

Agentic AI and Observability: Let the System Work for You

cisco Live !

Luca Relandini
Principal Architect, Observability

Leila Reyhani
Solutions Architect, Observability

Webex App

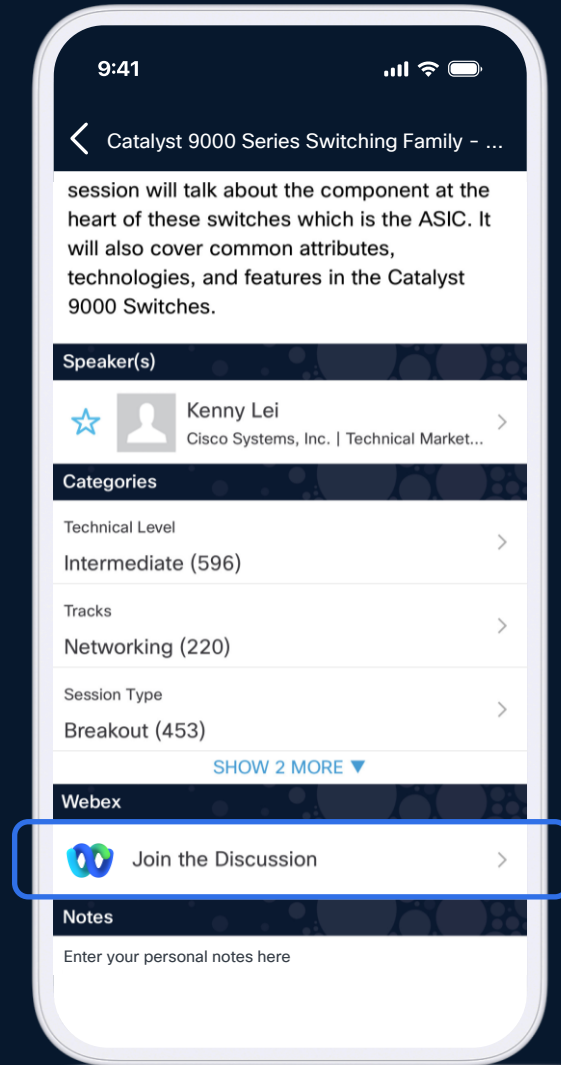
Questions?

Use Webex App to chat with the speaker after the session

How

- 1 Find this session in the Cisco Events 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 February 27, 2026.



Agenda

- 01 Evolution of AI
- 02 Agentic architecture
- 03 Need for visibility and security
- 04 Agentic Operations
- 05 Cisco solutions



Evolution of AI



Chatbots

Humans
talk to AI



Agentic

Workflows get
automated



Physical

Anything that
moves has AI



AI will make our world of **8B** people
feel like one with the capacity of **80B**

AI is rewriting the rules

Apps can now be
written with little
human involvement

AI agents will perform
basic troubleshooting
and fixes, autonomously

AI apps require
new forms
of telemetry

All of this has **massive implications**
for your technology architectures

AI is forcing a rethink of the tech stack

>80%

not AI ready—need
hardware acceleration

Network

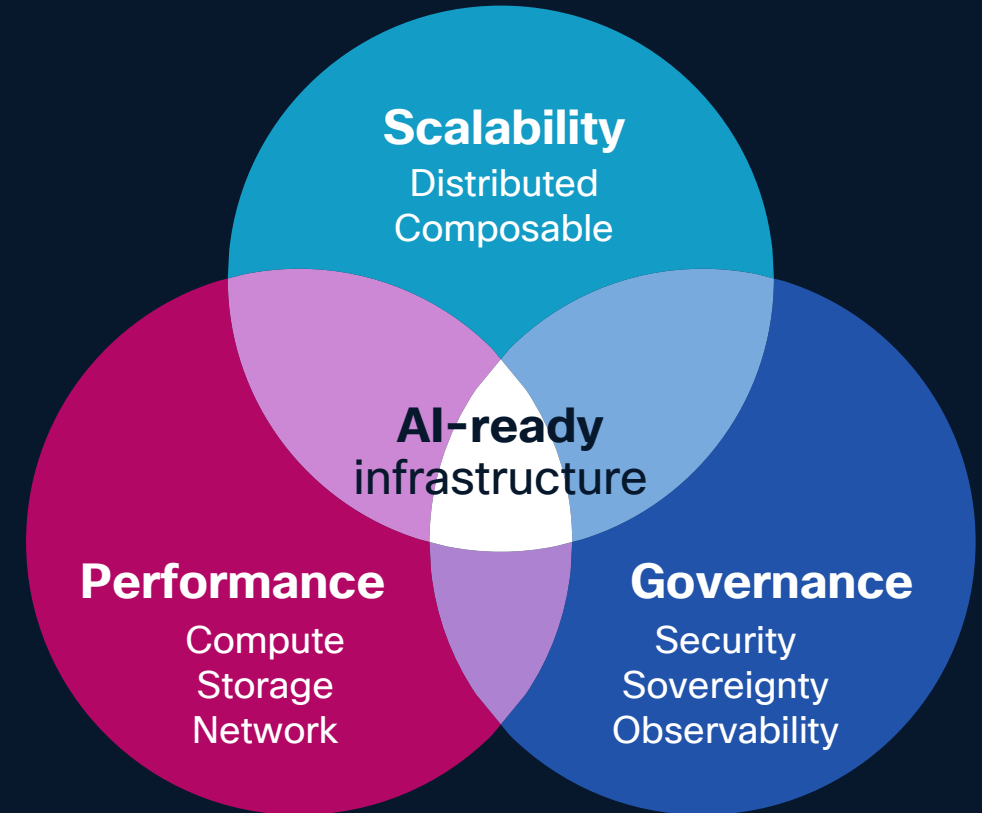
speed, fabric designs, and
performance

>48%

of network, compute, and
storage infrastructure is
10+ years old

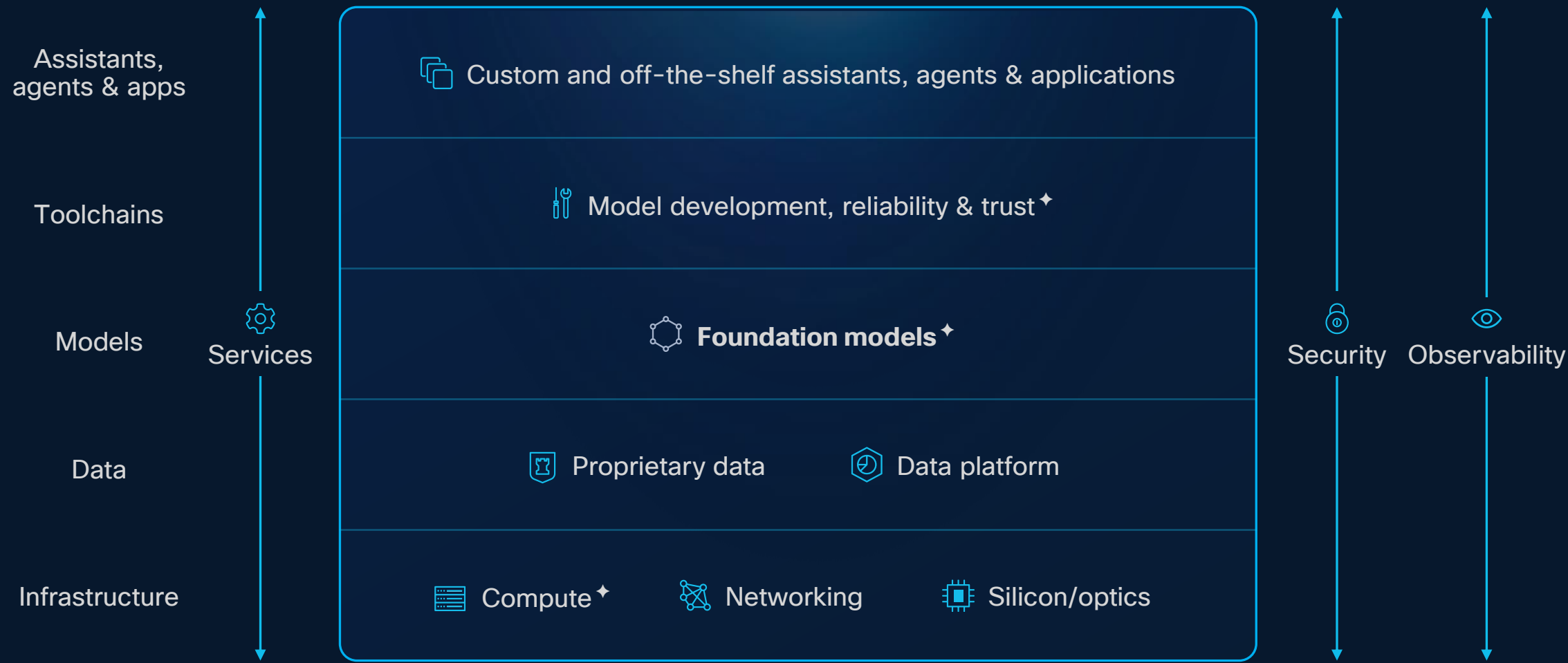
40%

increase in security
breaches year over year—
due to old infrastructure



Sources: IDC; Gartner; Forrester; TechRadar; CIO.com; Accenture Tech Vision 2025

Cisco supports customers across the AI stack



Current Cisco coverage

✦ With partners



AI-Ready Data Centers



Future-Proofed Workplaces

← Secure Global Connectivity →



Digital Resilience



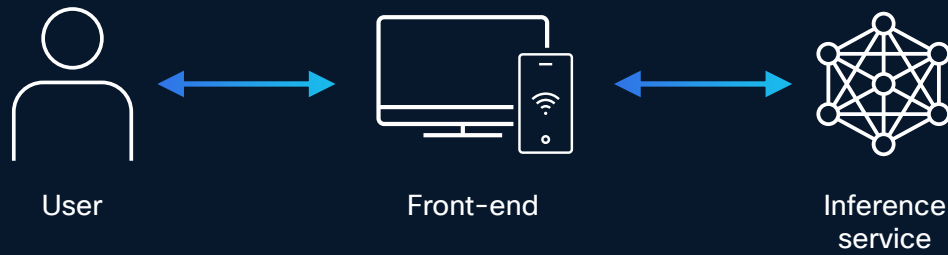
Accelerated by Cisco AI



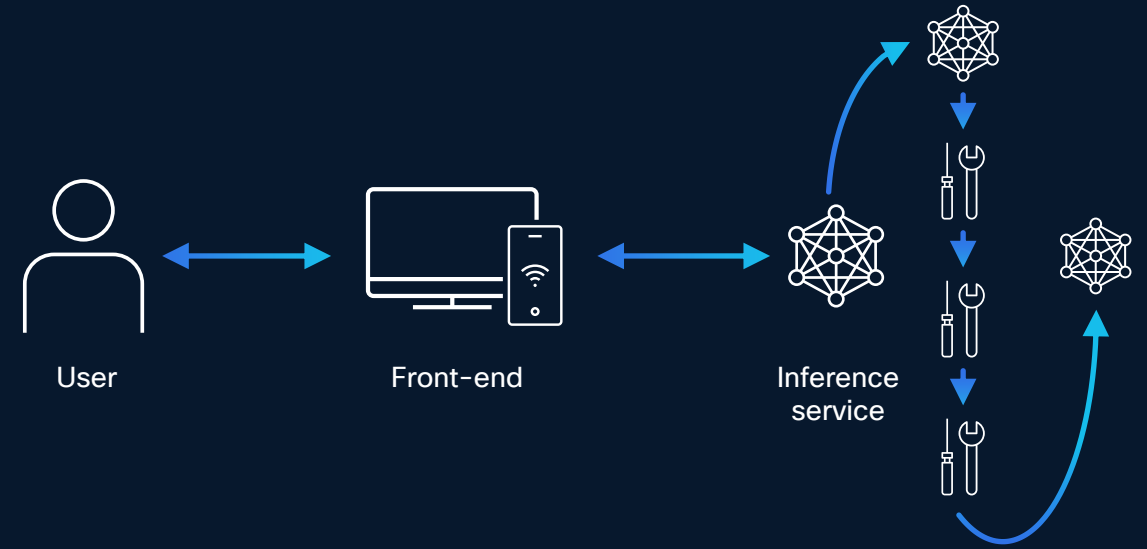
Agentic Architecture

Defining Agentic AI

Beyond chatbots and analytics – to automation and goal-seeking AI



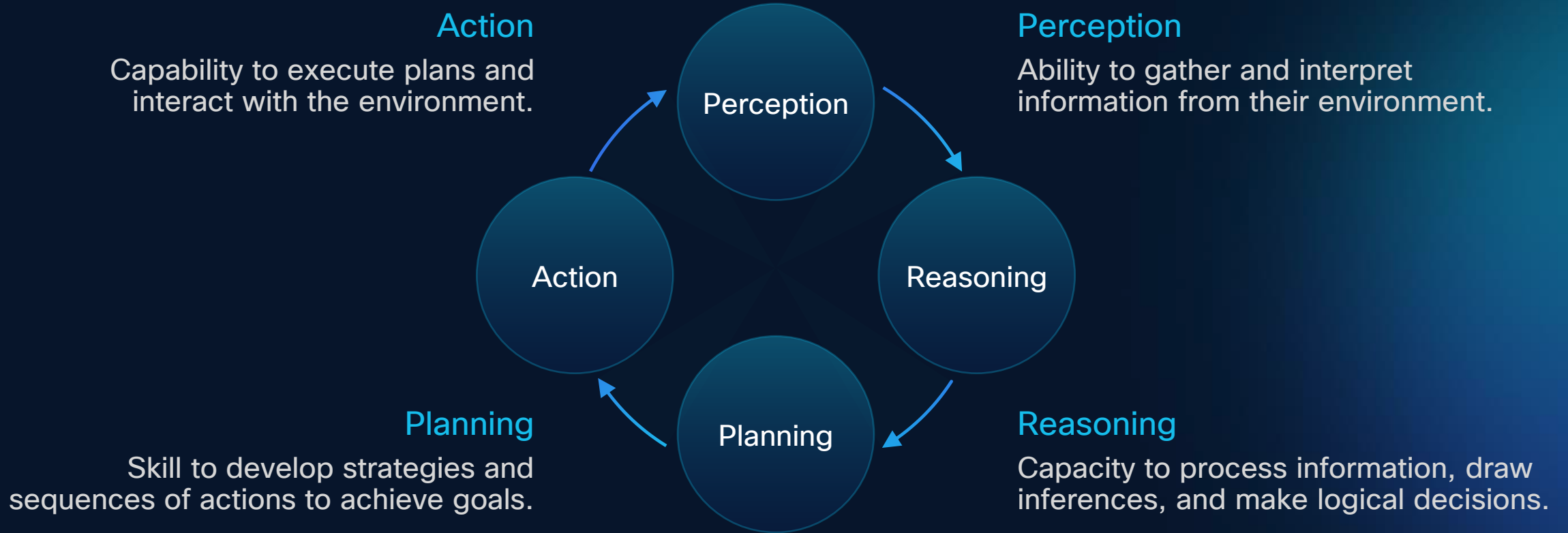
Traditional AI/ML



Agentic AI

The power of autonomy

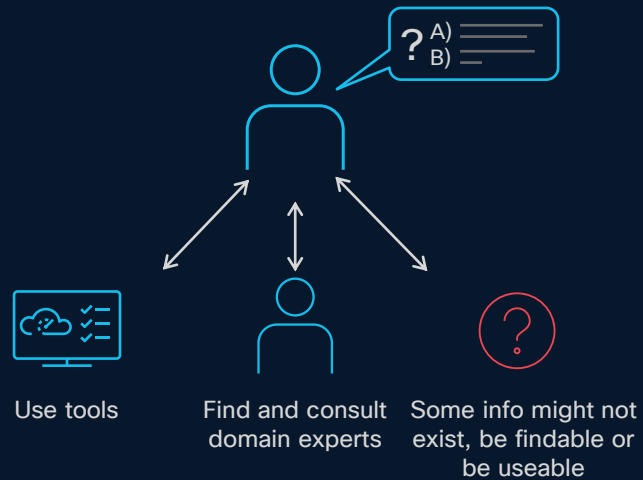
Understanding Agentic AI's core traits



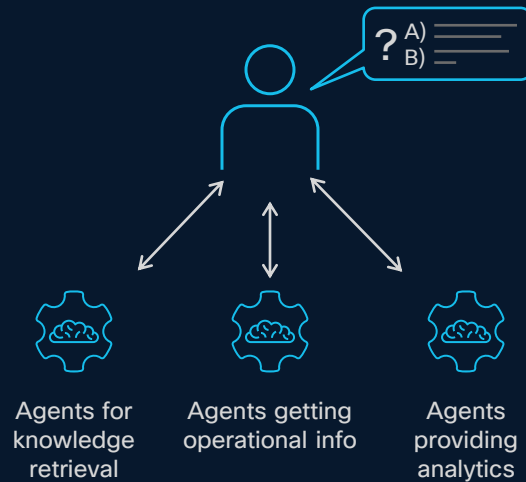
Embracing AI agents of the future

Agentic multi-agent system and questions of control

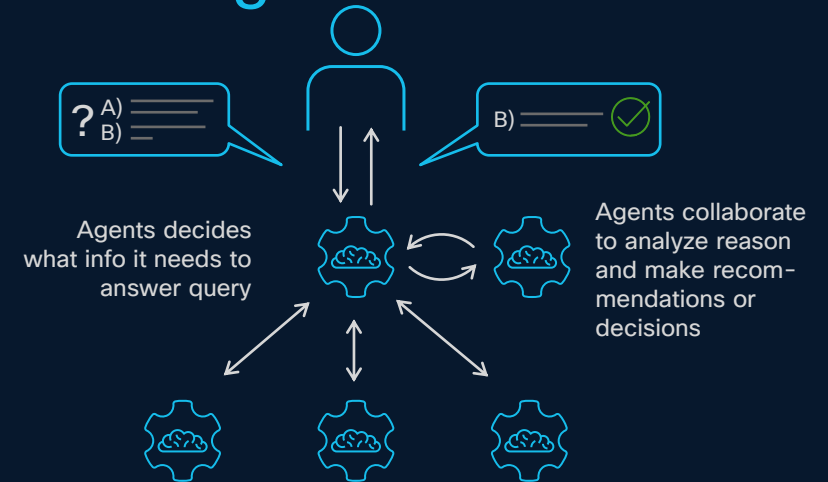
Yesterday



Single Agent



Multi Agent



Agentic computing needs a collaboration framework

Agents

- come from different **vendors**
- use different development **frameworks**
- run on different **clouds**
- have different **access** controls
- belong to different **organizations**
- take on different **profiles** and personas

Why standardize agent communication?

Cisco's leadership in connecting agents

Google's Agent-to-Agent
(A2A) protocol

Anthropic's Message/Model
Context Protocol (MCP)

Cisco's Internet of Agents
(IoA) architecture

Emergence
of AI agents



Organizations use many AI “agents” (LLMs, tools, services) that need to collaborate. Without common standards, every new agent integration requires custom adapters.

Integration
bottleneck



Ad-hoc communication leads to complexity and slows innovation.
A heterogenous mix of agents without standards creates “agent sprawl” and chaos.

Need for an
“HTTP for
agents”



Industry is coalescing around open protocols so that any AI agent can talk to any other seamlessly – enabling cross-platform workflows and reducing silos.



An open-source collective for inter-agent collaboration

The AGNTCY is where we are building the **Internet of Agents** to be:

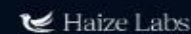
A diverse, collaborative space to innovate, develop, and maintain software components and services for agentic workflows and multi-agent software.

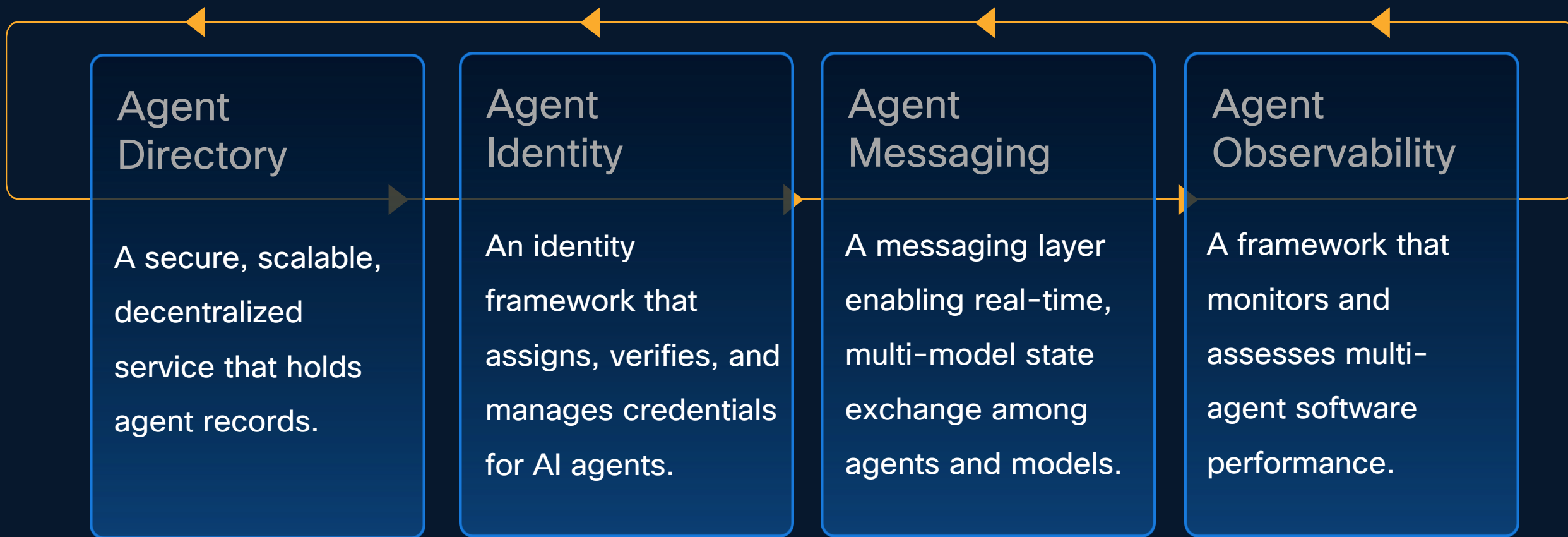
A proud project of
THE LINUX FOUNDATION

TECHNICAL STEERING COMMITTEE



MEMBERS AND CONTRIBUTORS





Multi-Protocol Support: AGNTCY building blocks are interoperable with A2A and MCP

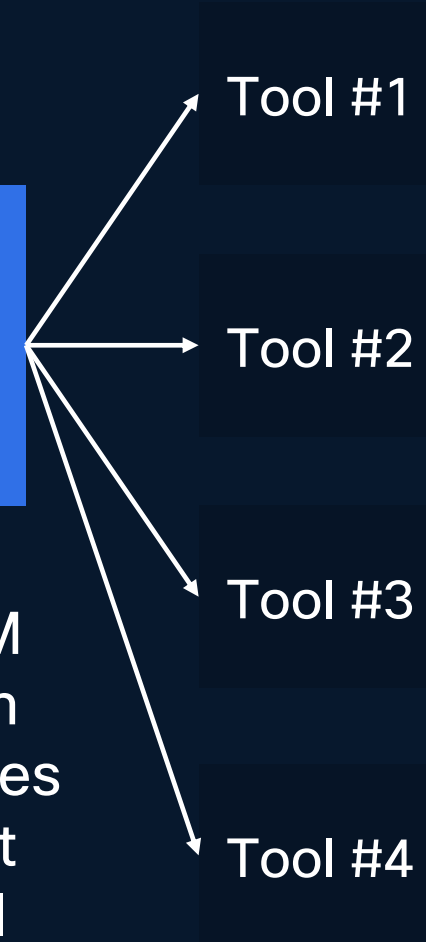
The Rise of Agentic Systems



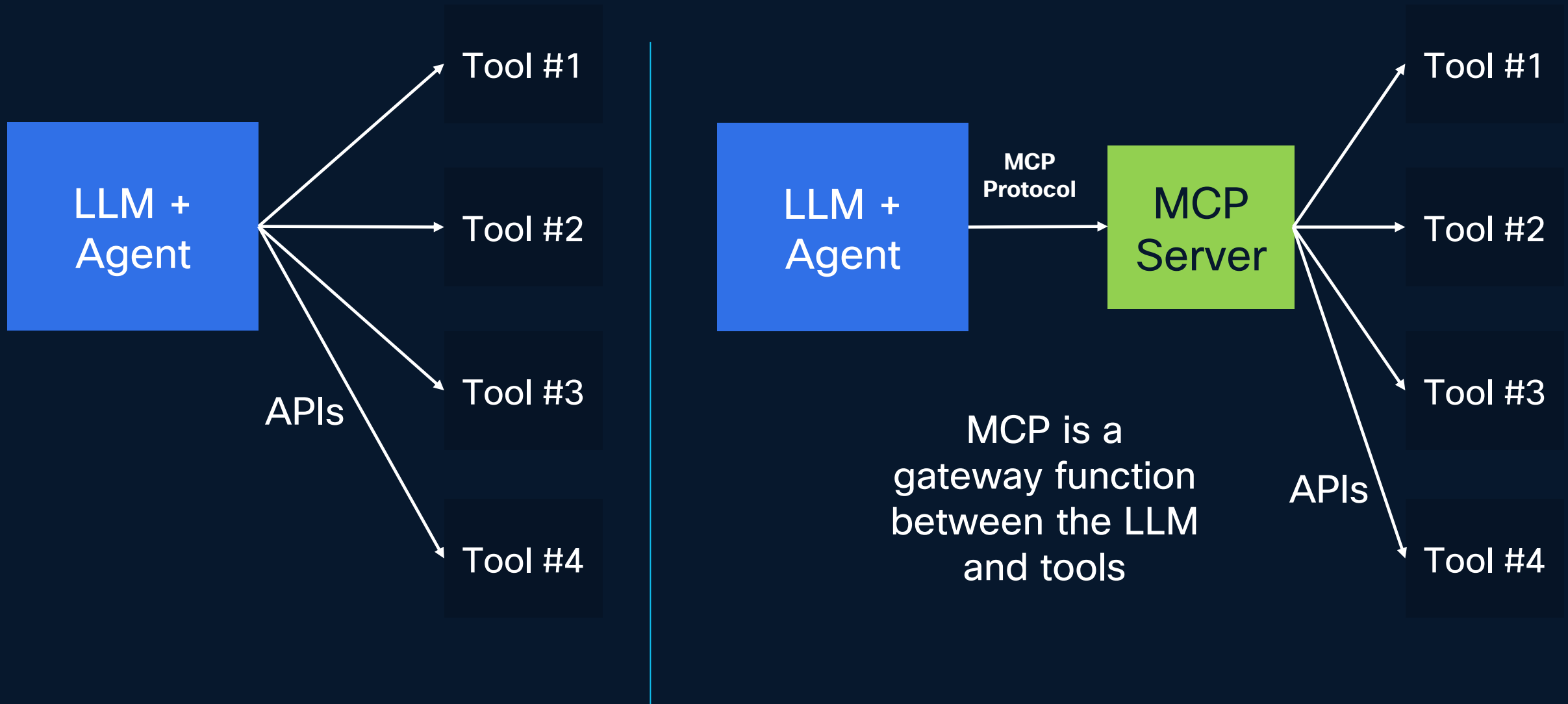
An LLM by itself
doesn't really do
anything except
generate text



When an LLM
interacts with
tools it becomes
powerful, but
relies on API
consistency

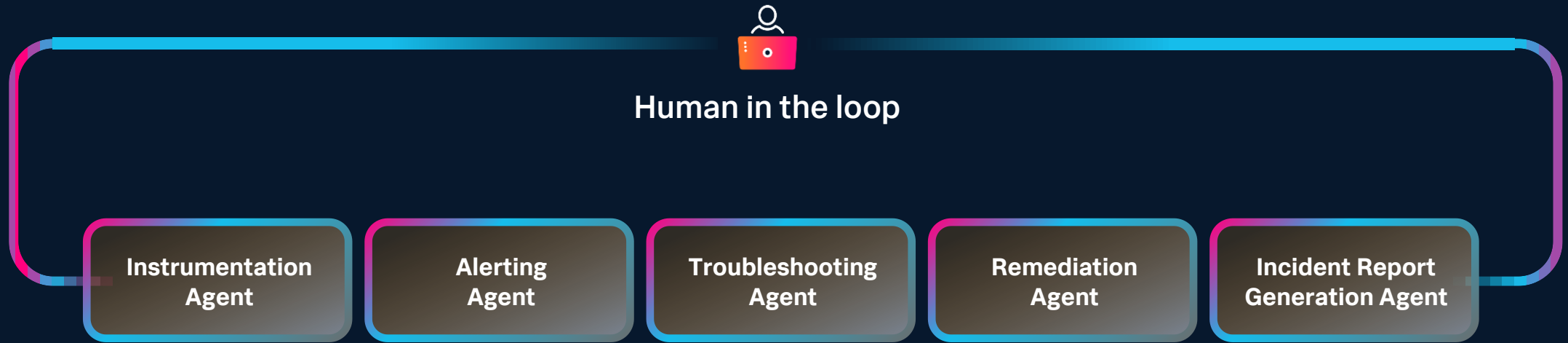


The Agentic Evolution to MCP

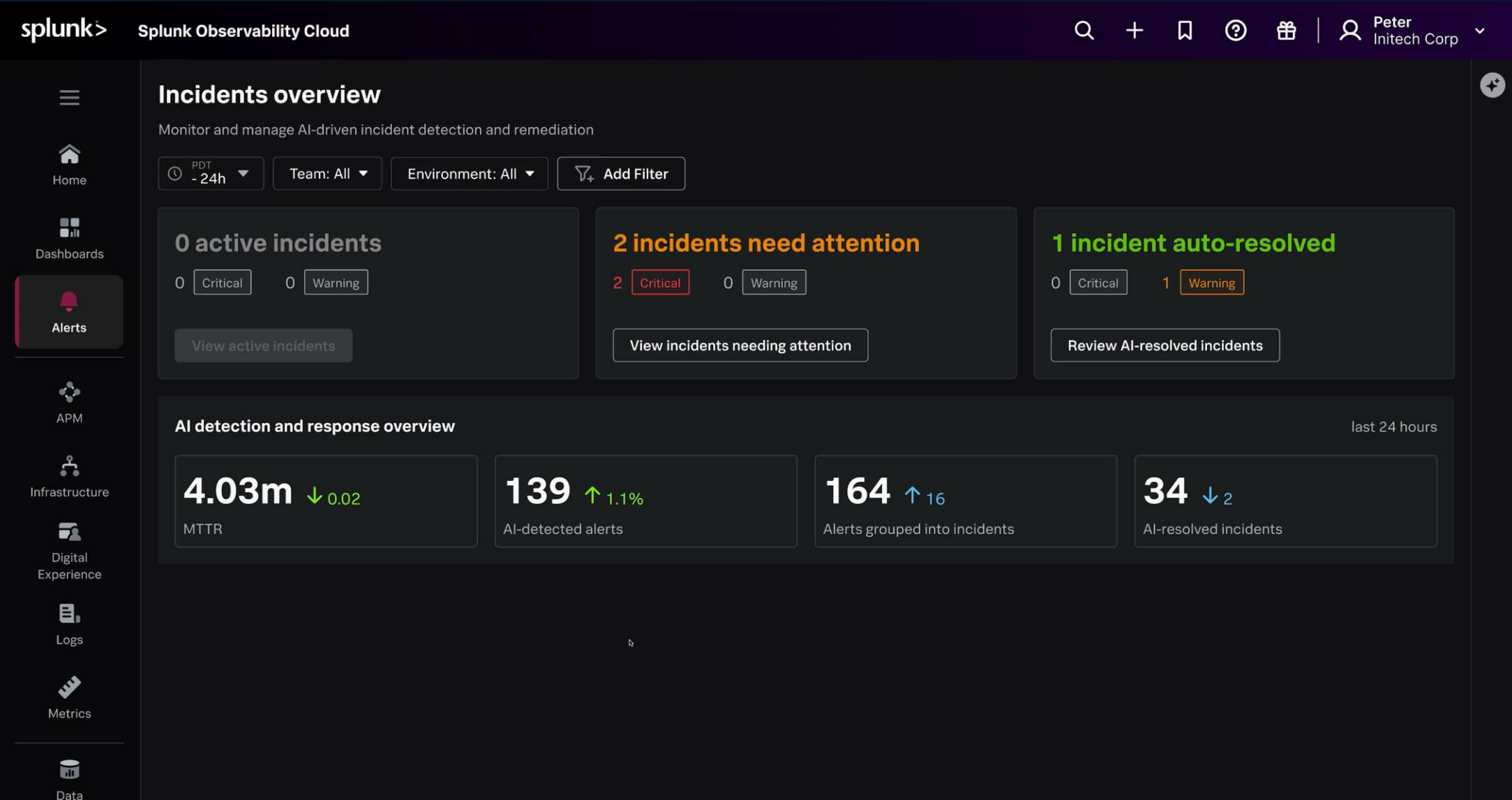


Observability for AI, and AI for Observability

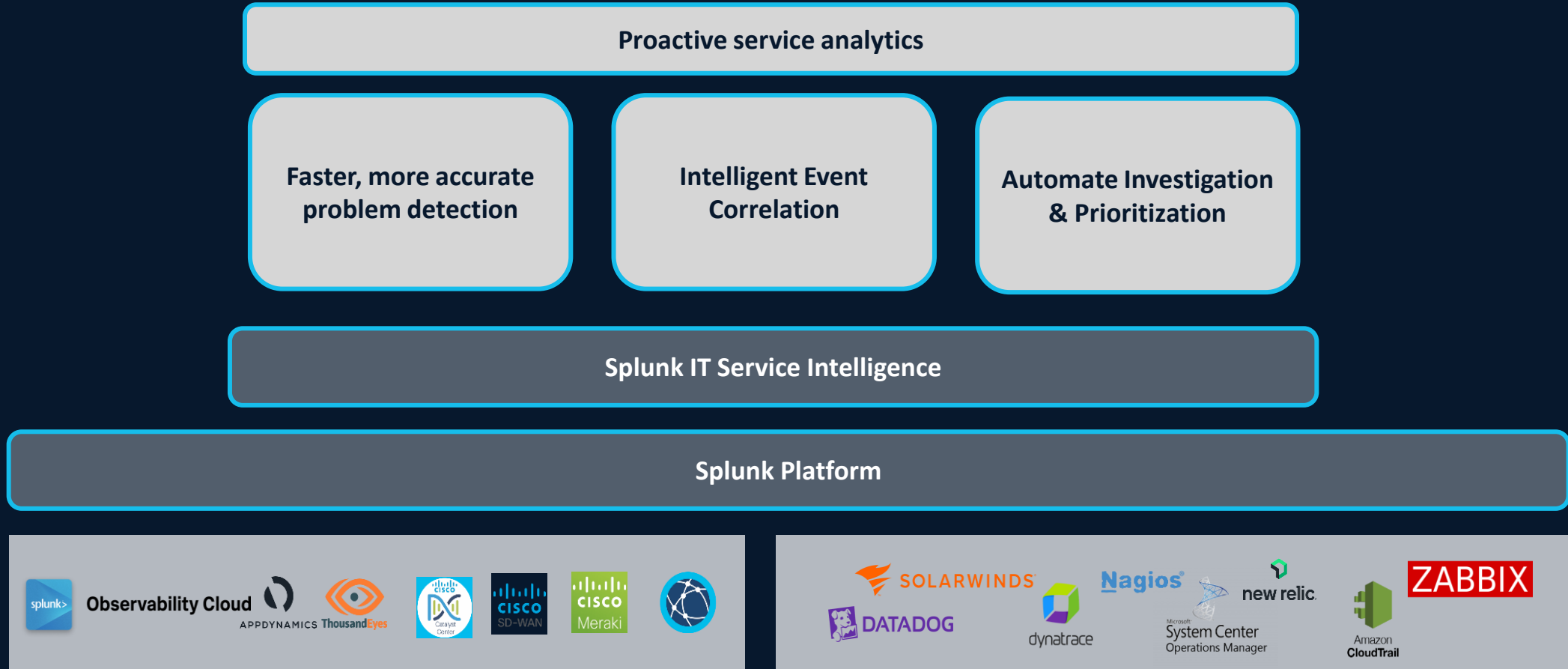
Fix and prevent issues with Agentic Ops



Autonomous agents built in Splunk Observability Cloud

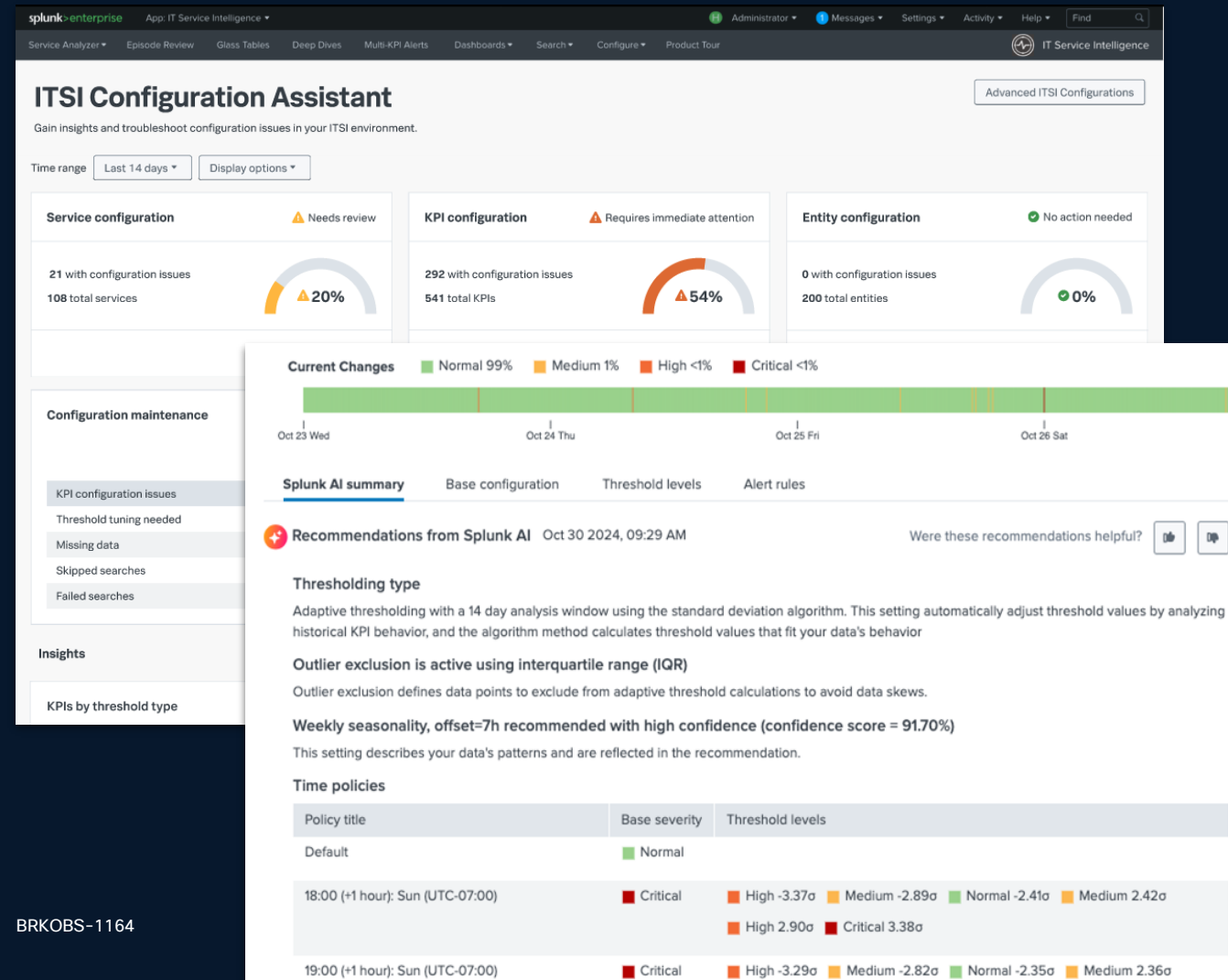


Embedded AI for seamless incident management & prevention in Splunk ITSI



Prescriptive Guidance with AI/ML in ITSI

- ML-Assisted Thresholds for OOTB guided approach to set up and identify anomalies on KPIs
- Drift Detection to proactively detect long trending anomalies as signals of performance degradation
- Configuration Assistant to signal when to optimize thresholds
- Ability to automate training and saving of ML-Assisted thresholds
- Entity-level thresholding recommendations to configure threshold monitoring dynamically



End-to-end AI-driven Incident Management in ITSI

Event iQ for dynamic correlation and Episode Summarization for incident insights + RCA

Incident Creation - Event iQ correlates alerts into actionable episodes that update dynamically as the situation develops:

Automated and tunable alert grouping policy generation incl. field prioritization

Dynamic, real-time alert aggregation based on common fields

The screenshot displays the 'Episode Review' interface in Cisco ITSI. At the top, it shows '1244 episodes' with filters for 'Last 24 hours' and 'Add Filter'. A search bar is also present. Below the filters, a table lists various episodes with columns for severity, count, owner, and severity. The episodes include 'Web Login Response Test - web-p...', 'Host Available - Ping -', 'Alarm: Webserver Port', 'Alarm: Host Available -', 'Alert Group: acme co', 'Boutique Mobile App', 'Splunk TShirt Compai', 'Boutique Cart Service', 'DemoSupport - BT Lo', and 'Boutique Checkout' h.

The detailed view for the 'Boutique Cart Service' episode shows an AI analysis titled 'AI Analysis: The Auth Service experiencing intermittent failures'. The analysis includes a summary, suspected root cause, impact, and a timeline of key alerts.

Summary
The Auth Service has been experiencing intermittent failures starting Sept 12th, 2024, at approximately 10:00 AM.

Suspected root cause
A recent update to the Auth Service API (v1.3.0) deployed on Sept 12th at 9:30 AM may have caused the issue, as the failures began shortly after the deployment. The update included changes to the authentication flow and API rate limiting.

Impact
Unable to log in [User Management Service](#), [Payment Service](#)

Timeline of key alerts

Time(GMT-7)	Key alerts	Source
10:02 AM	Authentication timeouts detected in Auth Service.	Nagios
10:05 AM	503 Service Unavailable errors in User Management and Payment Services.	Datadog
10:10 AM	Login failures reached 30% of normal traffic.	Splunk logs correlation search
10:15 AM	User session timeouts on login requests.	Datadog

> What we've checked

Introducing the AI Assistant

Monitoring

Troubleshooting

Automated Configuration

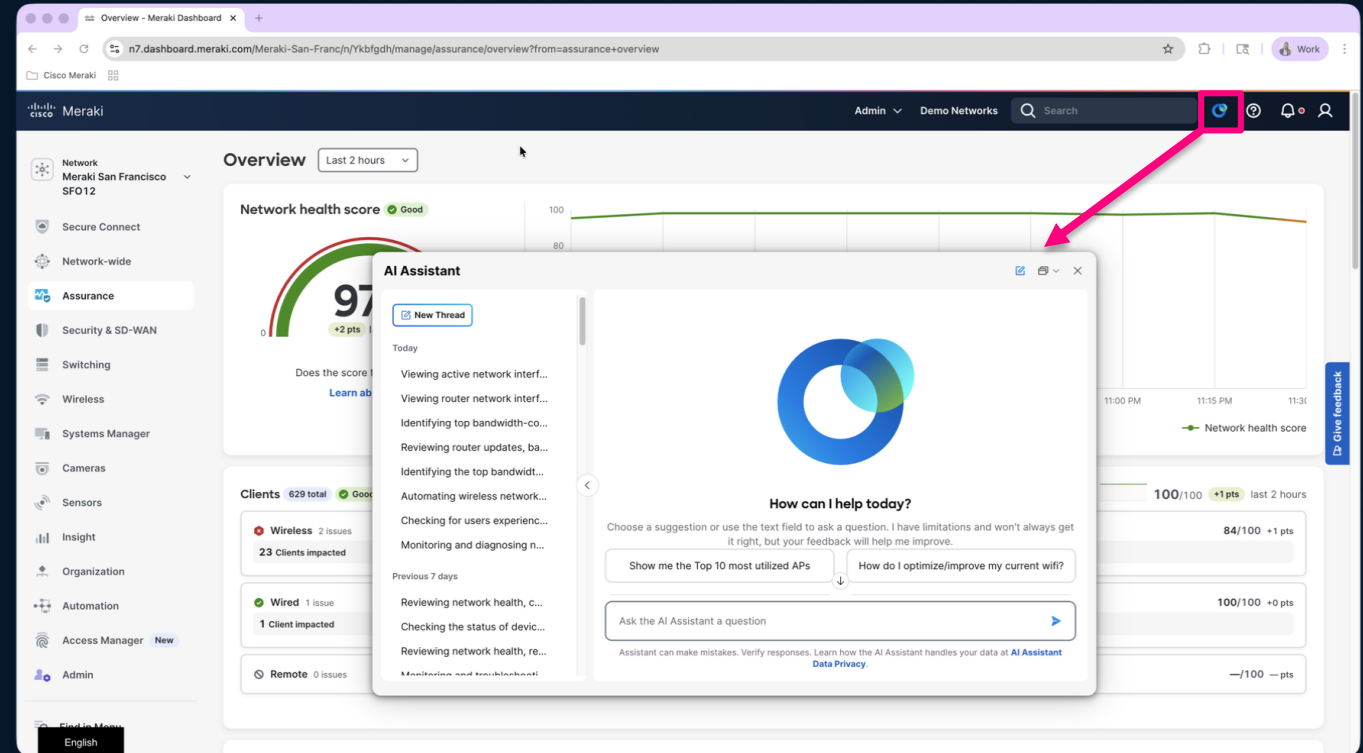


What is the AI Assistant?

The Cisco AI Assistant is a Generative AI solution designed for simplified, expedited network operations:

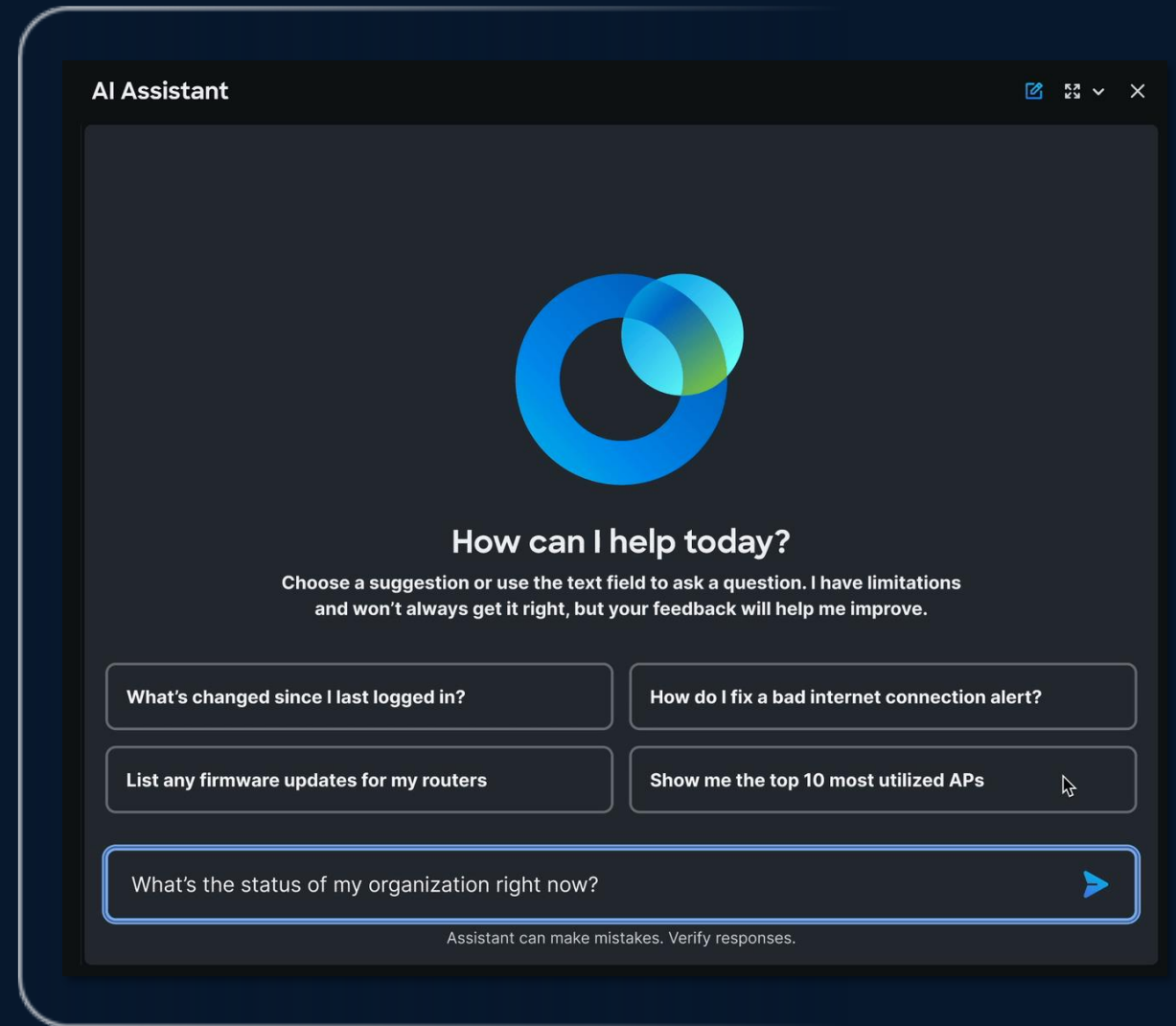
- Users make requests with natural language
- AI Assistant interprets request, determines relevant data, and gathers it
- Data is fed to the LLM engine for response generation
- Natural language response and relevant tables or graphs are generated

The AI Assistant is launched via dashboard (see below):



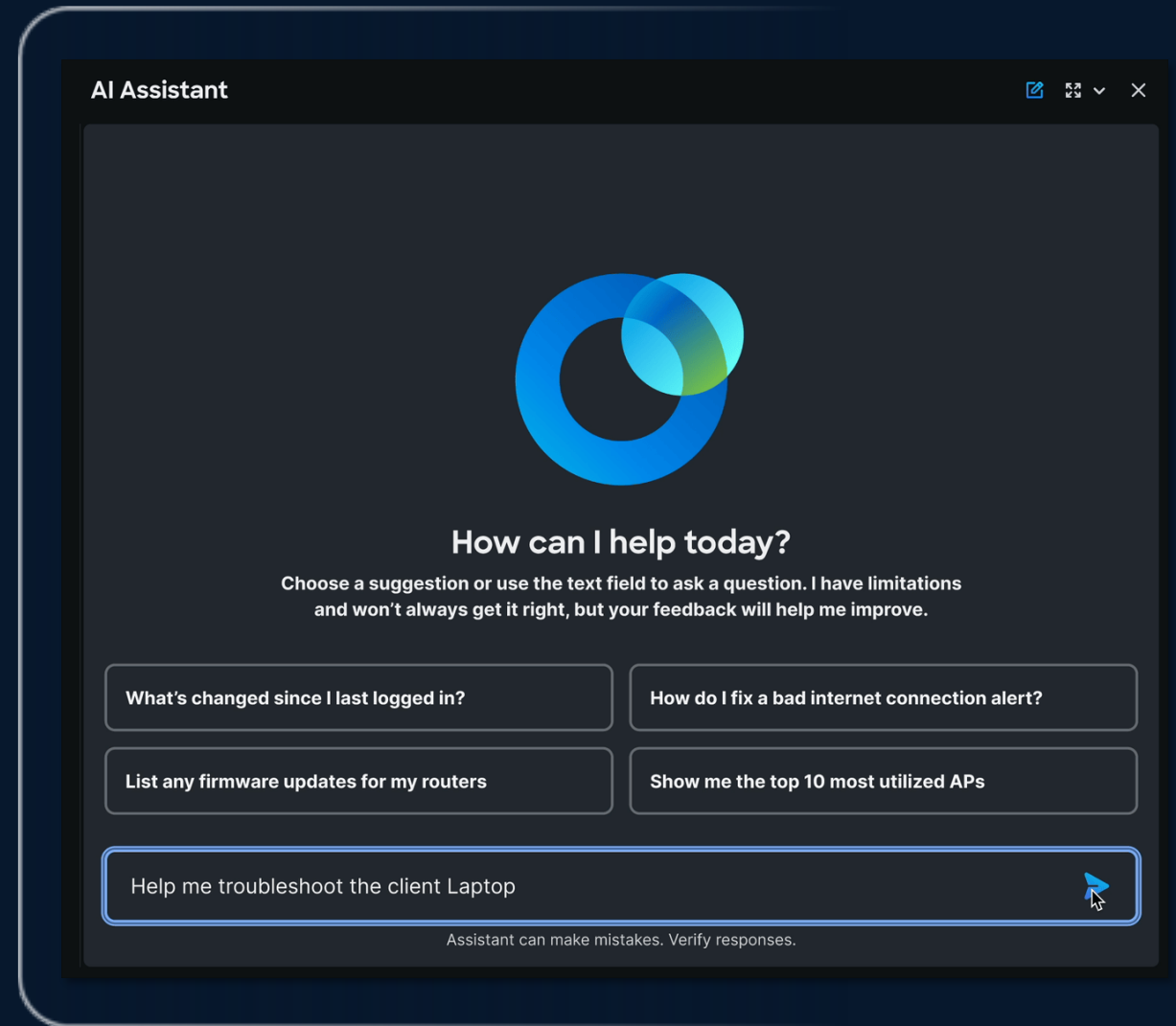
End-to-end monitoring, visibility & assurance

- Provides network health across clients, LAN, WAN, applications
- Visibility for owned and unowned networks
- Instantly shows impacted clients, devices, & applications
- Smart alerts with intelligent RCA, prioritized by severity level



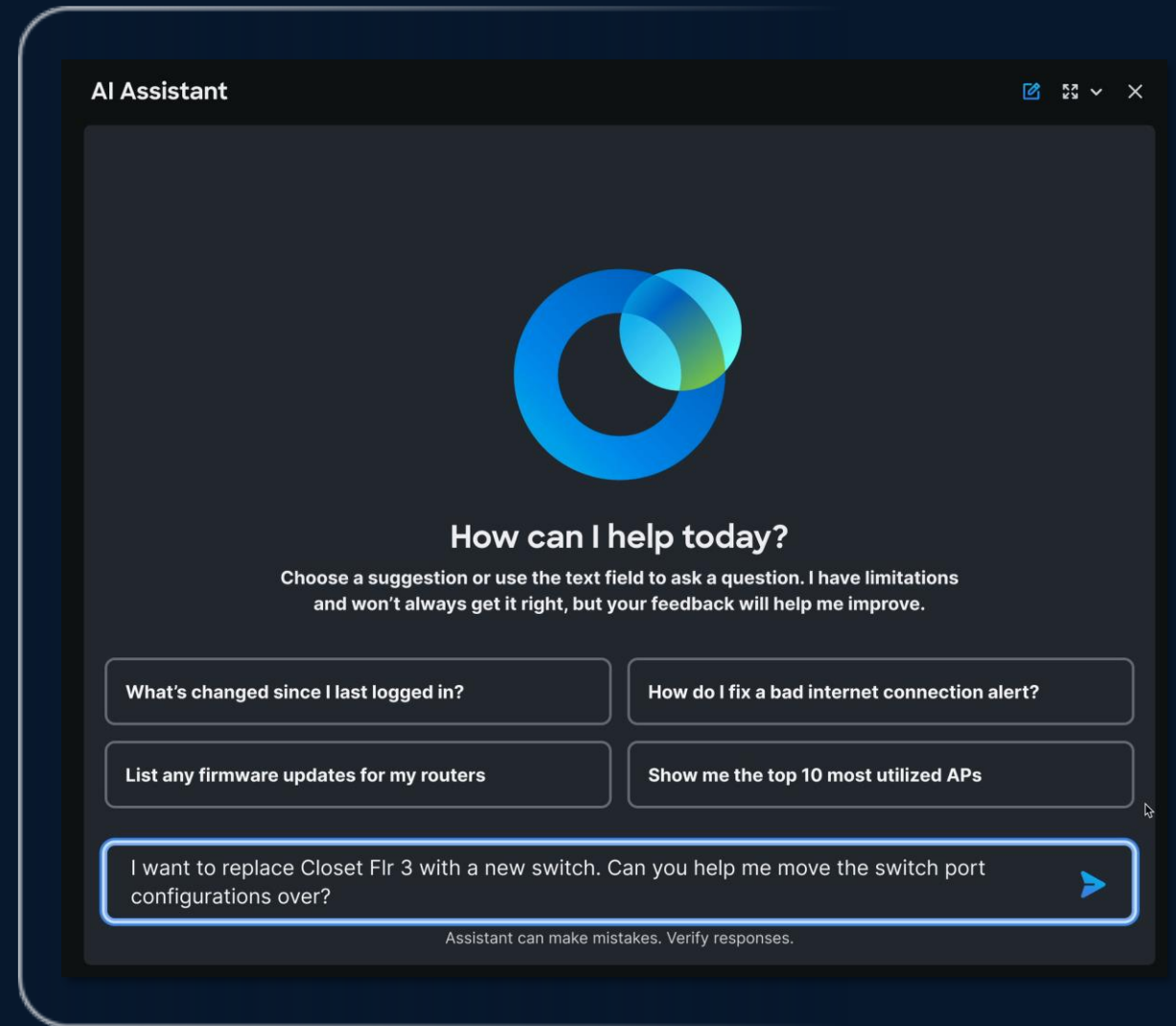
Rapid troubleshooting and Root Cause Analysis

- Identifies the root cause of an issue
 - not just the alert
- Provides full-stack visibility
- Pinpoints the issue and provides summarizations and step-by-step fixes



Automated configurations, IT workflows, and fixes

- Request automatic network changes with simple, conversational commands
- AI Assistant & Workflows integration results in streamlined, automated operations
- Cuts downtime and eliminates manual errors



AI Assistant demo in Splunk Observability Cloud (3 min),



Leila Reyhani



Leila Reyhani

AgenticOps

AI Ops

AgenticOps

Cross-domain

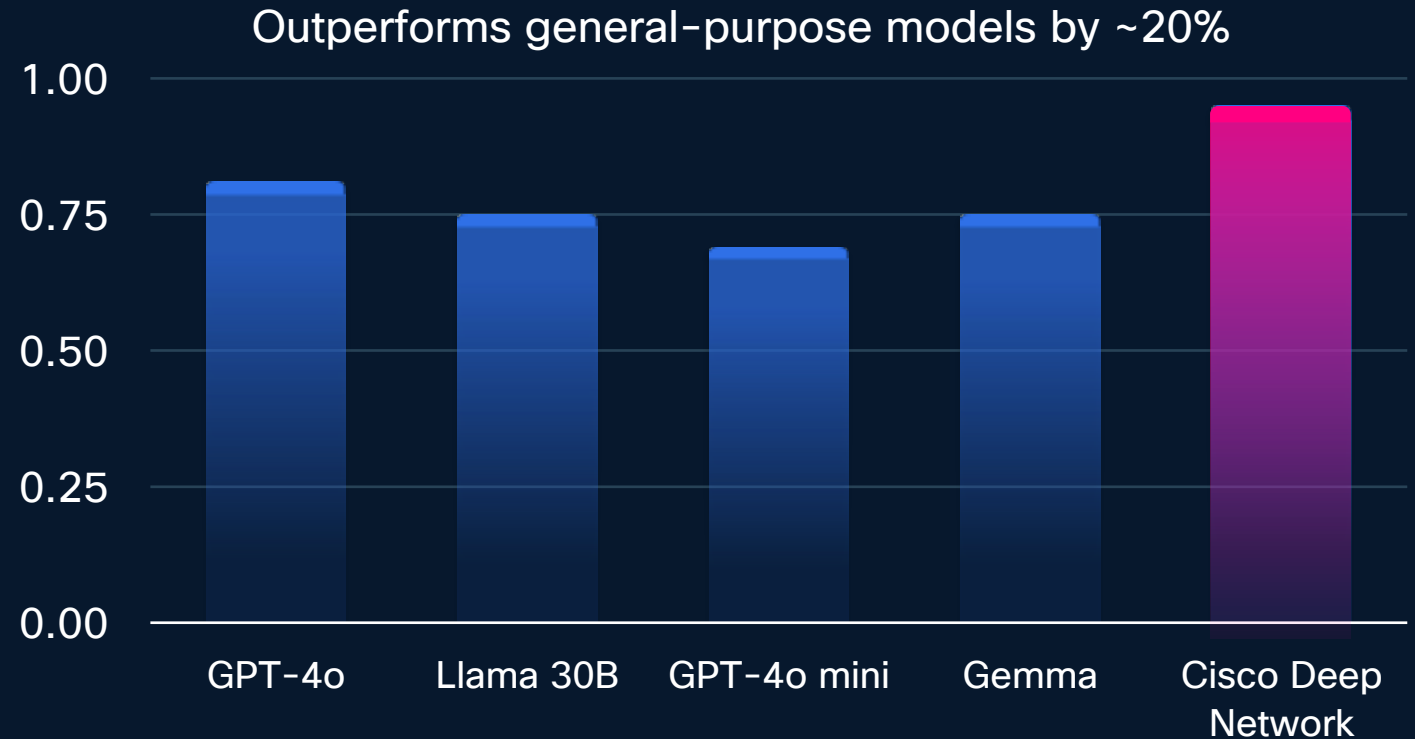
Multiplayer

Purpose-built models

Deep Network Model

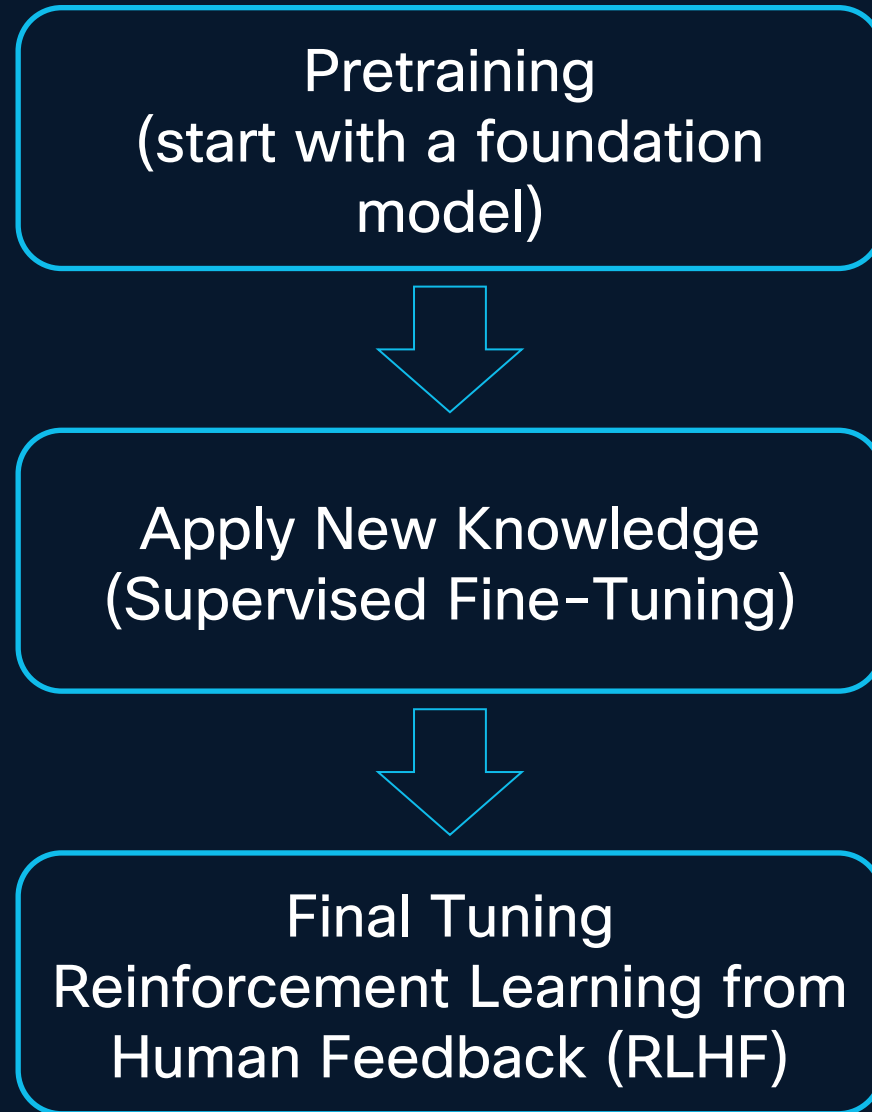
Purpose-built for networking, expert accuracy

- More precise reasoning for troubleshooting, configuration, and automation
- Fine-tuned on 40+ years of expertise and expert-vetted for accuracy
- Evolves with live telemetry and real-world Cisco TAC and CX insights



Accuracy on CCIE-style multiple choice questions (590-question benchmark), May 2025

Supervised Fine-Tuning (SFT) Steps



Demo

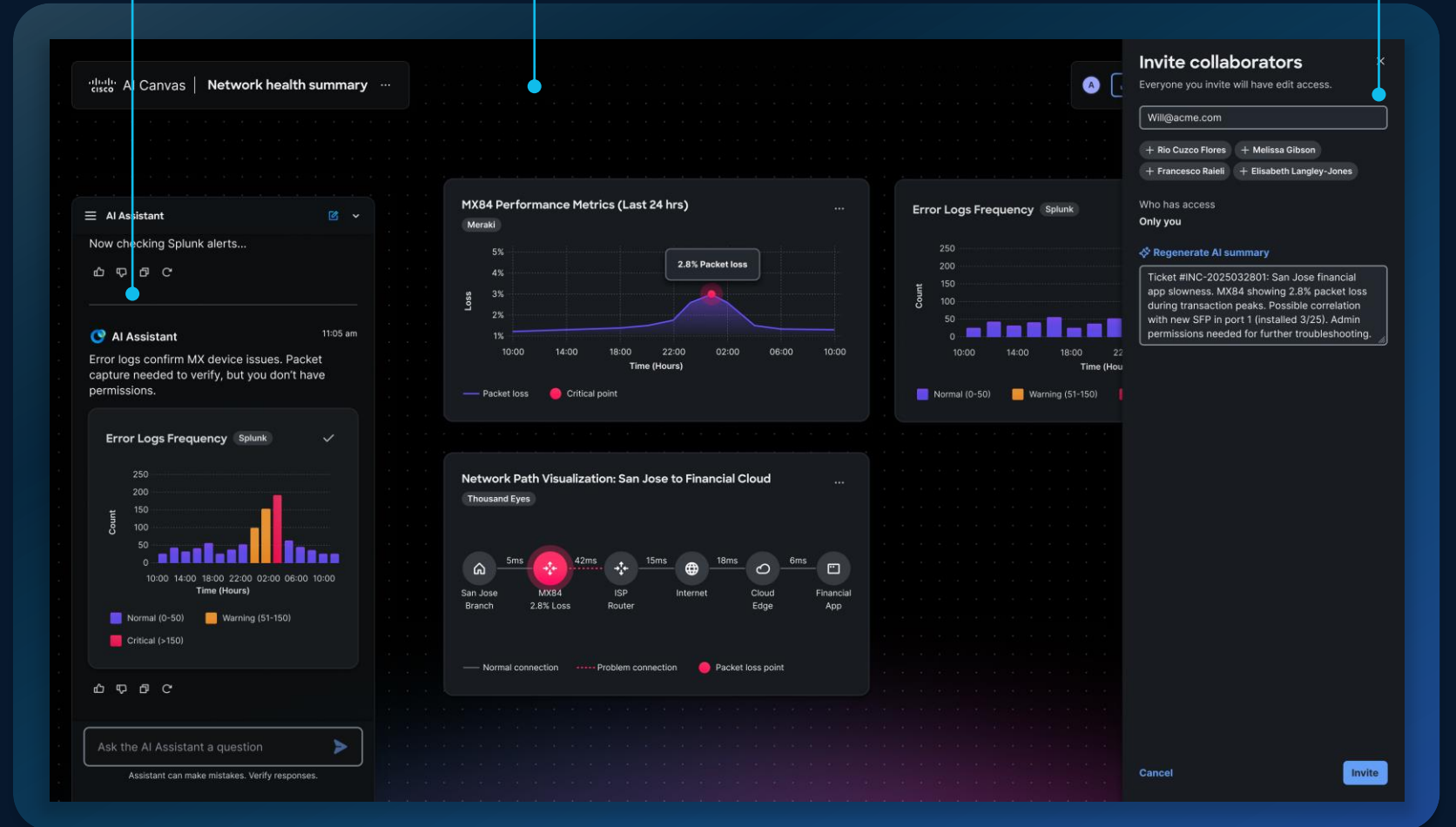
Introducing AI Canvas

- Single canvas for cross domain troubleshooting
- Generative UI with reasoning built-in
- Keeps NetOps, SecOps, IT and execs on the same page

AI Assistant

Shared Workspace

Users

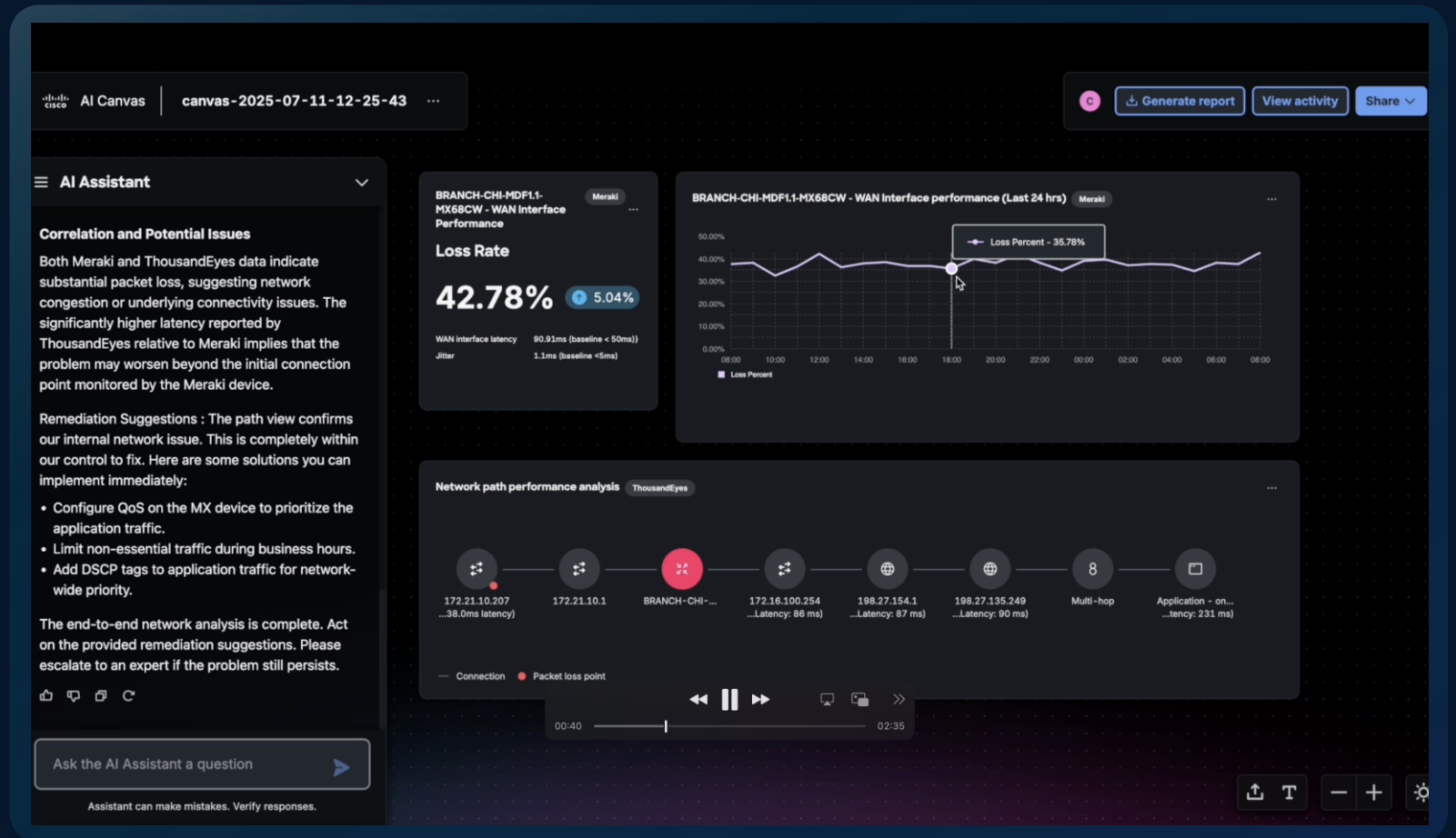


Investigate an open ticket from ServiceNow

The AI Assistant guides interactions with the network in natural language

AI Canvas reads the ticket and reasons how to solve the problem

Collects information from tools like ThousandEyes, Meraki, and Splunk to troubleshoot the issue

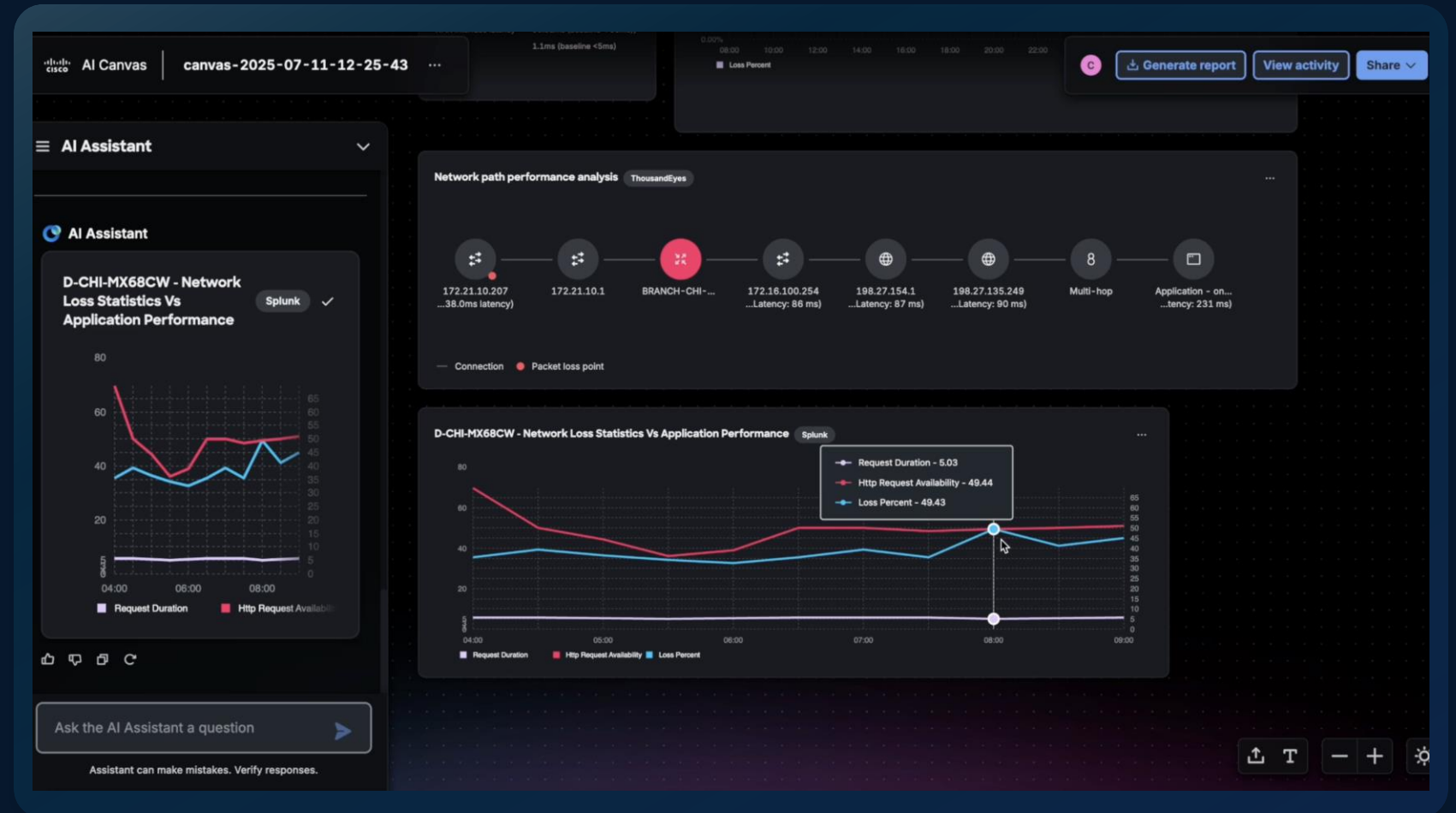


Troubleshooting and Execution Across Multiple Domains

Unifies telemetry across Cisco networking, security, cloud, apps

AI correlates and diagnoses across the full stack

Executes agentic workflows across domains



Collaboration Across Multiple Teams and Users

One shared workspace for NetOps, SecOps, IT, and execs

Drag and drop findings to persist and share with the team

AI evolves Canvas with every question, insight, and action



AI Canvas Demo

 AI Canvas

canvas-2025-07-11-12-25-43

...



Generate report

View activity

Share

Ask the AI Assistant a question

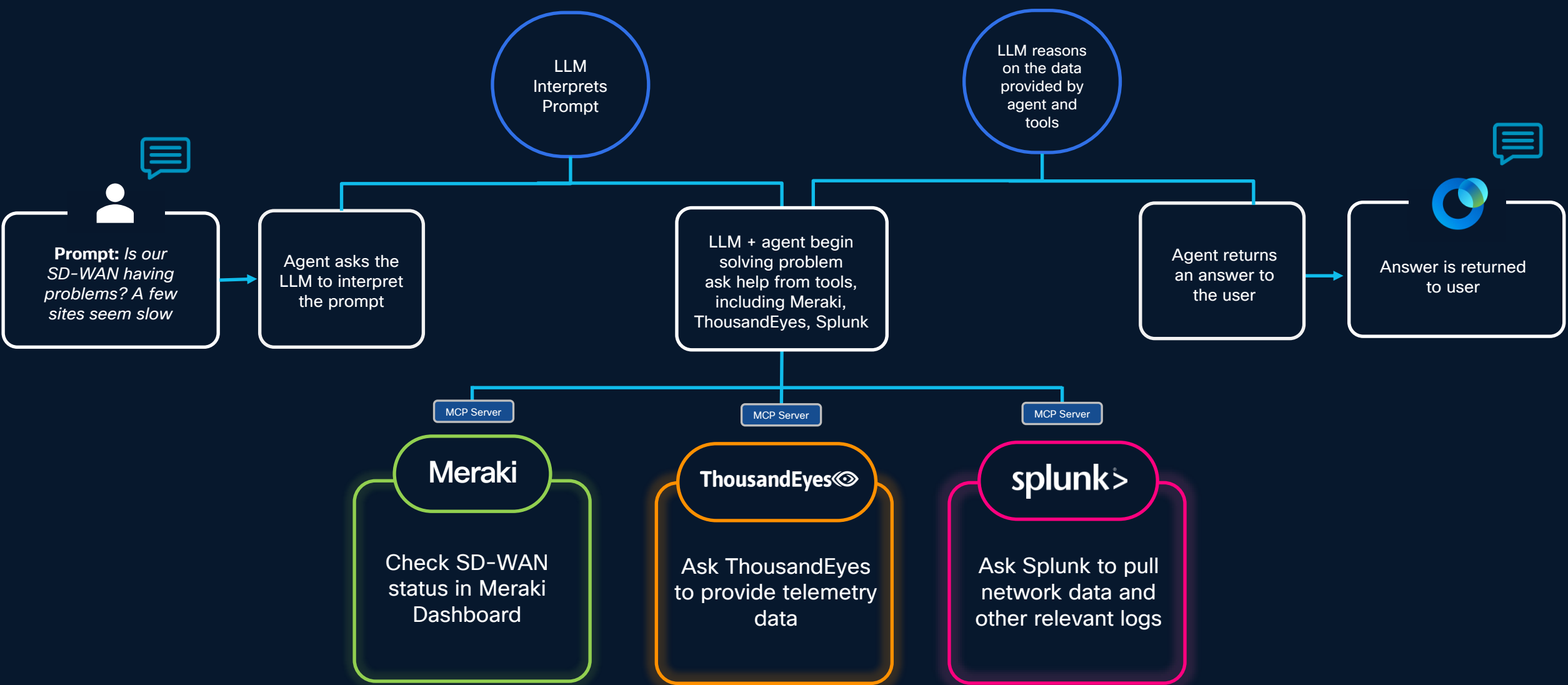
Assistant can make mistakes. Verify responses.

T

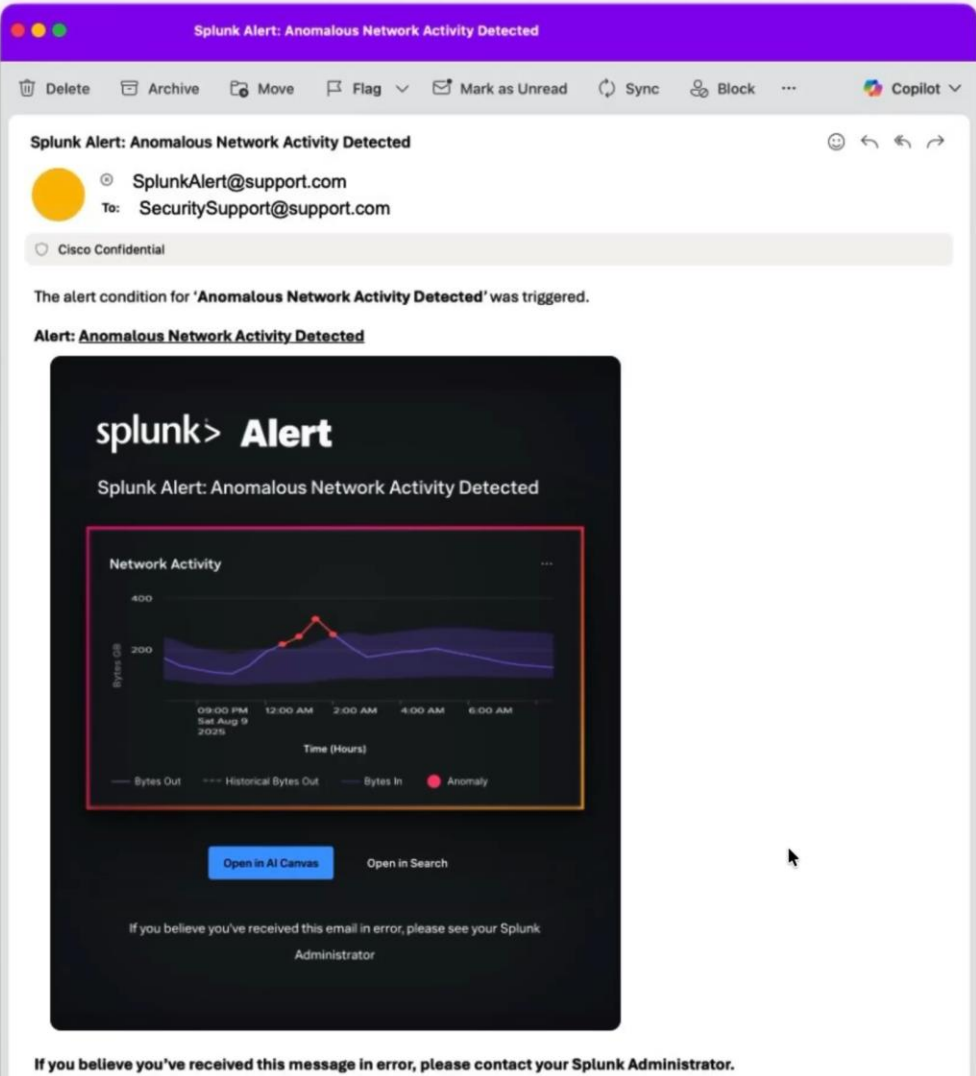
-

+

An AgenticOps Example



AI Canvas in Splunk

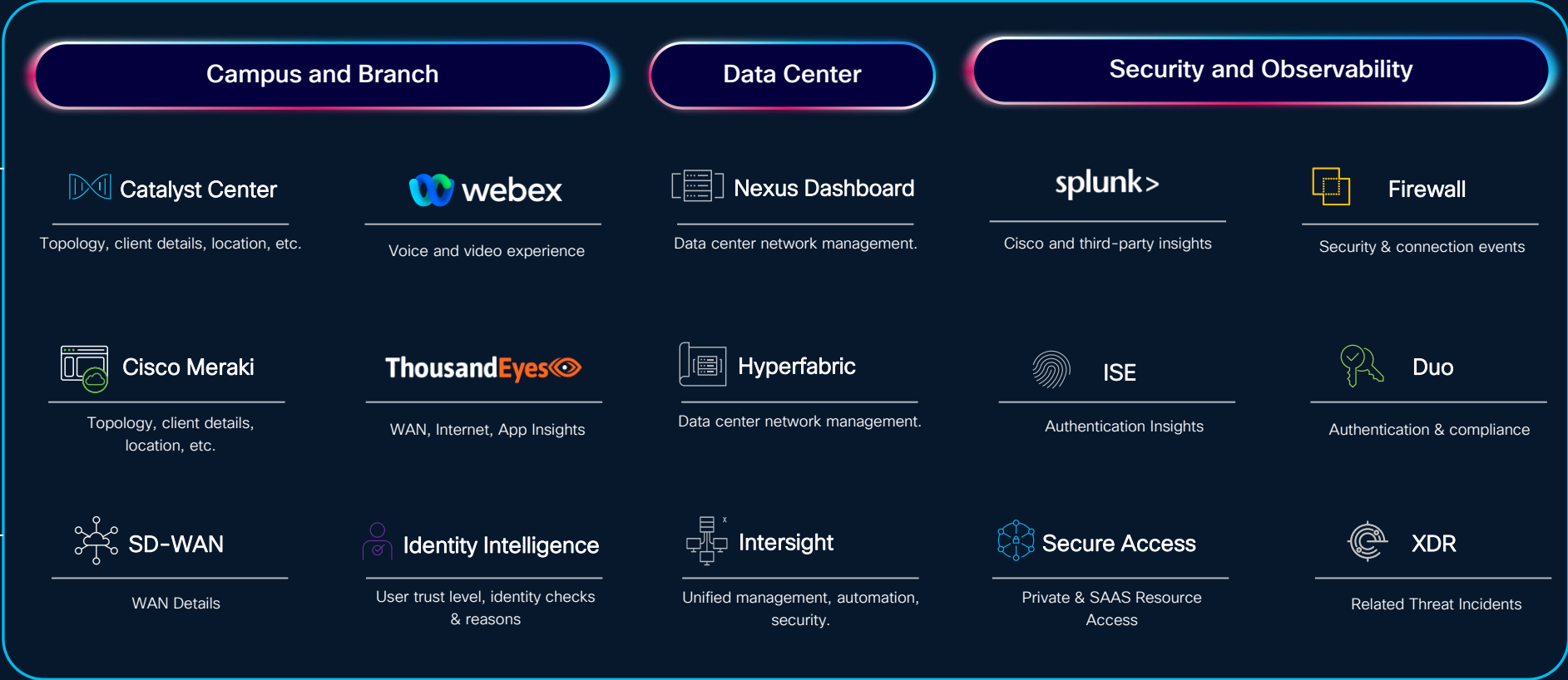
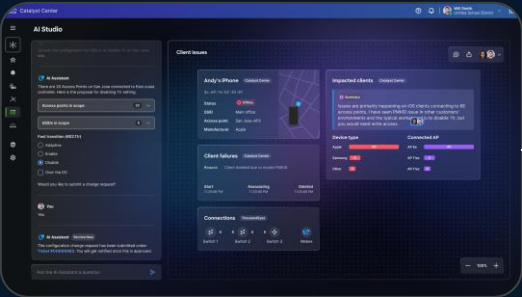


AgenticOps Across IT

AI Assistant



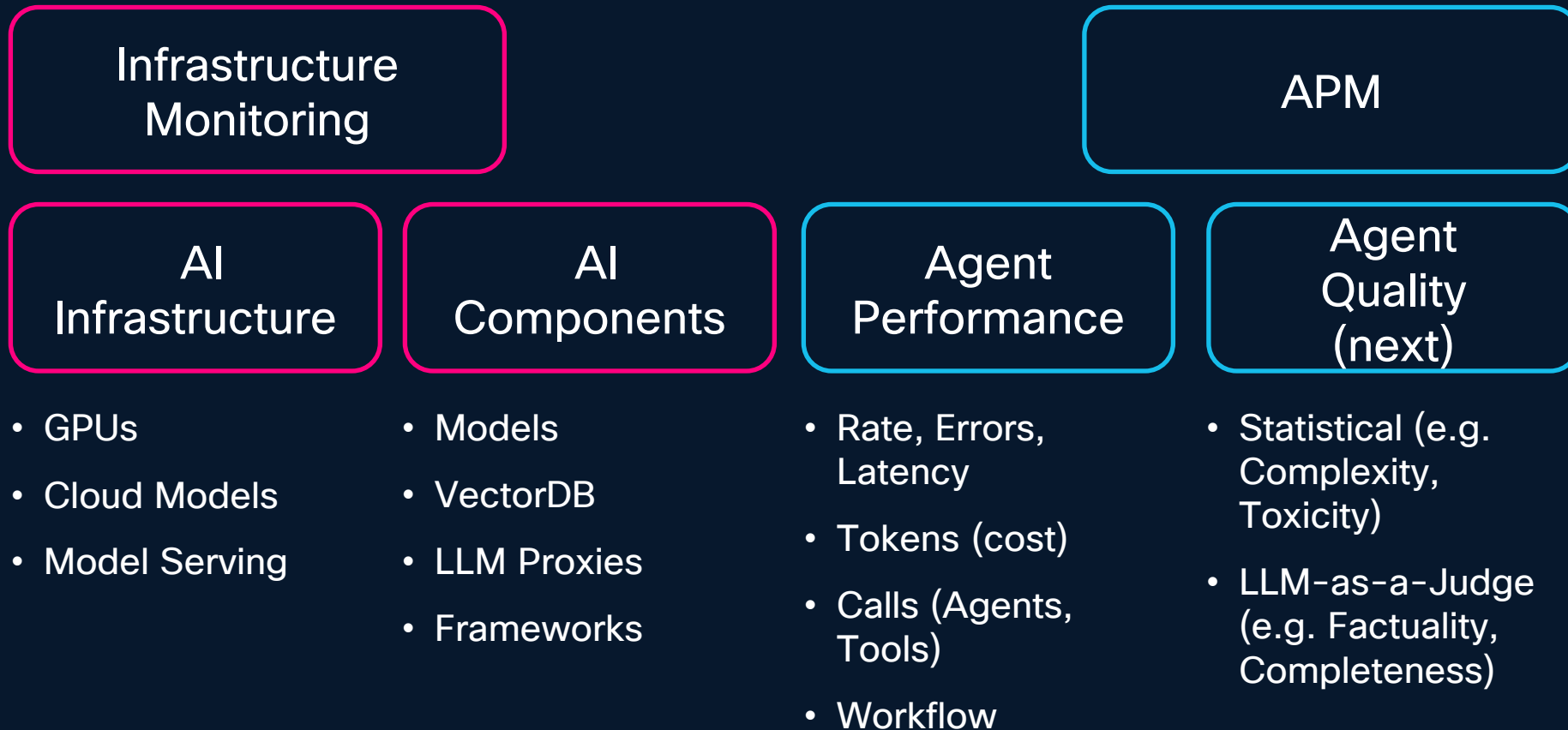
AI Canvas



Observability for AI: Software and Infrastructure Monitoring for the AI full stack

Observability for AI

Splunk Observability Cloud optimized for Agentic AI



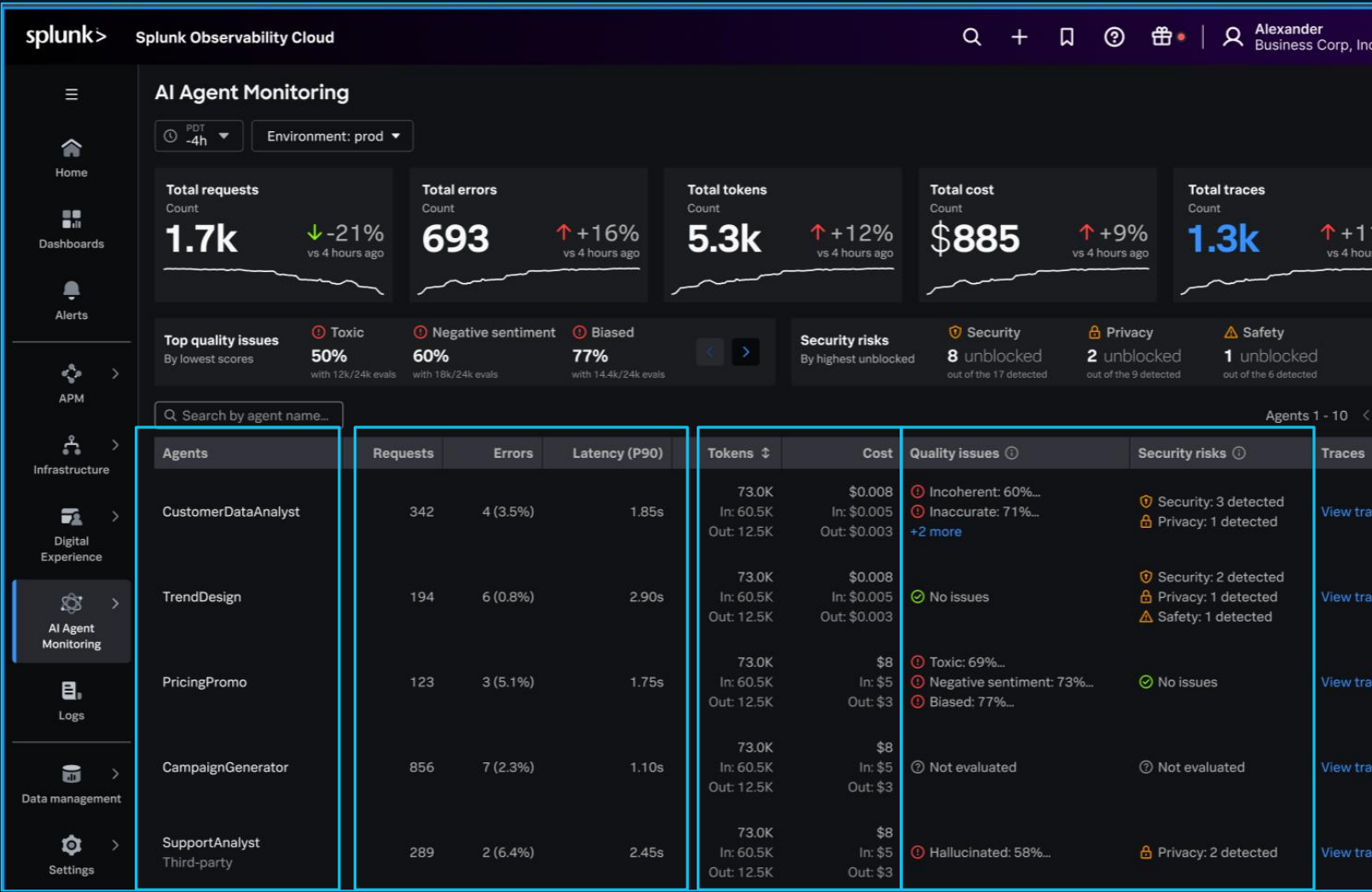
Monitoring AI Agents

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

Find root cause of accuracy, efficiency, and bias issues

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

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



- ☰
- 🏠

Home
- 📊

Dashboards
- 🔔

Alerts
- 🔗

APM
- 🏗️

Infrastructure
- 👤

Digital Experience
- 🧠

AI Agent Monitoring
- 📄

Logs

AI Agent Monitoring

🕒 PDT -4h ▾

Environment: prod ▾

Application: FlashSalesApp ▾

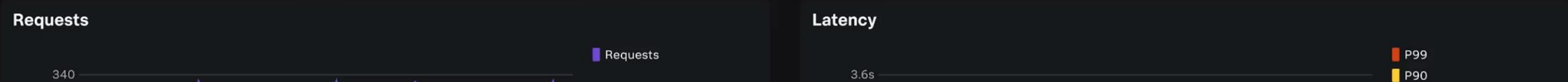
Agents

🔍 Search...

< Prev Next > Showing 5 of 17

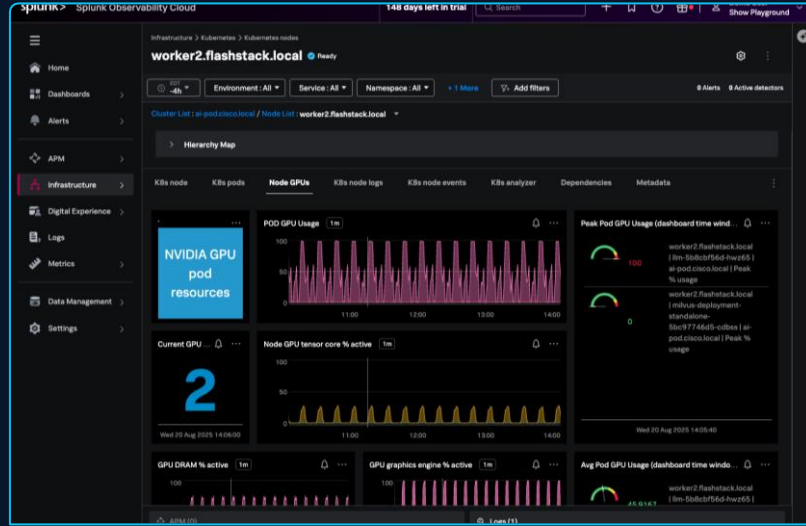
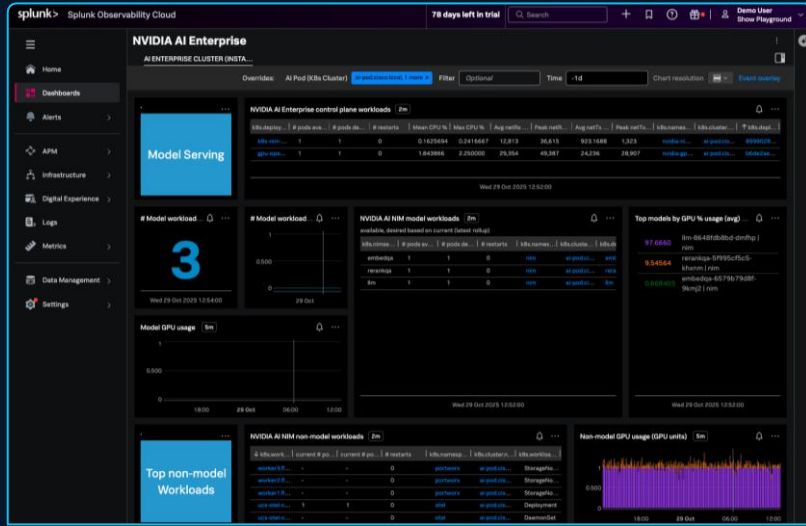
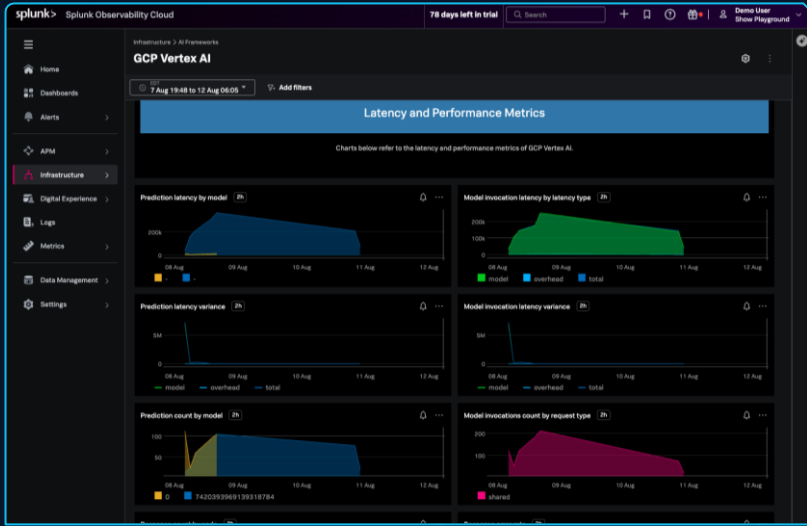
Agent	Requests	Latency (P90)	Errors	Quality	Security Risks	Tokens	Conversations
CustomerDataAnalyst	342	1.85s	4 (3.5%)	🟢 Healthy	No risk	📈 847K	57
TrendDesign	194	2.90s	6 (0.8%)	🟢 Healthy	No risk	📈 623K	89
PricingPromo	123	1.75s	3 (5.1%)	🟢 Healthy	No risk	📉 445K	120
CampaignGenerator	856	1.10s	7 (2.3%)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

Performance



Monitor Model Provider Platforms, Cloud/SaaS and Local

Monitoring performance, usage (cost) and availability wherever your models run



Cloud Hosted

- Externally Hosted Models
 - Foundation (e.g. Bedrock)
 - Custom (e.g. SageMaker)
- KPIs: Performance/Usage
- Key Providers: AWS, GCP, Azure, OpenAI

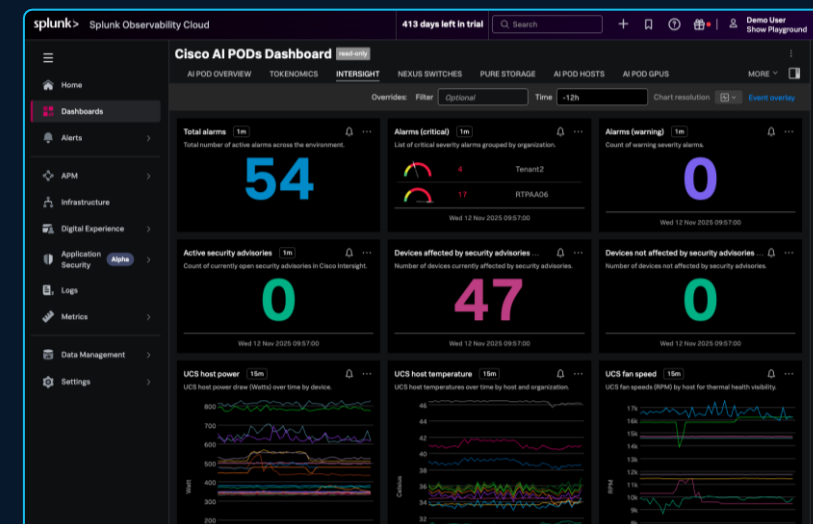
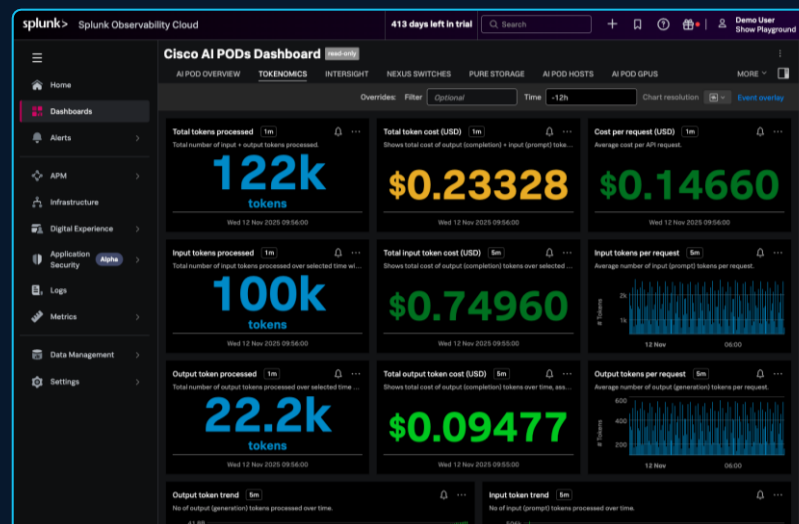
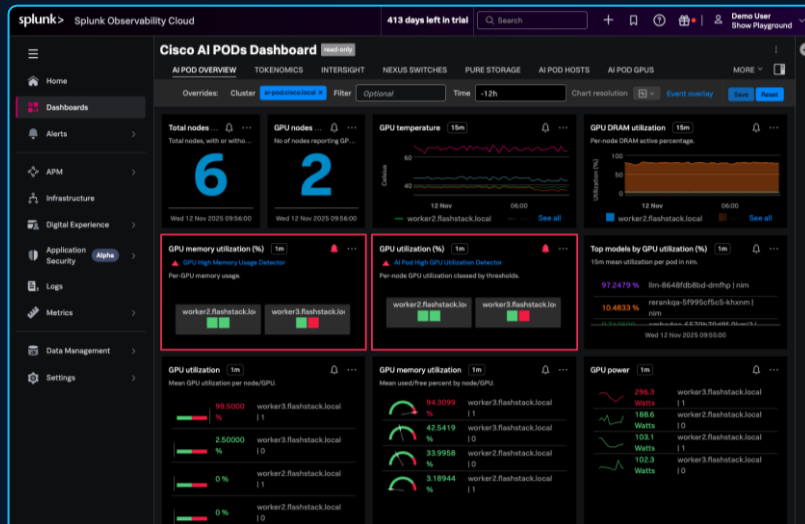
Model Serving Platforms

- Packaged + Custom Trained Models
- KPIs:
 - Performance / Usage
 - Resource Utilization / Contention
- Key Platforms: NVIDIA AI Enterprise, Ray, vLLM

Kubernetes Inferencing Clusters

- Infrastructure for Large-Scale Inferencing / Model Serving
- GPU internal details (queues, i/o, memory, power, ...)
- Node / Namespace / Workload / Pod centric views
- Key Platforms: NVIDIA DCGM, Model Serving

Monitoring Cisco AI Pods



Summary

Optimized for Platform SRE

Integrates Performance from Core Components:

- NVIDIA GPU
- Cisco UCS
- Cisco Nexus
- Storage (multiple options)
- Kubernetes Context

Tokenomics (Usage / Cost)

- Predefined views for cost attribution
- Tokens Measure Usage for each model / service / workload / namespace – relative measure
- Power consumption provides the absolute basis of costs

Component-Specifics

Individual Component Detail

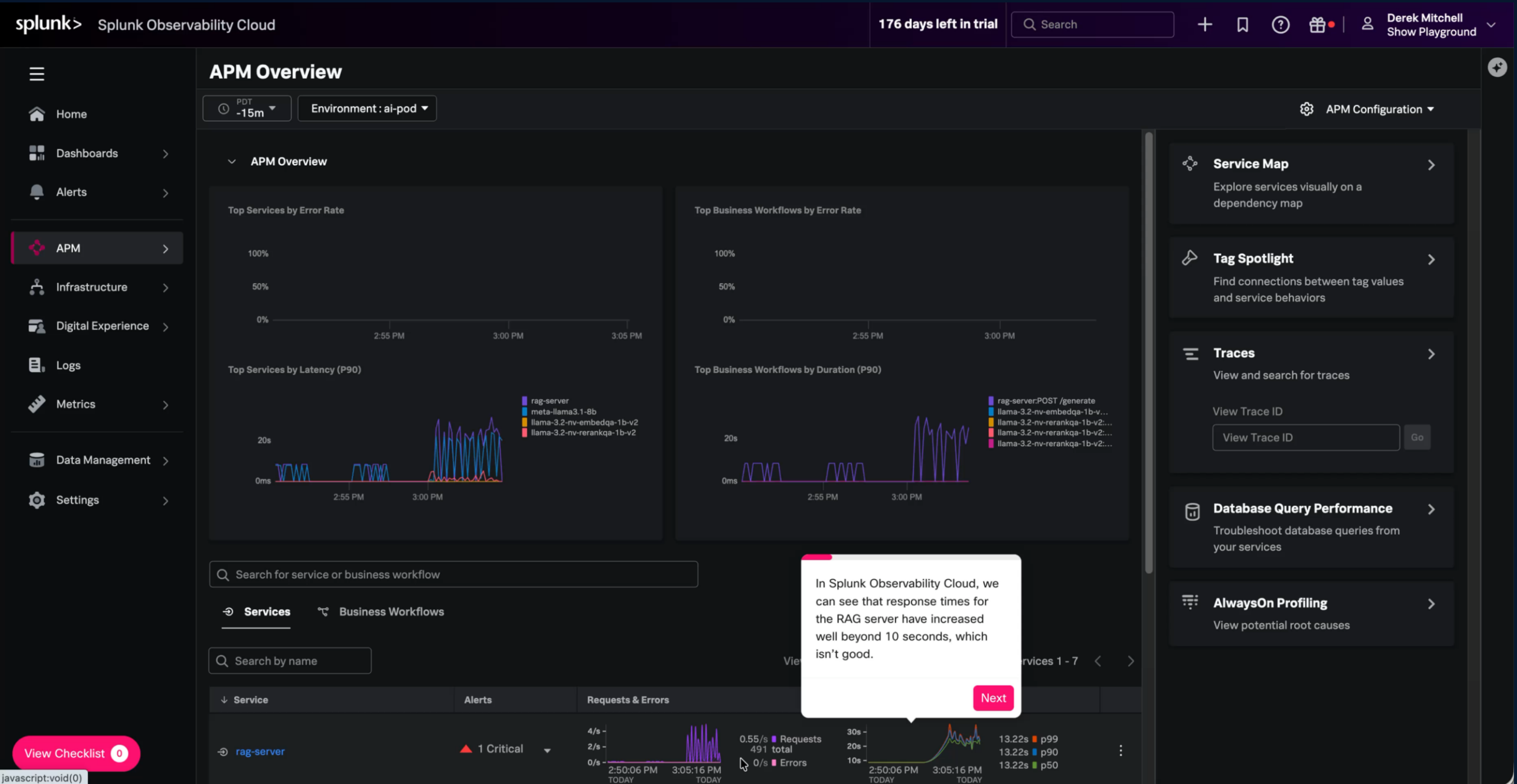
HW:

- NVIDIA GPUs
- UCS / Intersight + Nexus
- Storage / Hosts / Kubernetes

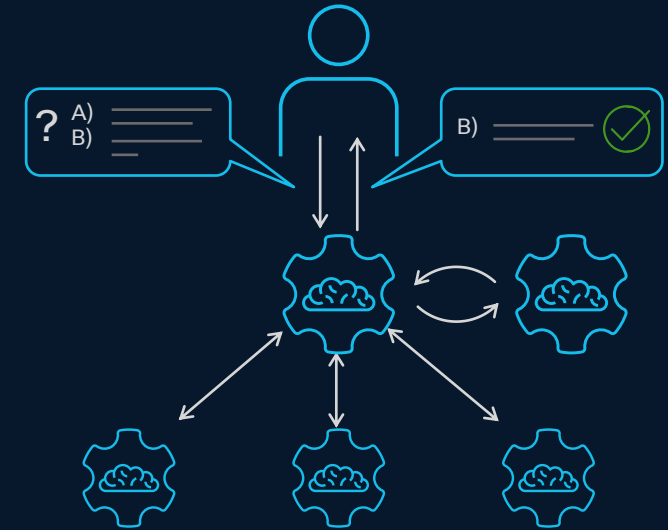
SW:

- NVIDIA AI Enterprise + NIMS
- Vector DB (e.g. Milvus)
- LLM, client app, etc.

LLM monitoring and AI Pod observability - Demo



Wrapping it up (let the agents do the job)



- Agents can do the job faster than humans
- Agents enable LLM to use tools, with human in the loop
- Agentic Operations rely on agents orchestrating specialised agents
- Splunk Observability uses agents for Alerting, Troubleshooting, Remediating, and Reporting
- AI Canvas supports cross domain, collaborative, agent-driven troubleshooting

Complete your session surveys



Complete your surveys in the Cisco Events App.



Complete a minimum of 4 session surveys and the overall event survey to receive a unique Cisco Live t-shirt.

(from 11:30 on Thursday, while supplies last)

Continue your education



Visit the Cisco Showcase for related demos



Book your one-on-one Meet the Engineer meeting

Visit the Technical Solutions Clinics to discuss your technical questions



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



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

Contact us at: Irelandi@cisco.com and Ireyhani@cisco.com

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

See a demo

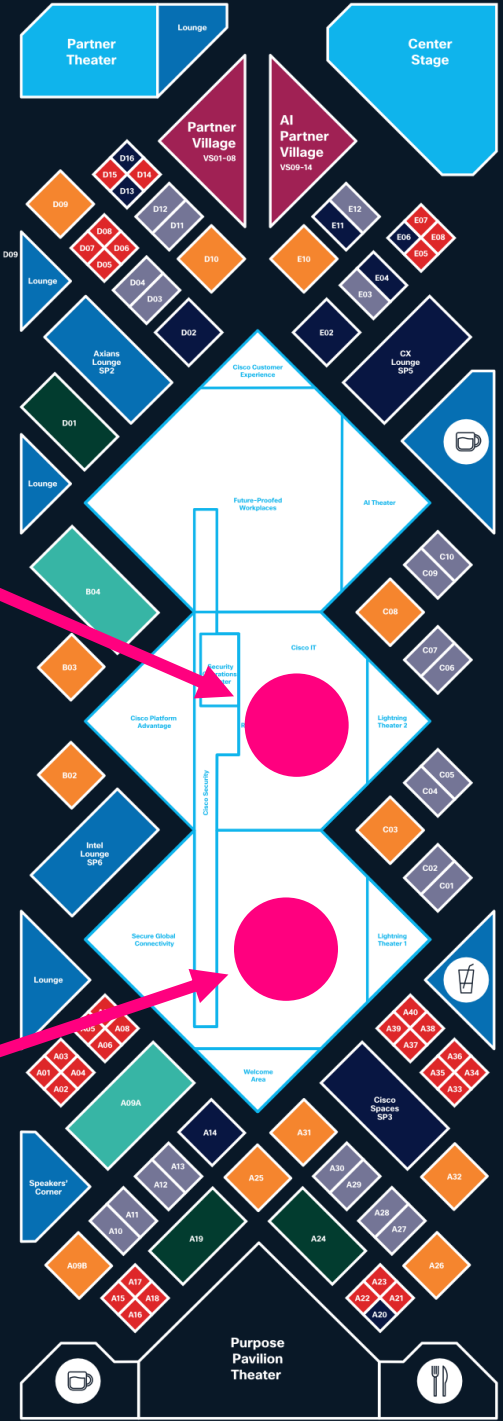


Digital Resilience

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



Thank you

CISCO Live !

