



# Harness a Bold New Era

Transform Data Centre and  
Service Provider Connectivity

**Kevin Wollenweber**

SVP and GM, Data Center and Provider Connectivity

**Jeremy Foster**

SVP and GM, Compute

**cisco** *Live!*

#CiscoLiveAPJC

# One Cisco

Future-proofed workplaces

Digital resilience

AI-ready data centers

Collaboration

Security

Observability

Networking

Data sources



AI



Infrastructure for AI

Data center | Mass-scale network



# Infrastructure for AI

Data center

| Mass-scale network



## Manufacturing

Predictive maintenance  
Quality control  
Demand forecasting



## Public sector

Smart cities  
Security and safety  
Services improvement



## Retail

Personalization  
Inventory optimization  
Sales forecasting



## Financial services

Fraud detection  
Risk assessment  
Trading

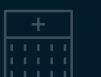


## Healthcare

Diagnosis  
Drive-thru optimization  
Patient support

## Education

Learning & teaching  
experiences  
Smart & secure facilities



Cloud & Enterprise IT

Networking team

Line of business

**AI practitioner**



# Every organization's AI approach and needs are different

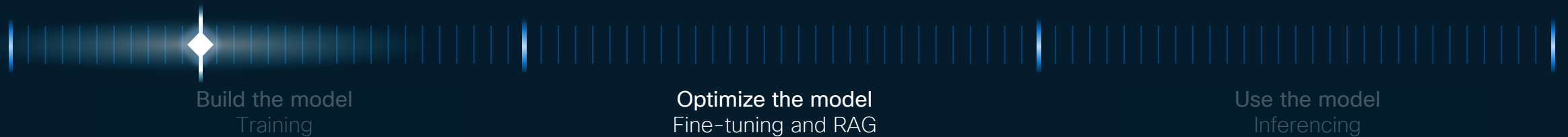


Build the model  
Training

Optimize the model  
Fine-tuning and RAG

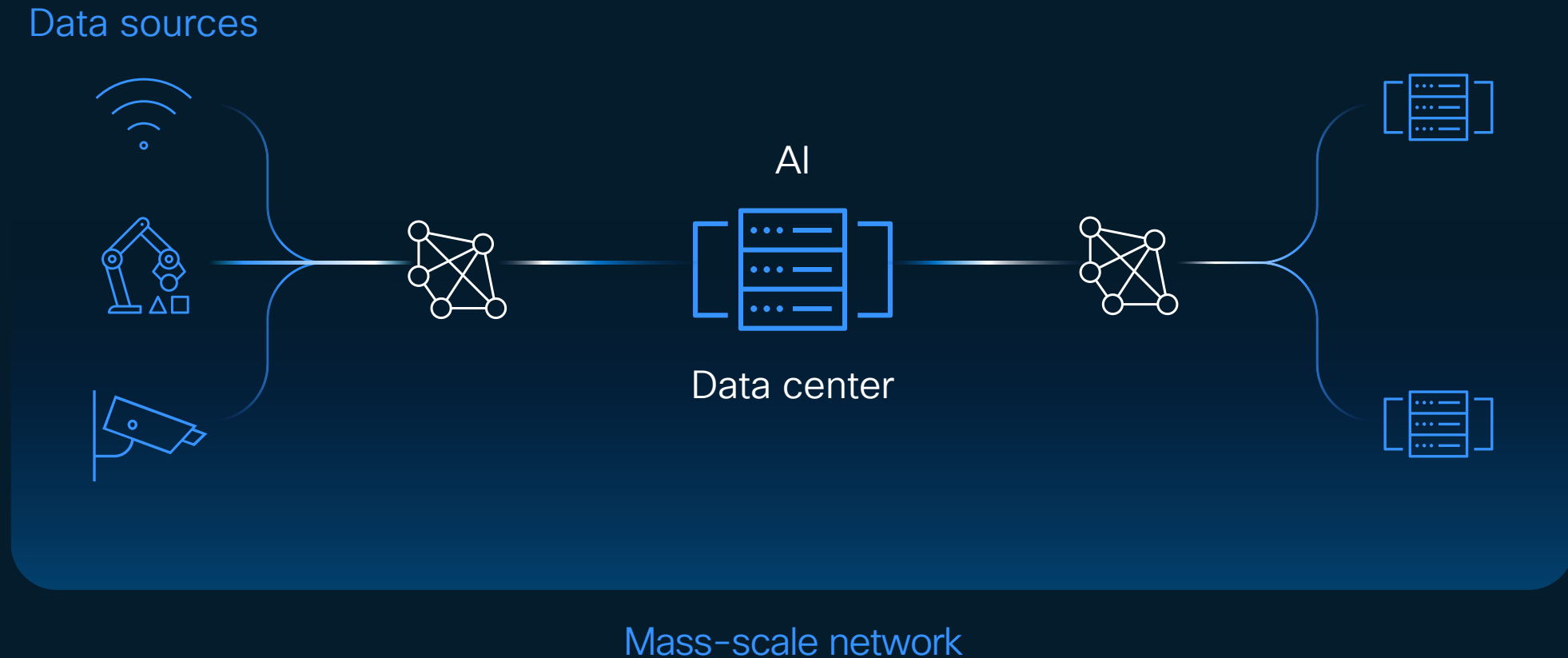
Use the model  
Inferencing

# Every organization's AI approach and needs are different



# Data centers are the Gold Rush of 2025

## Infrastructure for AI





## Webscale

Core and Optics

Business agility



## AI-native operations

Automation that predicts and prescribes

Operational efficiency



## Data Center Interconnect (DCI)

Networking across data centers

Scalability

FIRESIDE CHAT



Ben  
O'Shea

General Manager, Transformation & Cloud Platform  
Aussie Broadband

cisco *Live!*





Data Center Networking



Silicon One



Optics



Unified Compute



Data Center Networking



Data Center Networking



Unified Compute

# Building blocks



## Cisco Data Center Networking

**Nexus 9000 Series: 400G/800G**

Innovations in Cisco Silicon and Optics

**Nexus Hyperfabric**

**Nexus ONE solution**

**Validated Designs**

Web and  
Mass-scale  
Networking

Cisco 8000

Cisco Silicon and Optics

SONiC

Crosswork Automation



## Cisco Unified Compute

**Cisco UCS AMD M8 Portfolio**

Cisco UCS Intel M7 Refresh

**Cisco UCS X-Series Direct**

**Cisco Compute Hyperconverged  
with Nutanix**

Cisco Intersight Advanced Monitoring

Cisco Intersight Sustainability Dashboard

New Nvidia GPU Support

**Validated Designs**

**CISCO** Live!

#CiscoLiveAPJC

© 2024 Cisco and/or its affiliates. All rights reserved. Cisco Public

# Adding new building blocks to our AI Infrastructure portfolio

For data-intensive use cases like model training and deep learning

Orderable January

## Nexus 9364E-SG2

51.2 Tbps of bandwidth

Up to 64 ports of 800G

Two models: QSFP-DD and OSFP ports



Bringing high-density GPU servers to the Cisco UCS family

Orderable Now

## UCS Accelerated UCS C885A M8

NVIDIA HGX supporting:

8 NVIDIA H100 or H200 GPUs, or AMD MI300X GPUs

2 AMD 4<sup>th</sup> Gen/5<sup>th</sup> Gen EPYC Processors

# Introducing cloud-based network fabric as a service

## Cisco Nexus Hyperfabric

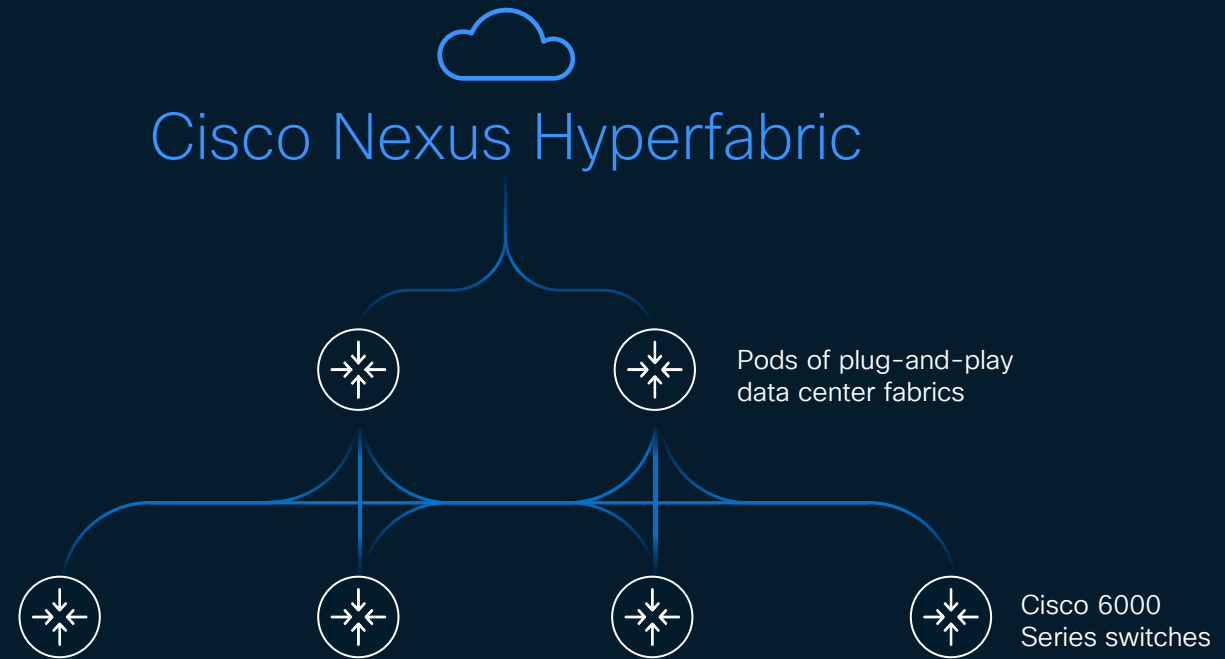
Cloud-managed  
data center network

Outcome driven results

Cloud-manage operations

Easy enough for IT generalists

Orderable January



Built on Cisco Silicon One and Optics innovations

Initial use cases: Greenfield | Mid-size and edge data centers | Service provider PoPs





Cisco Nexus Hyperfabric  
Beta Customer



# Cisco AI Infrastructure Solutions

Cisco Nexus  
Hyperfabric AI

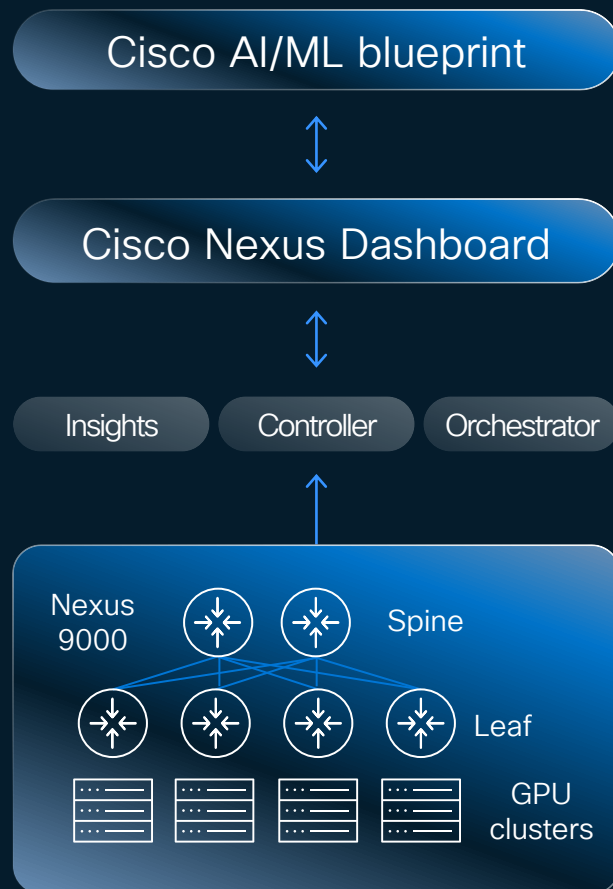
Orderable Q4FY25

Data Center  
Networking  
Blueprint &  
CVDs for AI

Available now

AI  
PODs

Orderable this week



## OUTCOMES



**Best performing**  
AI/ML networks, focus  
on app performance



**Intelligent buffering,**  
low latency, telemetry /  
visibility, RoCEv2



**Dynamic congestion**  
avoidance for  
various workloads



**One IP / Ethernet**  
network vs. dedicated  
front-end / back-end



**Lower TCO**



**Validated designs**  
for network and  
ecosystem partners

# Solutions to simplify and automate AI Infrastructure

Faster time  
to value

Less  
risk

More  
predictability

Deeper  
resilience



NUTANIX



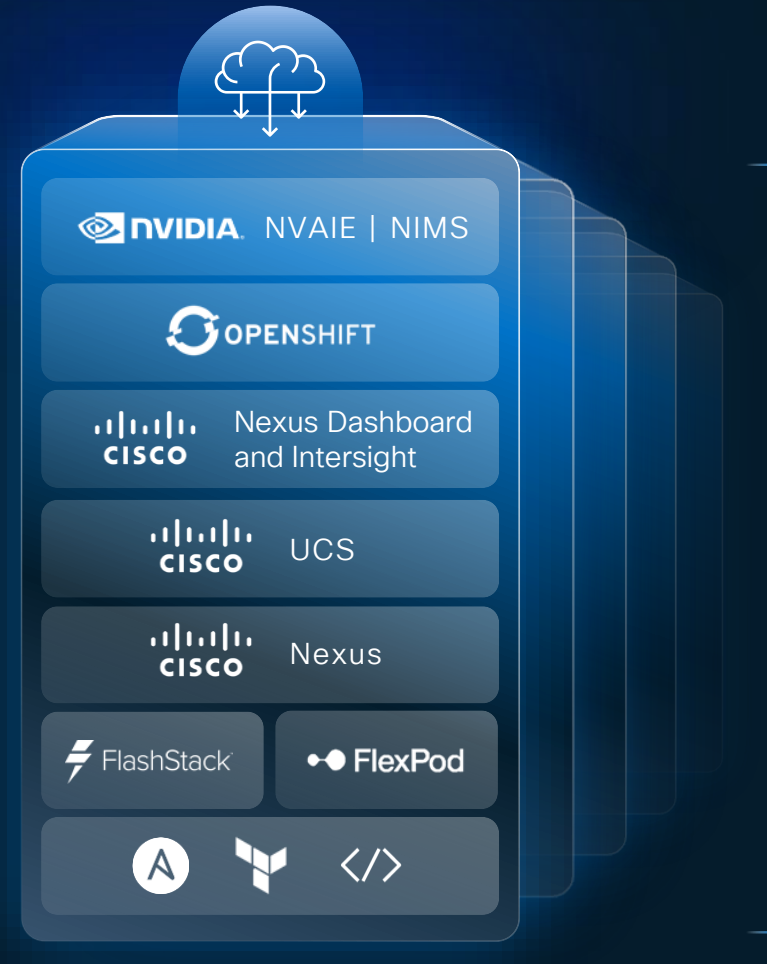
CLOUDERA

Build the model  
Training

Optimize the model  
Fine-tuning and RAG

Use the model  
Inferencing

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



**Edge Inferencing**  
(7B-13B Parameter)

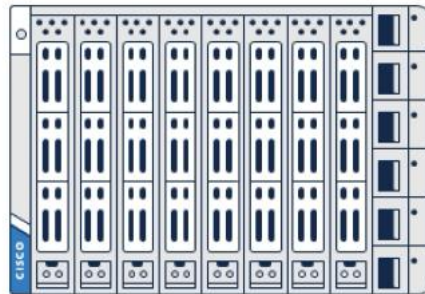
**RAG Augmented Inferencing**  
(13B +Parameter)

**Large Scale RAG Augmented Inferencing**  
(70B+ Parameter)

**Large Inferencing Cluster**  
(Inferencing Multiple Models)



NVAIE | PYTORCH | LLAMA



ARRAY 1

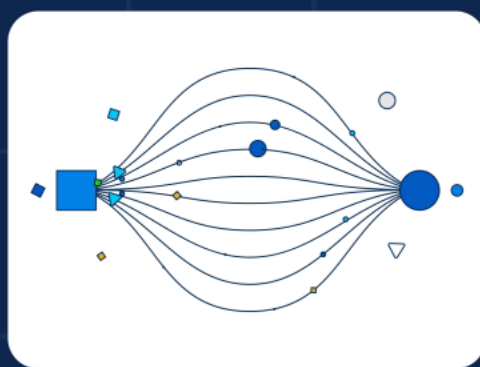
ARRAY 2



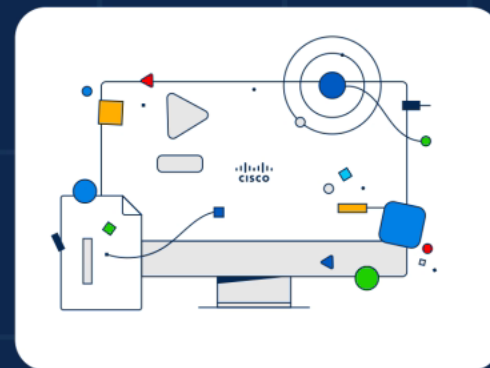
## Build



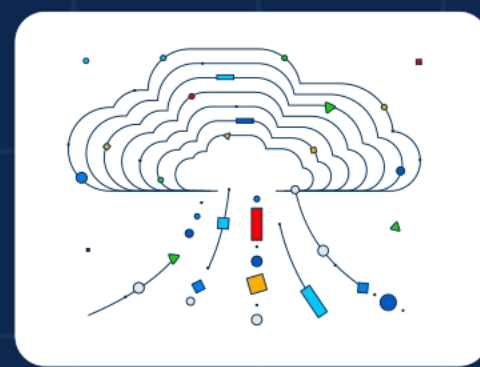
## Deploy



## Visibility



## Security





[Back](#)[Save as draft](#)[Next](#)

- General
- Platform Selection
- Blueprints Selection
- RAG Augmented Inferencing
- System Configuration
- Optional Settings
- 5 Summary

## Summary

Verify details of the blueprints and policies, resolve errors if any, and create workloads.

### General

Name  
**NVAIE-Inferencing-Workload**

Organization  
**default**

Tags  
-

Description  
**AI POD Inferencing Workload**

Platform Selection

**UCS X Series**

**General** System Configuration Optional Settings

Platform  
**VMware vSphere 8**

Kubernetes  
**Red Hat OpenShift Container Platform4.14.8**

Storage  
**Array-1**

GPU Operator  
**NVIDIA GPU Operator 23.9.1**

Inference Container  
**NeMo Framework Inference 23.10**

### Overview



- Data Network
- Storage Network
- Fabric Interconnect
- X210c 1TB w/NVAIE Stack
- 2x NVIDIA L40S GPU
- Empty Slot

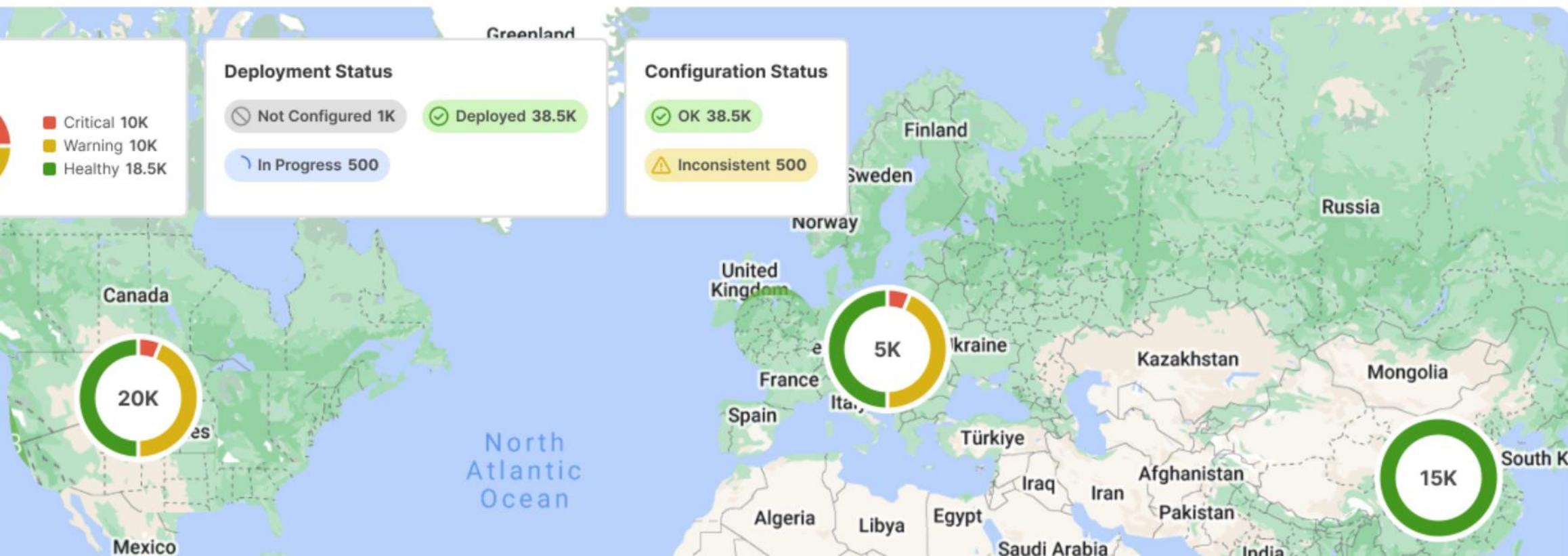
Cancel

Back

Save as draft

Create





ch

14

18

29

ection Vulnerability

Acknowledge

Summary

Details

Affected Devices

egrated Management Controller (IMC) Software could allow an user to a malicious web page.

tion of parameters in HTTP requests. An attacker could exploit this t to an affected system, which could cause the web interface of a malicious URL. This vulnerability is known as an open redirect



Infrastructure  
to power AI



Services to accelerate  
the value of AI



Security for AI,  
AI for security



Data to drive insights  
and context



Software to  
unlock productivity

Only Cisco brings  
together **networking,**  
**compute, security, and**  
**observability** to power  
AI-ready data centers

# One Cisco