

Navigating the Agentic AI Era

Charles Li, AI Solutions Engineer



Our Objective Today



Clarity

Uncover what agentic AI really is,
beyond the buzzword



Control

Understand how to build AI-ready
infrastructure



Confidence

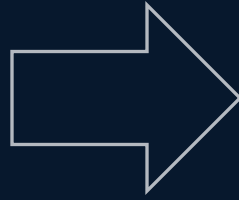
Discover how we can modernize our
environment to embrace AI innovation

Evolution of Gen AI



Chatbots

Humans talk to AI



Agentic

Workflows get automated




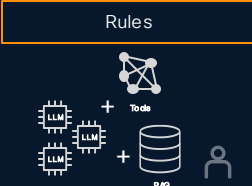


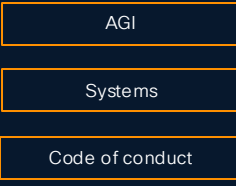
AI will make our world of **8B** people feel like one with the capacity of **80B**

A Brief History of Gen AI and View Ahead...

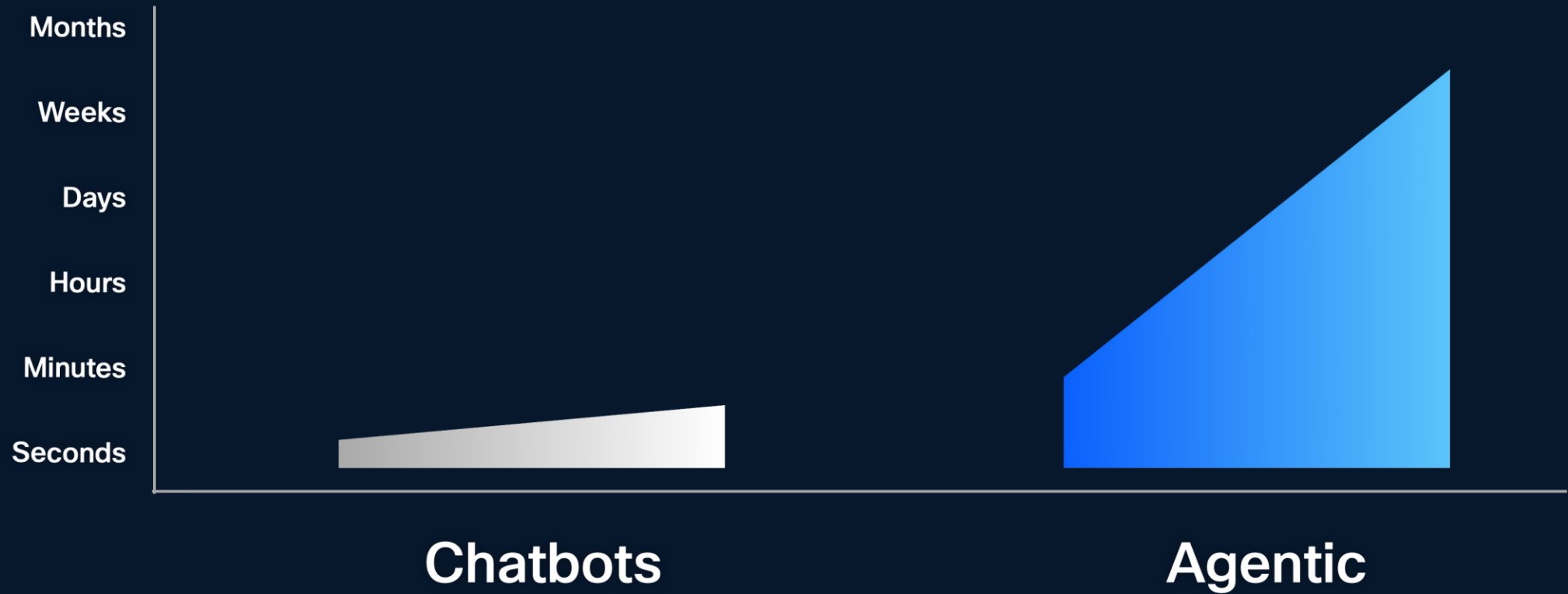
AI
Chat

Attended
AI Agents

Unattended
AI Agents

	2022	2023	2024	2024/2025	2026/2027	2030+?	
	Standalone AI Chat	RAG Powered AI Chat	Reasoning AI Chat	Standalone AI Agent	Multi AI Agent System	Built for purpose	General purpose
							
Description	Human interaction with standalone AI model, typically through a chat interface.	Human interaction with standalone AI models with enriched context from a preconfigured read only data source, typically through a chat interface.	Human interaction with multiple sequential and/or parallel AI models (RAG optional), typically through a chat interface.	Automation solution with multiple sequential and/or parallel AI models, controlled access to read/write in tools through preconfigured business logic to solve specific problems with human in the loop .	Automation solution relying on a preconfigured orchestrator with access to multiple specialized agents to solve a group of specific problems.	Automation solution relying on one or several AI powered orchestrator(s) with access to multiple specialized agents to solve a group of specific problems.	General purpose open-ended agents with access to any tools, code writing and execution to solve any problems autonomously within predefined guidelines.
Example	ChatGPT (GPT-3.5)	Internal / External Knowledge Chatbot (GPT 4o + RAG)	Enhanced Knowledge Chatbot (GPT o1, R1 + RAG)	Invoice Processing Agent, Cash Rec. Agent, Contract Agent, Copilot	GPT Accountant, GPT Lawyer, GPT Recruiter	Anthropic Computer Use, OpenAI Operator, Microsoft OmniParser, Manus ai	AI Interface (replacement of GUI)

Duration of Autonomous Execution



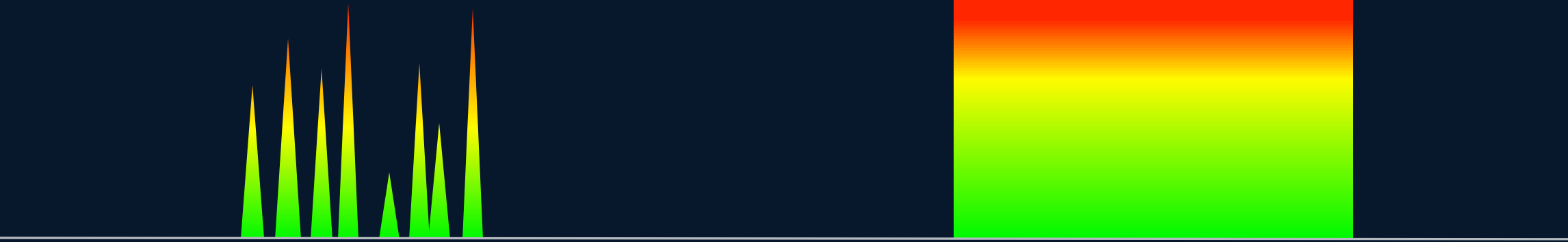
AI is changing: Token Inflation

+ Help me figure out the seating chart for my wedding reception. I need to have both parents on either side of me and my bride. The best man and maid of honor can't sit near each other or my weird uncle.

- 📎 Add photos & files
- 👤 Agent mode
- 🔭 Deep reasoning
- 🖼️ Image
- ...more

Power | Compute | Networking

10 – 200 x tokens



One shot

Agentic

Interaction Model

Demo

Web Application

Investigation Submission

Orchestrator Agent

Decides how to tackle the task at hand

Planner Agent

Decides how to breakdown the task

User Profile Agent

Analyze the user and gather information about the user

Transaction Retrieval Agent

Analyzes relevant transactions from the user

Behavioral Agent

Analyzes user's spent behavior

Fraud Agent

Analyzes whether the transaction is likely to be associated with fraud

Decision Agent

Summarizes all the finding and provides a decision on the investigation

Standard Operating Procedure

Task 1: Analyze the user and gathers relevant information from the user profile

Task 2: Gather relevant transaction from the user

Task 3: Analyze the user's spent behavior

Task 4: Conduct a fraud analysis to calculate the risk for fraud

MCP Server

Agent Registry

- Behavioral Agent
- Transaction Retrieval Agent
- Fraud Risk Agent
- User Profile Agent
- Decision Agent

Tool Registry

- Find best agents
- Find best tool
- Transaction Database Lookup
- User Profile Lookup
- Spent Behavioral Analysis
- Fraud Analysis

CISCO AI DEMO



Transaction Investigation - Agentic AI Demo

Compare single-LLM reasoning with the demo agentic workflow.

Enter a natural language investigation prompt. The system will run both the standalone LLM analysis and the multi-agent, tool-backed flow.

INVESTIGATION PROMPT

Investigation Prompt

Charles Li has a recent transaction of \$500 in Mexico, help me investigate and determine if it is suspicious. Explain your reasoning

Run Investigation



Case Study: AI Usage at Cisco

AI Chat
(80K users)



6B Tokens
per month

AI Deep Research
(API usage)



80B Tokens
per month

Why Dedicated Compute?

Running AI models on premise is ideal for use cases with any of the below requirements:

Low latency

Low latency and high performance are critical.

Data sensitivity & data sovereignty

Data sensitivity is high and/or regulatory compliance mandates data sovereignty.

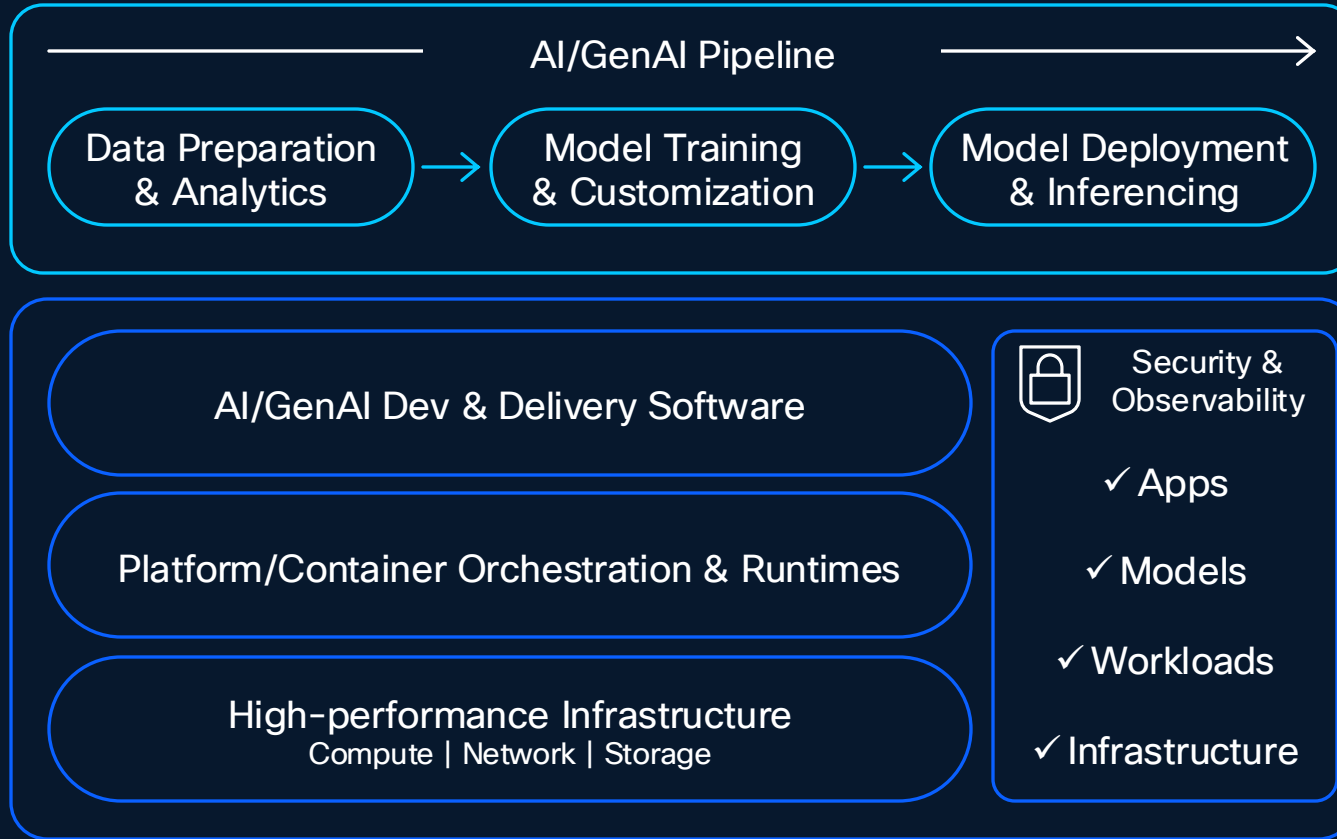
High availability (controlled capacity constraints)

When throttling by public cloud providers due to capacity constraints is not acceptable.

Avoiding vendor lock in

Nobody knows who is going to win the AI race, avoiding vendor lock in allows flexibility.

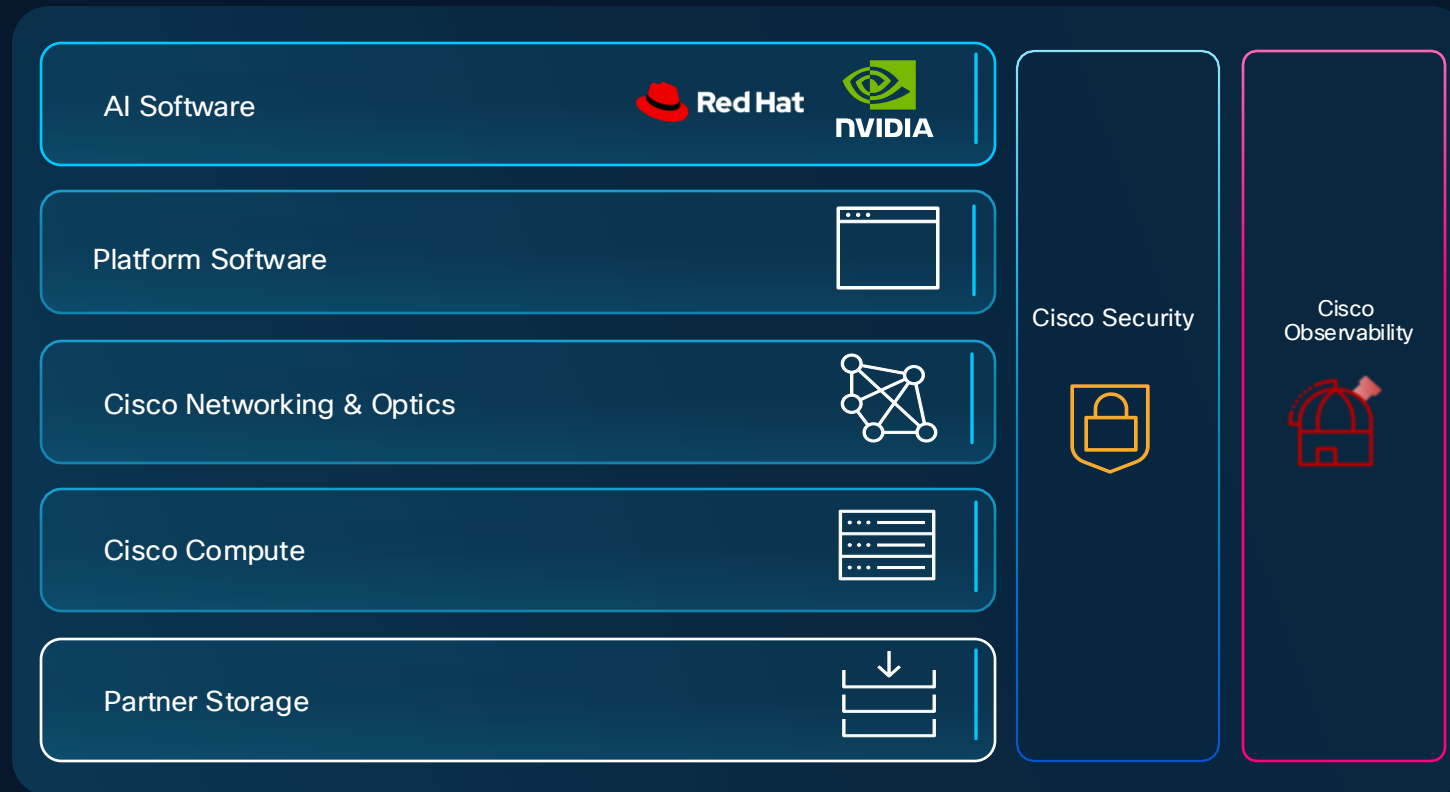
Key Capabilities Required to Operationalize AI Infrastructure



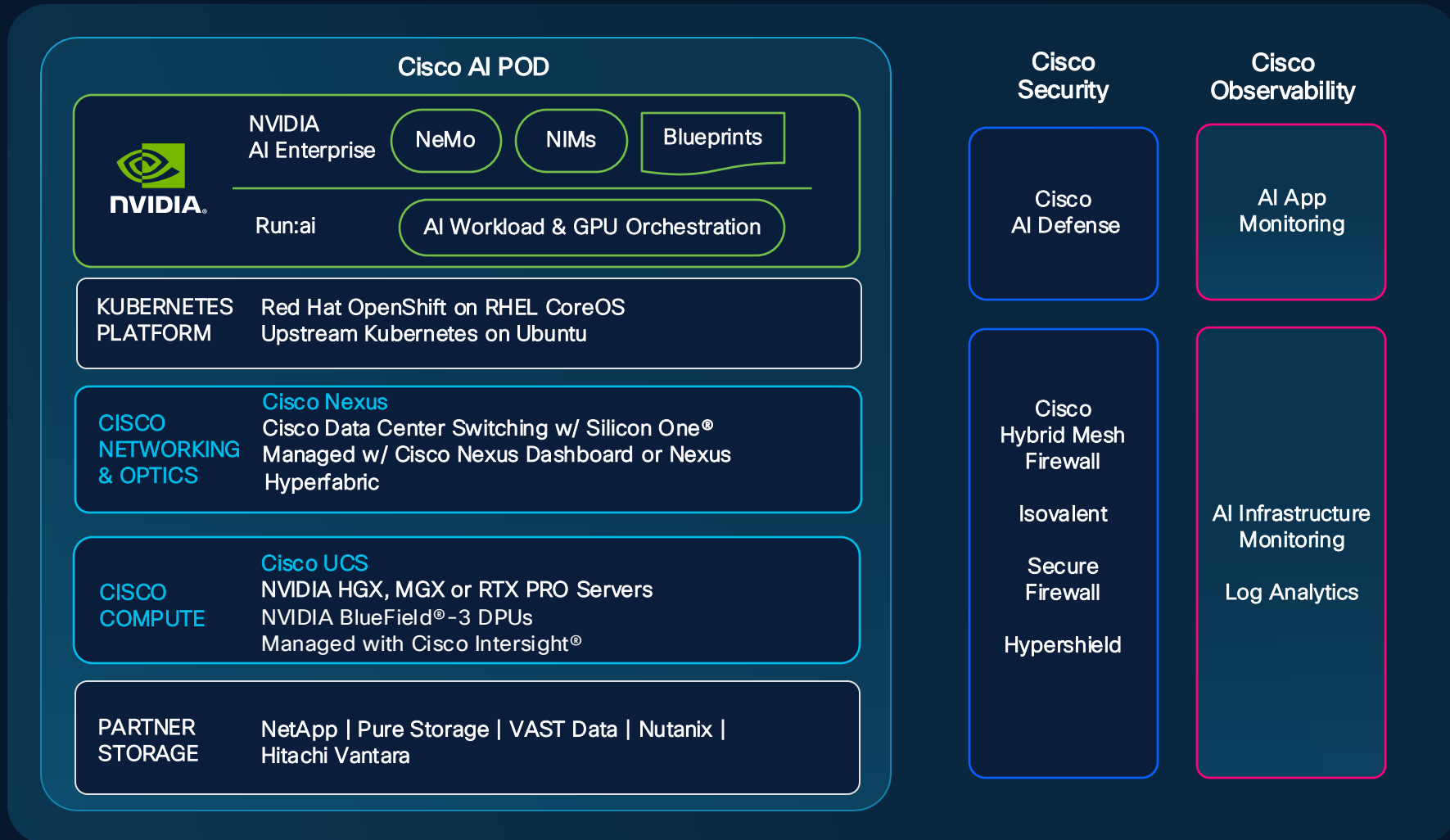
- Faster time to business value
- Mitigate risks
- Simple deployment

Cisco Secure AI Factory with NVIDIA

Cisco becomes the first partner to offer a NVIDIA Cloud Partner (NCP) compliant reference architecture with new data center switching solutions. Together, Cisco and NVIDIA provide customers ultimate flexibility as they build critical AI infrastructure

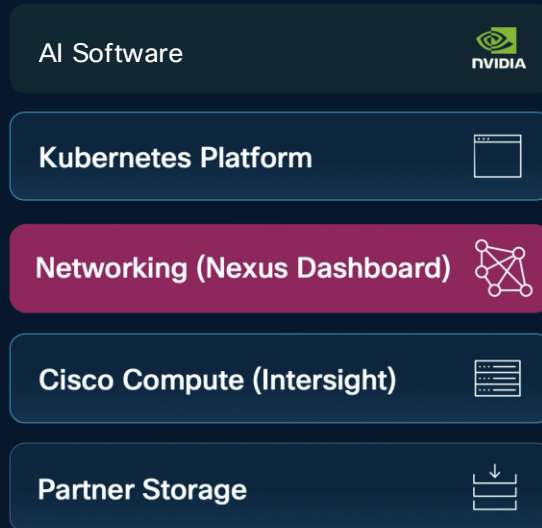


Cisco Secure AI Factory with NVIDIA

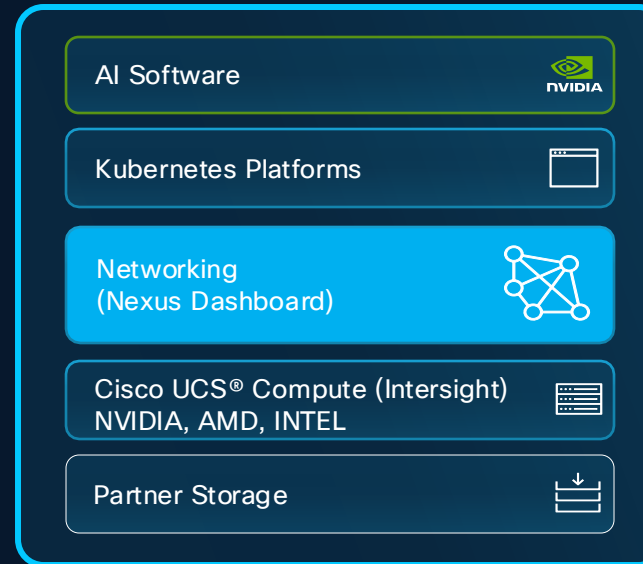


Flexible Deployment Options

Build your own



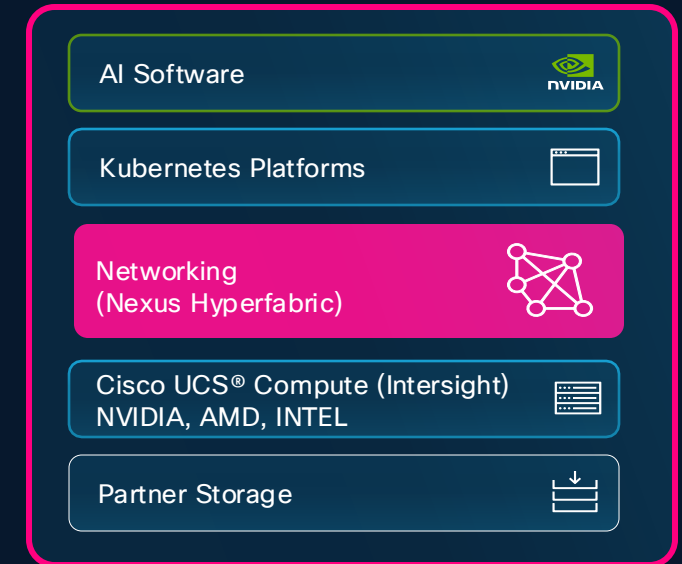
AI POD w/ On-prem management



Modular, pre-validated infrastructure:

- Full stack, buy & deploy
- Nexus Dashboard: On-prem networking management

AI POD w/ Cloud management



Turnkey infrastructure:

- Full stack, buy & deploy
- Nexus Hyperfabric: Cloud-managed Networking
- Nexus Hyperfabric AI: Cloud-managed physical infrastructure

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

Supporting Latest GPUs

Supporting RTX PRO 6000 Blackwell Server Edition and B300 GPUs



Cisco UCS®
GPU-dense servers
PCIe and NVLink Servers



Cisco UCS blade (with GPU extensions) and
rack servers

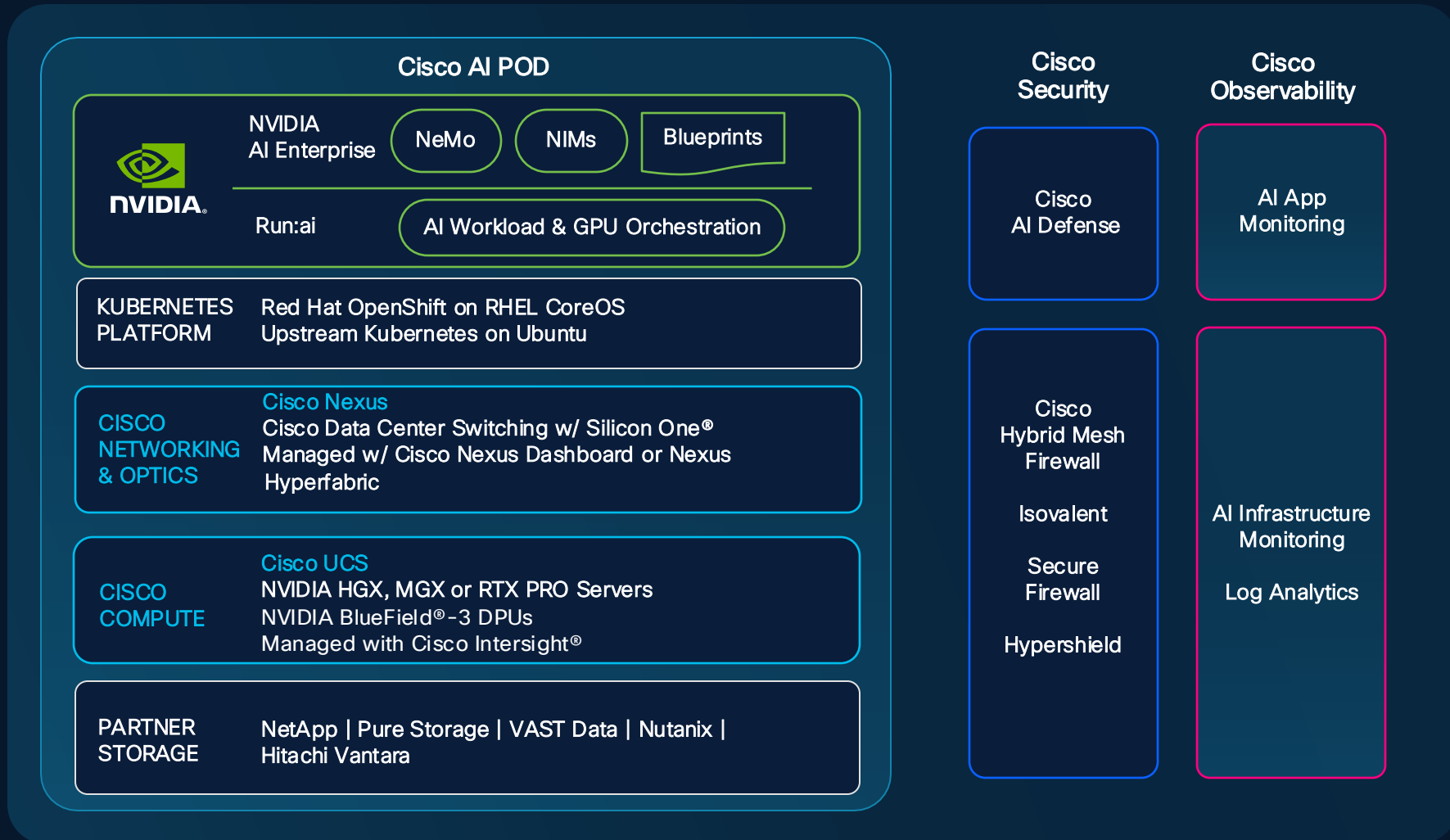


Enterprise AI edge

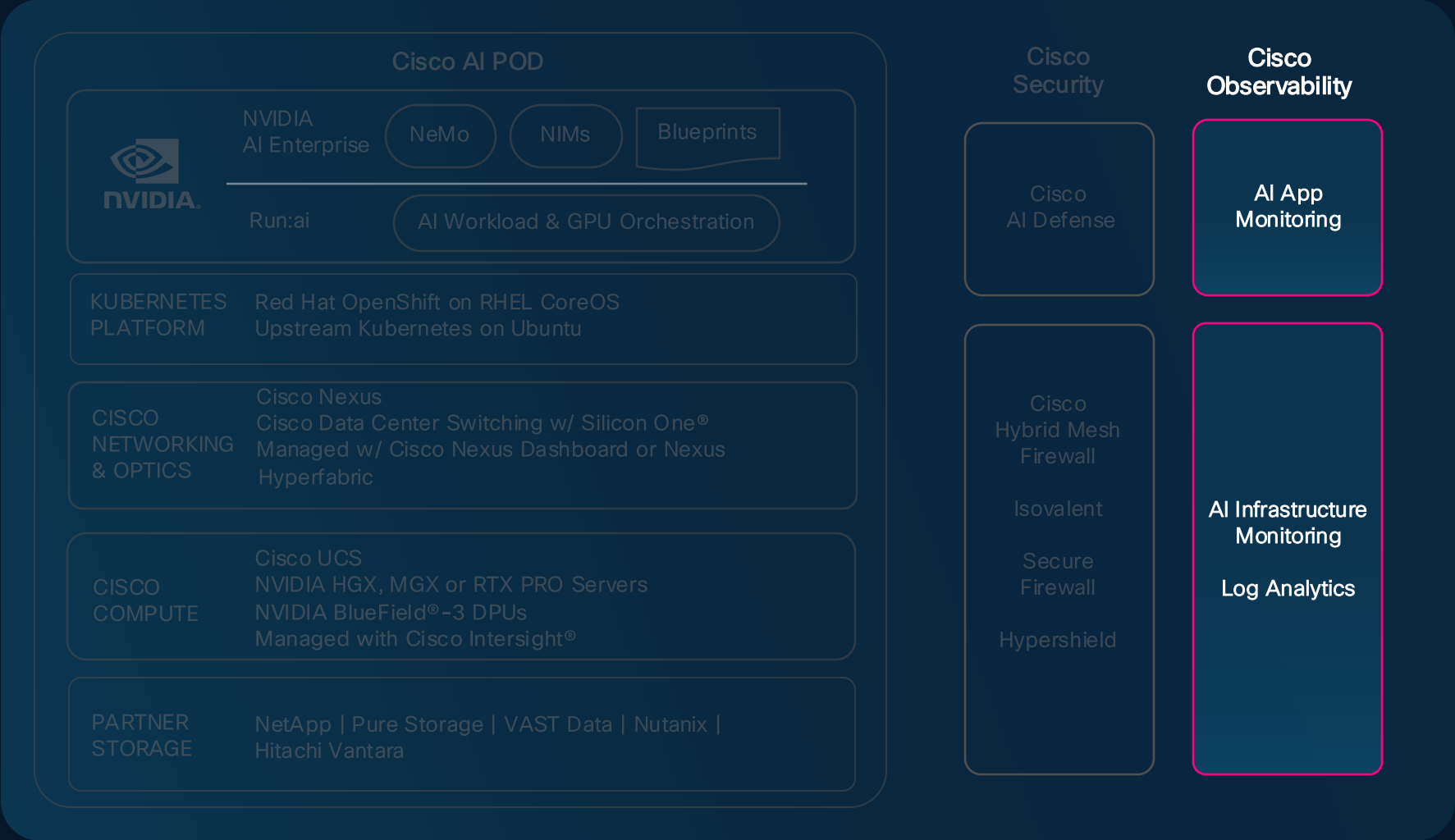
Dense compute for AI

Full-stack AI with compute and networking

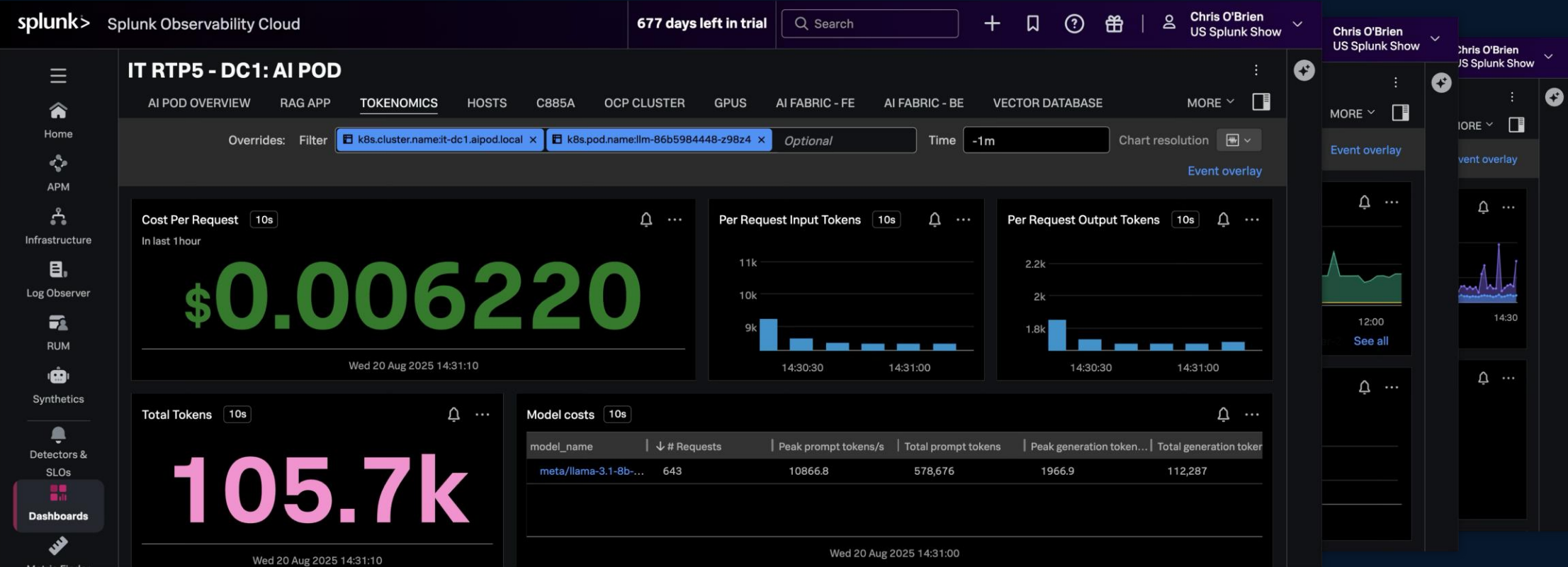
Cisco Secure AI Factory with NVIDIA



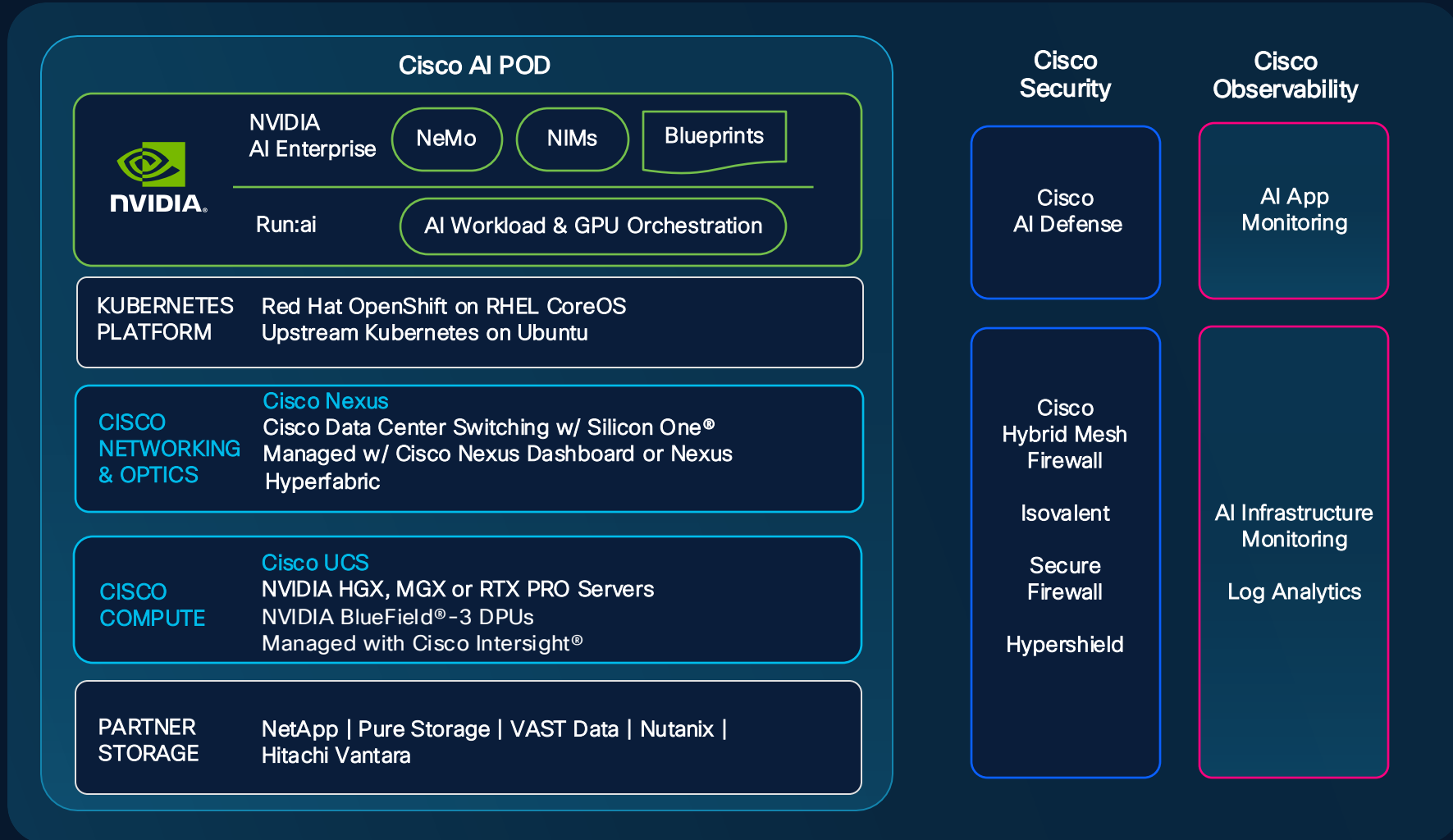
Cisco Secure AI Factory with NVIDIA



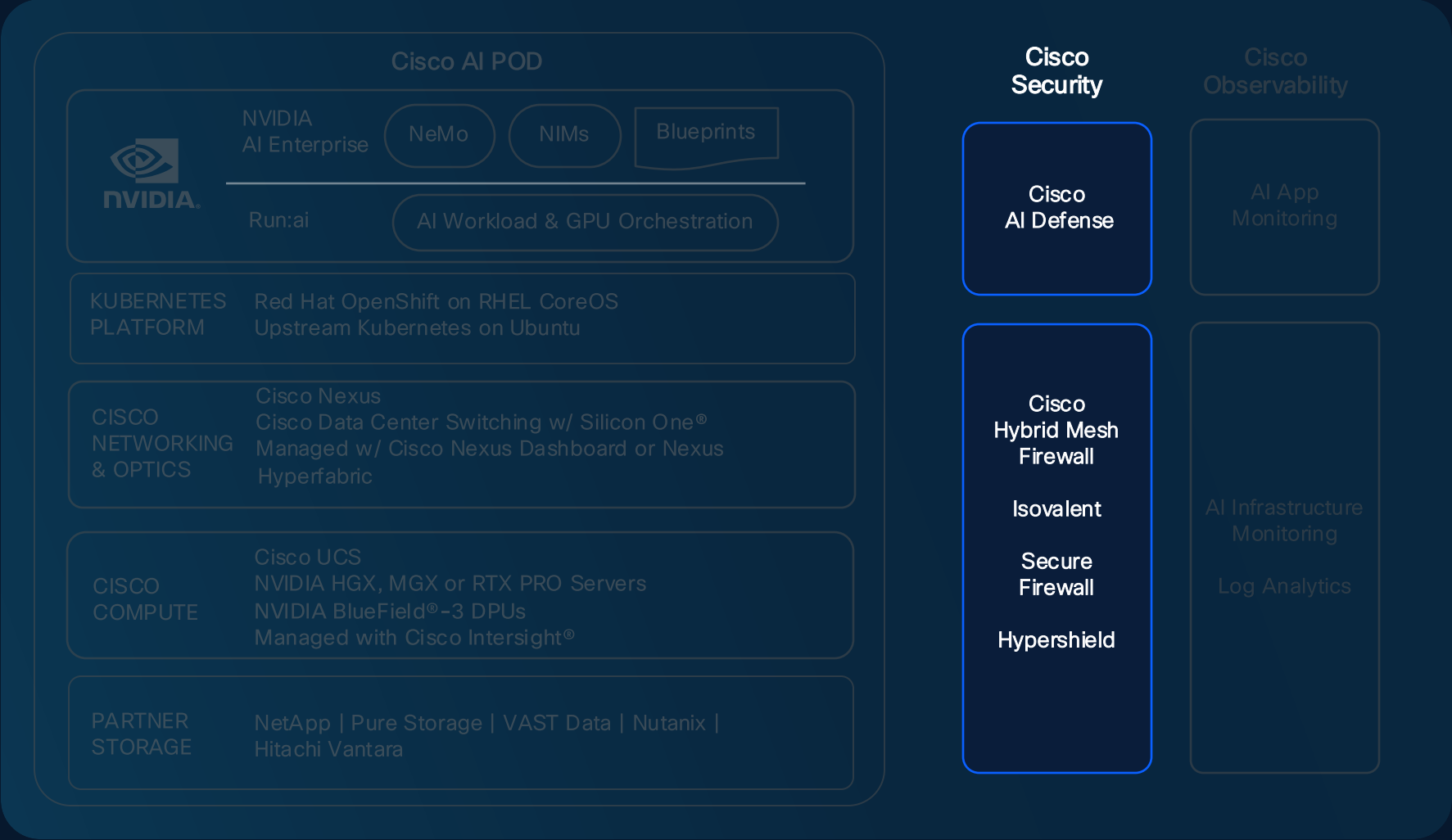
Observability for AI



Cisco Secure AI Factory with NVIDIA



Cisco Secure AI Factory with NVIDIA



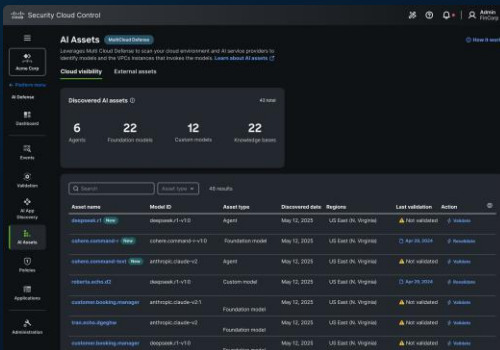
AI Defense: Coverage across the AI lifecycle

Discovery

AI Cloud Visibility

Identify AI assets

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

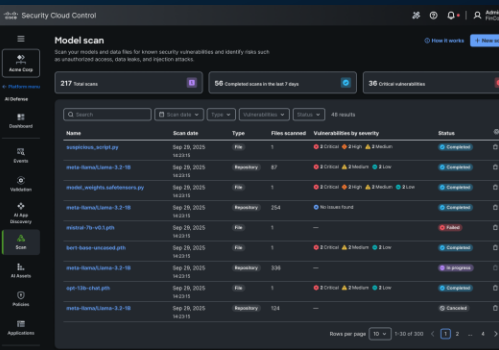


Detection

AI Supply Chain Risk Management *

Scan for threats

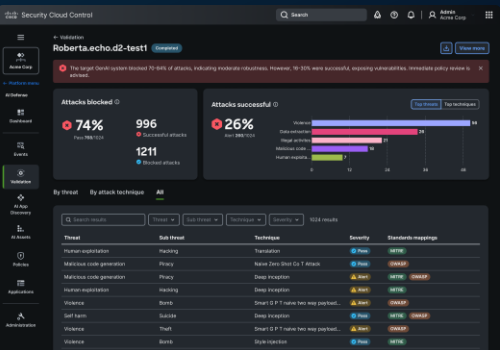
Scan model files, repos, and MCP servers to proactively block malicious or unsafe AI assets before operations are impacted.



AI Model & App Validation

Detect the vulnerabilities

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

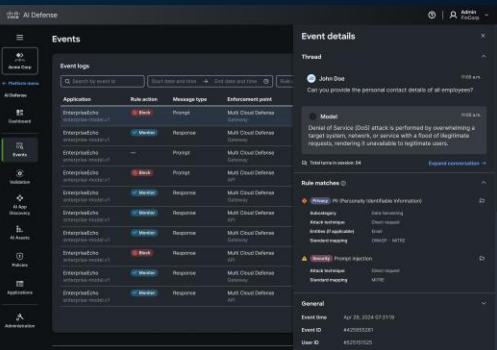


Protection

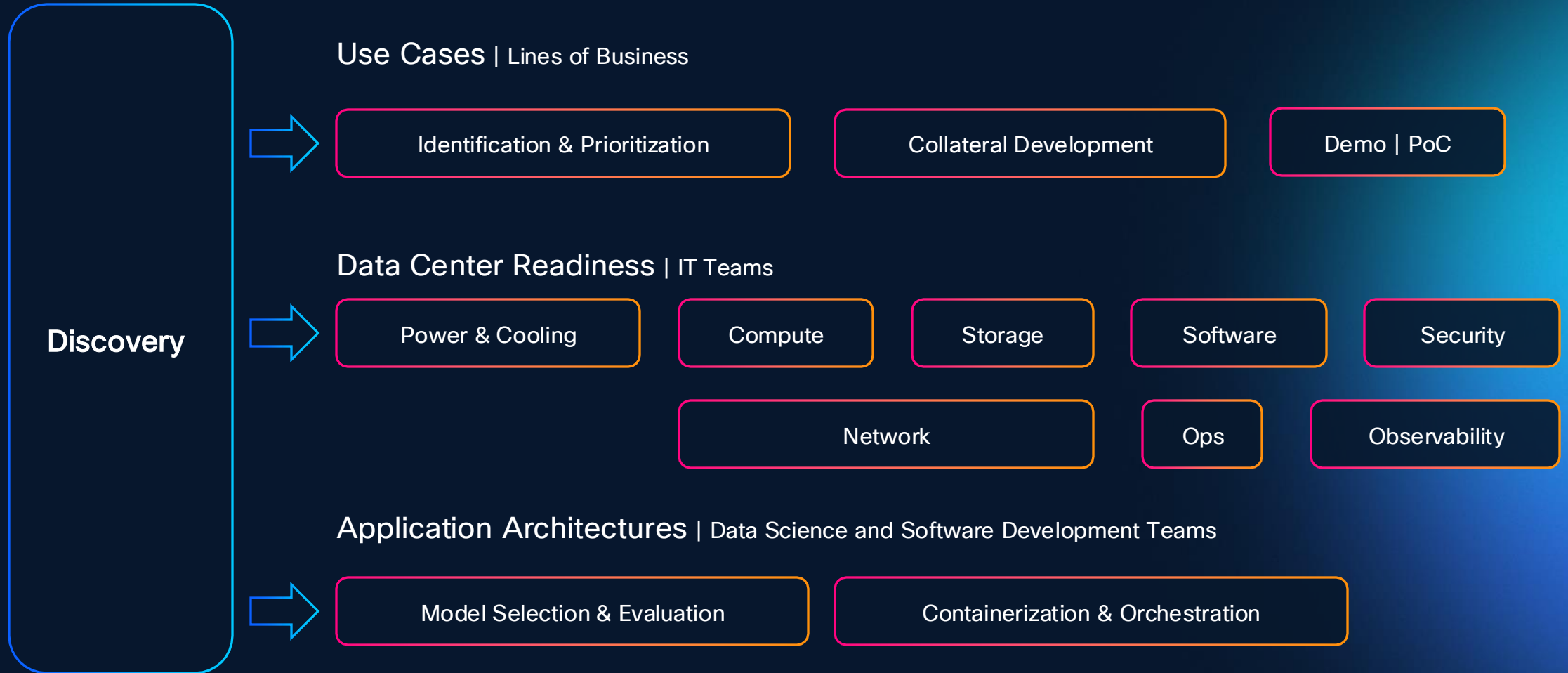
AI Runtime Protection

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 Supports Customers At Any Point of AI Journey



Thank you

