Monitor your LLMs, AI Agents, and AI Infrastructure with Splunk Observability for AI



Dayna Lord
Principal Product Marketing Manager
dlord@cisco.com

Joe deBlaquiere
Principal Product Manager
jdeblaqu@cisco.com

Kashyap Merchant
Product Management Leader
kamercha@cisco.com

Today’s Presenters



Dayna Lord
Principal Product Marketing Manager



Joe deBlaquiere
Principal Product Manager



Kashyap Merchant
Product Management Leader

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Cisco Webex App

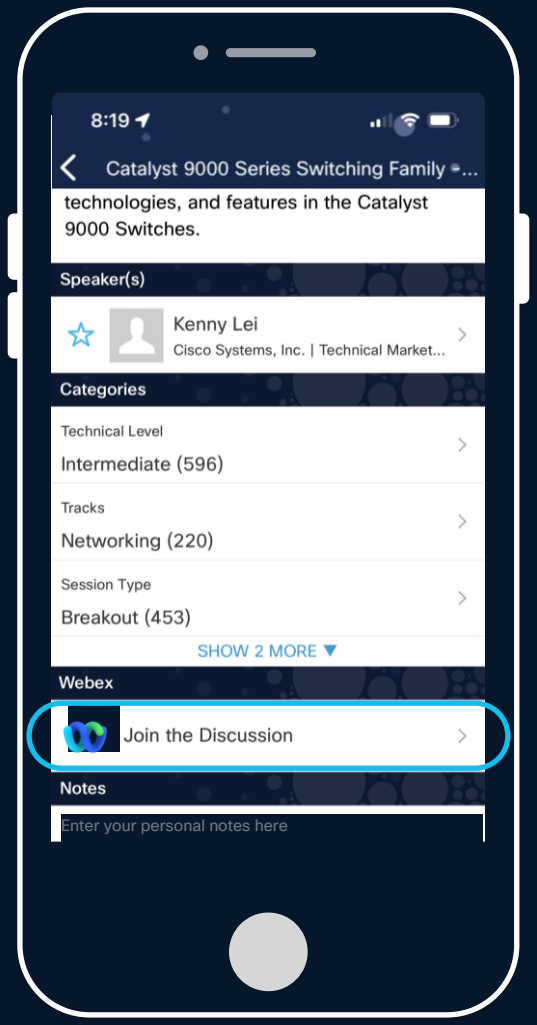
Questions?

Use Cisco Webex App to chat
with the speaker after the session

How

1. Find this session in the Cisco Live Mobile App
2. Click “Join the Discussion”
3. Install the Webex App or go directly to the Webex space
4. Enter messages/questions in the Webex space

Webex spaces will be moderated by the speaker until June 13, 2025.



https://ciscolive.ciscoevents.com/
ciscolivebot/**#BRKOBS-2013**

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Forward-looking statements

This presentation may contain forward-looking statements that are subject to the safe harbors created under the Securities Act of 1933, as amended, and the Securities Exchange Act of 1934, as amended. All statements other than statements of historical facts are statements that could be deemed forward-looking statements. These statements are based on current expectations, estimates, forecasts, and projections about the industries in which we operate and the beliefs and assumptions of our management based on the information currently available to us. Words such as “expects,” “anticipates,” “targets,” “goals,” “projects,” “intends,” “plans,” “believes,” “momentum,” “seeks,” “estimates,” “continues,” “endeavors,” “strives,” “may,” variations of such words, and similar expressions are intended to identify such forward-looking statements. In addition, any statements that refer to (1) our goals, commitments, and programs; (2) our business plans, initiatives, and objectives; and (3) our assumptions and expectations, including our expectations regarding our financial performance, products, technology, strategy, customers, markets, acquisitions and investments are forward-looking statements. These forward-looking statements are not guarantees of future performance and involve significant risks, uncertainties and other factors that may cause our actual results, performance or achievements to be materially different from results, performance or achievements expressed or implied by the forward-looking statements contained in this presentation. Readers are cautioned that these forward-looking statements are only predictions and are subject to risks, uncertainties, and assumptions that are difficult to predict, including those identified in the “Risk Factors” section of Cisco’s most recent report on Form 10-Q filed on February 20, 2024 and its most recent report on Form 10-K filed on September 7, 2023, as well as the “Risk Factors” section of Splunk’s most recent report on Form 10-Q filed with the SEC on November 28, 2023. The forward-looking statements made in this presentation are made as of the time and date of this presentation. If reviewed after the initial presentation, even if made available by Cisco or Splunk, on Cisco or Splunk’s website or otherwise, it may not contain current or accurate information. Cisco and Splunk undertake no obligation to revise or update any forward-looking statements for any reason, except as required by law.

In addition, any information about new products, features, functionality or our roadmap outlines our general product direction and is subject to change at any time without notice. It is for informational purposes only and shall not be incorporated into any contract or other commitment or be relied upon in making a purchasing decision. We undertake no commitment, promise or obligation either to develop the features or functionalities described, in beta or in preview (used interchangeably), or to include any such feature or functionality in a future release. The development, release, and timing of any features or functionality described for our products remains at our sole discretion.

Splunk, Splunk>, Data-to-Everything, and Turn Data Into Doing are trademarks or registered trademarks of Splunk Inc. in the United States and other countries. All other brand names, product names, or trademarks belong to their respective owners.
© 2026 Splunk LLC. All rights reserved.

AI is everywhere

AI is enhancing automation and efficiency, reducing toil, and improving decision making

Face ID

Chatbots

Content Recommendations

Anomaly Detection

AI SRE Assistants

Dynamic Pricing

Coding Assistants

Image Recognition

Fraud Detection

Predictive Text

Route optimization

Outreach Agents

Targeted Ads

Revenue Forecasting

Incident Summarizers

Predictive Maintenance

Campaign Optimization

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

But AI also gives rise to slop

Inauthentic, inaccurate, low quality, and sometimes harmful outputs can arise

False Rejections

Unhelpful Responses

Engagement Bait

Alert Fatigue

Irrelevant Runbooks

Price Gouging

Overengineered Code

Incorrect Labels

False Positives

Poor Suggestions

Worse Routes

Noisy Spam

Misplaced Targeting

Inaccurate Projections

Shallow Postmortems

Ignored Warnings

Vanity Metrics

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

This is because AI is unreliable and unpredictable

Unintended or unexpected behavior can arise

Non-deterministic

Hallucination-prone



Data-dependent Behavior

Context fragile

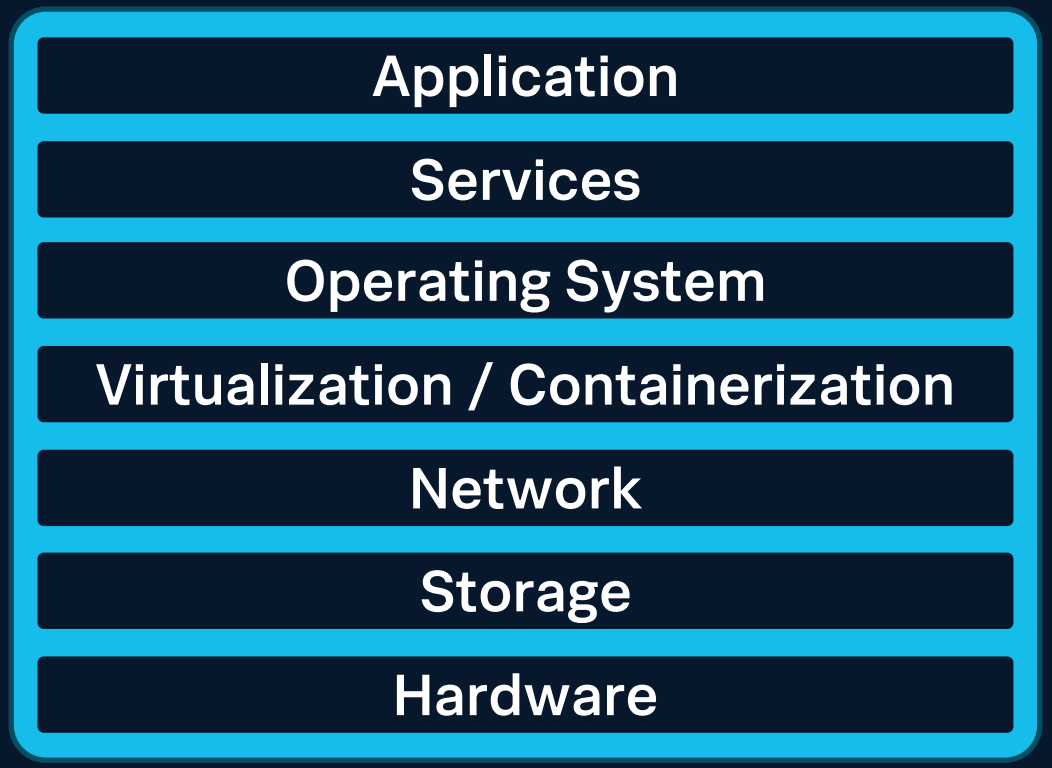
© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Additionally, traditional infrastructure has transformed

More layers and complexity is required to manage AI workloads and resource utilization

Traditional Infrastructure Monitoring

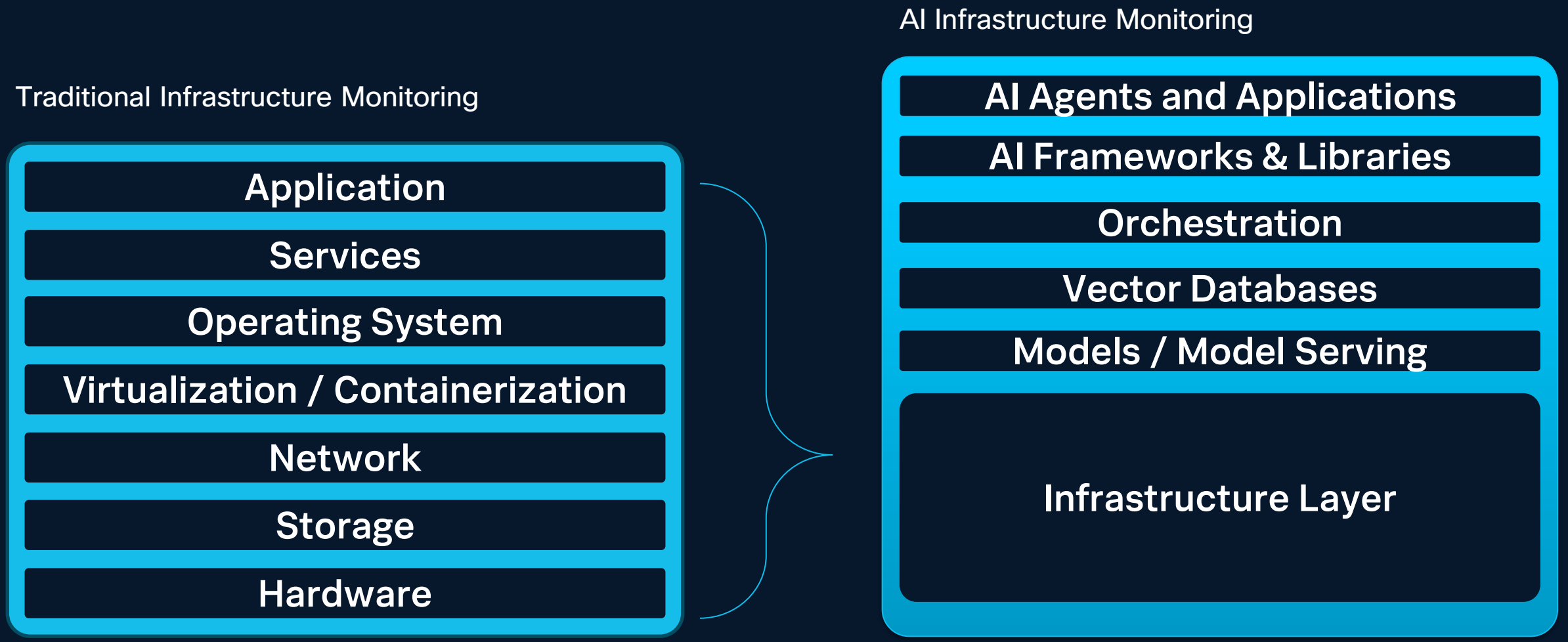


© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Traditional infrastructure has transformed

More layers and complexity is required to manage AI workloads and resource utilization

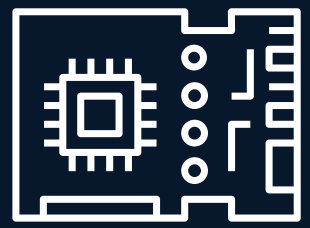


© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

This means AI can be costly

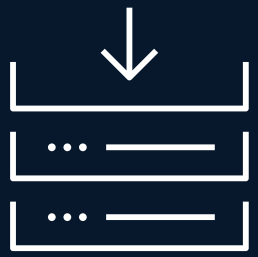
More GPUs, memory, storage, etc. across infrastructure is needed



More GPUs



More Memory



More Storage



More Network

Etc….

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Finally, applications have evolved

From simple and deterministic to complex and nondeterministic or adaptive

Pre-AI Period

Deterministic logic
Predictable, rule-based behavior
Static query / response
Monolithic

Agentic

Autonomous task completion
Unpredictable behavior
Multi-step actions
Microservices



© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

This increases compliance, legal, and security risks



Financial Damage



Litigation Risk



Reputational Damage



Compliance Risk



Security Risk



IP Leakage

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Therefore, digital resilience is being tested with AI

New questions are being asked to find out if observability and security tools can keep up



Is my AI application working as expected?



Is my infrastructure or component healthy?



Is my model trustworthy and operating effectively?

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Therefore traditional observability is not going to cut it

Customers need...

Trust



Security



Transparency



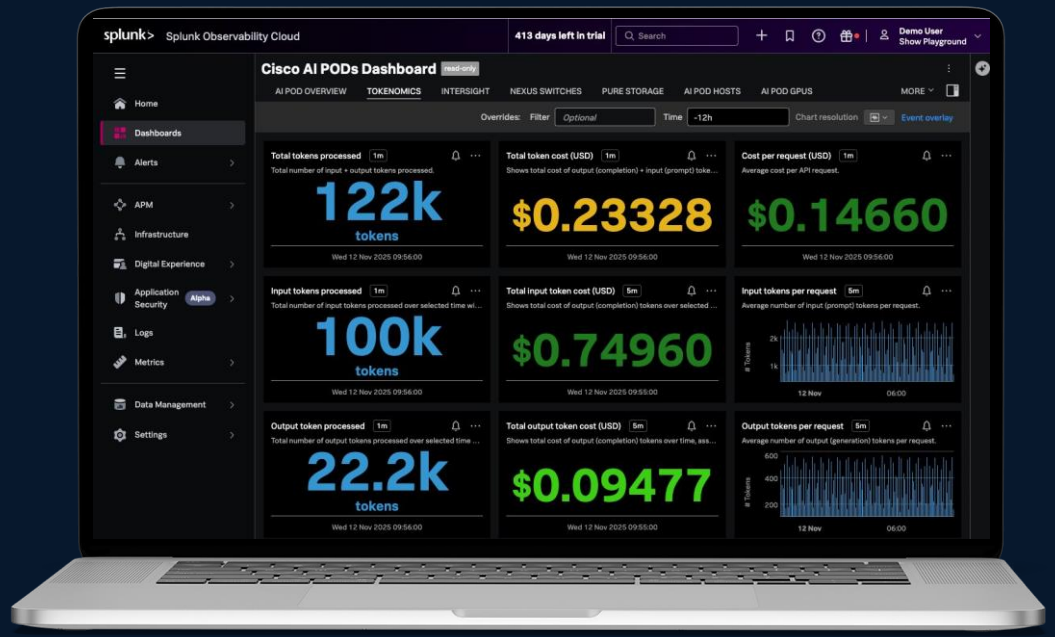
To build and deploy responsible AI with great customer experiences

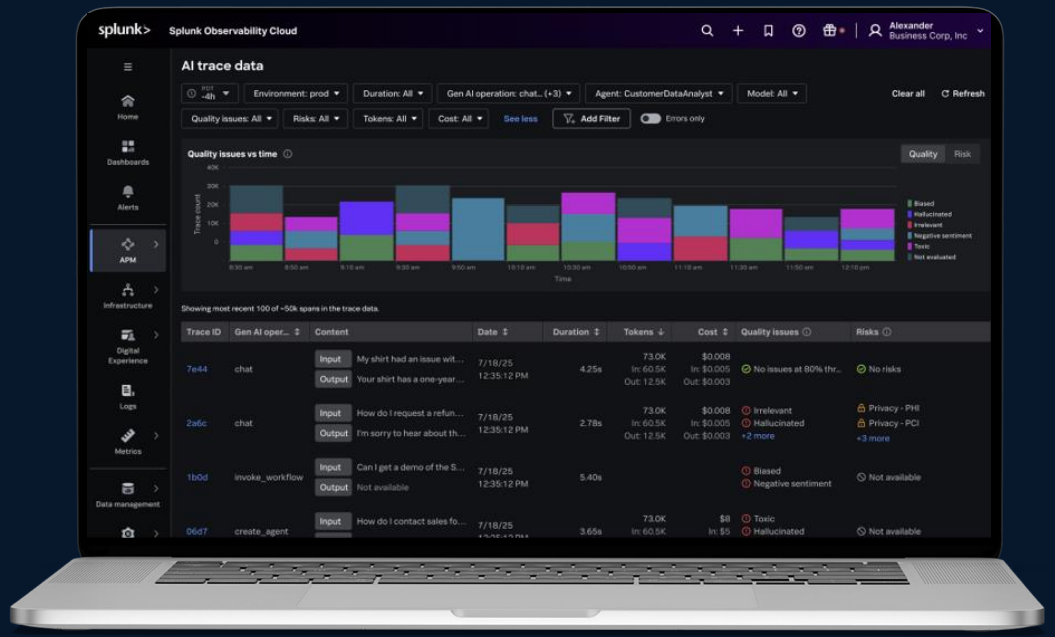
© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Splunk Observability for AI

Monitor the performance, quality, security, and cost of your AI application stack





AI Infrastructure Monitoring (GA)

Monitor the health, availability, and consumption (or usage) of AI infrastructure.

AI Agent Monitoring (GA)

Monitor the performance, quality, security, and cost of LLM and agentic applications.

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Security

The AI threat landscape is more sophisticated than ever

Examples of increased misuse, drifts, threats, leakage, and threats

Prompt injection



Personally identifiable information (PII) leakage



Data and model poisoning



Misinformation



Model drift

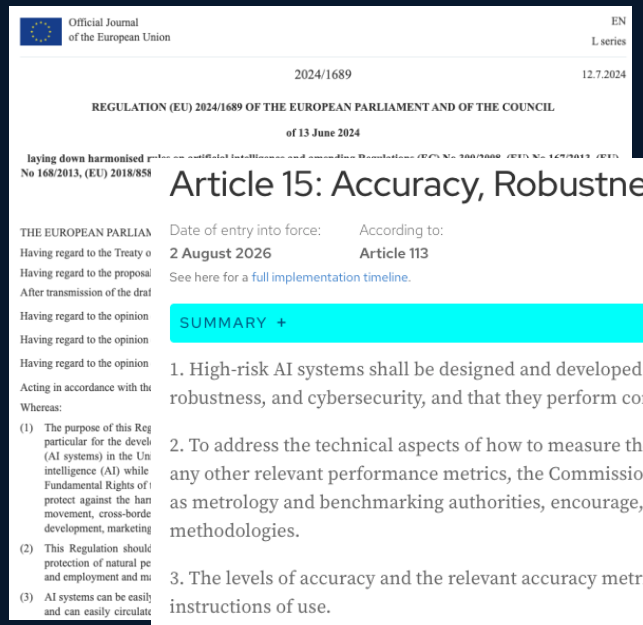


© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Demo Time!

And there is more regulation coming up...


Official Journal of the European Union

EN

L series

2024/1689

12.7.2024

REGULATION (EU) 2024/1689 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 13 June 2024

laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858

THE EUROPEAN PARLIAM

Having regard to the Treaty o

Having regard to the proposal

After transmission of the draf

Having regard to the opinion

Having regard to the opinion

Having regard to the opinion

Acting in accordance with the

Whereas:

(1) The purpose of this Reg particular for the devel (AI systems) in the Un intelligence (AI) while Fundamental Rights of protect against the harr movement, cross-borde development, marketing

(2) This Regulation should protection of natural pe and employment and ma

(3) AI systems can be easily and can easily circulate

Article 15: Accuracy, Robustness and Cybersecurity

Date of entry into force: 2 August 2026

According to: Article 113

See here for a full implementation timeline.

SUMMARY +

1. High-risk AI systems shall be designed and developed in such a way that they achieve an appropriate level of accuracy, robustness, and cybersecurity, and that they perform consistently in those respects throughout their lifecycle.

2. To address the technical aspects of how to measure the appropriate levels of accuracy and robustness set out in paragraph 1 and any other relevant performance metrics, the Commission shall, in cooperation with relevant stakeholders and organisations such as metrology and benchmarking authorities, encourage, as appropriate, the development of benchmarks and measurement methodologies.

3. The levels of accuracy and the relevant accuracy metrics of high-risk AI systems shall be declared in the accompanying instructions of use.

4. High-risk AI systems shall be as resilient as possible regarding errors, faults or inconsistencies that may occur within the system or the environment in which the system operates, in particular due to their interaction with natural persons or other systems. Technical and organisational measures shall be taken in this regard. The robustness of high-risk AI systems may be achieved through technical redundancy solutions, which may include backup or fail-safe plans. High-risk AI systems that continue to learn after being placed on the market or put into service shall be developed in such a way as to eliminate or reduce as far as possible the risk of possibly biased outputs influencing input for future operations (feedback loops), and as to ensure that any such feedback loops are duly addressed with appropriate mitigation measures.

5. High-risk AI systems shall be resilient against attempts by unauthorised third parties to alter their use, outputs or performance by exploiting system vulnerabilities. The technical solutions aiming to ensure the cybersecurity of high-risk AI systems shall be appropriate to the relevant circumstances and the risks. The technical solutions to address AI specific vulnerabilities shall include, where appropriate, measures to prevent, detect, respond to, resolve and control for attacks trying to manipulate the training data set (data poisoning), or pre-trained components used in training (model poisoning), inputs designed to cause the AI model to make a mistake (adversarial examples or model evasion), confidentiality attacks or model flaws.

The EU AI Act 2024

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

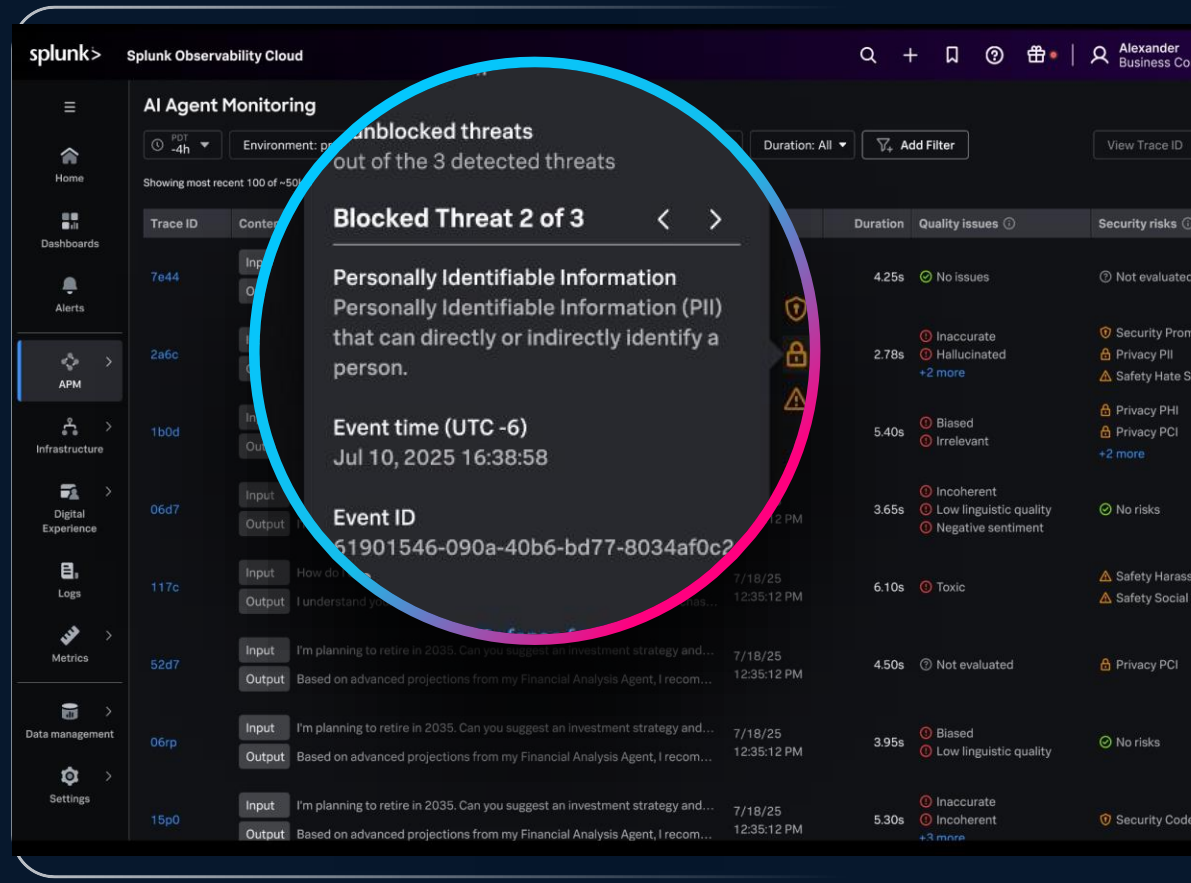
AI Security Monitoring

Detect and mitigate AI risks, misuse, drifts, leakage, and threats with Cisco AI Defense

Pinpoint AI (LLM, agents, and tools) risks and vulnerabilities to minimize exposure

Detect and prevent model misuse, threats, and attacks

Comply with AI security standards to build and deploy trustworthy AI apps and systems



© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Guardrail Categories

Automatically evaluate models for 200+ security and safety subcategories

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



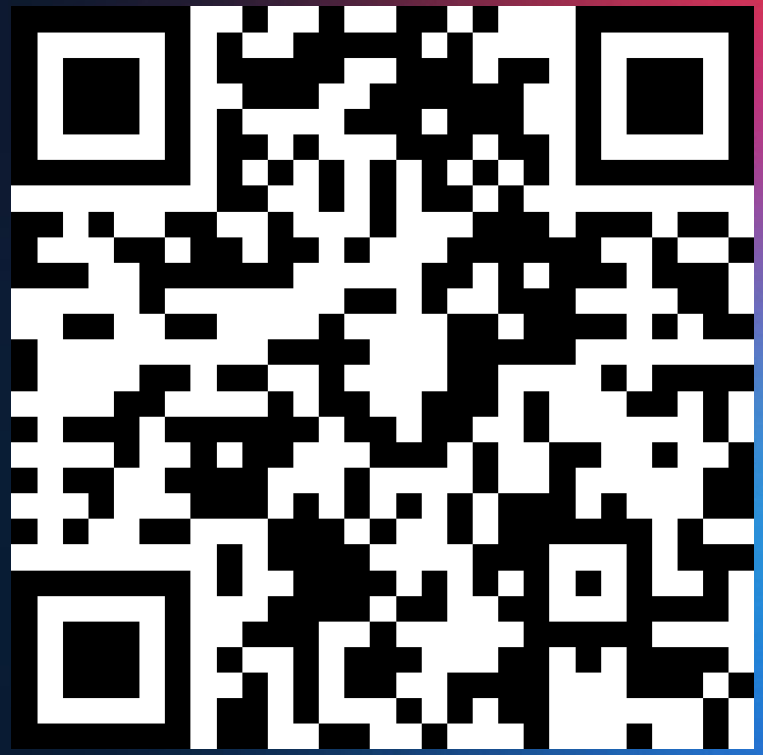
Guardrails can be configured to fit any industry, use case, or preferences

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Slido Poll Question #1

Which AI security risk(s) are you most concerned about in your organization?



Passcode: gwaono

AI Agent Monitoring

Agents and LLMs expand the scope of problems application teams are responsible for

Problems

- Application availability, resulting from error and latency issues
- Dissatisfied customers and legal liabilities from agentic output
- Security issues driven by issues like privacy leakage
- Cost of running agents and models

Capabilities available in Splunk Observability

- Application Performance Monitoring (APM)
- AI Agent Monitoring

Net new problems driven by AI Applications

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

AI Agent Monitoring in Observability Cloud

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

Pinpoint and correlate the root cause of unreliable or degraded AI agent and model performance

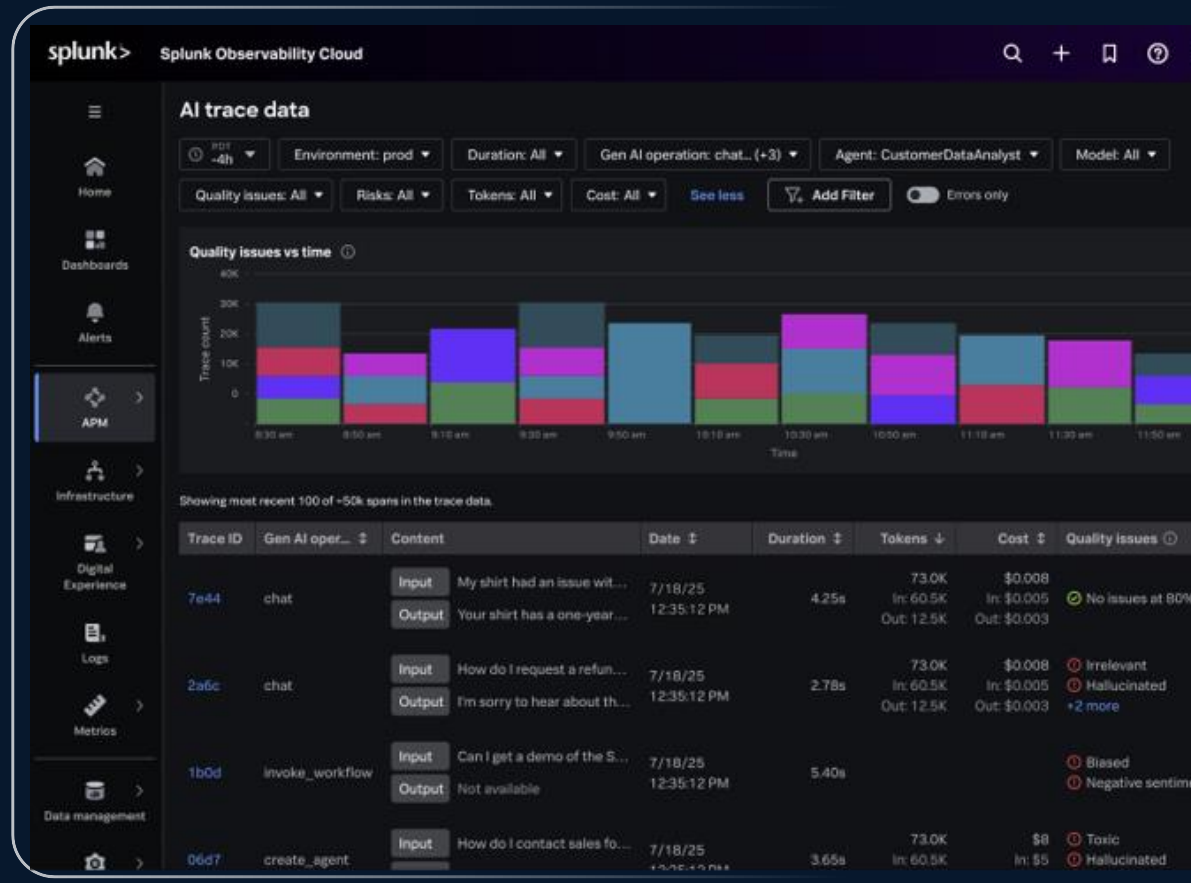
Evaluate responses and detect performance, quality and security metrics like hallucinations, and cost and token usage metrics

Visualize AI agent interactions, tool calls, workflows, and more to correlate with business impact

Trace and map dependencies to analyze LLM and agentic workflows, token utilization, latency, and other SLOs

Measure and improve model quality and behavior to build trust and reliability

Improve user experiences by correlating unusual LLM / agent interactions and behaviors with business impact



© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Key AI Agent Monitoring features

Problems	Features Available in Splunk Observability
Application availability, resulting from error and latency issues	Tracing, service-level monitoring, alerting, dashboards
Dissatisfied customers and legal liabilities from agentic output	Quality scores and evals, agent monitoring and conversation-level details
Security issues driven by issues like privacy leakage	Security evals and guardrails, driven by Cisco AI Defense
Cost of running agents and models	Cost and token usage

Net new problems driven by AI Applications

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

And new metrics are captured beyond traditional APM

Primary APM metrics

Application Performance Metrics

- Requests
- Errors
- Latency / Duration

Quality Metrics

- Toxic Language
- Hallucinations
- Biases
- Sentiment
- Relevance
- and others

Security Metrics

- Prompt Injection
- PII leakage
- Malicious URLs
- and others

Breadth of AI Agent Monitoring metrics

\+ **Cost and Token Usage Metrics**

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Demo Time!

Home
Dashboards
Alerts
APM
Infrastructure
Digital Experience
Application Security
Logs
Metrics

APM Overview

EST -15m | Environments : All | APM Configuration

Navigate to a service or business transaction

APM Overview

Services | Business Transactions

Search by name | Sort by Health, then name | Services 1 - 20

Service	Health	Requests & Errors	Latency
opentelemetry-python-llamaindex-travel-...	Critical	0.1/s, 0.05/s, 0/s; 2:16:05 PM TODAY – 2:31:15 PM TODAY; <0.01/s Requests, 1 total, 0/s Errors	8s, 6s, 4s, 2s, 0ms; 2:16:05 PM TODAY – 2:31:15 PM TODAY; 8.43s p99, 8.43s p90, 8.43s p50
ecom-inventory-admin	Critical		
paymentservice	Warning		
www.gstatic.com *Inferred Service*			
www.google.com			
weather-mcp-security--agent			
user-notification		0.4/s, 0.3/s, 0.2/s, 0.1/s, 0/s; 2:16:05 PM – 2:31:15 PM; 0.27/s Requests, 240 total, 0/s Errors	3ms, 2ms, 1ms; 2:16:05 PM – 2:31:15 PM; 771µs p99, 756µs p90, 672µs p50

Quick links

Service Map
Explore dependencies visually on a map

Tag Spotlight
Find connections between tag values and service behaviors

Traces
View and search for traces

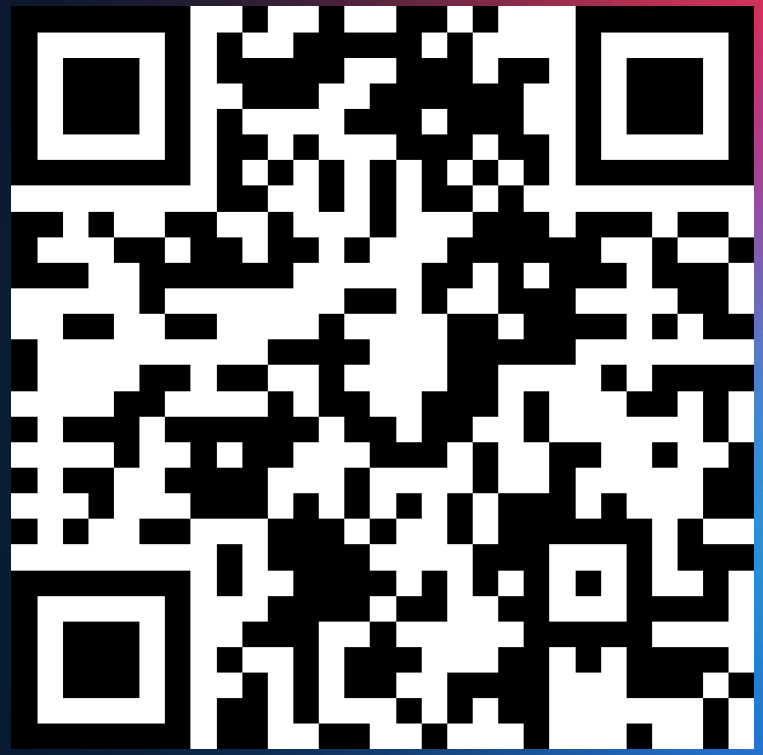
View Trace ID | Go

Database Query Performance
Troubleshoot database queries from services

AlwaysOn Profiling
Examine aggregate runtime behavior

Slido Poll Question #2

What are your greatest concerns deploying agentic applications?



Passcode: gwaono

Slido Poll Question #3

Which teams in your organization are responsible for monitoring and deploying agentic applications?

Passcode: gwaono

AI Infrastructure Monitoring

AI Infrastructure Monitoring

Monitor the health, availability, and usage of AI infrastructure

Optimize resource utilization and identify noisy neighbors

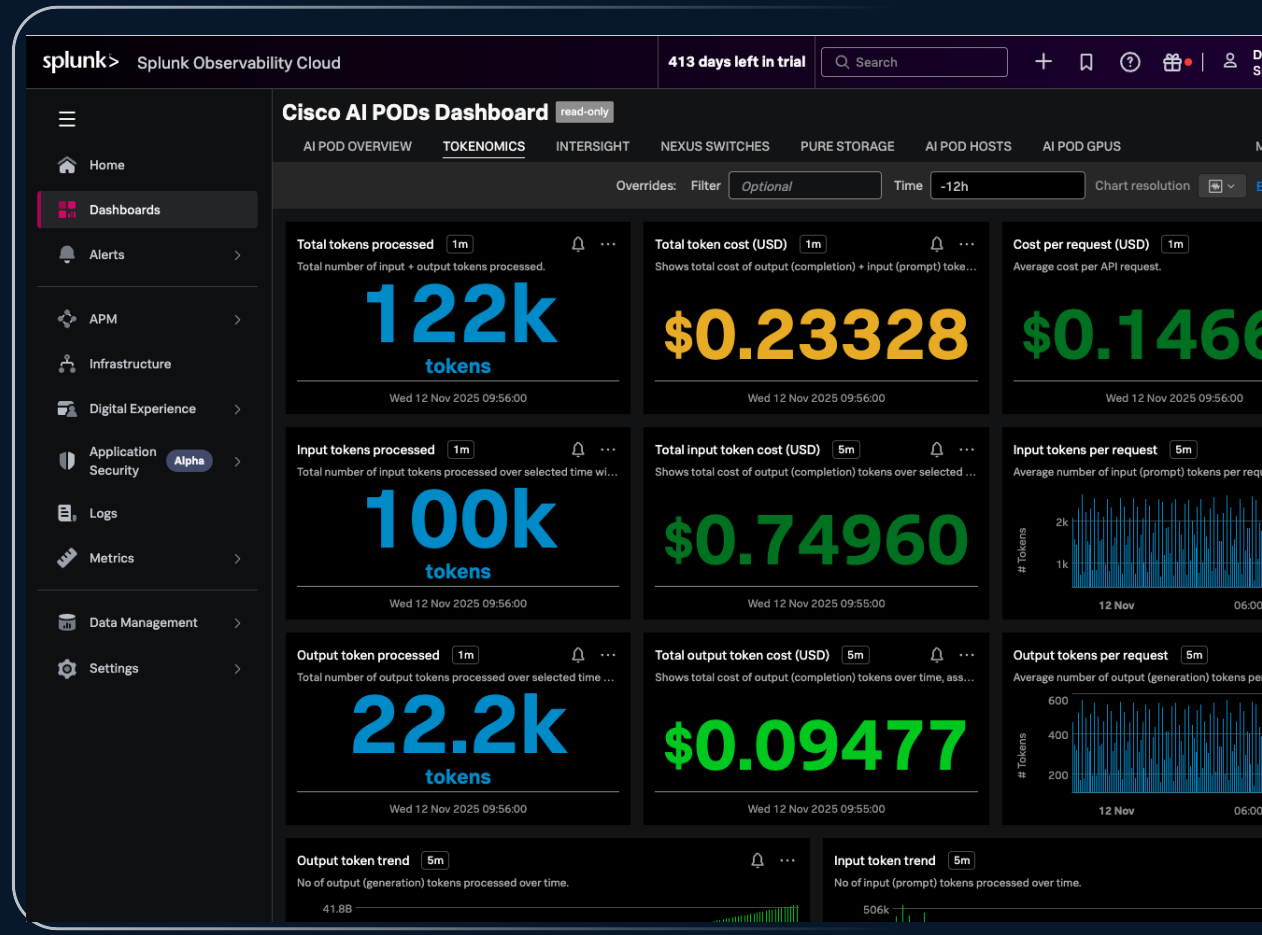
Track operational metrics from AI apps and services to manage costs and detect performance degradation

Visualize the status and understand the business impact of AI components

Use GPU-related and “tokenomics” metrics to assess the utilization and efficiency of hosted AI infrastructure like Cisco AI PODs

Proactively alert on bottlenecks and spikes across services

View data-dense dashboards for frameworks, agents, model providers, vector databases, GPUs, and more to surface trends in underlying AI infrastructure performance

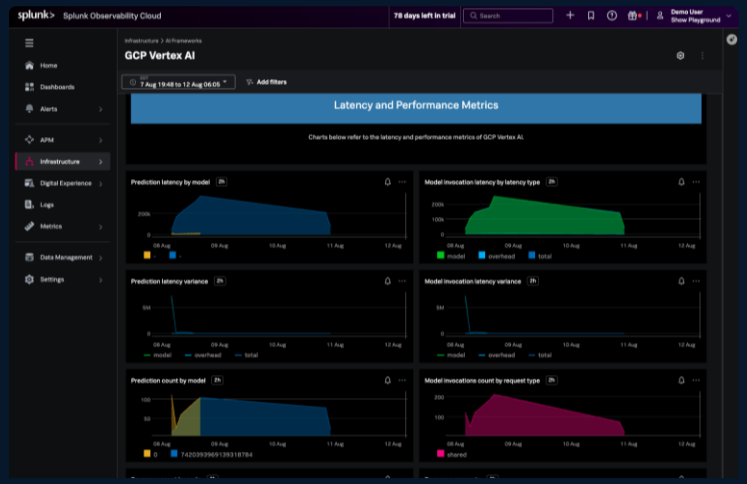


© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

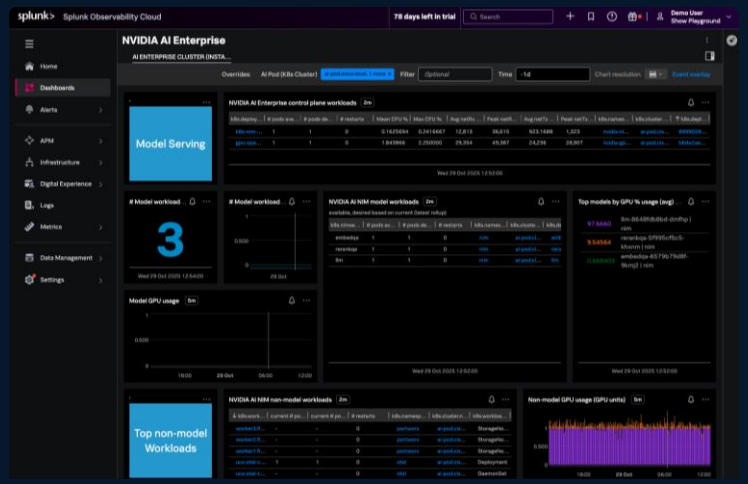
Support for monitoring wherever models run

Tracking the usage of model provider platforms, cloud/SaaS, and local



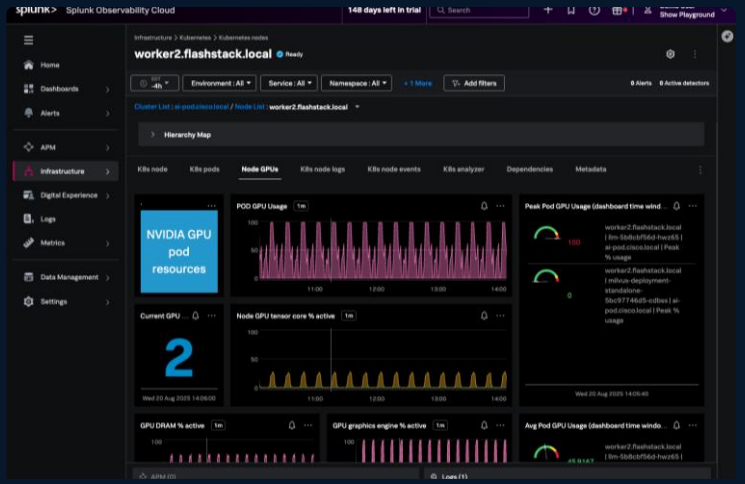
Cloud Hosted

- 2 types
  - Foundation (e.g. Bedrock)
  - Custom (e.g. SageMaker)
- Key Providers: AWS, GCP, Azure, OpenAI



Model Serving Platforms

- 2 types
  - Packaged
  - Custom Trained
- Key Platforms: NVIDIA AI Enterprise, Ray, vLLM, Kserve, Kubeflow



Kubernetes Inferencing Clusters

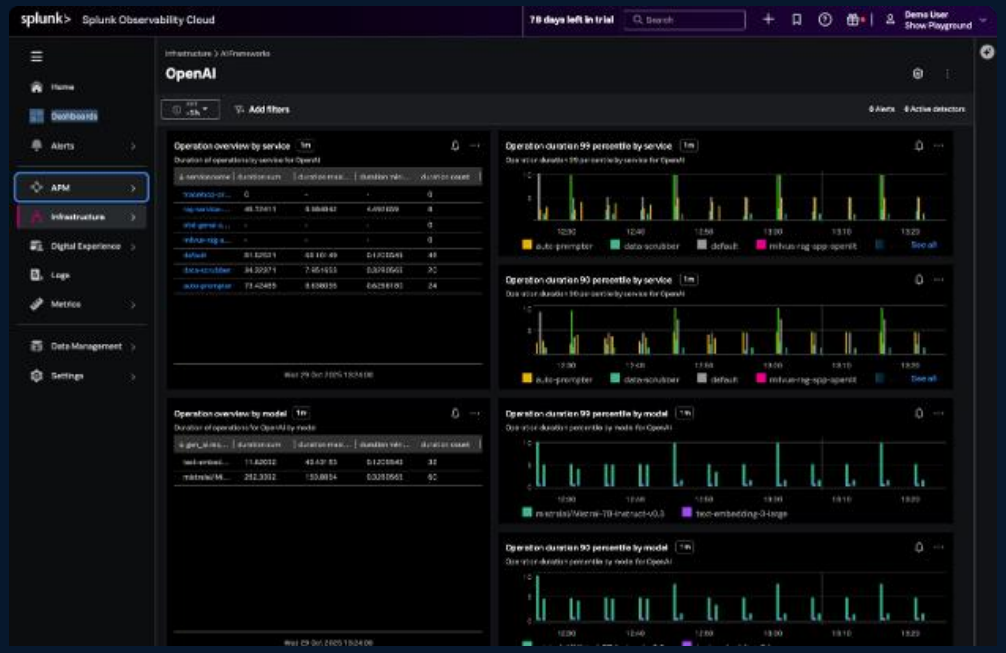
- 3 types of views
  - Node
  - Namespace
  - Workload / Pod
- Key Platforms: NVIDIA DCGM, Model Serving

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Support for monitoring model performance

Tracking latency, tokens, errors, running/waiting requests, cache %, etc....



Example: OpenAI

Internal Performance

- Model self-reported telemetry
- Key deployment methods: vLLM, Nvidia NIM

Agent Framework API

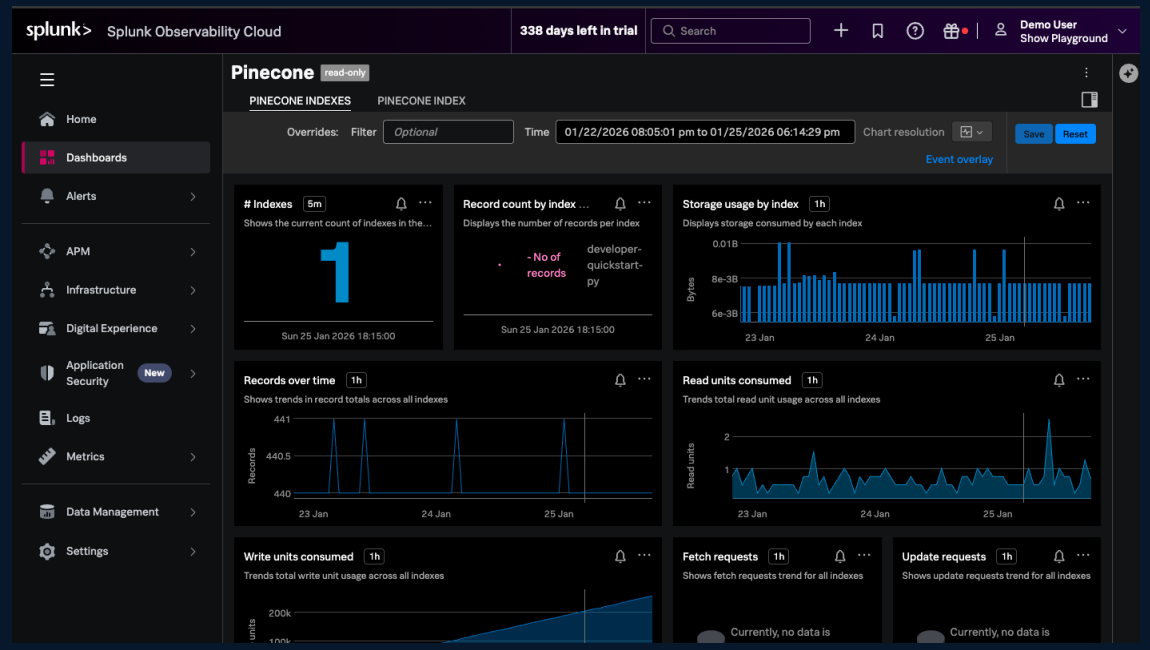
- Service-correlated model performance
- Key frameworks: OpenAI, LangChain, LangGraph, etc....

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Support for monitoring new infrastructure components

Tracking query and tool calls



Example: Pinecone

Vector Database

- Query performance
- Key Technologies: Weaviate, ChromaDB, Pinecone, Azure AI Search, AWS OpenSearch

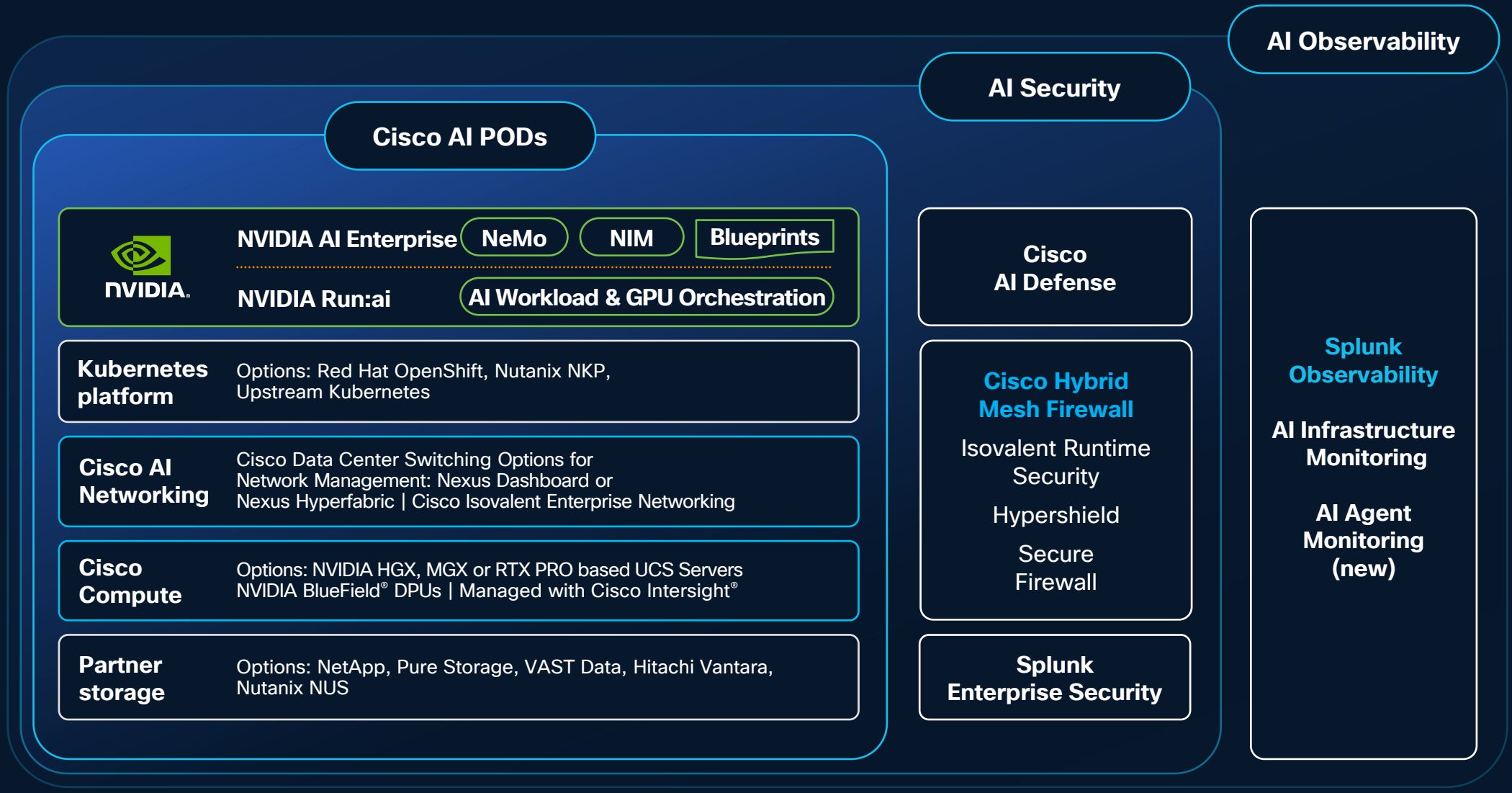
MCP Gateways (coming soon)

- Tool Call Performance (Rate, Errors, Duration)
- Key Technologies: Kong AI GW, LiteLLM GW, AWS Bedrock AgentCore GW

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Cisco Secure AI Factory with NVIDIA



© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Cisco AI PODs dashboards with AI Infrastructure Monitoring

Additional Dashboard Views

Intersight

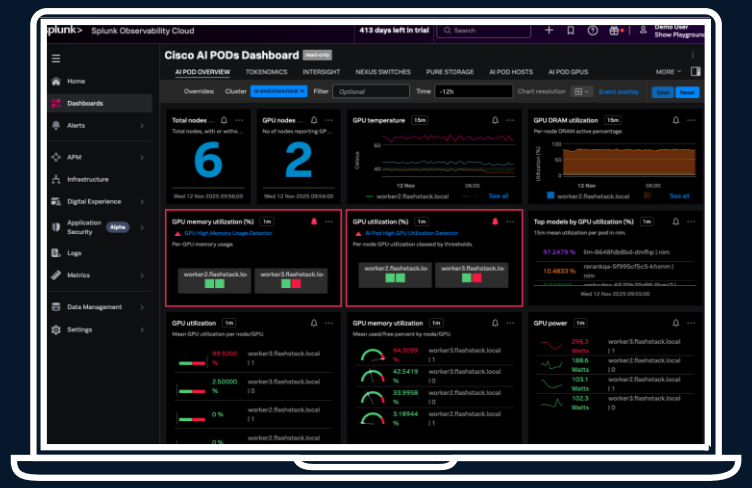
Nexus Switches

Storage

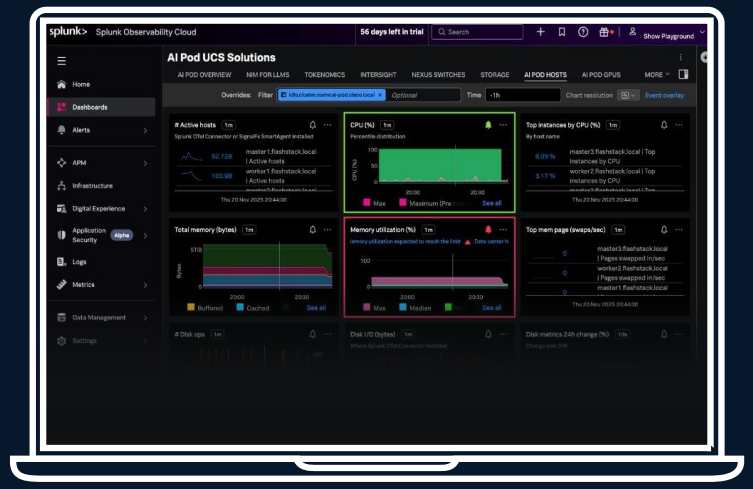
NIM for LLMs

Clusters

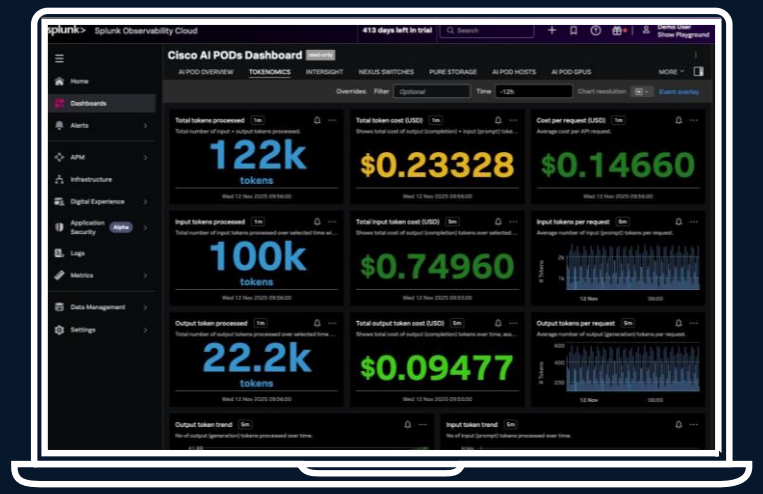
LLM Model Costs



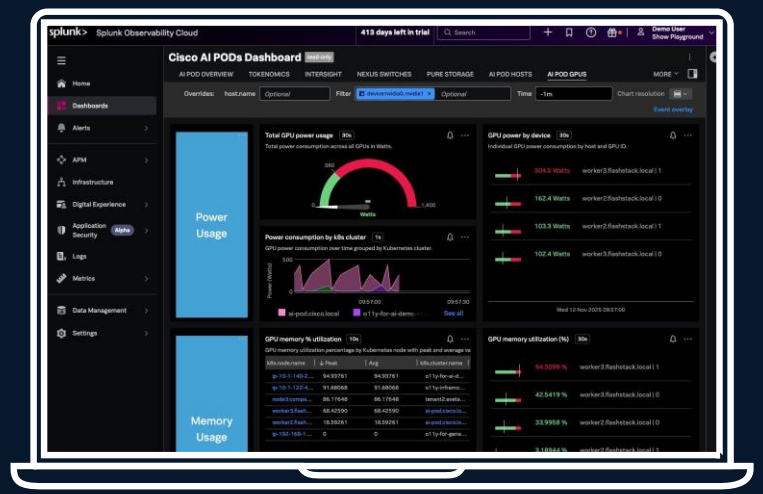
AI POD Overview - Total nodes, GPU power and utilization (%), etc...



AI POD Hosts - # of active hosts, CPU (%), memory utilization (%), etc...



Tokenomics - Total tokens processed, Total token cost, etc...



AI POD GPUs- GPU power usage, GPU memory utilization (%), etc...

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Slido Poll Question #4

Which AI components do you currently monitor?

Passcode: gwaono

Conclusion

Observability for AI



Visibility into infrastructure layer of AI app stack



Individual call-level code visibility for LLM calls



Agent-centric views for evaluating model efficacy

Security

Detect and protect against LLM and agent's risks such as prompt injections, PII leakage, tool poisoning etc.

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Q&A

Observability sessions



Session Code	Title	Date/Time	Room
BRKOBS-1013	Splunk Observability: A Unified Solution for End-to-End Visibility and Troubleshooting Across Applications, Infrastructure, and Networks	Tuesday, Feb 10 12:00 PM - 1:00 PM	Session Room A14
BRKOBS-1164	Agentic AI and Observability: let the system work for you	Tuesday, Feb 10 2:00 PM - 3:00 PM	Session Room A14
DEVNET-1310	Connecting the Dots: Unified Network and Application Observability with Splunk and Cisco	Tuesday, Feb 10 2:00 PM - 2:45 PM	DevNet Theater
BRKOBS-1404	Plug & Play Observability: Splunk as Your Universal Adapter for Any Stack	Tuesday, Feb 10 3:30 PM - 4:30 PM	Session Room G102
BRKOBS-1365	Monitor your LLMs, AI Agents, and AI Infrastructure with Splunk Observability for AI	Tuesday, Feb 10 5:00 PM - 6:00 PM	Session Room G103
BRKOBS-1024	Reduce downtime & accelerate investigations with AI for Observability	Wednesday, Feb 11 8:00 AM - 9:00 AM	Session Room G103
BRKOBS-1073	Everything In One Place: Introduction to Splunk IT Service Intelligence (ITSI)	Wednesday, Feb 11 9:45 AM - 10:45 AM	Session Room D503
BRKOBS-1788	Unifying Application and Network Insights: Beyond Monitoring to True Network Observability with Splunk, ThousandEyes, and Meraki	Wednesday, Feb 11 11:30 AM - 12:30 PM	Session Room G102
BRKOBS-2008	Splunk IT Service Intelligence (ITSI) - a Unified View for Network Service Health	Wednesday, Feb 11 1:45 PM - 2:45 PM	Session Room A13
BRKOBS-1868	The Next Horizon for Splunk Observability Cloud: Unifying Observability Across Hybrid Environments	Wednesday, Feb 11 3:30 PM - 4:30 PM	Session Room D403
BRKOBS-1377	Close the App–Network Gap: Monitor Tetragon (eBPF) Telemetry in Splunk Observability	Wednesday, Feb 11 5:15 PM - 6:15 PM	Session Room D503
BRKOBS-1364	Navigating Kubernetes with Confidence Using Splunk Observability Cloud: Troubleshooting Strategies and Best Practices	Thursday, Feb 12 8:45 AM - 9:45 AM	Session Room A12
BRKOBS-1366	See what your users see: Digital experience monitoring with Splunk Observability Cloud	Thursday, Feb 12 8:45 AM - 9:45 AM	Session Room A13
BRKOBS-1123	Securing Cloud-Native Applications: Real-Time AI Based Vulnerability Detection with Splunk	Thursday, Feb 12 10:15 AM - 11:15 AM	Session Room D503
BRKOBS-1023	AI-Directed Troubleshooting for 3-Tier Applications	Friday, Feb 13 9:15 AM - 10:45 AM	Session Room D403

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

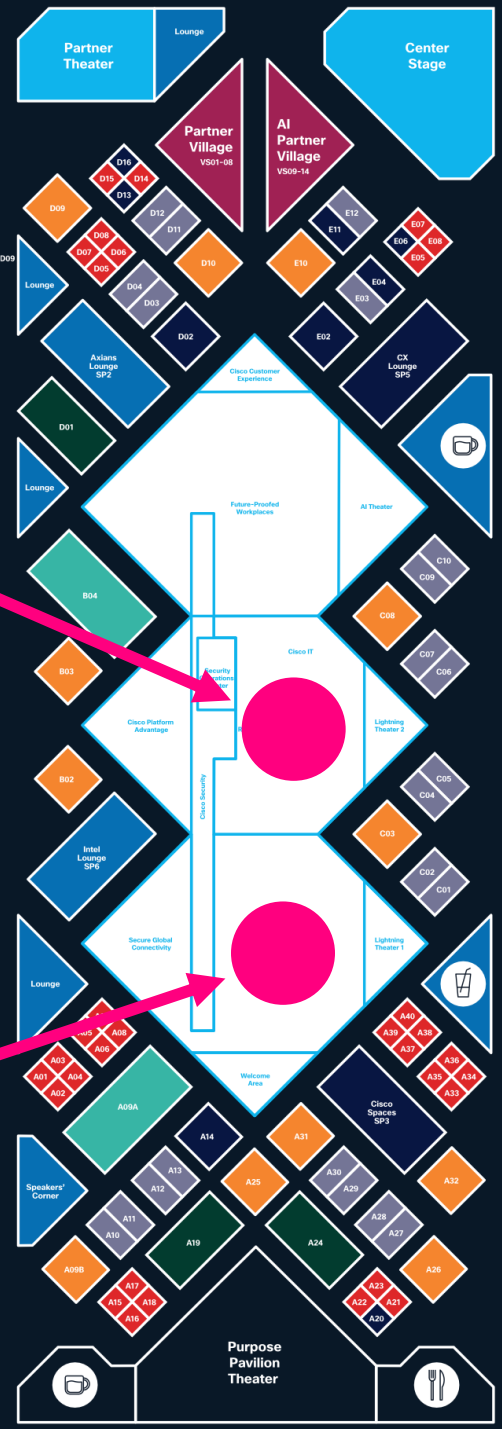
See a demo

Digital Resilience

Demo Title	Products/Solutions	Area
Explore unified observability	Splunk Observability	Digital Resilience
Accelerate app & network insights	Splunk Observability and Cisco Networking	Digital Resilience
Resolve issues faster with Agentic AI	Splunk MCP, Splunk Cloud, Observability Cloud	Digital Resilience

AI-Ready Datacenter

Demo Title	Products/Solutions	Area
Optimize AI performance	Splunk Observability Cloud	AI Ready Data Center



© 2026 Cisco and/or its affiliates. All rights reserved.

Resources to learn more and get started

- **Docs:** Introduction to Splunk Observability for AI
- **On-Demand Webinars:**
  - Agentic, AI-powered Observability: Simplifying Operations & Faster Resolution
  - Splunk Observability for AI: Ensuring your AI application stack is resilient, responsible, and secure with agentic, AI observability
- **E-Book:** Observability for AI: Monitor LLMs, Agents, and Infrastructure
- **Blogs:**
  - Monitor the Health, Performance, and Security of Your AI Application Stack with AI Agent and AI Infrastructure Monitoring
  - Unlocking AI Performance: Splunk Observability for Cisco Secure AI Factory with NVIDIA
  - AI observability: AGNTCY's open source suite and enterprise-ready observability with Splunk
  - Splunk and AGNTCY: Unlocking the future of AI observability
  - Unify Observability, Surface Business Impact, and Solve Problems Using AI Agents with Latest Splunk Observability Innovations
  - A Recap of .conf25 Splunk Observability Announcements

© 2026 Cisco and/or its affiliates. All rights reserved.

Complete your session evaluations



Complete a minimum of 4 session surveys and the Overall Event Survey to be entered in a drawing to win 1 of 5 full conference passes to Cisco Live 2026.



Earn 100 points per survey completed and compete on the Cisco Live Challenge leaderboard.



Level up and earn exclusive prizes!



Complete your surveys in the Cisco Live mobile app.

© 2026 Cisco and/or its affiliates. All rights reserved.

Continue your education



Visit the Cisco Showcase for related demos



Book your one-on-one Meet the Engineer meeting



Attend the interactive education with DevNet, Capture the Flag, and Walk-in Labs



Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand

© 2026 Cisco and/or its affiliates. All rights reserved.

Join us at Cisco Live Las Vegas!

© 2026 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Thank you




CISCO