




AMD DPU-Accelerated Networking for AI and Services in High-Performance Data Centers

Subtitle goes here

Shane Corban – Senior Director of Product Marketing
Network Technology Solutions Group, AMD
PARDCN-1333

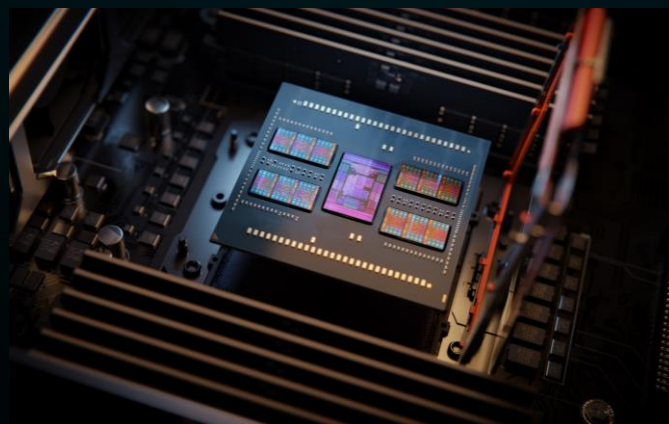


DPU-Accelerated Networking for AI and Services in High- Performance Data Centers

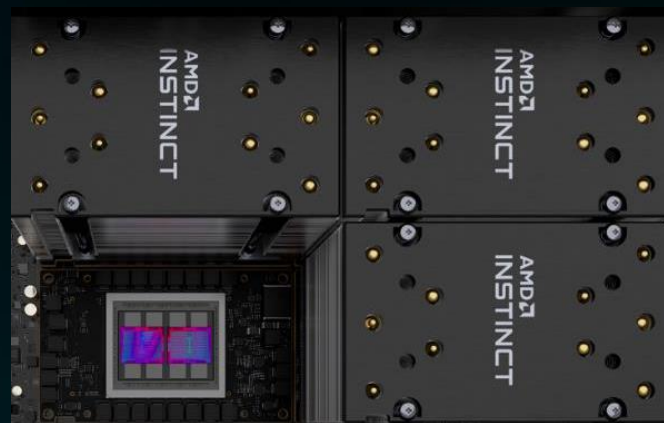
Shane Corban
Senior Director of Product Marketing
Network Technology Solutions Group, AMD

The Most Demanding Data Center Workloads

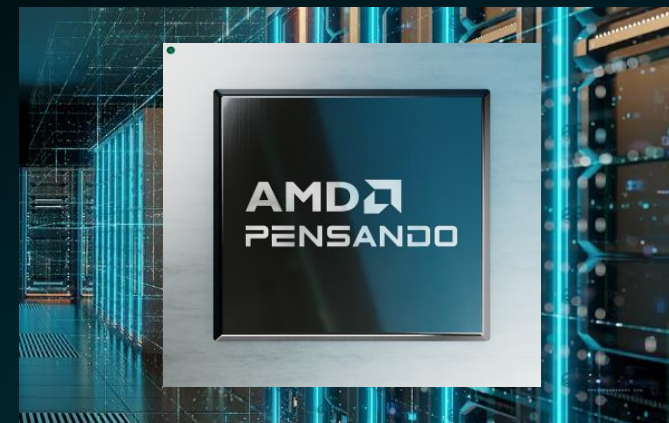
Requires leadership CPU, GPU and networking



AMD
EPYC



AMD
INSTINCT



AMD
PENSANDO

High performance AI systems for data center

AI Networking

Front-end Network

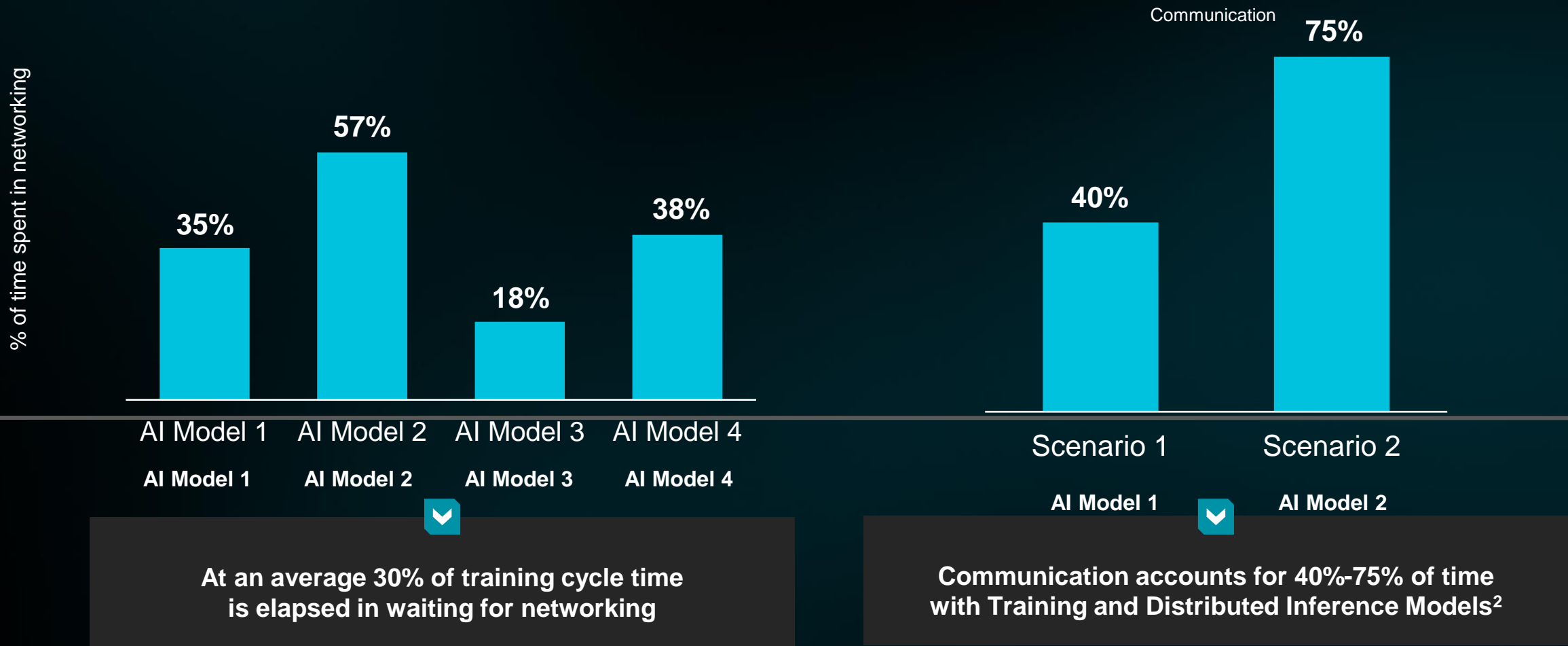
- ✓ Faster data ingestion
- ✓ Secure multi-tenant access
- ✓ Storage acceleration
- ✓ Zero CPU overhead

Back-end Network

- ✓ Scalability
- ✓ Resiliency
- ✓ High network utilization

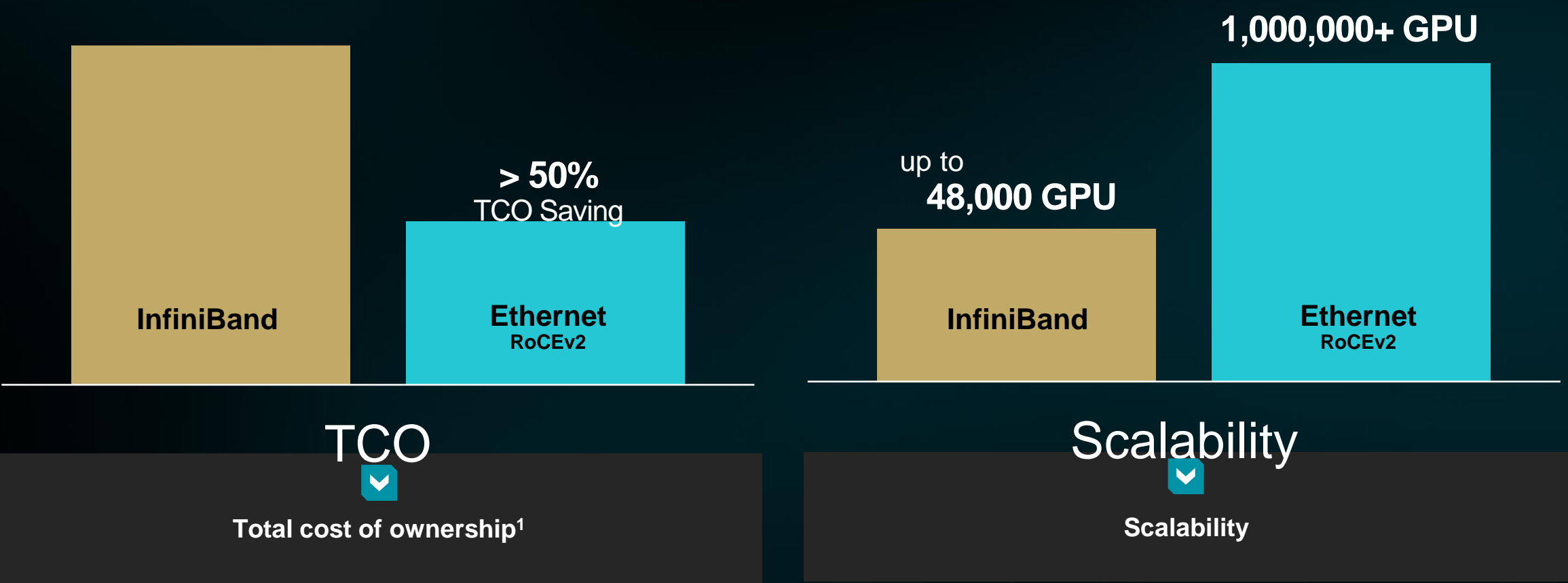


Network Challenges for AI Workloads



Sources: 1) 2022 OCP Keynote by Alexis Bjorlin, VP at Meta, 2) Computation vs. Communication Scaling for Future Transformers on Future Hardware, <https://arxiv.org/pdf/2302.02825>.

Ethernet is Always the Preferred Choice



Sources: 1) 650Group Datacenter AI Networking and Server SmartNIC Forecast Reports 2Q24.

AMD Pensando™ Pollara 400

“Pollara” - Industry’s first UEC ready NIC for AI backend networking

Availability Q1 CY25

Programmable
hardware pipeline

Up to 6x
performance boost for
RDMA

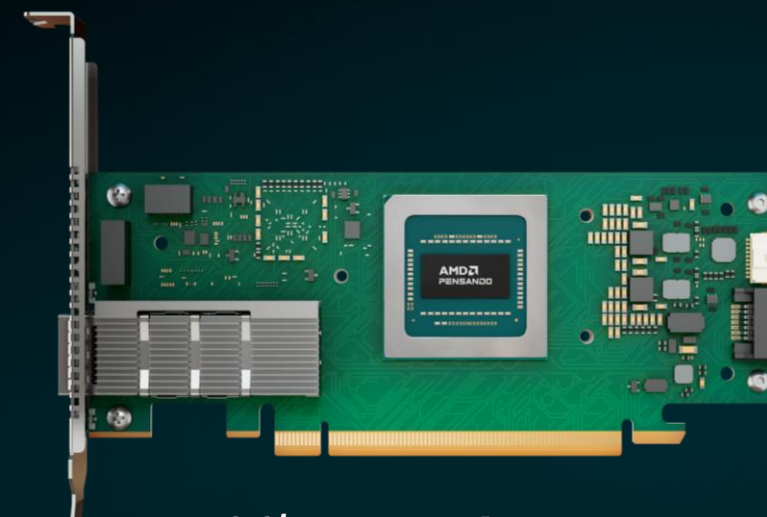
400 Gbps

Open ecosystem – any network (Lossy/Lossless)
UEC ready RDMA
Reduction in job completion times
Maximize Utilization for AI Backend Networks

Programmable
RDMA transport

Programmable
congestion control

Communication
library acceleration



Ultra Ethernet
Consortium

UEC Ready RDMA Enhancements **Outperform RoCEv2**

6x faster
message completion time

5x faster
collective completion time



Intelligent packet spray and in-order message delivery



Path aware congestion avoidance



Selective retransmission and fast loss recovery

* Reference: [STrack: A Reliable Multipath Transport for AI/ML Clusters](#), July 2024

AMD Pensando™ Salina 400

“Salina” 3rd Generation DPU - Best DPU for evolving AI front-end networks

Availability Q2 CY25

AMD Pensando™

#1 DPU for Hyperscalers

400G

PCIe® Gen 5 2x400GE

232 P4 MPU

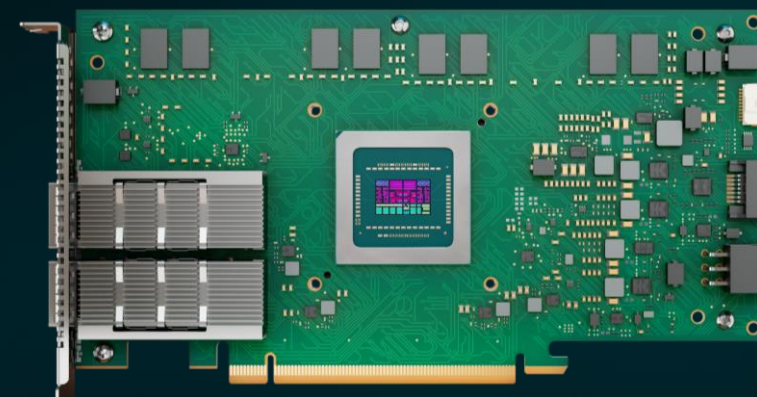
Multi-Services

2x DDR5

102GB/S Memory Bandwidth
Up to 64 GB DDR

16 N1

ARM Cores



Software
defined
network

Stateful
firewall

Encryption

Load
balancer

Network
address
translation

Storage
offload

Nexus 9300 DPU + Hypershield Enabled Platforms

From bolt-on to embedded services
into the data center fabric



Simplicity

- Combines networking and security into a single platform
- Reduces the need for dedicated hardware appliances
- Streamlines architecture for optimal traffic flow

Efficiency

- Improves Total Cost of Ownership
- Reduces power, cooling and space requirements
- Augments visibility with rich telemetry

Extensibility

- Futureproof data center architecture
- Activates stateful services when needed
- Creates a platform for innovation

Cisco Nexus 9000 Services Accelerated Switches

Best of breed platforms for DC enterprise & SP services

Nexus 9324C-SE1U (April 2025)



24-port 100G

Cloud Edge, Zone-Based FW & DCI

800G Services Throughput

4.8T Silicon One + 4 AMD DPU

Cisco 9348Y2C6D-SE1U (July 2025)



48-port 25G, 6-port 400G, 2-port 100G

DC Leaf/Access

800G Services Throughput

4.8T Silicon One + 2 AMD DPU



Benefits

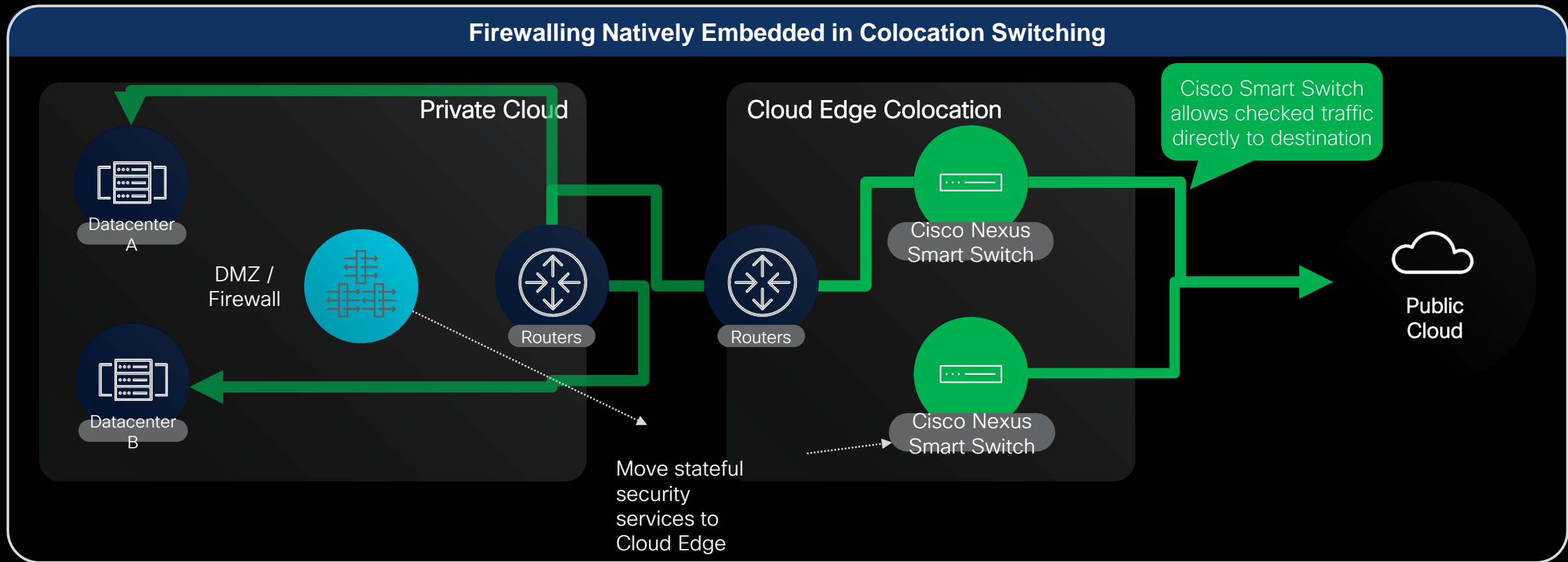
Policy scale with
stateful services

Pervasive application
traffic visibility

Zero trust security for
all workloads

TCO
reduction

Use Case #1: Securing Cloud Edge at On-Ramp



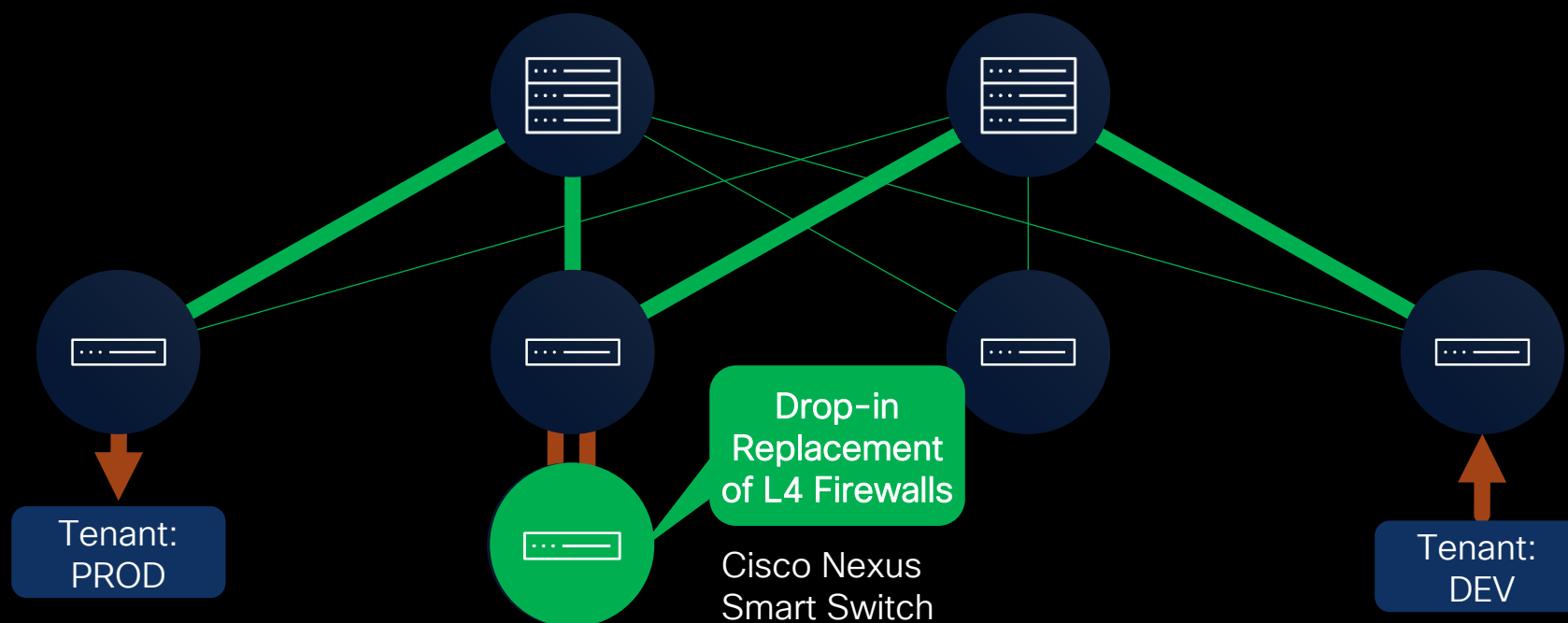
Main Benefits

- Simplifies Hybrid-Cloud connectivity w/o DMZ traffic tromboning
- Eliminates DMZ single point of failure
- Lossless HA and state failover for Active/Active firewall
- Hypershift ensures consistent security policy and policy testing
- Reduces cost with scale-out vs. firewall appliance scaling up

Use Case #2: Zone-based Segmentation

Securely connect security zones and business entities (tenants, VRFs) within and across fabric

Zone-based segmentation in DC with Cisco Smart Switch

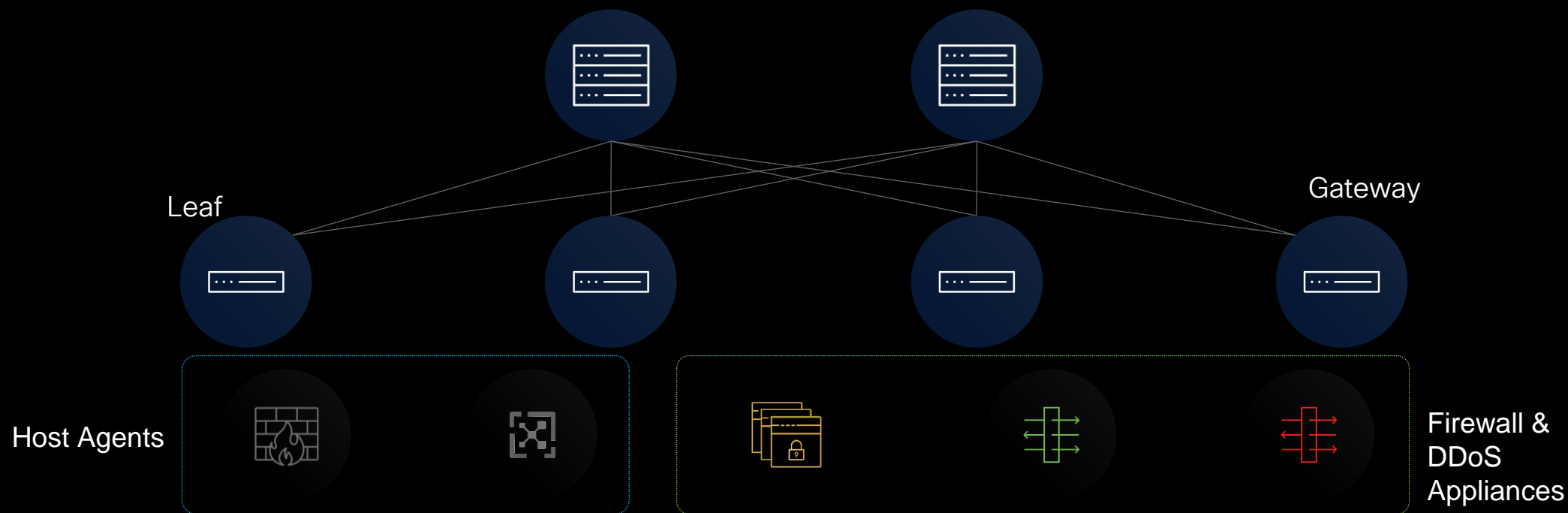


Main Benefits

- Starting point for extensible, consistent policy architecture across zones and workloads
- Stage policies before applying to production using Hypershield's dual-data plane

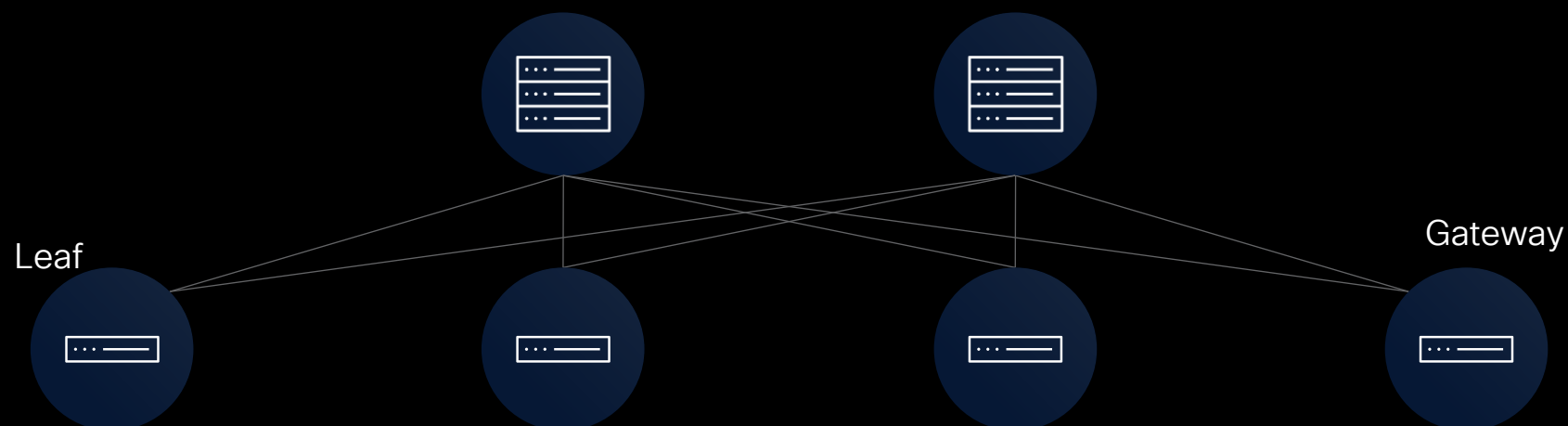
Use Case #3: Top of Rack Segmentation & Enforcement

From: Mixed Solutions, enforcement blindspots



Use Case #3: Top of Rack Segmentation & Enforcement

From: Mixed Solutions, enforcement blindspots

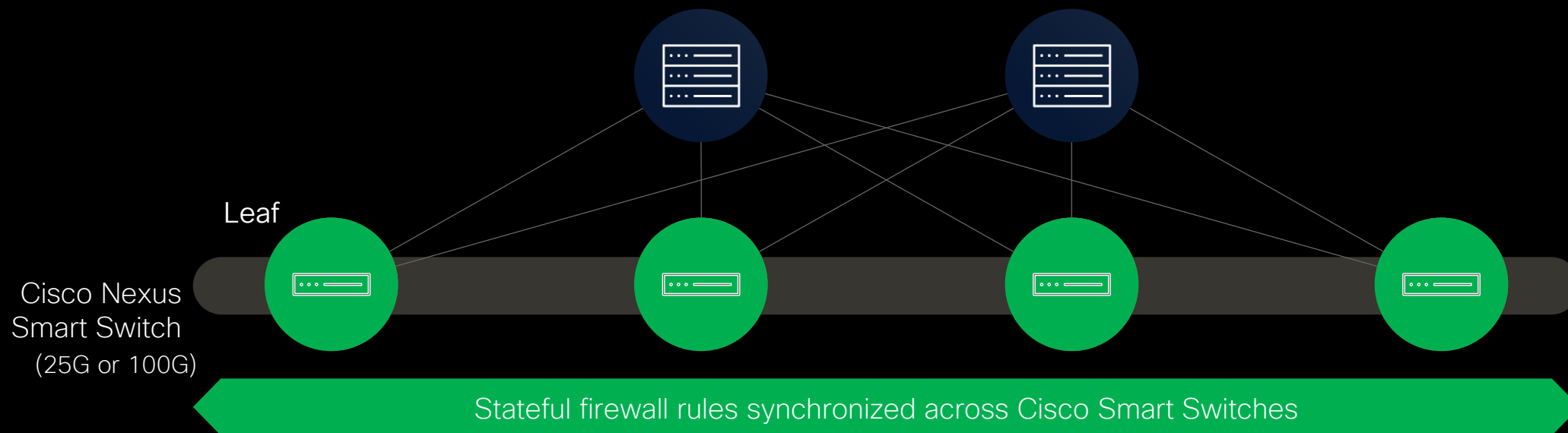


Challenges

- Multiple, disparate solutions - increased risk of error, inconsistent enforcement, & security breaches
- Point solutions hinder implementing holistic zero trust

Use Case #3: Top of Rack Segmentation & Enforcement

To: Pervasive East-West autonomous segmentation



Main Benefits

- Distributed stateful segmentation and L4 enforcement in every port
- Simple redirect policy (e.g. vrf or vlan) from network to within the switch
- Policy testing before deploy and firewall load updates
- Supports all workloads | Lower TCO

Cisco 400G Services Accelerated Switch

Cisco Silicon One Q200L

Supports 28x400G QSFP-DD

1.6T of DPU Services

- Using 8 AMD Elba DPUs on 4 SLED modules

SONiC Operating System

- SONiC instance on switch & DPU SLED

Upgraded System CPU & Memory

Services performance

- 24M CPS, 400M PPS

Cisco 8102-28FH-DPU-O



8K-DPU400-2A

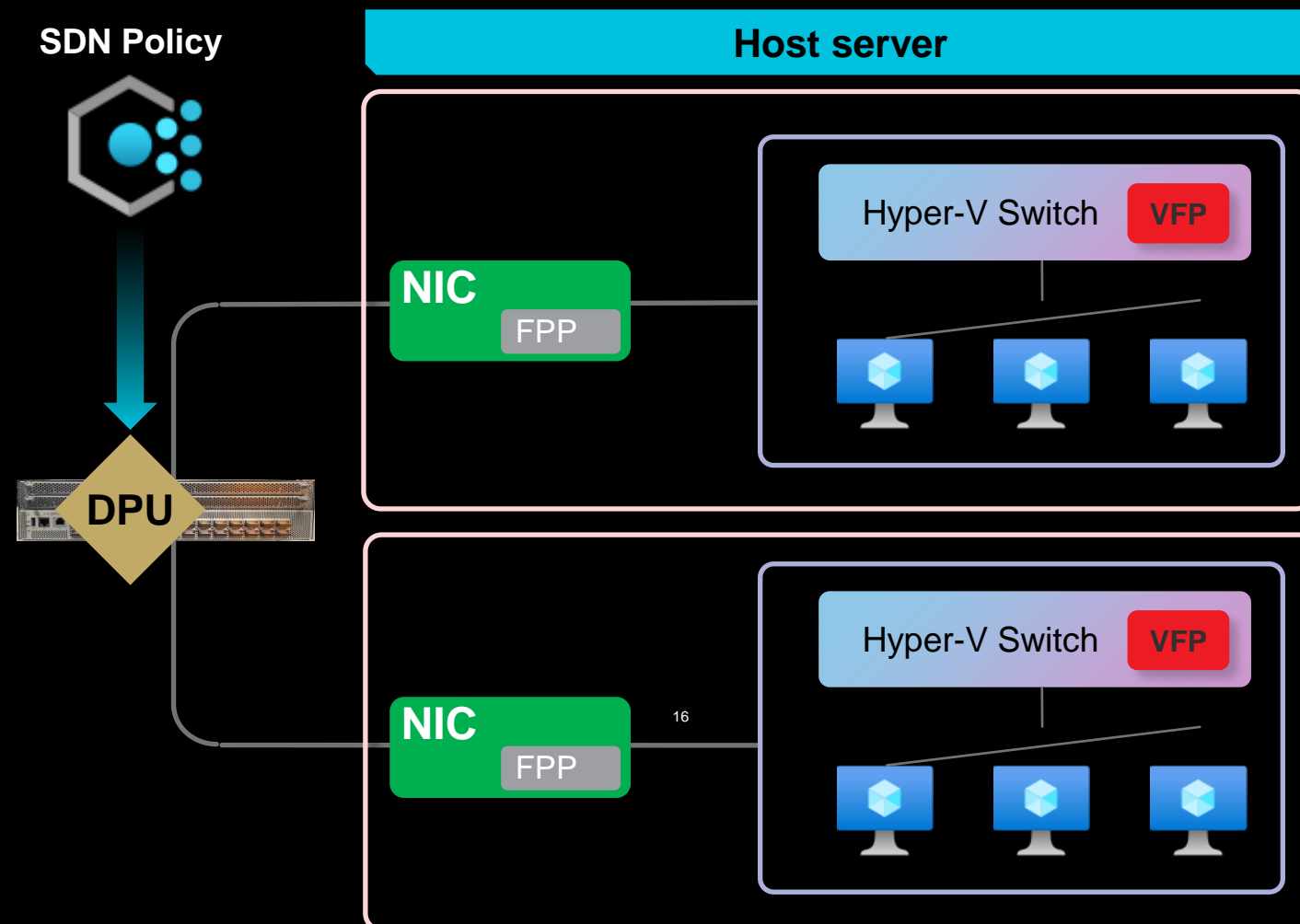


Enhancing SDN Services at Azure on Cisco Smartswitch

- 1 Offload network policy processing to DPU on Smartswitch
- 2 Complex policies are rendered in DPU
- 3 Service growth enabled through hardware scale in tier 1

24+
Million CPS

400+
Million PPS



Additional Reference Information

AMD Networking Landing Page:

- <https://www.amd.com/en/products/accelerators/pensando.html>

AMD Pollara Blog:

- <https://community.amd.com/t5/corporate/transforming-ai-networks-with-amd-pensando-pollara-400/ba-p/716566>

AMD STRACK Research Paper:

- <https://arxiv.org/pdf/2407.15266>

Cisco Nexus 9300 Smartswitch Platforms:

- <https://blogs.cisco.com/datacenter/fortify-your-data-center-with-new-cisco-n9300-series-smart-switches>



Thank You

Webex App

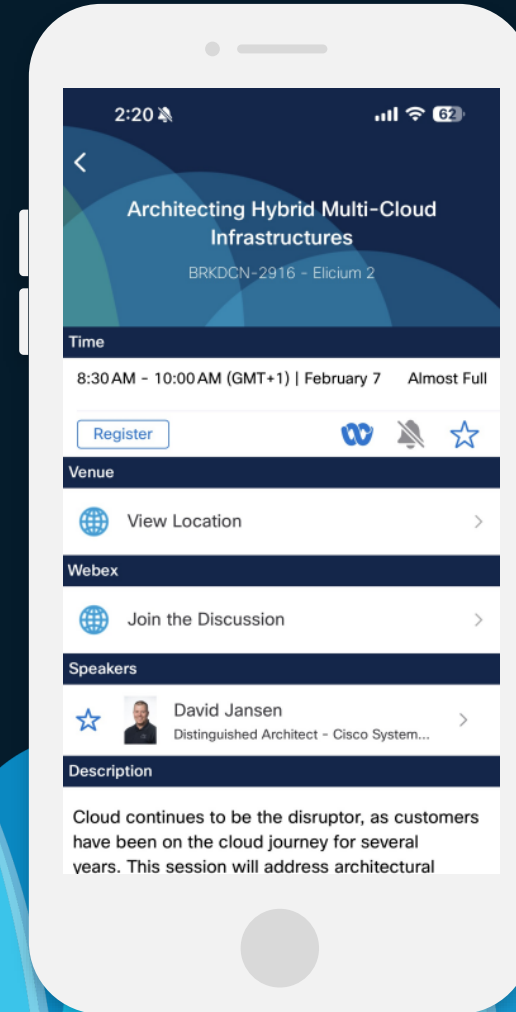
Questions?

Use the Webex app to chat with the speaker after the session

How

- 1 Find this session in the Cisco Events 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 February 28, 2025.



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.



Thank you

CISCO *Live!*

GO BEYOND

A series of overlapping, elongated oval shapes in various shades of blue, ranging from light sky blue to deep navy blue, positioned on the right side of the image.