Cisco Secure AI Data Center Reference Architecture

Ylva Fonselius-Bonne
Account Executive – Cloud & AI

AI use cases across industries

Knowledgebase copilots

AI assistants

Content and code generation

Text | Images | Video | Code

Visual computing

Digital twins | Video analytics
Imaging & diagnostics

Language translation

Multilingual real-time communication

Detection & prediction

Forecasts
Anomalies | Insights

Virtual agent

Specialized domain specific chatbots

AI implementation at JP Morgan

By treating data as an asset, JP Morgan has programmed its AI innovations to do wonders. They can now analyze complex patterns, forecast trends, and optimize investment strategies faster and more efficiently. Some of the most popular AI applications/platforms are listed below.

IndexGPT, AI-driven investment advisory

JPMorgan Chase has developed IndexGPT, an AI-powered tool for thematic investing.

Amazon announces 3 AI-powered innovations to get packages to customers faster

Amazon's latest AI advances in delivery location accuracy, product demand prediction, and intelligent robotics will help benefit customers and employees alike.

Walmart is transforming retail operations with AI tools for store associates, including real-time translation, GenAI-powered assistants and AR inventory systems, to boost efficiency, improve customer service and support career growth.

Walmart has unveiled a new suite of AI tools for its store associates, part of a wider push to integrate AI in retail operations at scale. The initiative aims to improve workflows, enhance service and create more rewarding roles for its 1.5 million-strong workforce.

Available through the Walmart associate app, the tools are designed to remove friction, simplify tasks and increase efficiency on the shop floor.

"AI is a key enabler in improving how we work, and we believe its full potential is unlocked only when paired with the strengths of our people," said Greg Cathey, Senior Vice President, Transformation and Innovation. "When you put intuitive, accessible technology into the hands of millions of associates, the impact isn't incremental – it's transformational."

© 2025 Cisco and/or its affiliates. All rights reserved.

Challenges with AI projects delays time to value realization

Security vulnerabilities

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

Performance bottlenecks

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

Complex AI infrastructure deployment

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

© 2025 Cisco and/or its affiliates. All rights reserved.

Introducing: Cisco AI PODs

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

Training

Optimization

Inference

AI PODs

BUILD TUNE DEPLOY

DATA PIPELINE

AI FACTORY

AI TOOLING

One solution for your AI infrastructure journey

CPU

GPU

ADVANCED SERVICES

ACCELERATED COMPUTE

KUBERNETES

ADVANCED NETWORKING

DATA STORAGE

OPERATIONS

Faster Project Deployments

Lower Risk

Resources Optimization

© 2025 Cisco and/or its affiliates. All rights reserved.

cisco

Cisco AI PODs

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

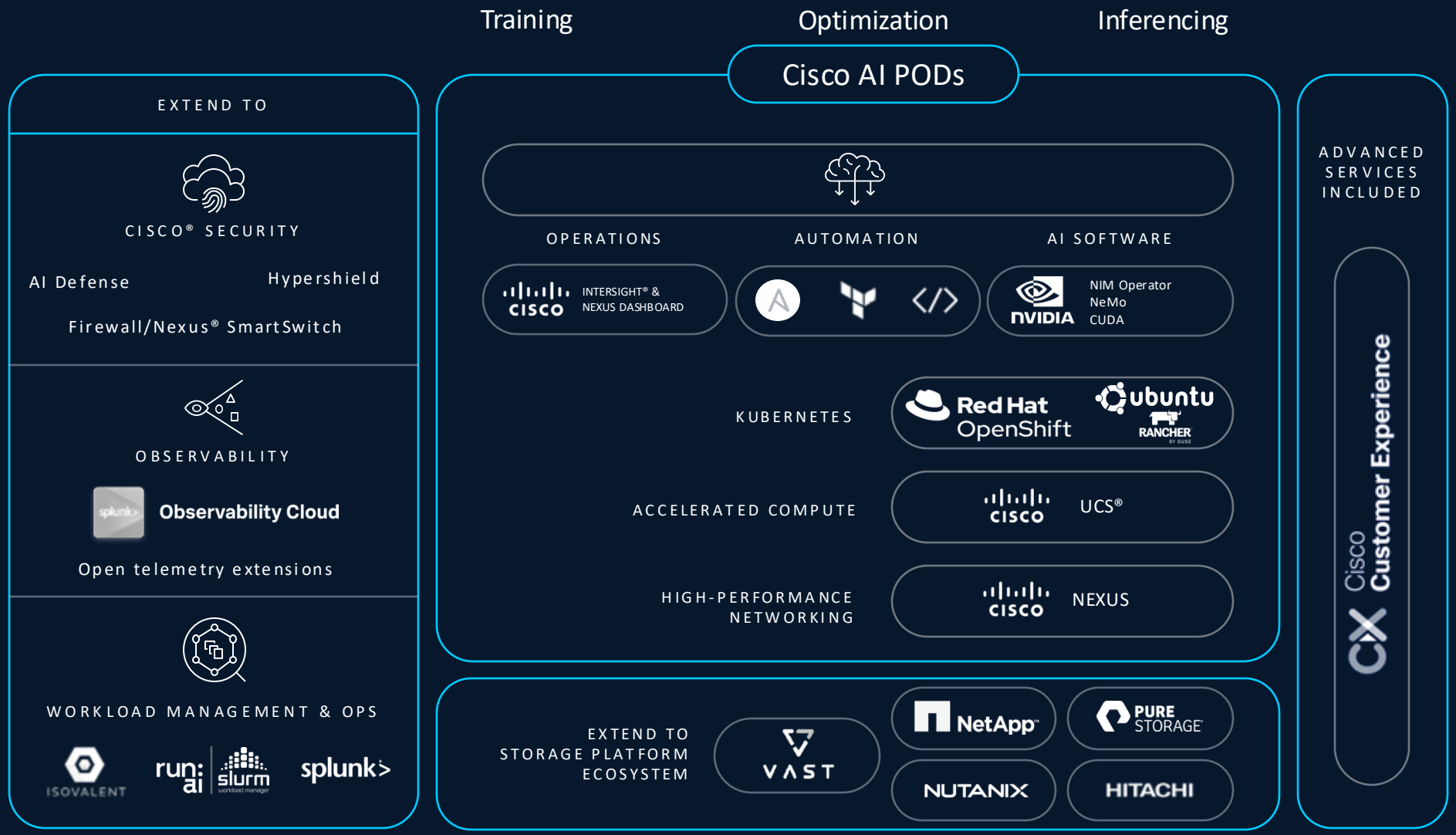
Deploy AI with confidence
Cisco CVD, NVIDIA ERA

Fully supported stack including Cisco and 3rd party components

Cisco CX Success Track

Orderable, use case driven AI-ready infrastructure stacks
Inferencing. Optimization. Training.

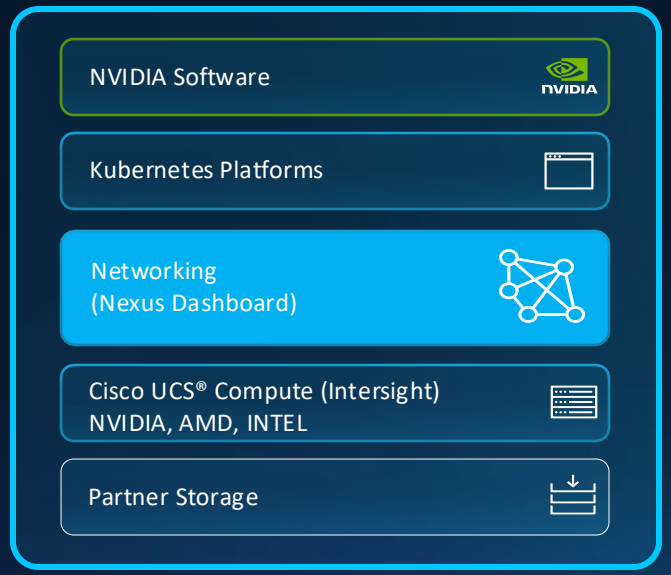
Incremental, atomic-level – or- fabric-based cluster scale



© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco AI PODs: Flexible Operating Models

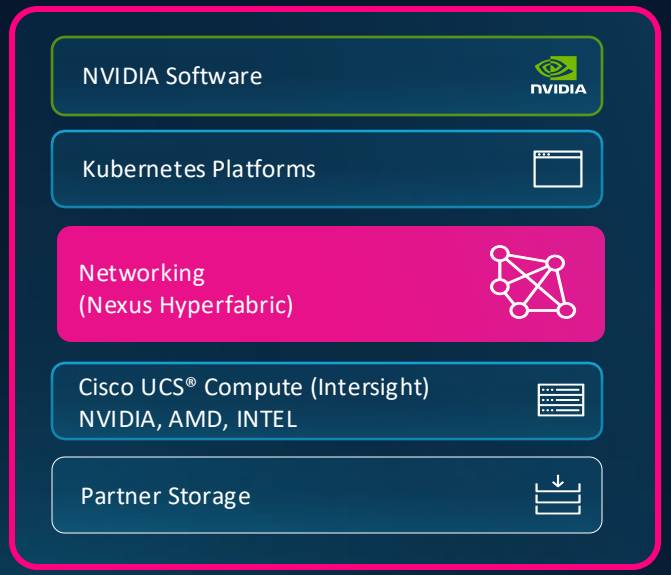
AI POD w/
On-prem management



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

© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco AI PODs

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

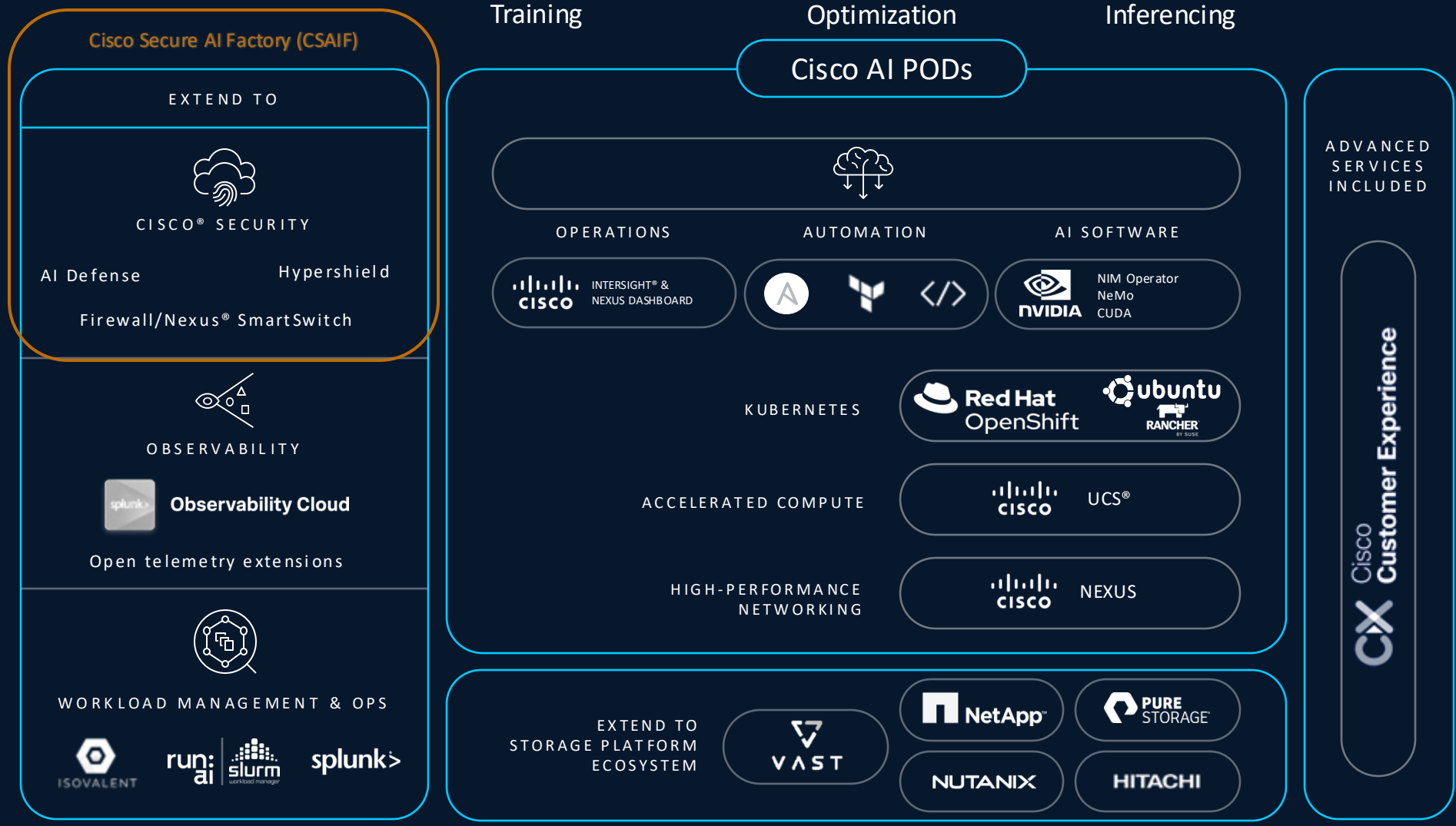
Deploy AI with confidence
Cisco CVD, NVIDIA ERA

Fully supported stack including Cisco and 3rd party components

Cisco CX Success Track

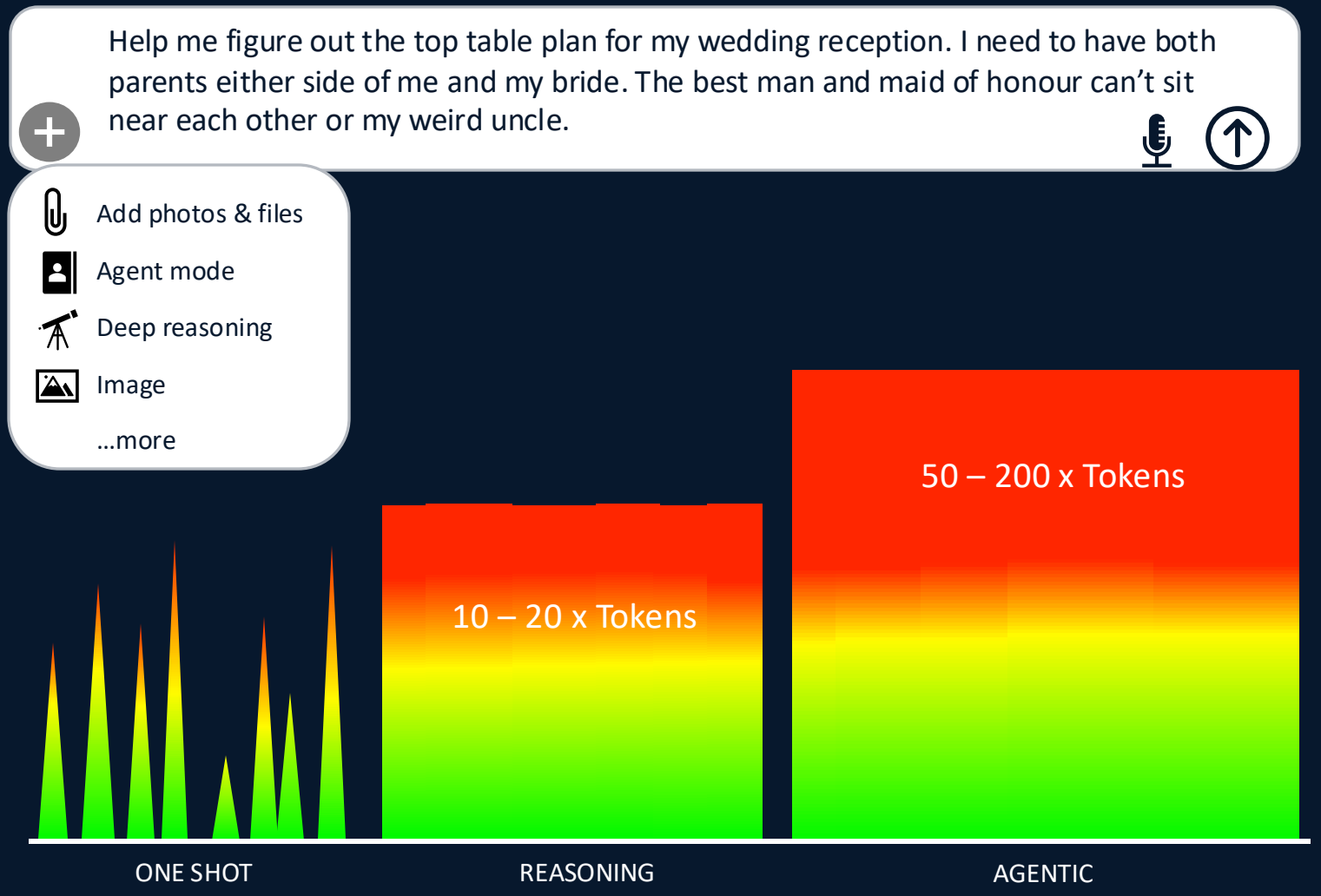
Orderable, use case driven AI-ready infrastructure stacks
Inferencing. Optimization. Training.

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



© 2025 Cisco and/or its affiliates. All rights reserved.

AI is Changing: Token Inflation

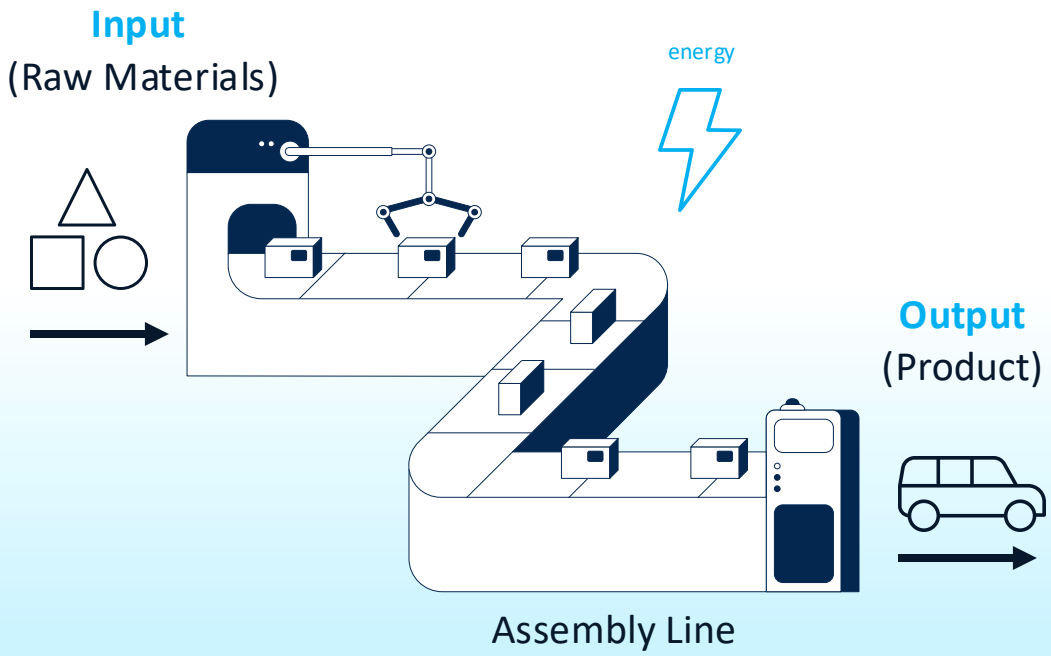


© 2025 Cisco and/or its affiliates. All rights reserved.

The Factory

Repeatable, scalable business capability

- Mass Production Efficiency
- Quality Control
- Supply Chain Integration

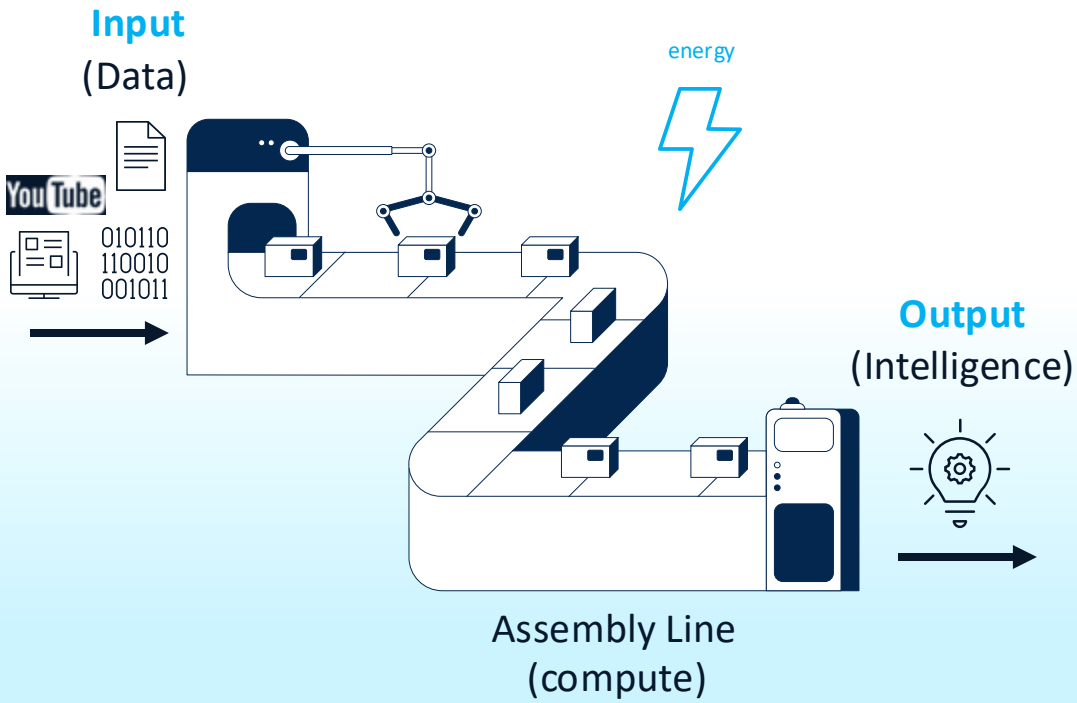


© 2025 Cisco and/or its affiliates. All rights reserved.

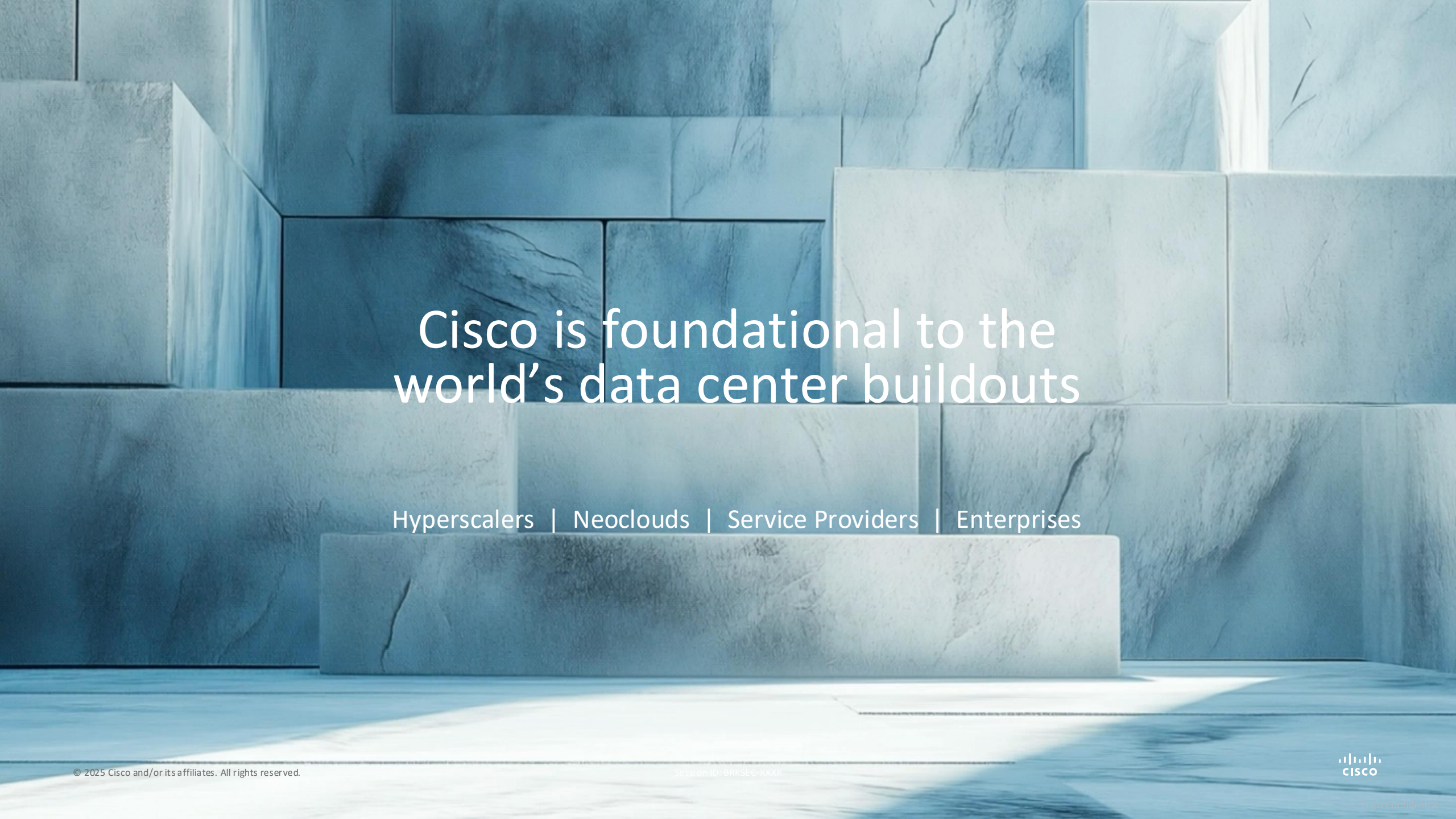
The AI Factory

Repeatable, scalable business capability

- Mass model/token production
- Model performance control
- Data and workflow integration



© 2025 Cisco and/or its affiliates. All rights reserved.



© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

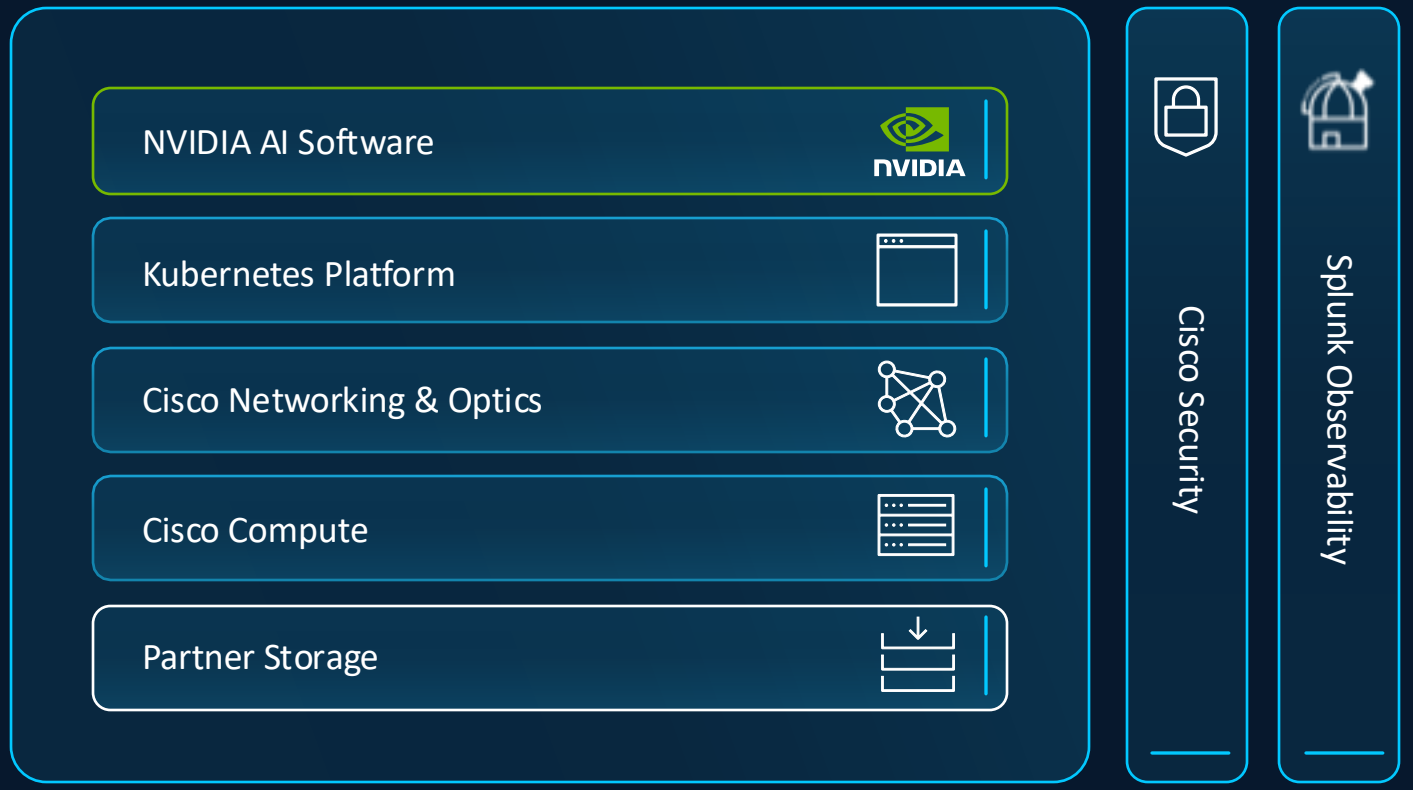
Cisco Secure AI Factory with NVIDIA

What is it?

Reference architecture

Validated solutions and turnkey offerings

Differentiated with Security and Observability



© 2025 Cisco and/or its affiliates. All rights reserved.

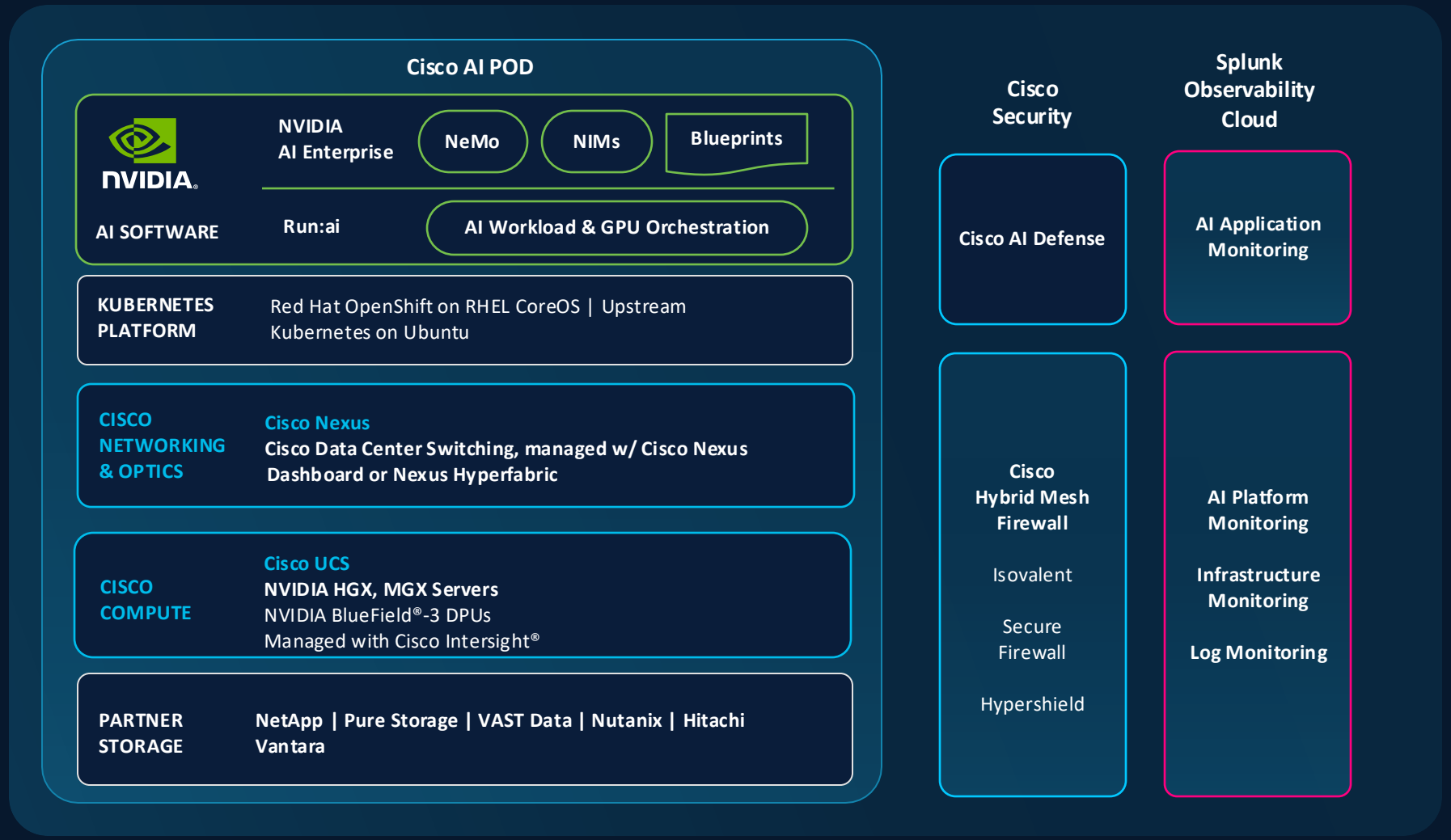
What’s the risk?

AI applications are complex and non-deterministic

© 2025 Cisco and/or its affiliates. All rights reserved.

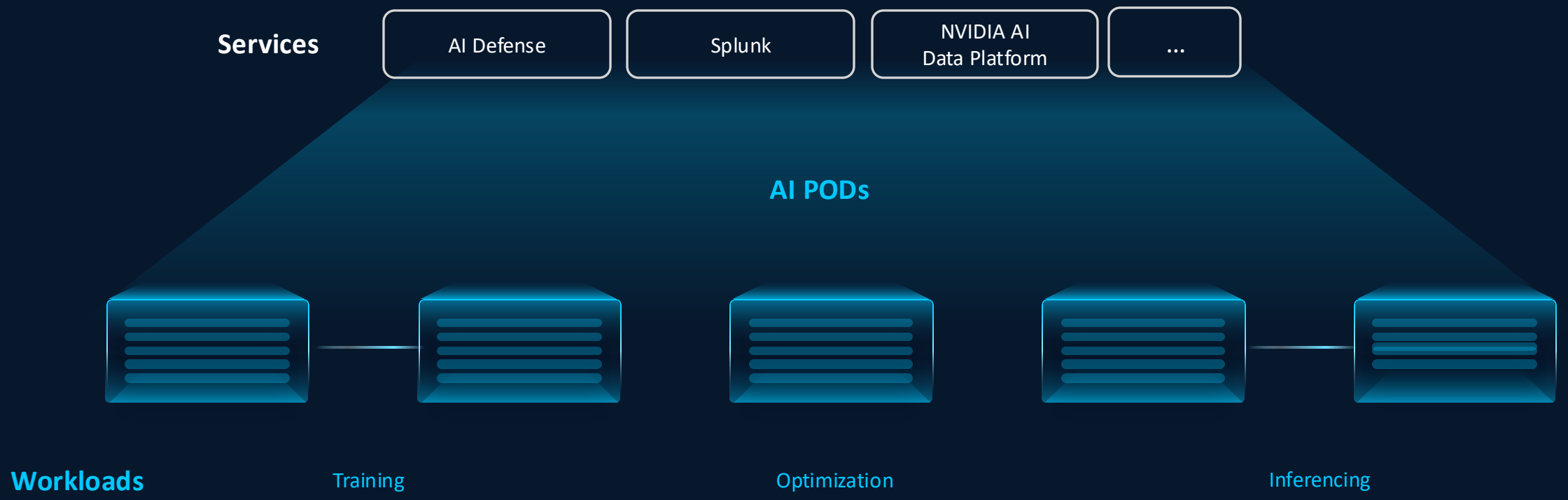
Cisco Secure AI Factory with NVDIA

Under the covers



© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Secure AI Factory with NVIDIA AI PODs



© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Differentiation

The Security

Security-first architecture enables safe enterprise AI

The Network

High-performance integrated AI networking enables efficient model training and inferencing

The Assurance

Pre-validated AI infrastructure stack with flexible deployment options improves data scientists and developer productivity

© 2025 Cisco and/or its affiliates. All rights reserved.

Thank you