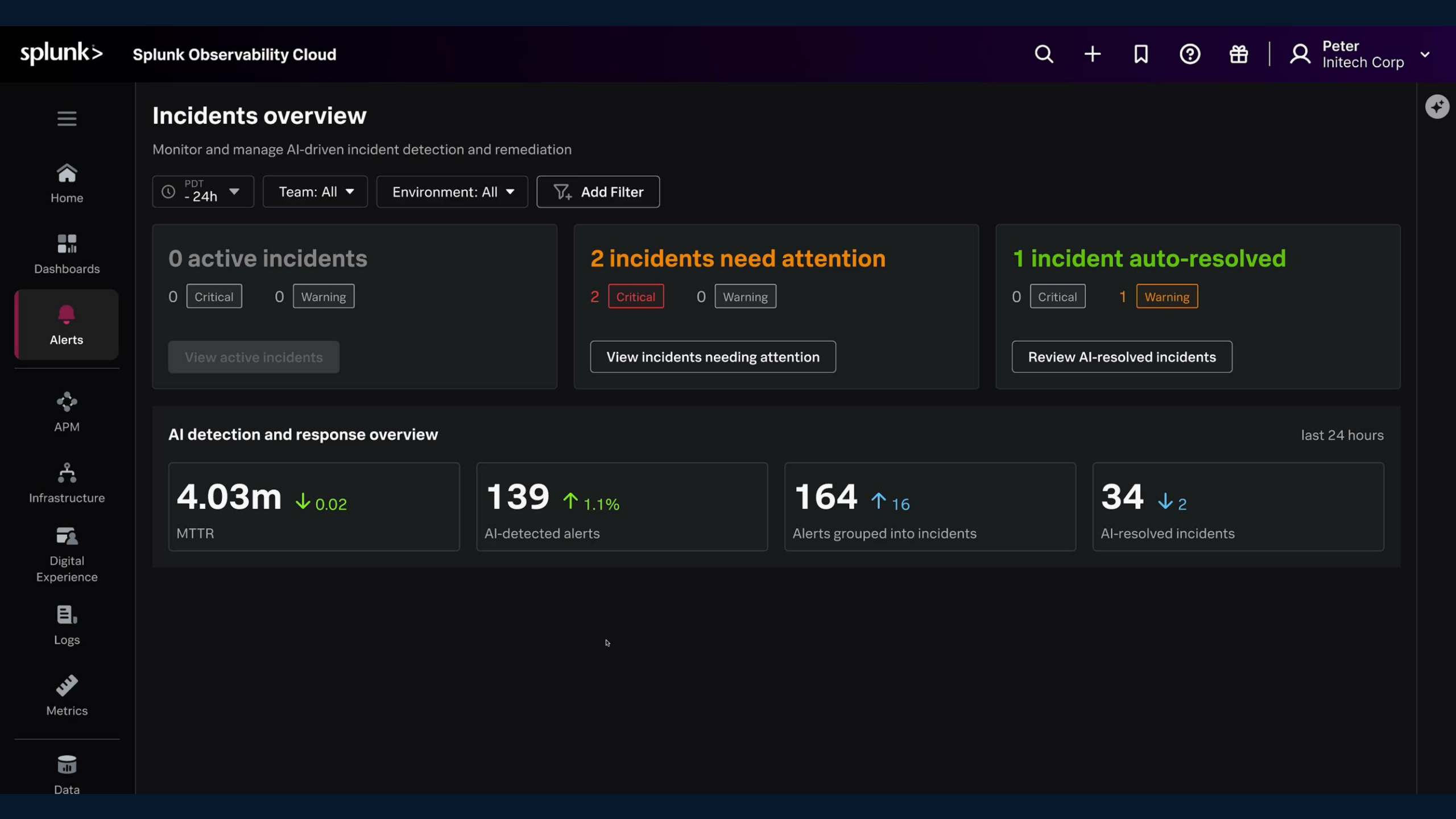
Proactively fix issues with AI agents

Pull metrics, logs, traces to root cause

Make everyone an expert



Demo


splunk> Splunk Observability Cloud
Peter
Initech Corp
Incidents overview
Monitor and manage AI-driven incident detection and remediation
PDT
- 24h
Team: All
Environment: All
Add Filter
0 active incidents
0 Critical
0 Warning
View active incidents
2 incidents need attention
2 Critical
0 Warning
View incidents needing attention
1 incident auto-resolved
0 Critical
1 Warning
Review AI-resolved incidents
AI detection and response overview
last 24 hours
4.03m ↓ 0.02
MTTR
139 ↑ 1.1%
AI-detected alerts
164 ↑ 16
Alerts grouped into incidents
34 ↓ 2
AI-resolved incidents
Home
Dashboards
Alerts
APM
Infrastructure
Digital Experience
Logs
Metrics
Data

Agentic Observability

Fix and prevent with AI agents

Observe AI agents and infrastructure

Unify observability and show business impact

ALPHA
November '25

AI Agent Monitoring

Monitor the quality, accuracy, security, and cost of LLM and agentic applications

Pinpoint the root cause of accuracy, efficiency, and bias issues

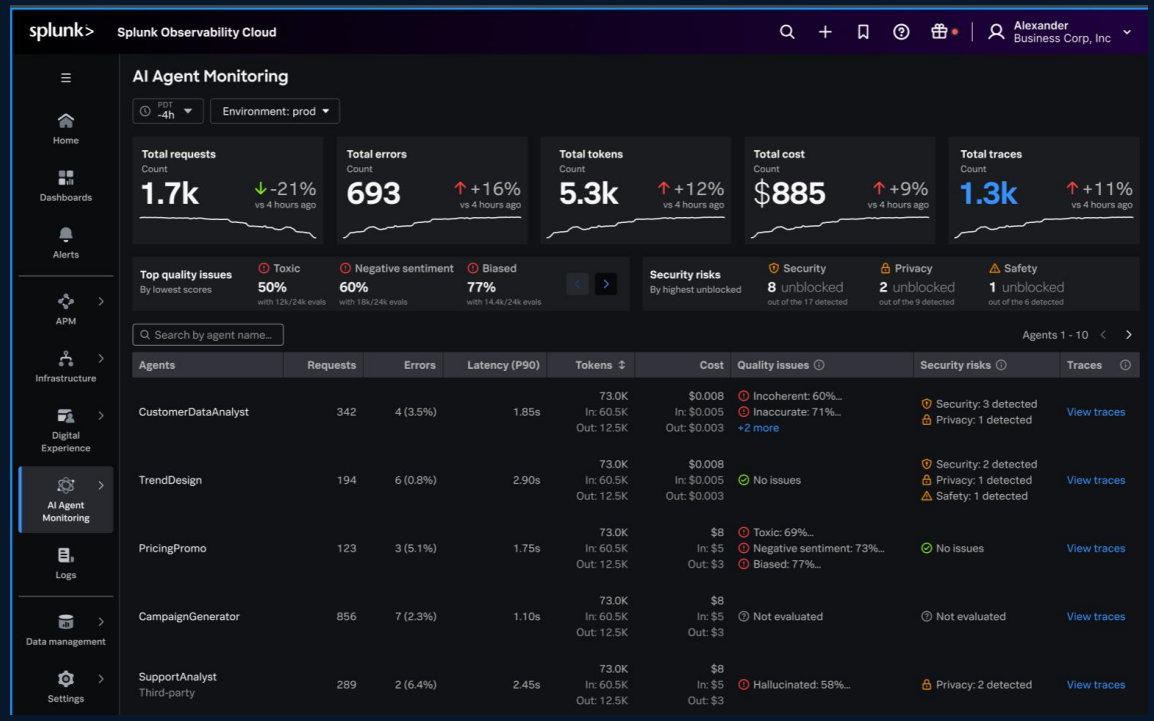
Reduce LLM chain errors and failures to build responsible and reliable AI, increase customer trust and satisfaction, and reduce costs

Visualize AI components and services to improve business performance

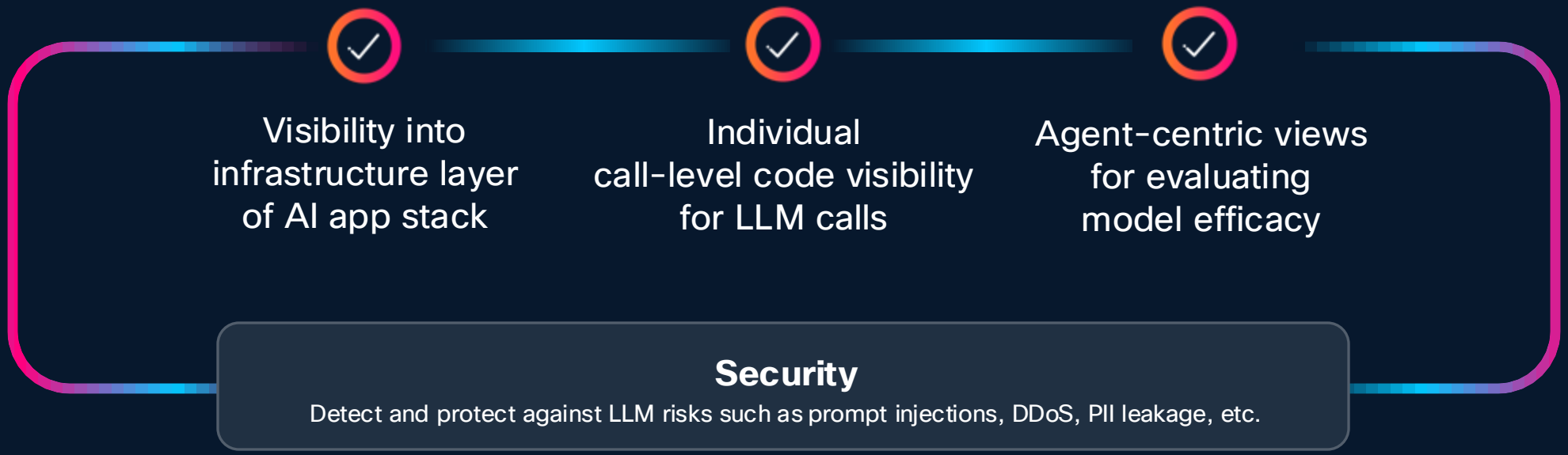
Trace and map dependencies to analyze LLM and agentic workflows, token utilization, latency, errors, and other service level objectives (SLOs)

Measure, secure, and improve user experiences with LLMs

Reduce model and data drift by identifying unusual activities, patterns, outliers, behaviors, hallucinations, and AI security risks (with Cisco AI Defense) in user interactions with LLM agents and models



Observe AI Agents



Demo

Home
Dashboards
Alerts
APM
Infrastructure
Digital Experience
AI Agent Monitoring
Logs

AI Agent Monitoring

PDT -4h | Environment: prod | Application: FlashSalesApp

Agents

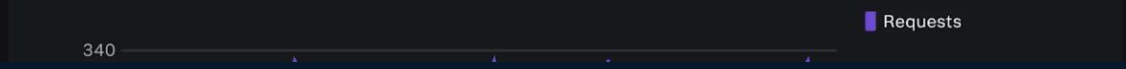
Search...

Prev Next Showing 5 of 17

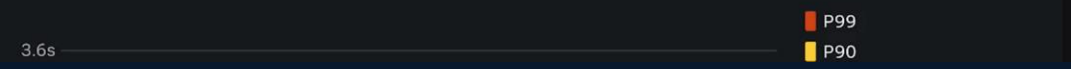
Agent	Requests	Latency (P90)	Errors	Quality	Security Risks	Tokens	Conversations
CustomerDataAnalyst	342	1.85s	4 (3.5%)	Healthy	No risk	↑ 847K	57
TrendDesign	194	2.90s	6 (0.8%)	Healthy	No risk	↑ 623K	89
PricingPromo	123	1.75s	3 (5.1%)	Healthy	No risk	↓ 445K	120
CampaignGenerator	856	1.10s	7 (2.3%)	Unhealthy	No risk	↓ 298K	125
SupportAgent Third-party	289	2.45s	2 (6.4%)	Healthy	No risk	↑ 267K	248

Performance

Requests



Latency



Agentic Observability

Fix and prevent with AI agents

Observe AI agents and infrastructure

Unify observability and show business impact

GA September '25

AI Infrastructure Monitoring

Monitor the health, availability, and consumption of AI infrastructure

Optimize resource utilization and identify noisy neighbors

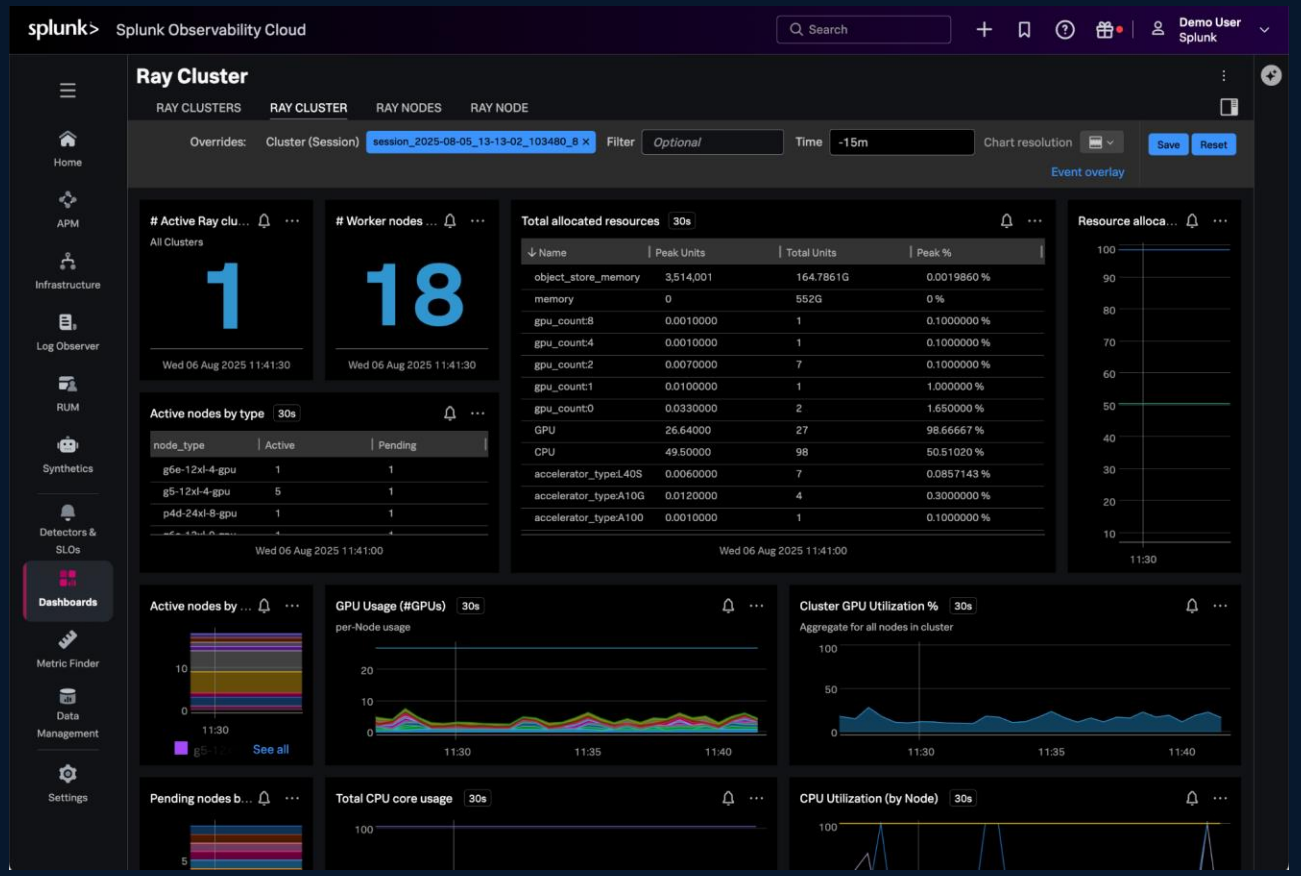
Track operational metrics from AI apps and services to manage costs and detect performance degradation

Visualize the status and understand the business impact of AI components

Isolate problems in transaction queues, locks, and threads, and correlate them with business health, usage trends, patterns, and outliers

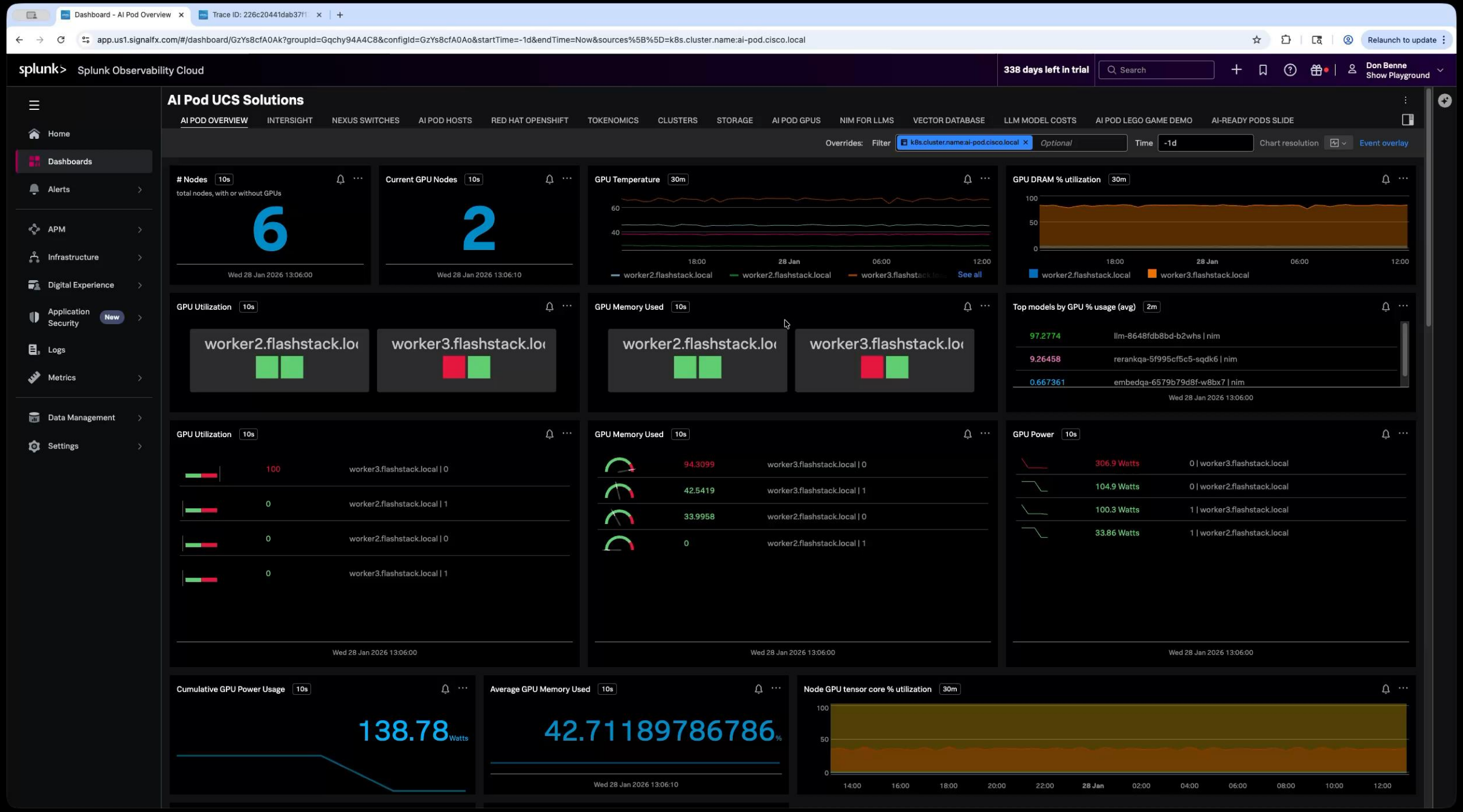
Proactively alert on bottlenecks and spikes across services

View data-dense dashboards and detectors for individual and aggregate frameworks, agents, model providers, vector databases, GPUs, and more

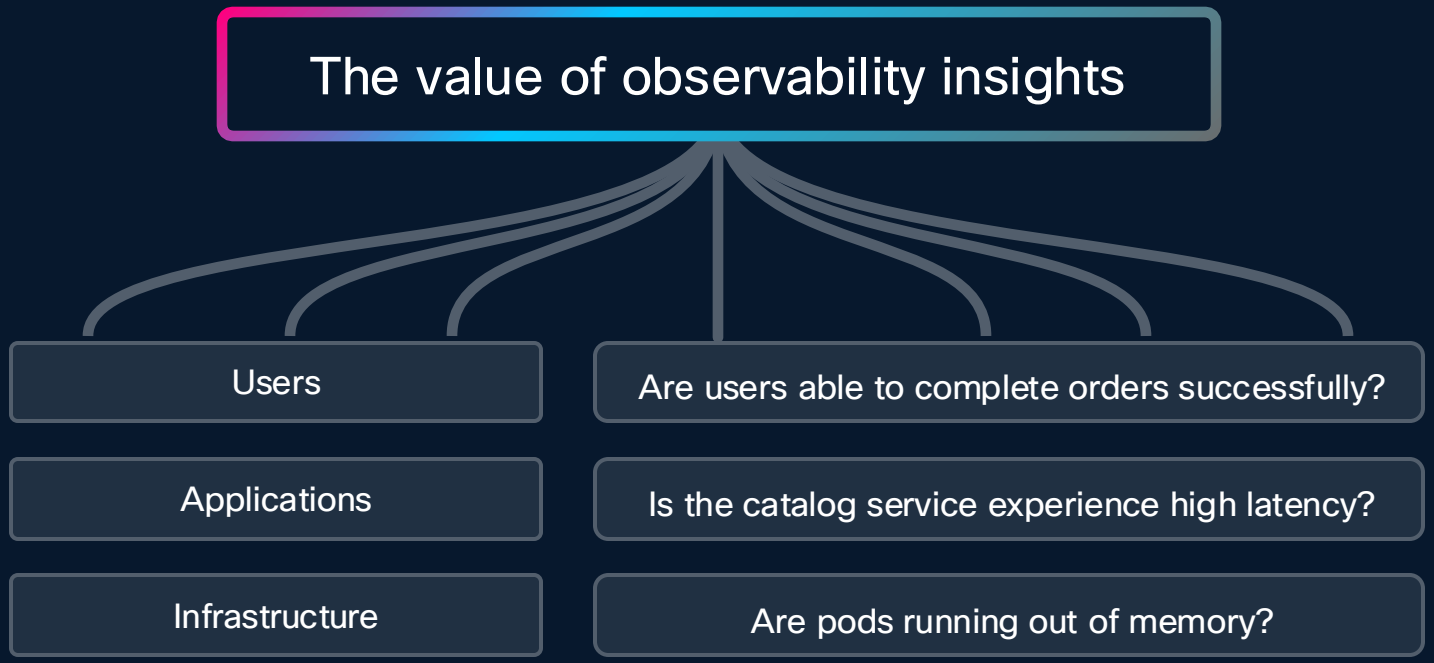


Presenter: Joe deBlaquiere

cisco


Dashboard - AI Pod Overview
Trace ID: 226c20441dab37f1
app.us1.signalfx.com/#/dashboard/GzYs8cfA0Ak?groupId=Gqchy94A4C8&configId=GzYs8cfA0Ao&startTime=-1d&endTime=Now&sources%5B%5D=k8s.cluster.name:ai-pod.cisco.local
Relaunch to update
splunk> Splunk Observability Cloud
338 days left in trial
Search
Don Benne
Show Playground
Home
Dashboards
Alerts
APM
Infrastructure
Digital Experience
Application Security
New
Logs
Metrics
Data Management
Settings
AI Pod UCS Solutions
AI POD OVERVIEW
INTERSIGHT
NEXUS SWITCHES
AI POD HOSTS
RED HAT OPENSHIFT
TOKENOMICS
CLUSTERS
STORAGE
AI POD GPUS
NIM FOR LLMS
VECTOR DATABASE
LLM MODEL COSTS
AI POD LEGO GAME DEMO
AI-READY PODS SLIDE
Overrides:
Filter
k8s.cluster.name:ai-pod.cisco.local
Optional
Time
-1d
Chart resolution
Event overlay
Nodes
10s
total nodes, with or without GPUs
6
Wed 28 Jan 2026 13:06:00
Current GPU Nodes
10s
2
Wed 28 Jan 2026 13:06:10
GPU Temperature
30m
60
40
18:00
28 Jan
06:00
12:00
worker2.flashstack.local
worker2.flashstack.local
worker3.flashstack.local
See all
GPU DRAM % utilization
30m
100
50
0
18:00
28 Jan
06:00
12:00
worker2.flashstack.local
worker3.flashstack.local
GPU Utilization
10s
worker2.flashstack.lo
worker3.flashstack.lo
GPU Memory Used
10s
worker2.flashstack.lo
worker3.flashstack.lo
Top models by GPU % usage (avg)
2m
97.2774
llm-8648fdb8bd-b2whs | nim
9.26458
rerankqa-5f995cf5c5-sqdk6 | nim
0.667361
embedqa-6579b79d8f-w8bx7 | nim
Wed 28 Jan 2026 13:06:00
GPU Utilization
10s
100
worker3.flashstack.local | 0
0
worker2.flashstack.local | 1
0
worker2.flashstack.local | 0
0
worker3.flashstack.local | 1
Wed 28 Jan 2026 13:06:00
GPU Memory Used
10s
94.3099
worker3.flashstack.local | 0
42.5419
worker3.flashstack.local | 1
33.9958
worker2.flashstack.local | 0
0
worker2.flashstack.local | 1
Wed 28 Jan 2026 13:06:00
GPU Power
10s
306.9 Watts
0 | worker3.flashstack.local
104.9 Watts
0 | worker2.flashstack.local
100.3 Watts
1 | worker3.flashstack.local
33.86 Watts
1 | worker2.flashstack.local
Wed 28 Jan 2026 13:06:00
Cumulative GPU Power Usage
10s
138.78 Watts
Average GPU Memory Used
10s
42.71189786786 %
Wed 28 Jan 2026 13:06:10
Node GPU tensor core % utilization
30m
100
50
0
14:00
16:00
18:00
20:00
22:00
28 Jan
02:00
04:00
06:00
08:00
10:00
12:00

Unify Observability and Show Business Impact



Splunk Observability

Meeting customers where they are

Traditional environments

Unified observability experience

- APM
- Infrastructure monitoring
- Digital experience monitoring
- Business insights
- Application security
- Observability for AI
- Network observability

Splunk Platform

- Federation
- Log analytics
- Event intelligence

Cloud native environments

GA

Splunk ITSI & Cisco Enterprise Networking

Enterprise Network Monitoring for branch & campus to quickly pinpoint site & device issues in Cisco networks

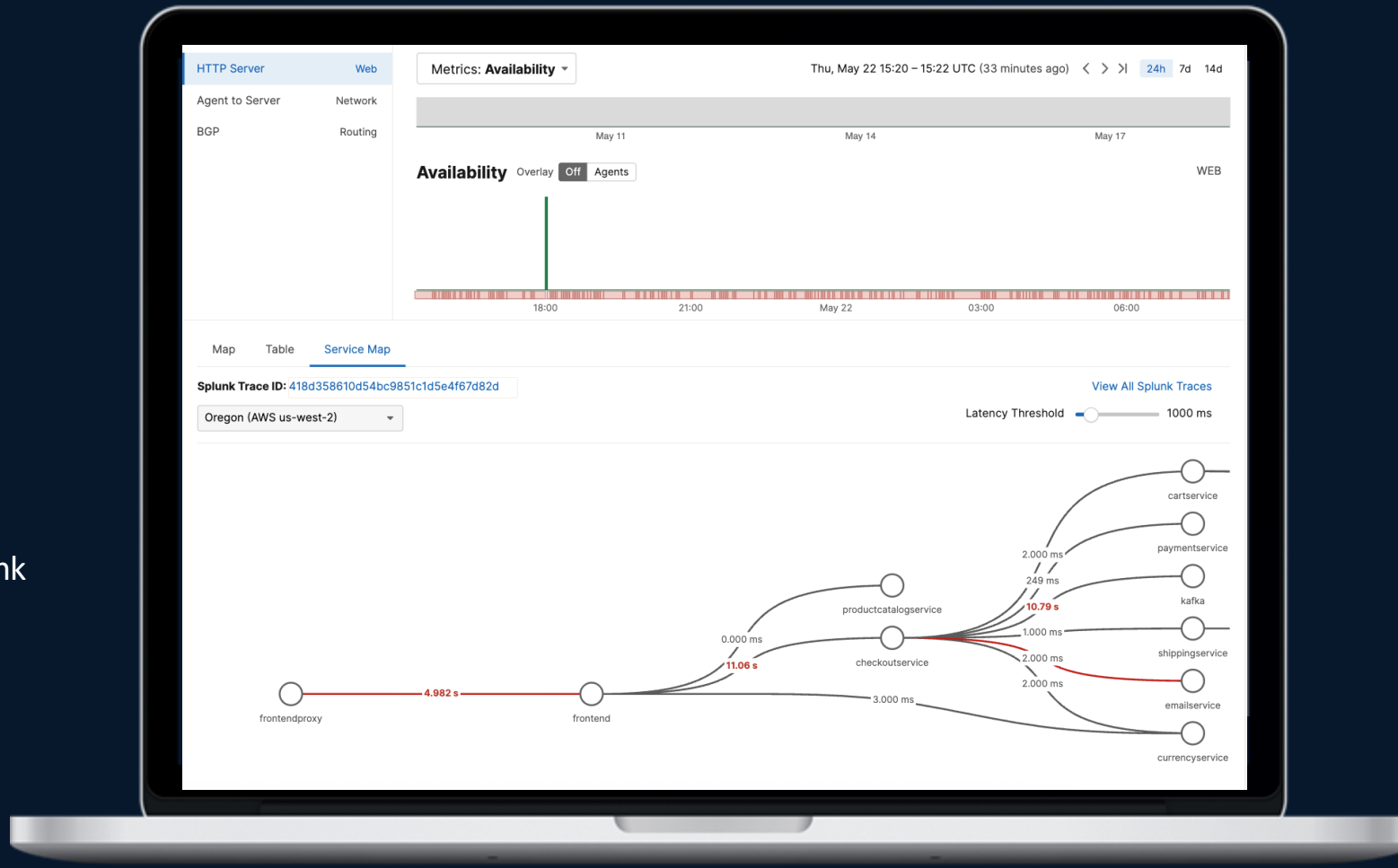
- ITSI content packs for Catalyst Center and Meraki
- Cross-domain correlation for reduced alert noise and domain isolation
- Out-of-the-box topology to measure the health of a location (e.g. retail store) and isolate problematic devices
- Device alert import, normalization, deduplication, and correlation logic
- Insights for problem troubleshooting (e.g. recent configuration changes)
- In-context guidance into Catalyst Center & Meraki to take action on devices

GA

Splunk Observability Cloud APM & ThousandEyes

End-to-end visibility across ThousandEyes tests and Splunk APM traces for faster MTTI & MTTR

- Quickly diagnose failing ThousandEyes test
- Understand if slowness is from the network or application layer
- View trace topology and key intra-service metrics in context of the ThousandEyes test
- Easily set up tests for services instrumented in Splunk Observability Cloud from the ThousandEyes UI

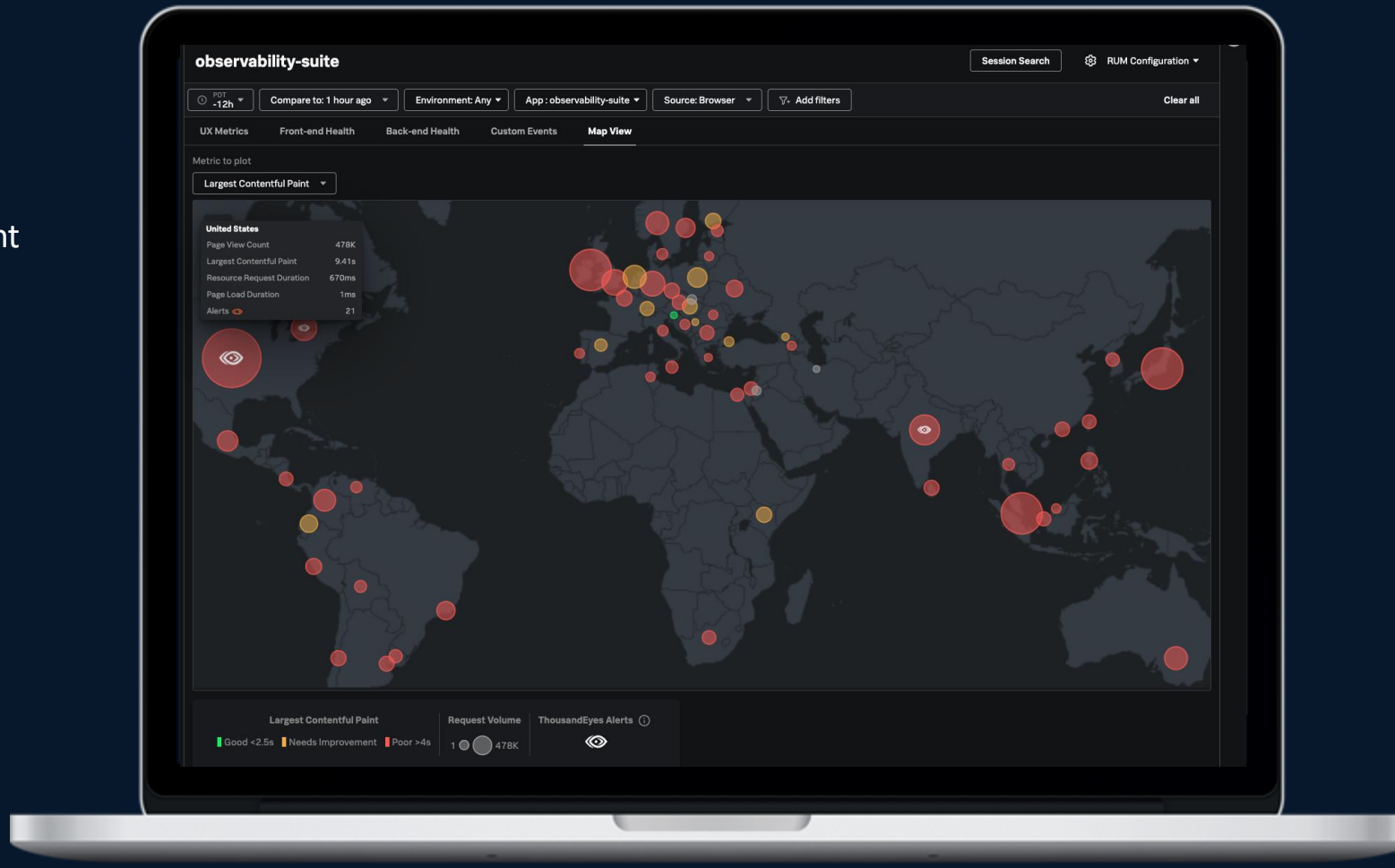


Alpha, GA in Oct 2025

Splunk Observability Cloud RUM & ThousandEyes

End-to-end visibility from user to network

- Visualize global user experience and quickly highlight impacted regions with Geo Maps
- Overlay real user monitoring and ThousandEyes network data to pinpoint the problem domain
- Troubleshoot deeper with seamless drill-downs



Appendix

CISCO Connect

Thank you


CISCO