Modernizing Cisco's Data Center for AI

CENCOM-1025

Modernizing Cisco's Data Center for AI

Danny McGinniss
VP, Product Management, Compute BU
CENCOM-1025

We all know that AI is
disrupting the market

The pace of AI innovation is staggering

AI challenges you to deploy new technology **faster than ever before.**

Why is AI so complex in business?

You are redesigning the car and the road at the same time

So, we have to rethink everything

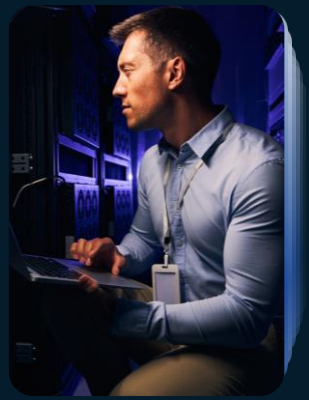
Security | Sustainability | Performance Demands | Power Constraints | Investment | Cooling Dilemma | Space Efficiency | Network Implications | Ecosystem

Every organization’s AI approach and needs are different

BUILD THE MODEL
Training

OPTIMIZE THE MODEL
Fine-tuning and RAG

USE THE MODEL
Inferencing



Enterprise IT

Networking team

Line of business

AI practitioner

cisco Live!

Security

Customer Experience

THE CISCO AI STORY

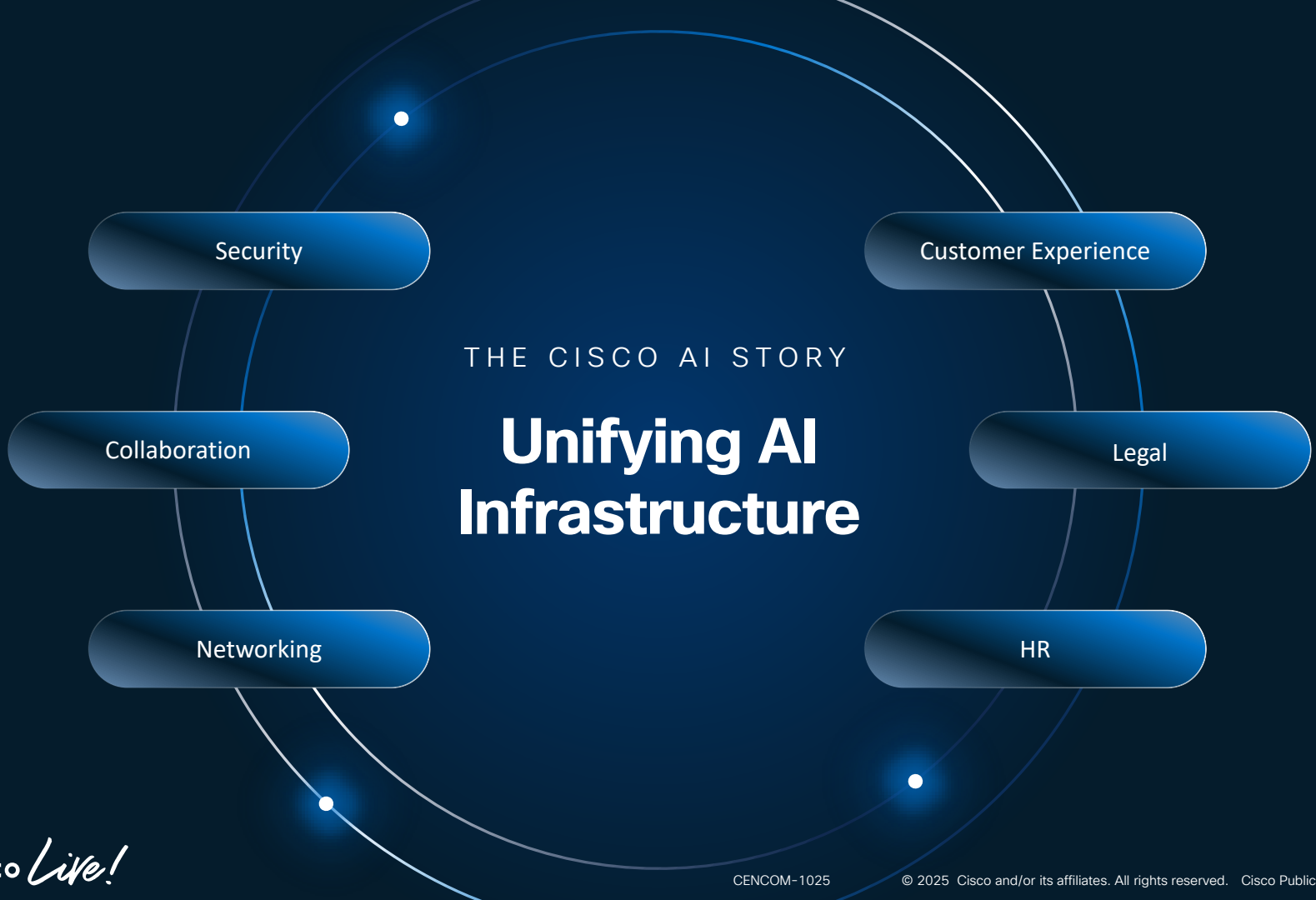
Organic AI Projects Across the Business

Collaboration

Legal

Networking

HR



cisco Live!

Benefits of Unified AI Infrastructure

Lower costs

Capex savings

Economies of scale

Savings on staffing

Decreased learning curve

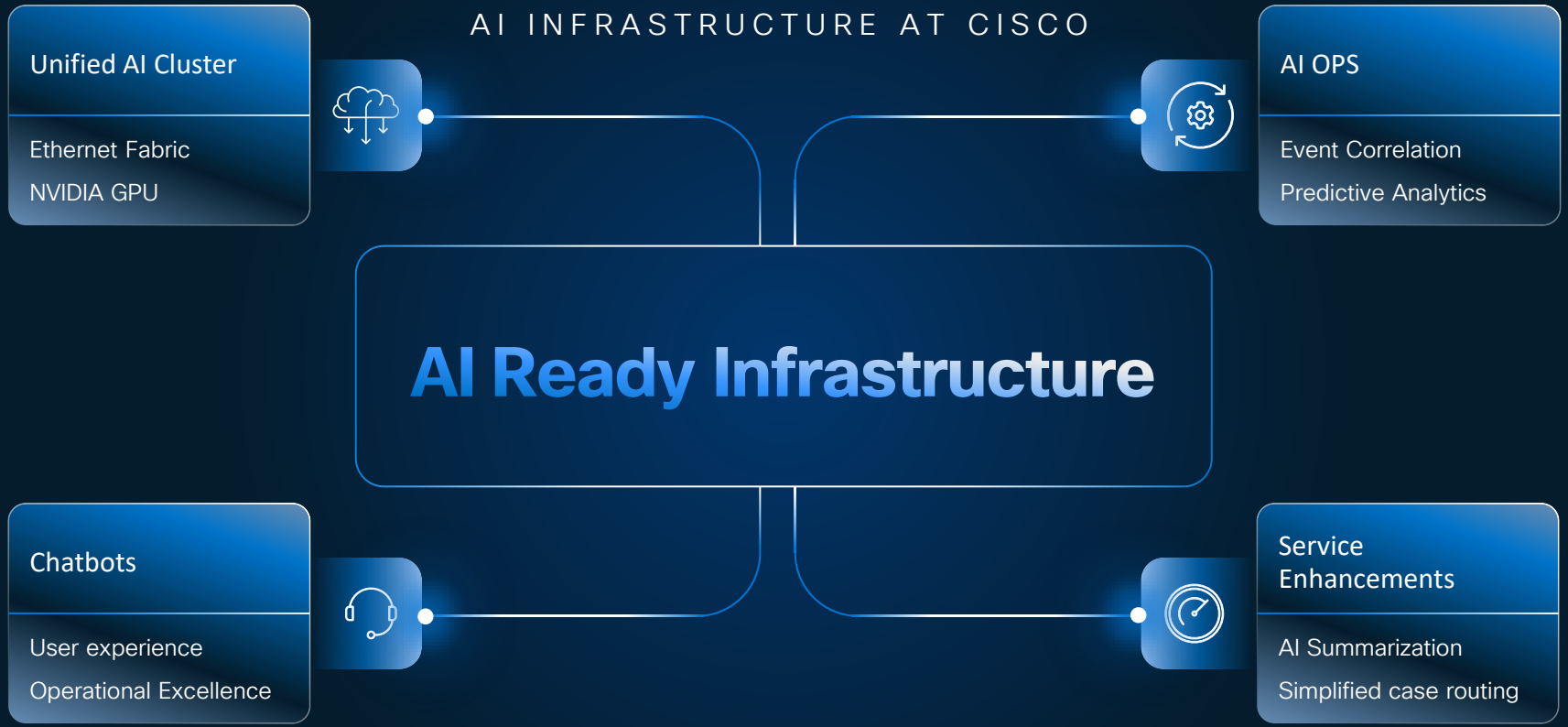
De-duplication of effort

Optimized use of capacity

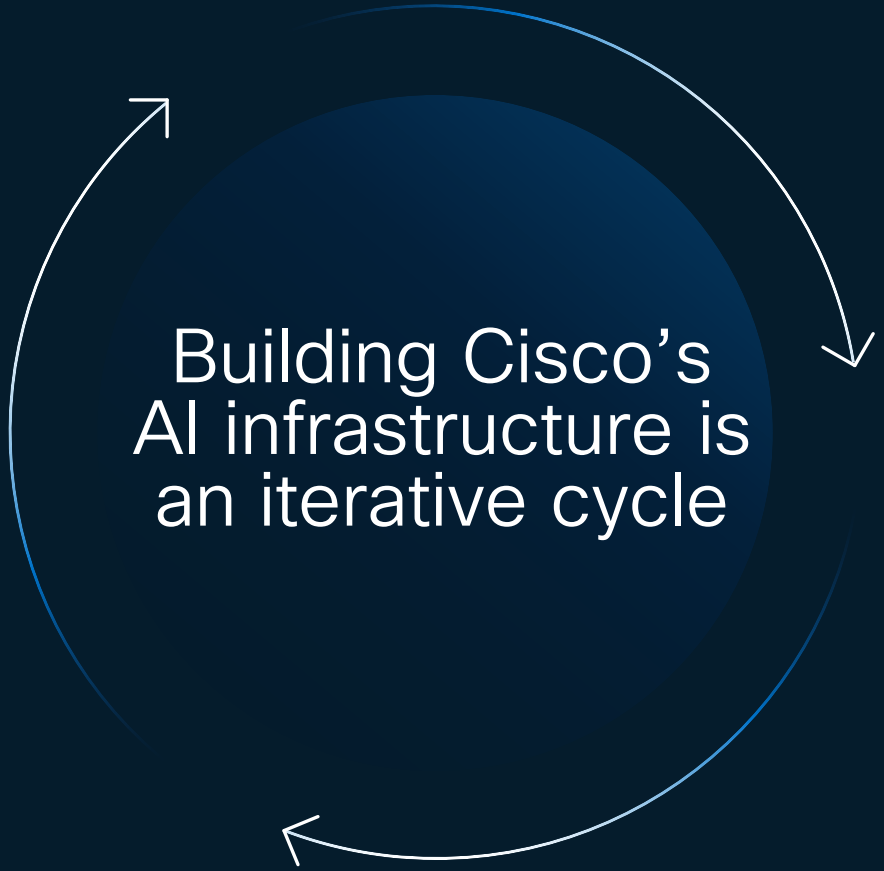
Enhanced security

Reduces the liability of shadow IT

AI INFRASTRUCTURE AT CISCO



Building Cisco's AI infrastructure is an iterative cycle



01

Learning

Internal/external research

DevNet

Training VODs and classes

Procured Lab Test Devices

Partnerships

02

Planning

Cost justification

Engagement with leadership

Architecture design

Ordering

03

Implementing & Using

Physical build

Testing & verification

Allocation of GPUs

Use case development

AI Cluster

First fabric built on Silicon One ASICs

32 PetaFLOPS AI performance per node

SONiC Operating System with Cisco 8000 Switches

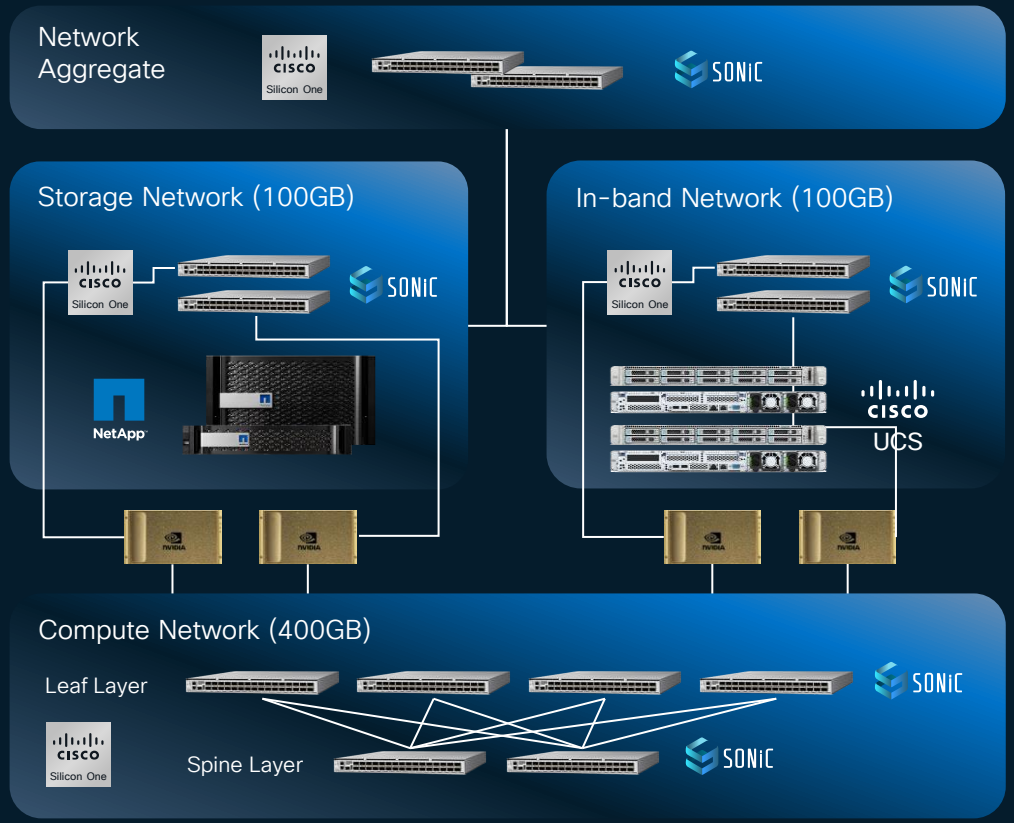
12.8T throughput per compute leaf

Ethernet-based architecture

400G non-blocking compute fabric

32 Nvidia compute nodes (8 GPUs each)

4 NetApp AFF A900 nodes



Our journey with AI for Human Resources

MARCH 2024

Pondering the question “How can we bring AI into People & Communities (P&C)?”

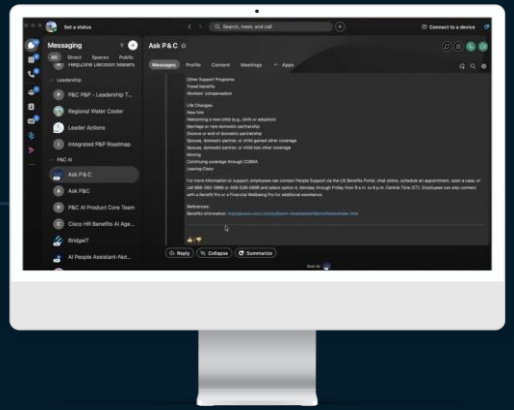
Partnership with Cisco IT for use of AI clusters to train the model

Team planning and development, with careful attention to restricted data

Testing and validation of Ask P&C

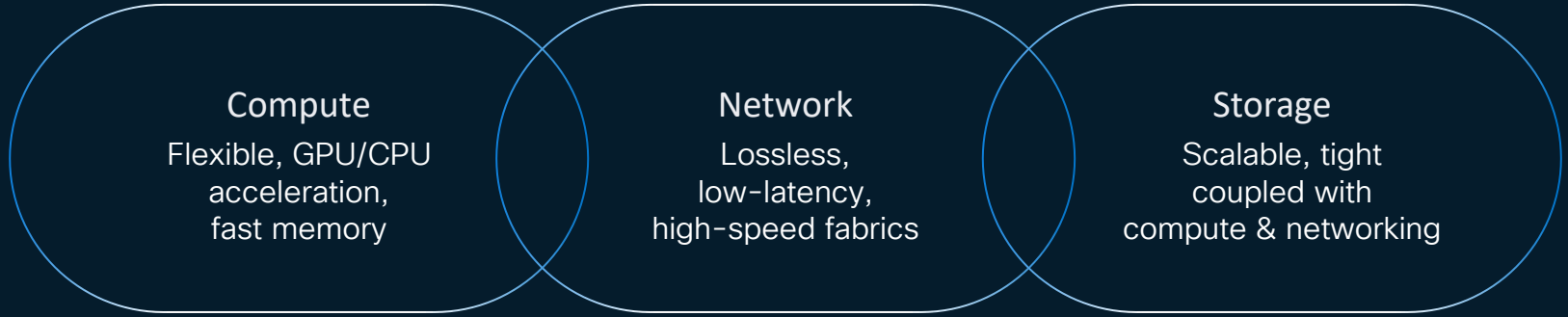
APRIL 2025

General availability of Ask P&C



The challenges we faced

INFRASTRUCTURE DEMANDS



AND BEYOND

Rapid growth in data volume and variety	Unfamiliar application stacks and new, complex infrastructure patterns	Insufficient IT automation and observability	Greater cybersecurity threats
New operational silos	Shortage of technical expertise	Disorienting AI hype	High entry cost and lock-in issues

What we learned along the way

Technology

- Use case definition and refinement
- Common operating model with automation
- Right-sized AI infrastructure
- Secure infrastructure and systems

People

- Learning and skills
- Curiosity
- Relationship building - get to know your lines of business
- Intrinsic motivation to be proactive and get started

Best-in-class

AI-ready infrastructure

Data Centre Networking

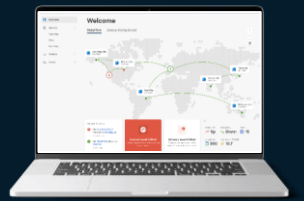
Silicon One

Optics

Unified Compute

Unified Compute

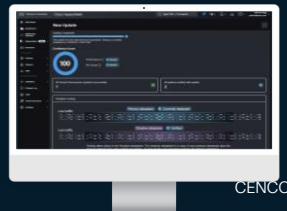
Unified Compute



Data Centre Networking

Data Centre Networking

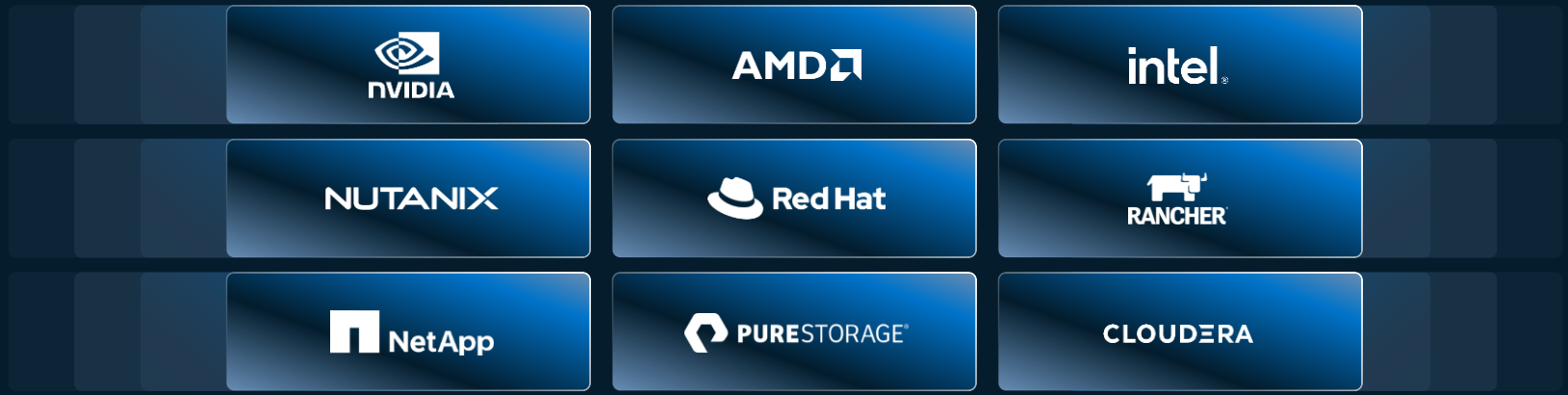
Security



Unified Compute

Fully Validated and Tested Designs

There's a simpler path to deploying and automating your AI infrastructure



Build the model
Training

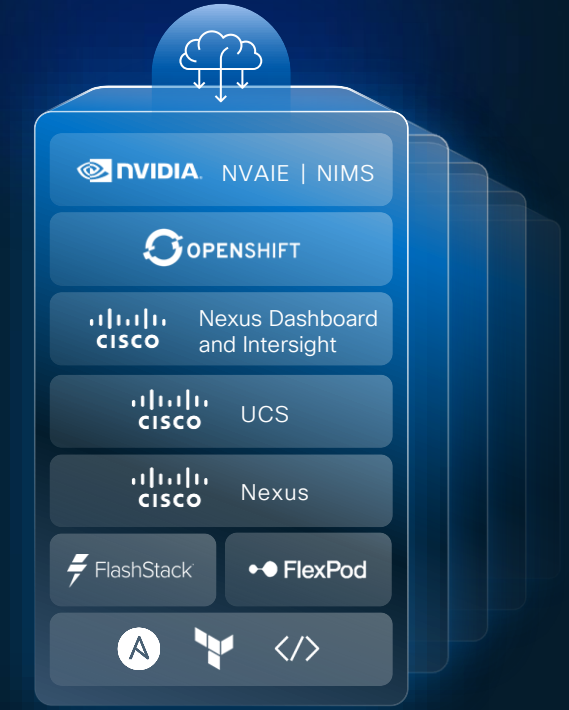
Optimize the model
Fine-tuning and RAG

Use the model
Inferencing

Cisco AI PODs

There's a better way to purchase infrastructure for your AI use cases

- Large language models
- AI tooling
- Kubernetes
- Operations
- Accelerated compute
- LAN and SAN networking
- Converged infrastructure
- Automation



As your needs become more specific so do our recommendations

AI Advisor

Helping you optimize your AI infrastructure

- Workload sizing
- GPUs recommendations
- Multiple LLM model support
- End-to-end guidance
- Bot assistance and BOM generation
- Expanded model selection

Data

Security for AI

AI-native Products

AI Infrastructure

Services

Only Cisco unifies networking, compute, security, and observability to deliver AI-ready data centers.

How to take action

1
Define your desired outcomes and use cases

Prioritize based on impact

Align your workforce to your AI priorities

2
Identify your infrastructure needs

Scan to use the AI Readiness Assessment

3
Make informed decisions about strategic investments

Leverage Cisco certifications and validated designs


cisco
Thank you
cisco Live!

Fill Out Your Session Surveys

Participants who fill out a minimum of 4 session surveys and the overall event survey will get a unique Cisco Live t-shirt.

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

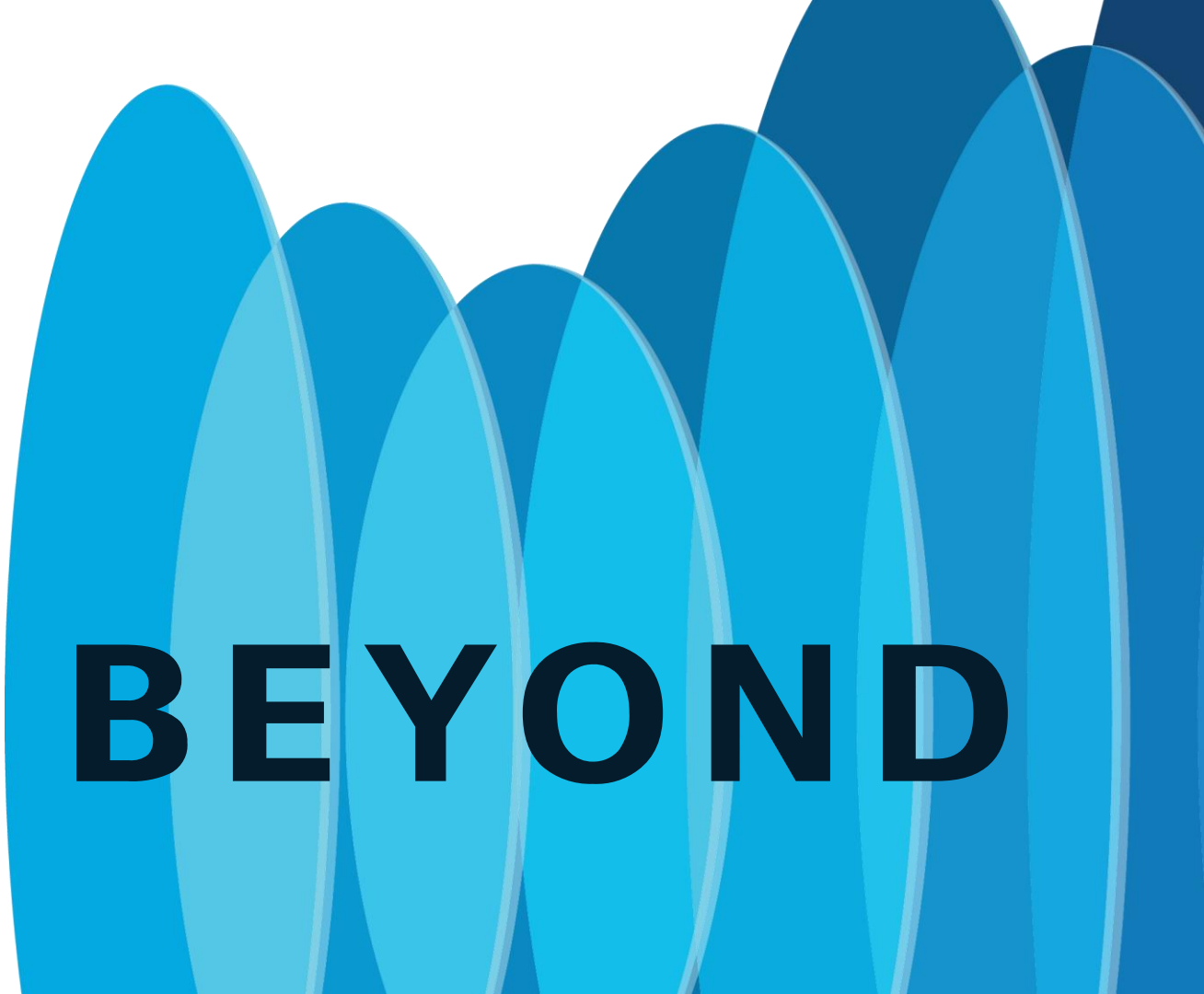
All surveys can be taken in the Cisco Events mobile app or by logging in to the Session Catalog and clicking the ‘Participant Dashboard’

Content Catalog

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 ciscolive.com/on-demand. Sessions from this event will be available from March 3.

cisco Live!



GO BEYOND