Unlocking the Power of Edge Computing and AI: Challenges, Use Cases, and Cisco's Edge Compute Solutions

Ronnie Chan
Leader, Product Management
Cisco Compute

BRKCOM-1009

Cisco Webex App

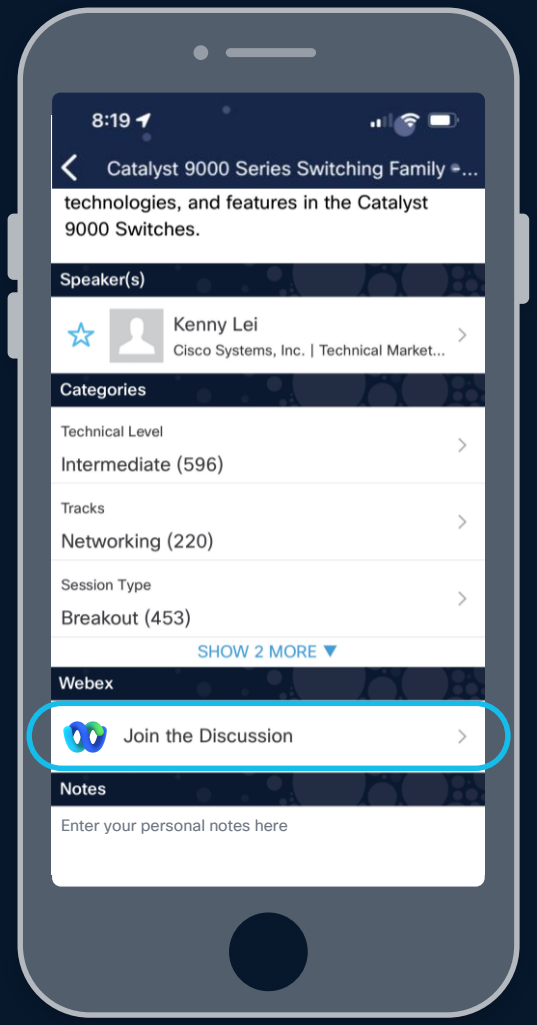
Questions?

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/#BRKCOM-1009

Once a upon a time, a grad student was thirsty...

David Nichols
Carnagie Mellon University

Source: https://www.ibm.com/think/topics/iot-first-device

... And the world's first IoT device was born!

About Me

- Product Leader for Edge AI, Cisco Compute
- Working with computing and edge for past 7 years
- Background in data center and cloud infrastructure, spanning storage, compute, HCI, networking and security

Agenda

Part I – What is and why Edge AI

Part II – Use Cases and Challenges of Edge AI

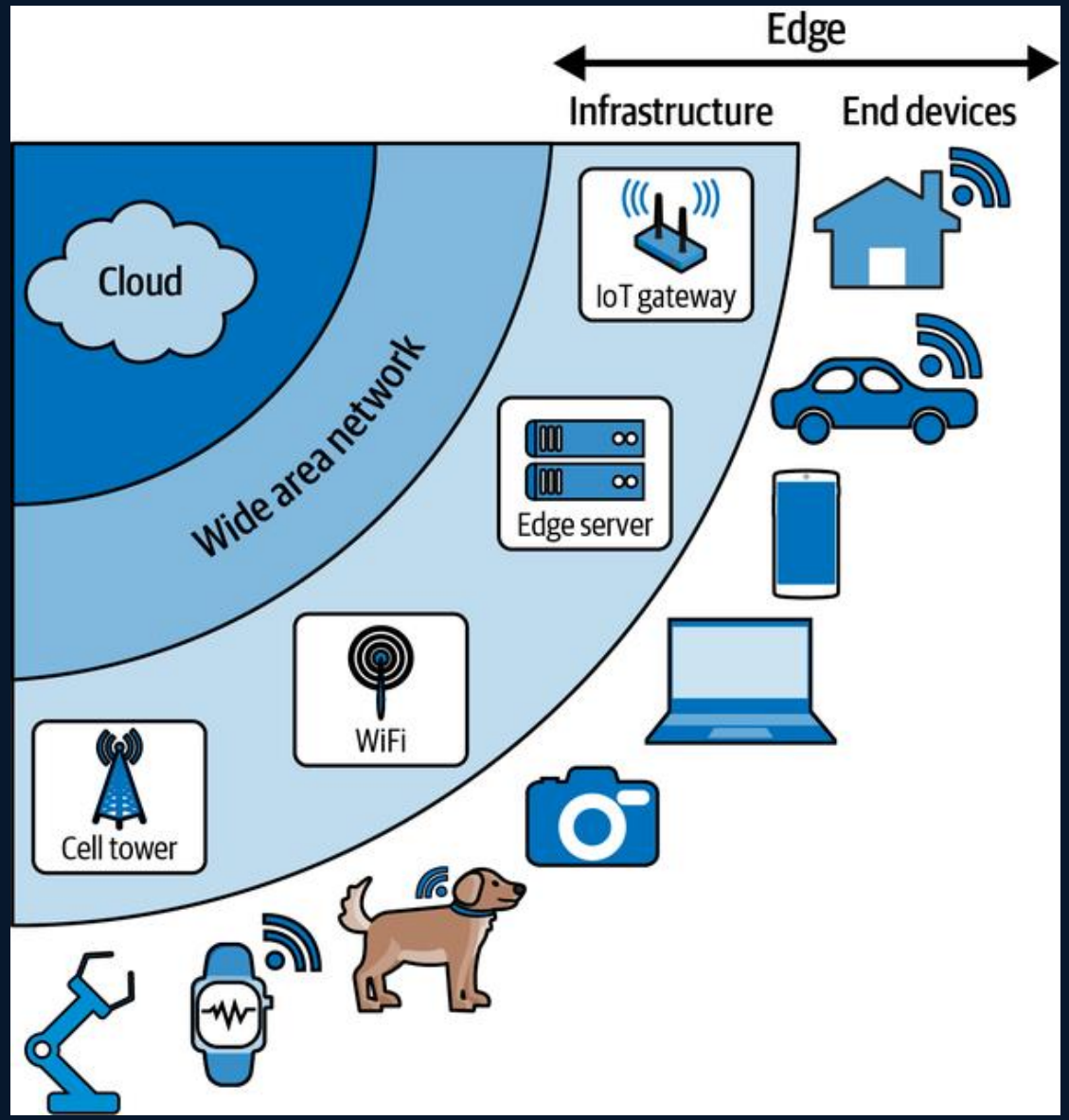
Part III – Cisco’s compute solutions for Edge & Edge AI

Part I

What is and why Edge AI

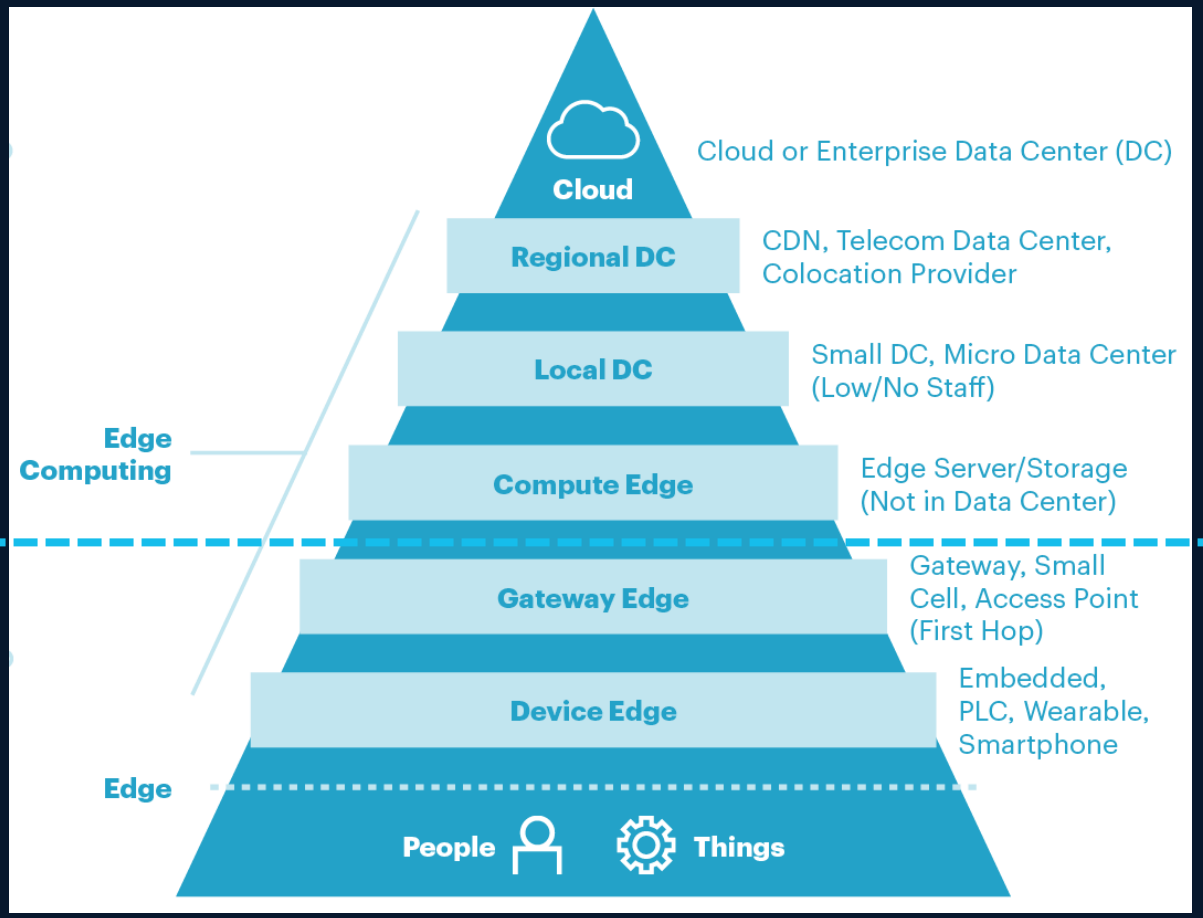
Today's network edge has grown in distance, proliferated in reach, and diversified in applications

While Cisco plays in multiple aspects of edge infrastructure, remainder of this session focuses on edge compute.



Source: Situnayake, D., & Plunkett, J. (2023). *AI at the Edge: Solving Real-World Problems with Embedded Machine Learning*. O'Reilly.

Segmenting Edge Computing by Deployment Spaces



Target edge computing segments for Cisco Compute solutions

Source: Gartner

Edge AI vs. Traditional AI

Inferencing is the norm while training on the edge is rare

Edge inferencing focuses on sensor data which can be noisy, voluminous, and hard to manage

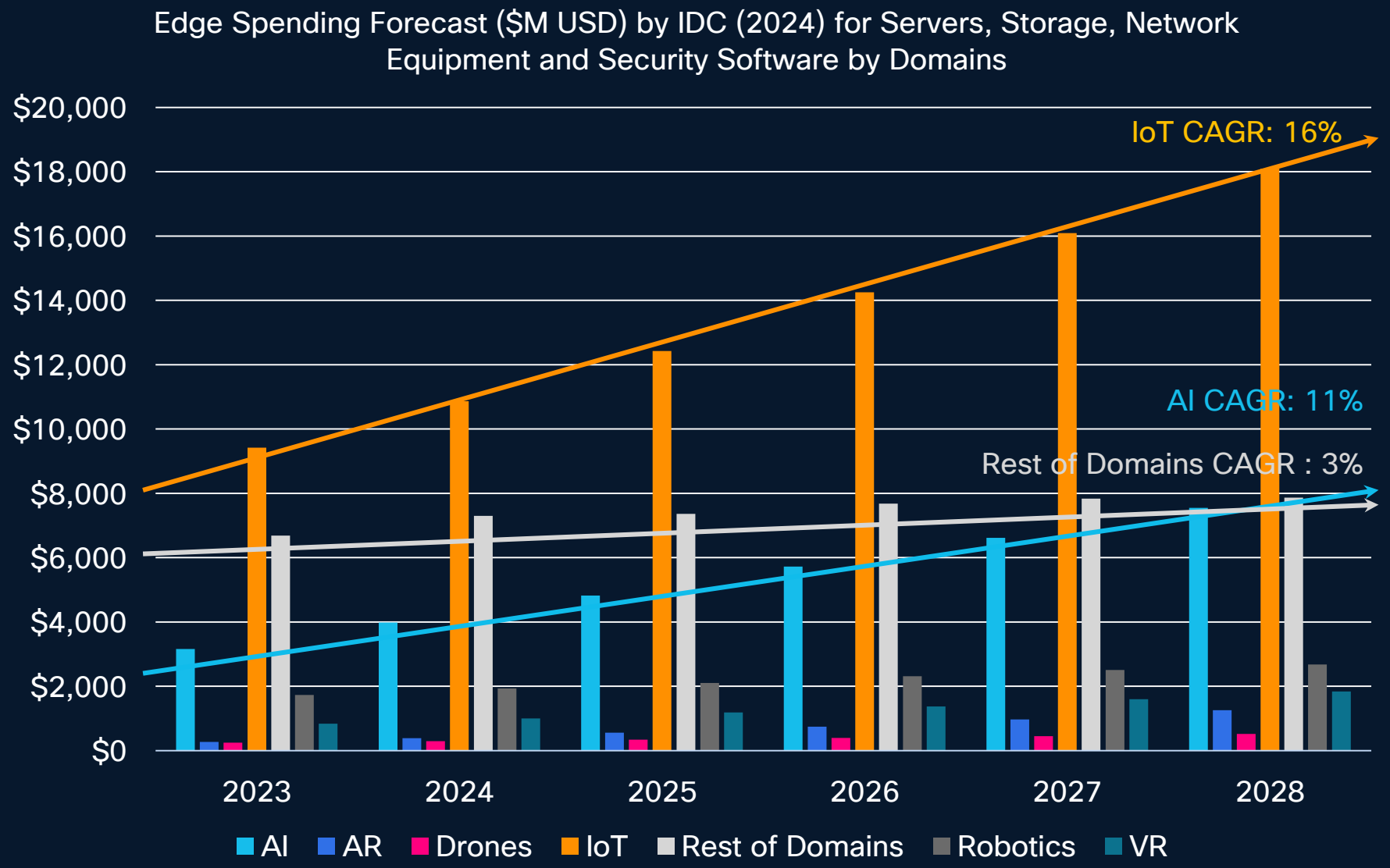
ML models used at the edge can be small and often focus on specific tasks

Compute devices used for edge inferencing is heterogeneous, including CPUs, SoC, GPUs & FPGAs

Good enough accuracy is often the goal

Source: Situnayake, D., & Plunkett, J. (2023). *AI at the Edge: Solving Real-World Problems with Embedded Machine Learning*. O'Reilly.

IoT & Edge Digitization Drive Spend, Edge AI Catching Up



Flexible Infrastructure to run non-AI and AI edge workloads

Industry use cases are accelerating the need for AI inferencing at the edge

Industry-specific use cases and requirements are being evaluated

Retail
Drive thru optimization

Manufacturing
Asset visibility and control

Financial
Financial crime/ fraud detection

Healthcare
Augmented diagnosis system

Accelerating the need for AI Inference and applications at the edge

Data sovereignty

Regulatory requirements and data sensitivity make cloud-centric architecture challenging, pushing AI workloads to the edge, enhancing data security by minimizing transport risks

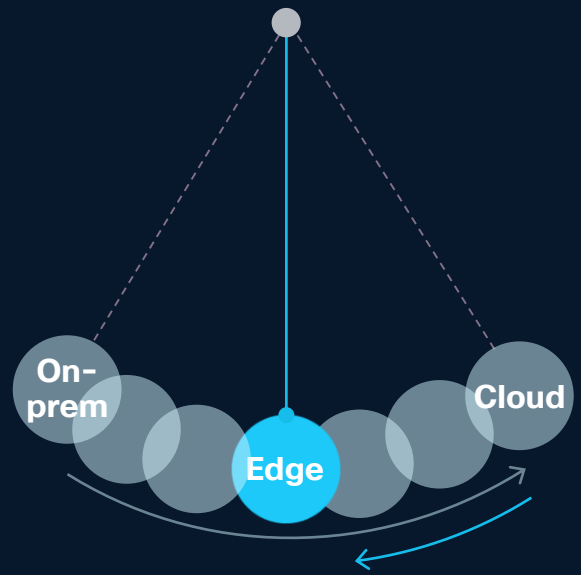
Latency considerations

On-device processing allows real-time analytics and decision-making, enhancing customer experience and supporting crucial AI applications like security and autonomous systems

Bandwidth needs

Local data processing enhances bandwidth efficiency while enabling offline functionality for devices

Creating a paradigm shift from the centralized cloud model to Edge AI



Optimized for being closer to use case

Part II

Use Cases and Challenges of Edge AI

Personalized Shoppers' Experience

- Merchandise Pickup
- Customized Offer
- More In-Store Purchases
- More Time in Store

- App, Loyalty Prgm.
- Inventory Mgmt.
- Digital Signage
- POS Systems
- Cameras, Sensors
- Reliable WAN
- Reliable LAN, WiFi
- Security
- SFF Compute
- Data Storage

Optimized Associates' Experience

- Personalized Customer Info
- Automated Inventory Mgmt.
- Traffic Flow Optimization
- Theft Detection and Prevention

- App, Loyalty Prgm.
- Inventory Mgmt.
- Digital Signage
- POS Systems
- Cameras, Sensors
- Reliable WAN
- Reliable LAN, WiFi
- Security
- SFF Compute
- Data Storage

Categories of Edge AI Applications

Computer Vision

- Count inventory in a shelf and places new orders automatically
- Monitor shipment for damage w/ smart packaging
- Identifying and tracking POIs w/ CCTV feeds

Operational Technology

- Monitor pipelines for signs of maintenance
- Quality control on a production line
- Picking items in warehouse w/ robots

Understanding People & Living Things

- Alerting workers presence of a hazard
- Identifying when an ICU patient deteriorates and alerts a health worker
- Counting number of people waiting for concessions and alerts more stalls to be opened

Generating or Transforming Signals

- Understanding a scene in a shot and add metadata
- Understanding handwritten medical charts & generating summaries
- Transcribing a spoken conversation for accurate notetaking

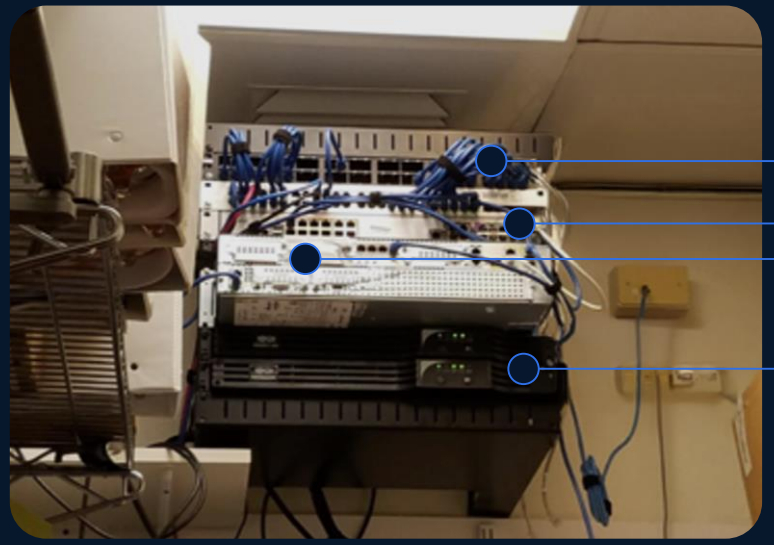
Is This a Good Fit for Edge AI? BLERP!

- **B**andwidth: not just bandwidth constraint, but energy cost of transmitting data
- **L**atency: can required response time tolerate the data RTT + processing time
- **E**conomics: connectivity, API calls, tokenization, subscriptions all cost $
- **R**eliability: what is the cost of an outage
- **P**rivacy: does data leaving the premises invoke potential privacy concerns

Source: Situnayake, D., & Plunkett, J. (2023). *AI at the Edge: Solving Real-World Problems with Embedded Machine Learning*. O'Reilly.

But Operationalizing Edge AI is Complex

Legacy edge infrastructure



Deployment inconsistency	Hard to deploy consistent, repeatable infrastructure and workloads across multiple sites
Product incompatibility	Manually planning upgrades across multiple sites can lead to interoperability issues and downtime
Environmental constraints	Power, cooling, space, and acoustic limitations may limit solution choices
Limited technical expertise	non-technical staff on site that impede the ability to onboard or troubleshoot systems seamlessly
Operational complexity	Operating infrastructure across numerous locations requires coordination across domain teams and can often leads to inefficiencies
Digitization of everything	Business impact of an outage is now more costly as critical operations are fully digitized

Part III

Cisco’s compute solutions for Edge & Edge AI

Cisco Compute: Edge Deployment Customer Examples

US national retailer, $25B sales, 700+ stores

- HCI on UCS C-Series, cloud-based management
- 400 sites, 800 nodes, running POS, File & Print for each store
- Enables repeatable consistent deployments, cost savings w/ remote operations

Multinational financial services company, $1.5T assets under mgmt.

- UCS X-Series at data center, X-Direct at remote site
- Unified architecture & unified management across DC and edge
- Enables small team to scale managing 1000+ servers centrally

Global leader in remote patient monitoring, $3B revenue

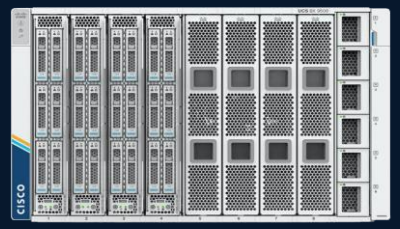
- Cisco-Nutanix edge compute solution on UCS C-Series
- Enables remote deployments of patient monitoring systems in hospitals
- Global deployment with managed services

Cisco UCS Compute Portfolio

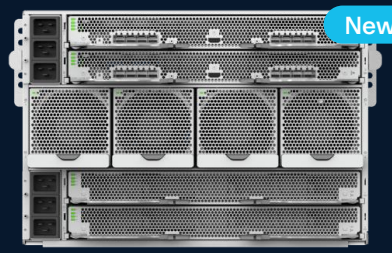
MAINSTREAM ENTERPRISE SERVERS

UCS X-Series X9508 Chassis

IFM Module



UCS X-Series Direct



UCS X210c M7

UCS X210c M8

UCS X410c M7

UCS B200 M6

UCS X215c M8

UCS C240 M8E3S
36 EDSFF E3.S1T

UCS C240 M8SX
28 HDD/SDD/NVMe

UCS C240 M8L
16 LFF + 4 SFF

UCS C240 M7SN
28 NVMe

UCS C240 M6S
14 SSD/HDD Media drive

UCS C240 M6N
14 NVMe Media Drive

UCS C220 M8E3S
16 EDSFF E3.S1T

UCS C220 M8S
10 HDD/SSD/NVMe

UCS C220 M7N
10 NVMe

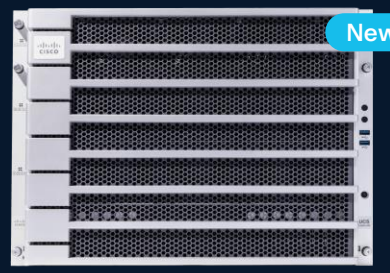
UCS C245 M8SX
28 HDD/SDD

UCS C225 M8S
10 HDD/SSD

UCS C225 M8N
10 NVMe

AI SERVERS

UCS C885A M8
8RU Dense GPU Server



UCS C845A M8
4RU MGX Server

Cisco Compute Solutions for Heavy Edge

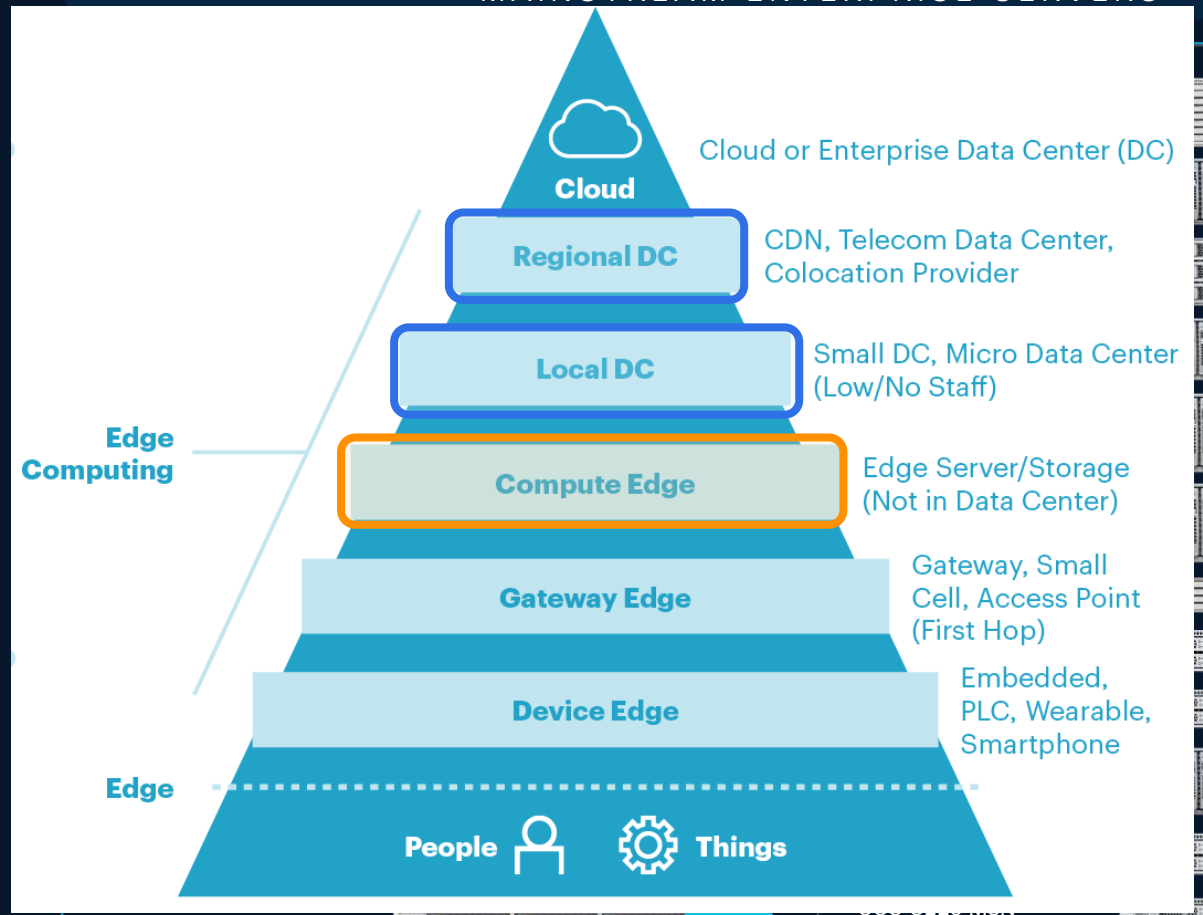
Cisco UCS Compute Portfolio

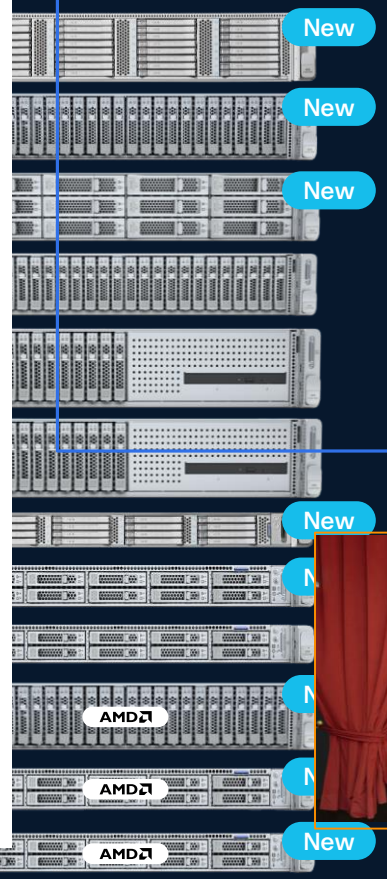
MAINSTREAM ENTERPRISE SERVERS

UCS X-Direct Modular Compute Syste

UCS C-Series 1U & 2U

AI SERVERS





UCS C885A M8
8RU Dense GPU Server

UCS X215c M8

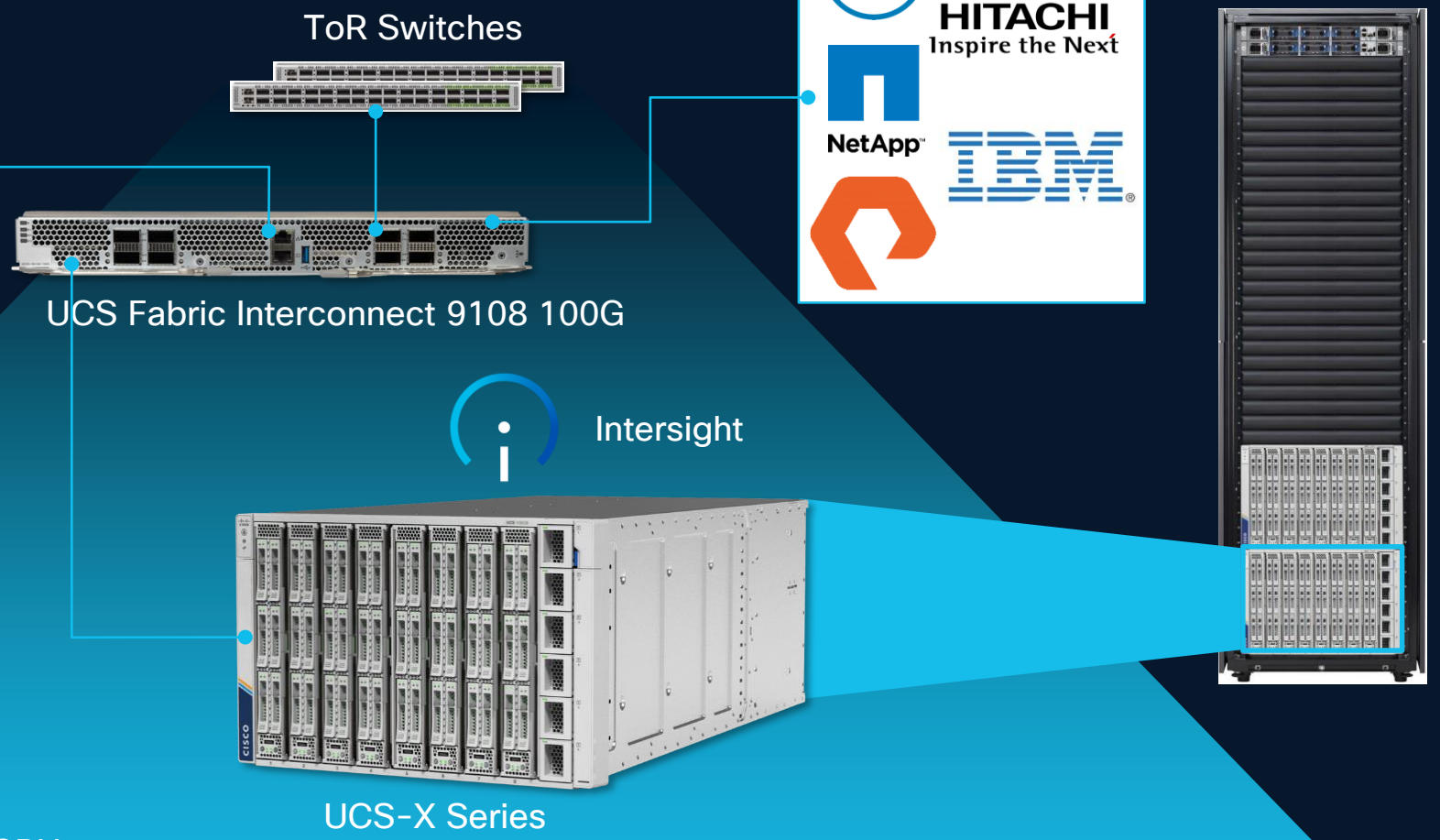
10 NVMe

Modernize at the edge

UCS X - DIRECT

- Unified architecture
- Unified fabric
- Flexible modular infrastructure
- Optimized for traditional and AI workloads*

*Support single and double-wide GPUs



*Post-FCS

UP TO 68% CapEx Savings

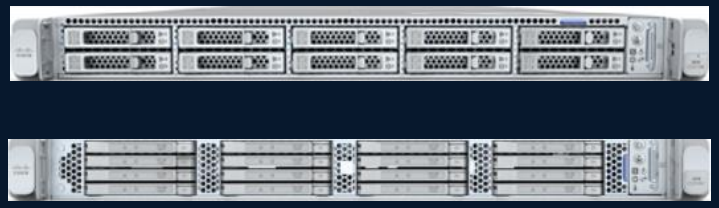
UP TO 64% Better Performance

UP TO 52% Lower Power

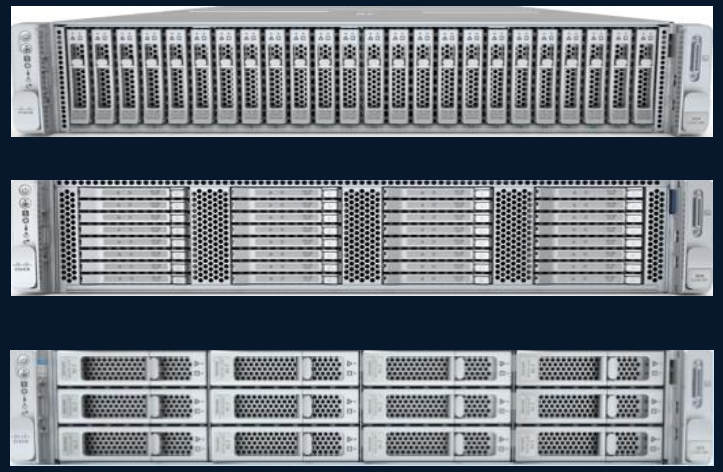
UP TO 50% More sustainable

Cisco UCS C-Series M8 Servers

Dense form factors for a wide range of workloads, including virtualization, web, databases, big data analytics, cloud, and bare-metal applications



Cisco UCS C220 M8



Cisco UCS C240 M8

- Up to 2x 6th Gen Intel® Xeon® processors
- Up to 8 TB DDR5 memory
- PCIe 5.0 options
- 10/25/40/50/100/200 mLOMs and VICs
- Up to 16 (C220) or 36 (C240) E3.S drives
- Up to 10 (C220) or 28 (C240) SFF SAS/SATA/NVMe drives
- Up to 16 LFF SAS/SATA/NVMe drives (C240)
- Single and double-wide GPUs

Simplified Orderability

AI PODs

Faster time to value with pre-configured bundles

Deploy AI with confidence

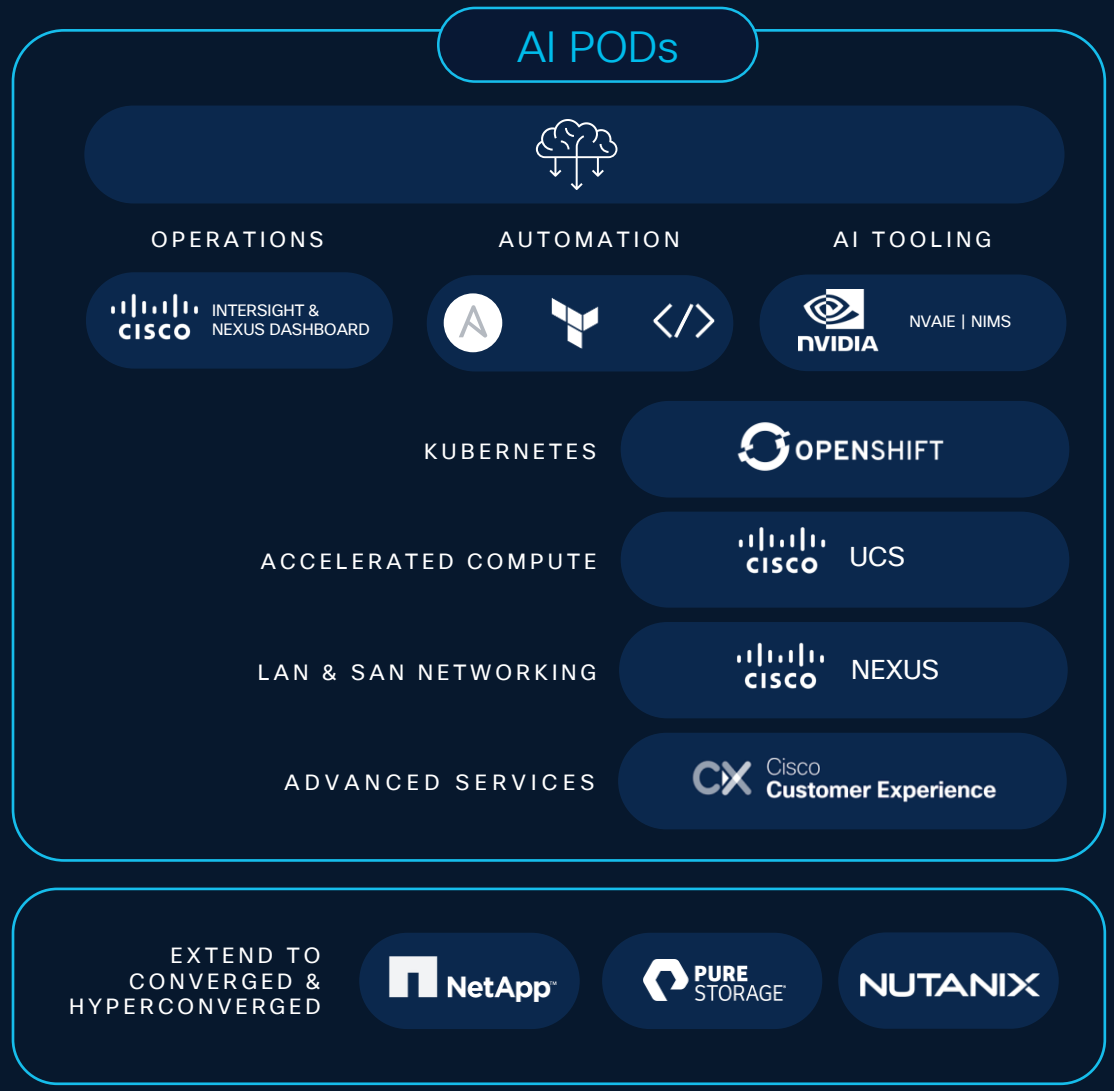
Orderable, validated AI-ready infrastructure stacks

Fully supported stack including Cisco and 3rd party components

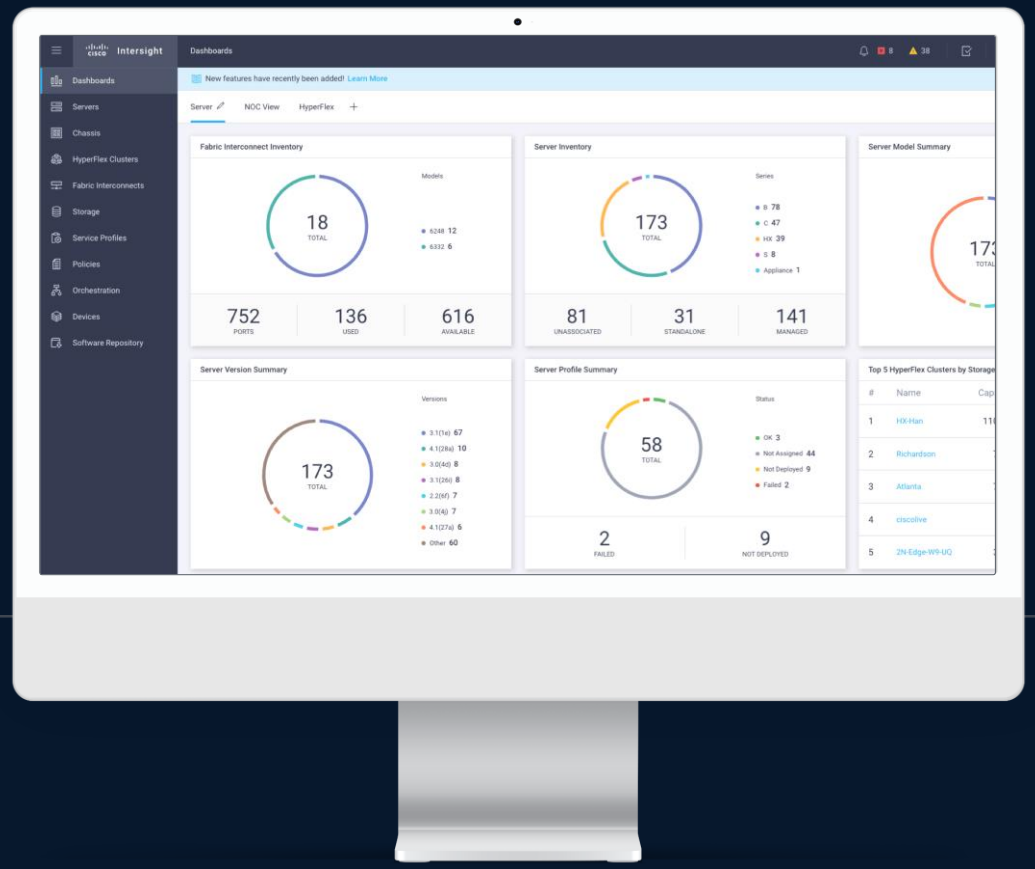
AI Advisor tool for configuration guidance

Coming Soon

Cisco AI-Ready Infrastructure Stacks



Work smarter and faster with a simplified, unified operating model



Cisco Intersight

See your global on-premises, cloud, and edge environments

Connect your infrastructure operations across compute and storage

Secure operations with built-in advisories and continuous risk mitigation

Automate deployments, configuration, workflows, and day-0 to day-N tasks

AI-driven capabilities in Intersight

Deliver predictive insights
HCL alerts, contract status

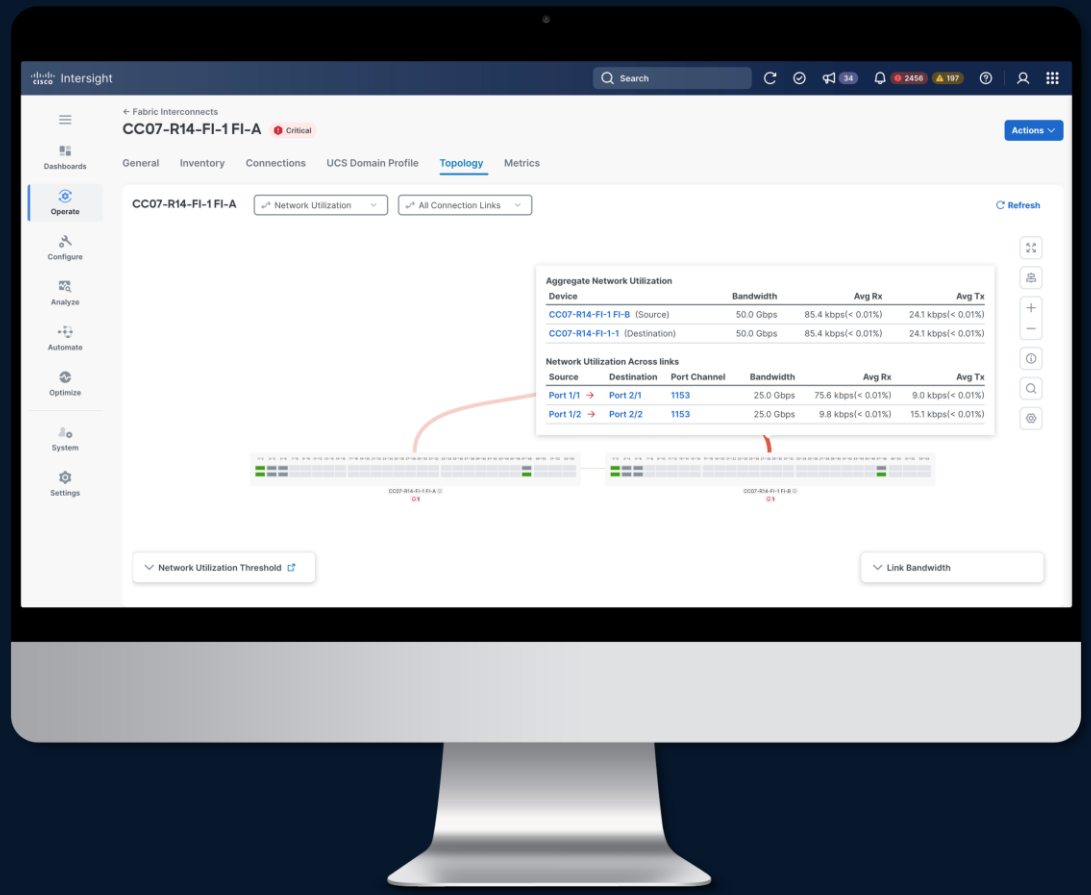
Predict potential failures
Enriched topology view with network bandwidth/utilization

Give contextual alerts, enhance security
Custom field notices, security advisories

Self-heal
Proactive RMAs

Automate routine tasks
Automatic log collection, Connected TAC

Deliver real-time and historical metrics
Analyze performance, troubleshoot, forecast and budget



Summary: Three Takeaways

The modern compute edge must support traditional and AI workloads, and a unified architecture provides flexibility while preventing architecture silos

SaaS-based unified management across simplifies operations across edge and data center with global visibility and control

Accelerate time-to-value and increase deployment confidence with Cisco Validated Designs and AI PODs, backed by Cisco full-stack support

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.

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

Contact me at: ronnchan@cisco.com

Thank you


CISCO