



Complexity to Clarity

Optimising Data Centers for the Future

Chris Gascoigne
Principal Solutions Engineer

March 2025



*Let's talk
about the
elephant in
the room ...*

“

“The server virtualisation market is facing its most significant disruption in over a decade. I&O leaders will be forced to question their underlying assumptions for current and future workloads”

– Gartner

THE NEW STACK

Why Broadcom Is Killing off VMware's Standalone Products

Jan 23rd, 2024 11:59am by Chris J. Preimesberger

ars TECHNICA

BIZ & IT TECH SCIENCE POLICY CARS GAMING & CULTURE

VMWARE, BUT BY BROADCOM —

VMware customers face uncertain future as Broadcom ends VMware partner programs

Only Broadcom's favorites will be able to sell VMware-related offerings.

SCHARON HARDING · 1/10/2024, 3:53 PM

THE CHANNEL CO.
CRN

Virtualization ▶

Broadcom Takes Top VMware Accounts Direct 'Effective Immediately'

BY O'RYAN JOHNSON
JANUARY 8, 2024, 8:42 AM EST

The Register

Dell kills sweetheart distribution deal with Broadcom's VMware

No longer willing to be a friend with benefits — perhaps thanks to Big B killing OEM licenses?

Simon Sharwood

Tue 30 Jan 2024 · 23:35 UTC

THE CHANNEL CO.
CRN

Virtualization ▶

Broadcom Is Making Shareholders Rich, Rivals Happy And VMware Partners Bitter

BY O'RYAN JOHNSON
JANUARY 28, 2024, 5:15 PM EST

Significant Broadcom Changes

Organisations already seeing 2x-6x increases at renewal



Ending perpetual
licenses



Transition to
subscription
licensing



Move from CPU to
Core based
licenses



Reduce portfolio
to 4 packages*,
enforced VCF for
“strategic”
customers



*Let's talk
about the
other elephant
in the room ...*



Artificial Intelligence

AI Changes **Everything**

\$15.7T

Potential contribution to
global economy by 2030

\$300B

Global spending on
AI by 2026

75%

Of large enterprises will rely
on AI-infused processes by
2026

“30% of GenAI Projects Will Be Scrapped by 2025 Due to Lack of ROI”

– Gartner



Improperly Sized Infrastructure

“Build it and they will come”

Over-sizing puts ROI at risk

Under-sizing puts adoption at risk

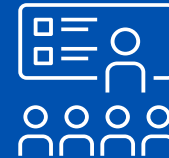


Overly Complex Infrastructure

Long “time to value”

High operational costs

Data challenges



Skills Gap

Delayed application delivery

“Science projects”



The background features a series of concentric, semi-circular patterns on the left side, composed of many thin, overlapping segments in shades of light blue, grey, and yellow. On the right side, there are several parallel, slightly curved lines in a light blue-grey color, creating a sense of depth and perspective. A solid black horizontal bar spans the width of the image, positioned in the lower-middle section, containing white and green text.

These disruptions are both drivers of
Data Centre Modernisation

Platform Modernisation Options

STAY

OPTION 1

vmware®

Modern 3-Tier
Architecture (CI)

Refresh to reduce
licensing costs

PURESTORAGE® NetApp® HITACHI

OPTION 2

vmware®

Adopt Hyperconverged
Infrastructure (HCI)

vmware® vSAN | NUTANIX

MOVE

OPTION 3

Alternative Approach

NUTANIX

RED HAT®
OPENSIFT

Azure Stack HCI

Bare Metal

OPTION 4

Public Cloud

NUTANIX

RED HAT®
OPENSIFT

Cloud Native

Public cloud isn't the easy option any more

*“You’re crazy if you don’t start in the cloud; you’re **crazy if you stay on it**”*

– *The Cost of Cloud, a Trillion Dollar Paradox*, Andreessen Horowitz

83%

of organisations plan to **repatriate workloads** to private cloud

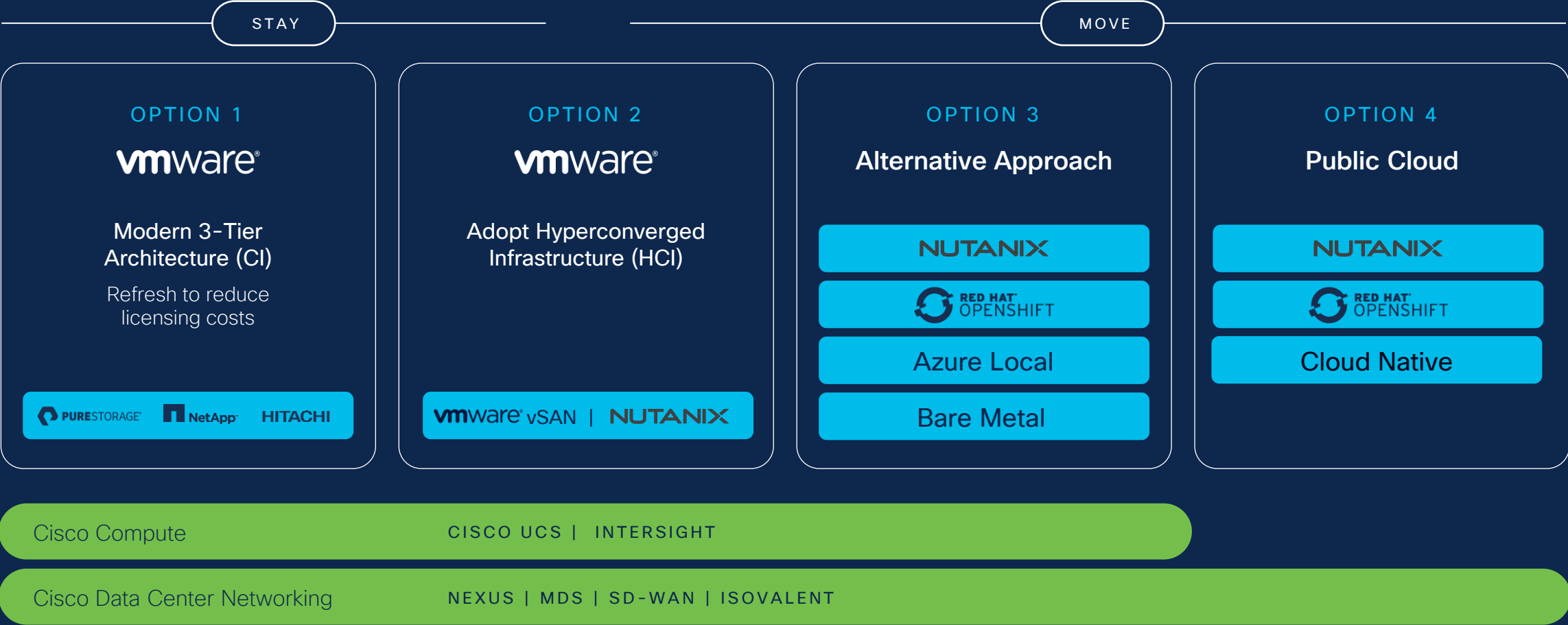
– Barclay’s CIO Survey 1H24

84%

of organisations list **managing cloud spend** in their top cloud challenges

– Flexera 2024 State of the Cloud Report

Common infrastructure platforms are essential



Blurring the line between rack and blade

HCI on our most sustainable, modular server yet

Modular, sustainable,
easily upgradeable solution

Higher performance,
smaller footprint

Powers modern apps
and traditional workloads

Fabric-based architecture with
unified fabric connectivity

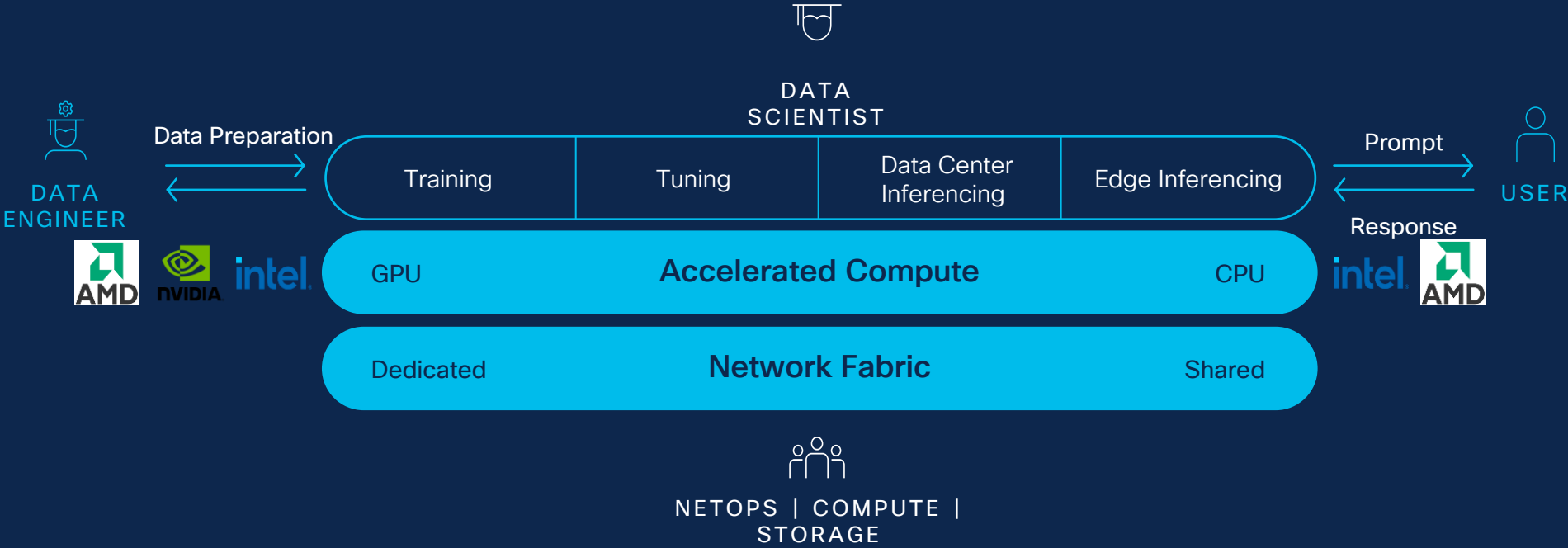
Up to: 2PB raw storage, 8 x H100 NVL GPU,
1376 cores, 64TB RAM per chassis

* With UCS M8 blades



Sources: Energystar.gov / EPEAT.net

AI Workloads are a Spectrum



Model performance tightly coupled to infrastructure

New Building Blocks for AI Infrastructure

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

Nexus 9364E-SG2

51.2 Tbps of
bandwidth

Up to 64 ports
of 800G

Two models:
QSFP-DD and OSFP ports



UCS C845A M8

NVIDIA MGX supporting:

2-8 NVIDIA
H100NVL/H200NVL/L40
S GPUs

2 AMD 5th Gen
EPYC Processors

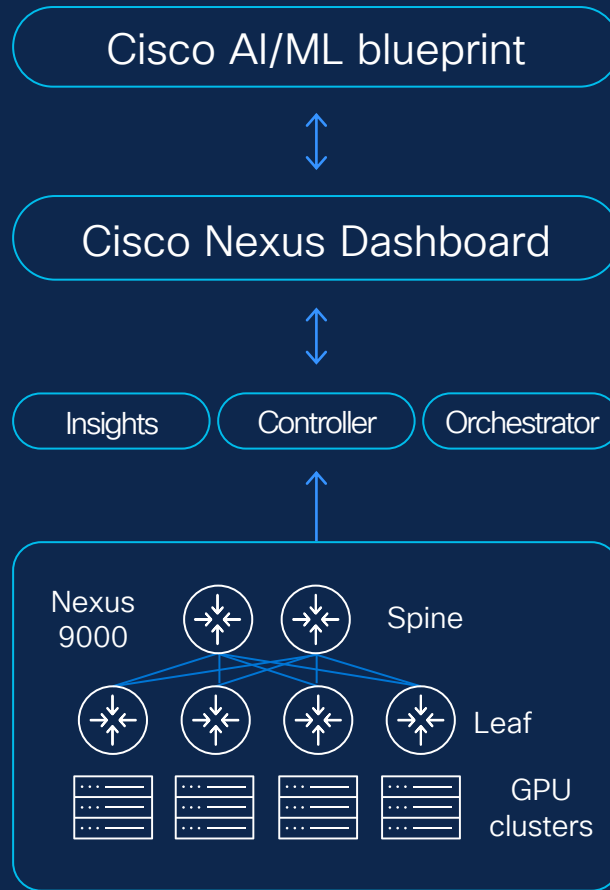
UCS C885A M8

NVIDIA HGX supporting:

8 NVIDIA H100 or
H200 GPUs, or AMD
MI300X GPUs

2 AMD 4th Gen/5th Gen
EPYC Processors

Blueprint for AI/ML Networking



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



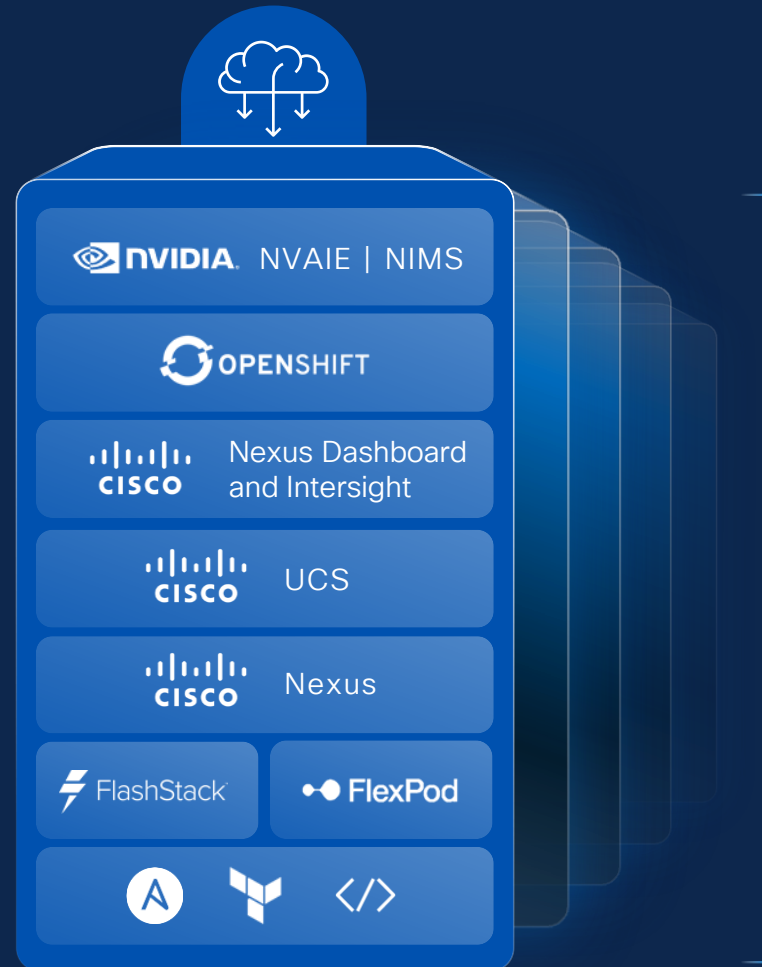
Lower TCO



Validated designs
for network and
ecosystem partners

Cisco AI Pods

- 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)

Infrastructure Modernization for AI

AI PODs

Typical use case	Data Centre and Edge Inferencing	RAG Augmented Inferencing	Scale Up for High Performance	Scale Out for Large Deployments	Roadmap
Sizing example	(Llama-2 7B GPT 2B)	(Llama-2 13B OPT 13B)	(Code Llama 34B Falcon 40B)	Multi-Model Deployments High Concurrency	
PID	UCSX-AI-EDGE	UCSX-AI-RAG	UCSX-AI-LARGERAG	UCSX-AI-LARGEINF	
Pod specifications	<div>1x x210C compute node</div> <div>2x Intel 5th Gen 6548Y+</div> <div>512 GB system memory</div> <div>2x 1.6 NVMe drives</div> <div>1x x440p PCIe</div> <div>1x NVIDIA L40s</div> <div></div>	<div>2x x210C compute nodes</div> <div>4x Intel 5th Gen 6548Y+</div> <div>1 TB system memory</div> <div>4x 1.6 NVMe drives</div> <div>2x x440p PCIe</div> <div>4x NVIDIA L40s</div> <div></div>	<div>2x x210C compute nodes</div> <div>4x Intel 5th Gen 6548Y+</div> <div>1 TB system memory</div> <div>4x 1.6 NVMe drives</div> <div>2x x440p PCIe</div> <div>4x NVIDIA H100 NVL</div> <div></div>	<div>4x x210C compute nodes</div> <div>8x Intel 5th Gen 6548Y+</div> <div>4 TB system memory</div> <div>8x 1.6 NVMe drives</div> <div>4x x440p PCIe</div> <div>8x NVIDIA L40s</div> <div></div>	

Performance and Scale →

Nexus Hyperfabric

Cloud-Managed Operations | Outcome Driven for IT Generalists | Modernise Infrastructure Lifecycle



Design



Order



Deploy



Validate



Monitor



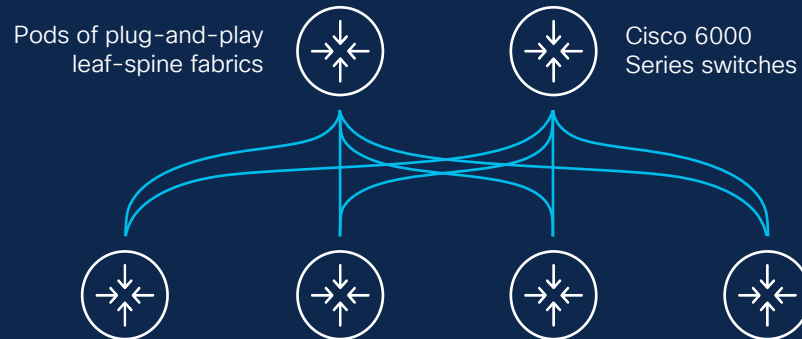
Upgrade



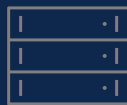
Collaborate

Available now

Cisco Nexus Hyperfabric

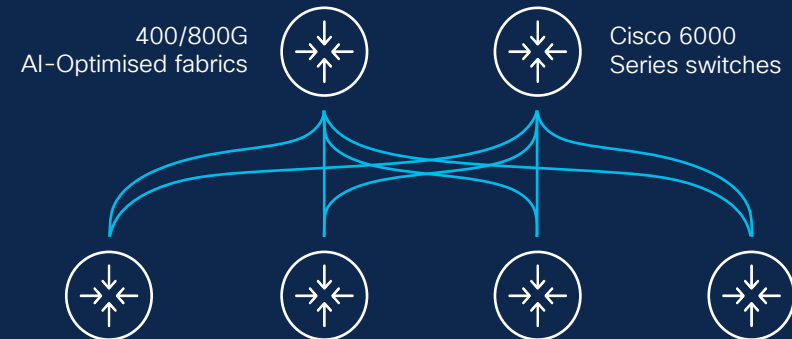


Any servers/
storage/etc



*Shipping mid-CY25

Cisco Nexus Hyperfabric AI*



Cisco UCS
c885A M8



NVIDIA GPU



NVIDIA DPU/NIC
BlueField-3



VAST Storage



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

Cisco Confidential

AVAILABLE NOW

Cisco Nexus Hyperfabric



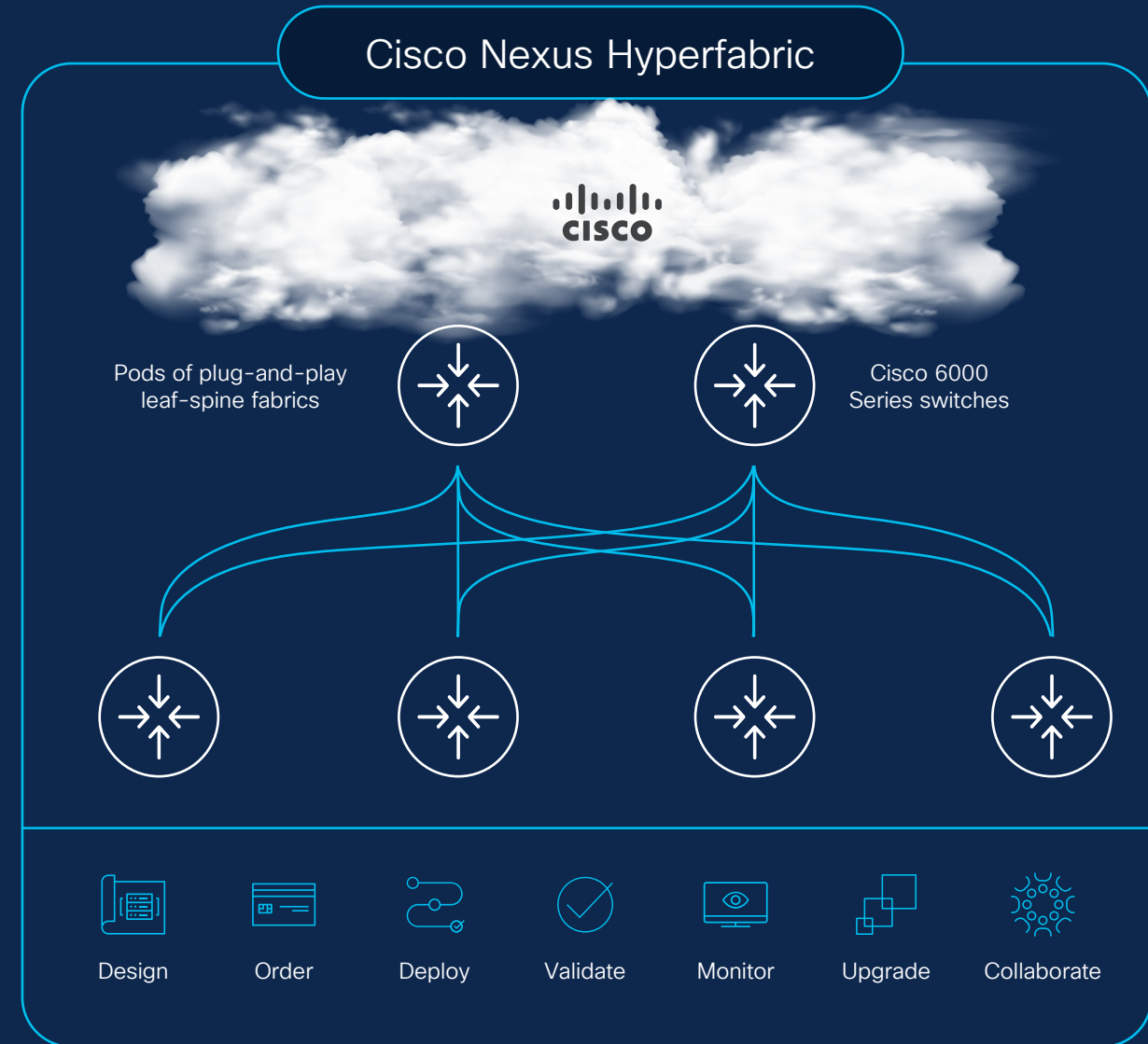
Design, deploy and operate on-premises fabrics located anywhere



Easy enough for IT generalists, application and DevOps teams



Outcome driven by a purpose-built vertical stack



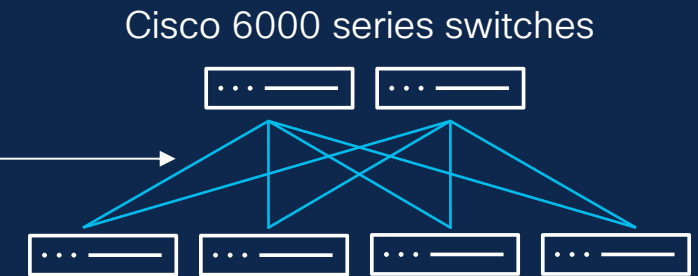
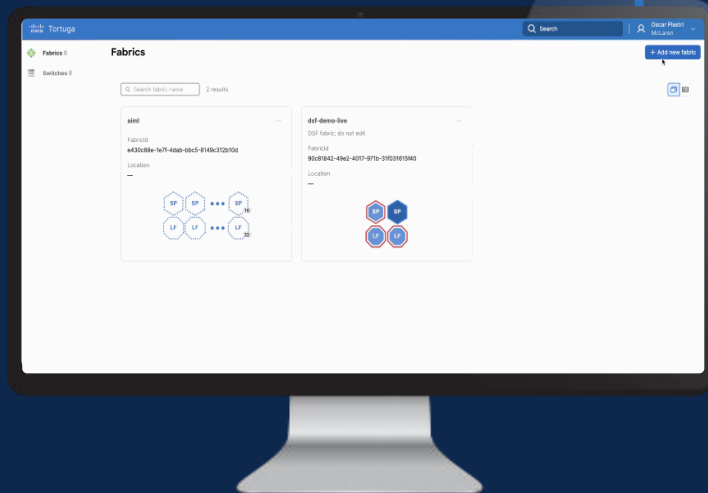
© 2025 Cisco and/or its affiliates. All rights reserved.

Cisco Confidential

Nexus Hyperfabric

How It Works

Plan, deploy, manage

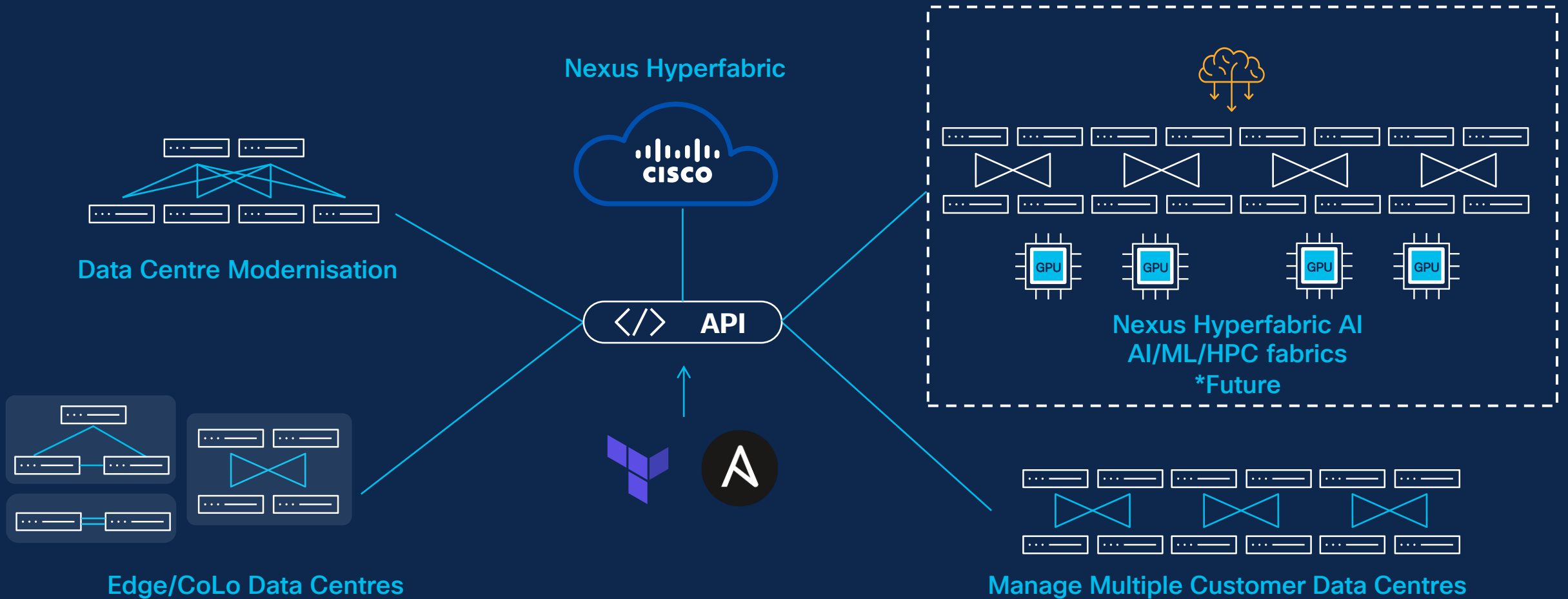


Plug-and-Play DC Fabrics
Self discovery / standards-based
Always-on telemetry
Assertion-based monitoring

Purpose-built for **predictable outcomes** optimised for ease of use

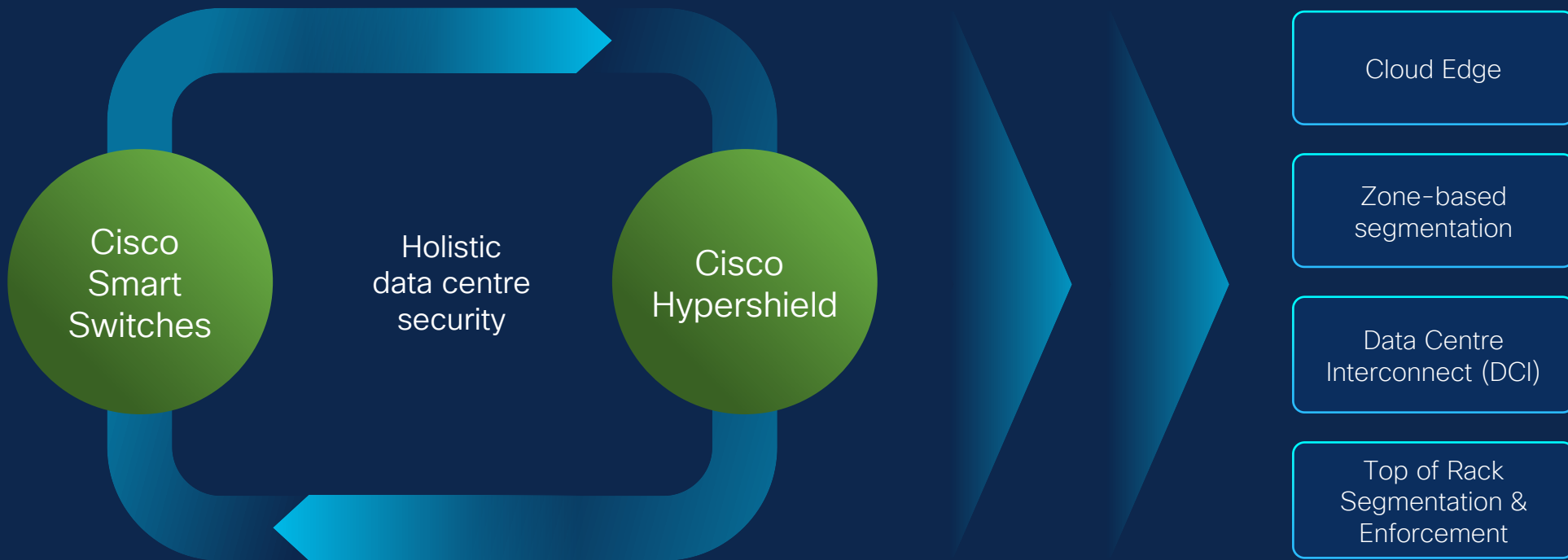
Nexus Hyperfabric

Use Cases



Security infused into the data centre fabric

Cisco Nexus 9300 Smart Switches



Where security meets the network

Simple insertion

- No more costly appliances
- Simple topology
- Very high performance

Autonomous policies and updates

- Test policy changes
- Continuous updates

Advanced protections

- Autonomous segmentation
- Comprehensive inputs for policy creation



Platforms for DC
Modernisation

AI-Ready
Infrastructure

Security Infused into
DC Fabric

Thank you
for
attending
our session



Visit our stand for expert advice and live demonstrations including:

1. Cisco Hypershield: Reimagining Security for the AI Era
2. Cisco Nexus Hyperfabric: A New Data Centre Experience
3. Webex Meetings, Microsoft Teams Meetings, Smart Workplace and Cisco Video Devices

Don't miss
our next
session –
on now



A New Era of AI-Native
Security for Data
Centers and Cloud

Dan Boucaut, Cyber Security
Private Sector Lead, Cisco

Breakout Room P10

