

# **Ethernet Fabrics for AI Clusters**

Powered by Cisco SiliconOne or Nexus for Ultra High Performance, Scalable and Non-Blocking Ethernet Fabrics

S. John Banner - Technical Solutions Architect Lieven Colman - Site Reliability Engineer BRKCOC-3005

cisco /

# Webex App

#### **Questions?**

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

#### How

- 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.

cisco / illa



# About your Speaker

- S. John Banner
- Solutions Architect
- Based out of Porto, Portugal
- Cisco employee since 2000
- 35 years in IT industry
- Focus on DC Networking

How I spend my free time:

- Reading
- Exploring Portuguese Culture
- Spending time with Family and Friends



# About your Speaker

- Lieven Colman
- Site Reliability Engineer
- · Based out of Brussels, Belgium
- Cisco employee since 2021
- 17+ years with Cisco IT services
- Focus on DC Networking

How I spend my free time:

- Competitive padel fanatic and casual cycler
- Extensive culinary time with friends
- Amateur binge watcher and gamer
- Explore domestic creativity







#### Agenda

- Importance of AI to Cisco
- The Business Priority
- Architecture and Design
- Implementation
- Day 2 Support
- Lessons Learned
- What's Next?

cisco / ille

# Why is Al Important?

- · Changing the world
- · It's everywhere
- New market opportunities
- Improving our products
- Empower our services



cisco ile

# **Business Priority**

cisco live!

# Ask from Leadership

- Deployment of AI Cluster with 256 Nvidia H100 GPUs
- Blueprint for Large-Scale AI Cluster with Cisco Ethernet in Partnership with Common Hardware Group
- Deployment of AI Cluster with 512 H200 GPUs for Business Applications
- Deployment of AI Cluster with 92 H200 GPUs in UCS885s for Cisco Internal Cloud

# **Deployment Objectives**

- Validate AI Infra design with Cisco Technology stack
  - Support Engineering & Business use cases for AI model training
- Enable LLM use cases across Cisco products
- Enable AI Applications for Business Use Cases
- Offer GPU as a service

#### Timeline Key Milestones



# Getting Started The clock is ticking

cisco ive!

# Where We Began

Learning and Upskilling the Team

- Internal/External Research
- DevNet
- Training VODs/Classes
- Procured Lab Test Devices
- Partnerships with Cisco and NVIDIA Engineering Leadership
- Identified SMEs to Learn and Train on Design and Architecture Components

#### Al Cluster Fabrics Scalable Dedicated Environments



#### Non-Blocking Lossless Fabric High Performance Network That Doesn't Drop Packets



- Full bandwidth host communication
- Equal bandwidth from leaf to host and spine
- Congestion management to prevent packet loss

# Al Cluster Architecture and Design

cisco ive!

# Foundational Principles for our Al Clusters



Focus on application performance



Intelligent buffering, low latency, telemetry / visibility



Dynamic congestion avoidance for various workloads



Dedicated front-end / back-end networks. Non-blocking fabric



Leverage Automation for Day 0 to 2 operations



cisco /

## **Network** Zones



## Inband Network

- Clients to access the Al cluster and schedule jobs
- 5 Management UCS hosts
- Simple Linux Utility for Resource Management (SLURM)
- Layer 2 between the DGX and UCS Managers
- Inband MGMT switches connected to Cisco Network
- Cisco 8102-64H-O switch,100Gig ports





cisco ile

## Storage Network

- Client access to periodic data snapshots
- Netapp Storage, A900 series
- Layer 2 network, 400Gig Links between storage switches
- 100Gig links to Netapp and DGX
- Cisco 8101-32FH-O

Future – Build L3 underlay with VxLAN overlay





cisco / ile

#### Collapsed In-Band and Storage Network

#### New Design on Nexus

- Client and Data Access
- 8 Management UCS Hosts
- Netapp Storage, A900 series
- VXLan over BGP, 400Gig Links between Leaves and 4 Spines
- 100Gig links to Netapp, DGX and UCS
- Nexus N9K-C9364D-GX2A and Nexus N9K-C93600CD-GX
- Entirely built using NDFC







# Compute Network

Original Design

- High-speed data transfer between GPUs
- RAIL optimized design with L3 connections
- 2x400Gig MLAG Leaf to Spine connectivity
- iBGP with Route Reflectors
- 400Gig non-blocking bandwidth per GPU
- Cisco 8101-32FH-O
  - Then 8122-64EH-O
  - Then N9K-C9364D-GX2A
- QoS with ECN and PFC
  - Then Distributed Scheduled Fabric
  - Then ECN, PFC and DLB

cisco live!





# Cisco IT's Original AI Cluster

- First fabric built on Silicon One ASICs
- Utilizing SONiC as the NOS
- · Ethernet-based architecture
- · 400G non-blocking compute fabric
- · 32 Nvidia compute nodes (8 GPUs each)
- 12.8T throughput per Compute leaf
- · 32 PetaFLOPS AI performance per node
- 4 NetApp AFF A900 nodes



cisco / ile

# Cisco IT's Second Al Cluster

- Utilizing NX-OS as the NOS
- Our first VXLan DC Fabrics
- NDFC and NDI for Fabric Build and Management
- Ethernet-based architecture
- 400G non-blocking compute fabric
- · 64 Nvidia compute nodes (8 GPUs each)
- 51.2T throughput per Back-End leaf
- · 32 PetaFLOPS AI performance per node
- 4 NetApp AFF A900 nodes



# Original Infrastructure Build



# Day1 Implementation





Bootup and Install SONiC



Network Configuration



cisco live

## Data Center Cabinet Layout

- Utilized existing cabinets, power, and cooling
- 32 DGX Units distributed for high power and cooling (10.2 kW per DGX)
- Direct connections with long patch cords to save time and cost
- No major power or cooling upgrades needed



cisco il

# Data Center Cabinet Elevation Layout

#### **Network Row Optimization**

- Centralized network and storage hardware
- Reduced cabling and patching complexity
- Ensured failover and power redundancy



#### **Compute Row Optimization**

- Distributed DGX hardware (~10 kW per cabinet)
  Newer DCs optimized for 25kW (2 Nvidia) per cabinet
- Maintained N+1 power and cooling redundancy
- Reduced cost and deployment time



## **Optics for Original AI Cluster**







QDD-400G-DR4-S

QSFP-100G-SR4-S QSFP-100G-SM-SR

MMS4X00-NS-FLT

Vendor	Cisco	Cisco	Cisco	Nvidia
Speed	400G	100G	100G	800G Duplex 400G-DR4
Fibre Mode	Single Mode	Multimode	Single Mode	Single Mode
Max Cable Distance	500m	70m	2000m	100m
Fibre Connector Type	MTP/MPO-12	MTP/MPO-12	Duplex LC	Duplex MPO-12
Use Case	Compute Fabric (Leafs/Spines)	Storage and Inband Mgmt	DCC	Compute Fabric

cisco iver

## Second Al Cluster Added







QDD-400G-FR4-S QSF

QSFP-100G-DR-S

QSFP-100G-FR-S

Vendor	Cisco	Cisco	Cisco
Speed	400G	100G	100G
Fibre Mode	Single Mode	Single Mode	Single Mode
Max Cable Distance	2000m	500m	2000m
Fibre Connector Type	Duplex LC	Duplex LC	Duplex LC
Use Case	Leaf to Spine links	UCS Mgmt Nodes	Nvidia Storage and In-Band Links

cisco ive!

# Cisco Network Gear for Original Al Cluster

	8101-32FH-O	8102-64H-O	93108TC-FX3H
ASIC	Silicon One	Silicon One	Cloud Scale
Rack Units	1RU	2RU	1RU
NOS	SONIC	SONIC	NX-OS
Ports	32xQDD	64xQSFP28	48x10GBASE-T
Total Throughput	12.8 Tbps	6.4 Tbps	1.8 Tbps
Use Case	Compute Fabric (leafs and spines) Storage Network	In-band Connectivity	OOB Mgmt
Port Speed	Four Hundred (FH)- 400Gb/s	Hundred(H)- 100 Gb/s	1/10Gb/s

cisco live!

# Cisco Network Gear for Second Al Cluster

20 mm	International Address	letter der		-	10000	and the second	terms in		and served	in second	1000	1000
			-									
do												
Share Trees												
	_		_					_				_
2005-31-01	THE OTHER	CHE C	= =	1 (1111)	(CITES	<u>1000</u>		988 (B			1111 C	
	and the second								1.5			

NOK-CO36/D-CY2A



NOK-CO3600CD-CX



NOK-CO3108TC-EX3H

	NOR OSOULD UNLA	Non OSCOUDD an	
ASIC	Cloud Scale	Cloud Scale	Cloud Scale
Rack Units	2RU	1RU	1RU
NOS	NX-OS	NX-OS	NX-OS
Ports	64xQSFP-DD	28xQSFP28, 8xQSFP-DD	48x10GBASE-T
Total Throughput	25.6 Tbps	12 Tbps	1.8 Tbps
Use Case	Back-End Fabric (leafs and spines) Front-End Spines	Front-End Leaves OOB Mgmt GWs	OOB Mgmt SWs
Port Speed	400Gb/s	100 Gb/s	1/10Gb/s

cisco live

# Configuring the SONiC Network

cisco live!

# Cisco 8000 Setup

- Bootup
  - o Transfer SONiC image
  - o Sonic Installer
  - o Reload
- Upgrade
  - o Backup Files
  - o Sonic installer
  - o Reload



cisco,

## Network Configuration

• Native sonic cli

#### Modify the config\_db JSON File

dmin@svlngen4-sonic-8k-1:~\$ sudo config interface ip add Ethernet16 172.20.12.1/30 admin@svlngen4-sonic-8k-1:~\$ sudo config interface ip add Ethernet24 192.168.20.1/24					admin@svlngen4-sonic-8k-1:/etc/sonic\$ pwd /etc/sonic admin@svlngen4-sonic-8k-1:/etc/sonic\$ ls   grep config_db.json config_db.json
admin@svlngen4-sonic-8k admin@svlngen4-sonic-8k admin@svlngen4-sonic-8k	-1:~\$ sudo config save -1:~\$ -1:~\$ show in interfac	e -y			<pre>admin@svlngen4-sonic-8k-1:/etc/sonic\$ grep -A 5 INTERFACE config_db.json "INTERFACE": {     "Ethernet16": {},     "Ethernet16 172.20.12.1/30": {},     "Ethernet24": {},     "Ethernet24 192.168.20.1/24": {} }.</pre>
Interface Master	IPv4 address/mask	Admin/Oper	BGP Neighbor	Neighbor IP	
					admin@svlngen4-sonic-8k-1:~\$ sonic-cfggen -j /etc/sonic/config_db.json
Ethernet16 Ethernet24	172.20.12.1/30 192.168.20.1/24	up/up up/up	8101–2 8101–2–vlan20	('default', '172.20.12.2') ('default', '192.168.20.2')	admin@svingen4-sonic-8k-1:~\$ admin@svlngen4-sonic-8k-1:~\$ sudo config save -y Running command: /usr/local/bin/sonic-cfggen -dprint-data > /etc/sonic/config_db.json admin@svlngen4-sonic-8k-1:~\$

cisco lite

## Network Configuration

sudo config portchannel add PortChannel0000 --min-links 2 sudo config portchannel member add PortChannel0000 Ethernet0 sudo config portchannel member add PortChannel0000 Ethernet8 sudo config portchannel add PortChannel0016 --min-links 2 sudo config portchannel member add PortChannel0016 Ethernet16 sudo config portchannel member add PortChannel0016 Ethernet24 sudo config portchannel add PortChannel0032 ---min-links 2 sudo config portchannel member add PortChannel0032 Ethernet32 sudo config portchannel member ; INTERFACE.json

2

3

4

5

6

7

8

9

10

11

12

14 15

16

sudo config portchannel add Por sudo config portchannel member a sudo config portchannel member ; sudo config portchannel add Por sudo config portchannel member a sudo config portchannel member ; sudo config portchannel add Por sudo config portchannel member a sudo config portchannel member ; sudo config portchannel add Por sudo config portchannel member a 13 sudo config portchannel member ; sudo config portchannel add Por<sup>.</sup> sudo config portchannel member a 17 sudo config portchannel member a 18

	T.4
	15
'INTERFACE": {	16
"PortChannel0000": {},	17
"PortChannel0000 10.2.20.2/30": {},	18
<pre>"PortChannel0016": {},</pre>	19 20
"PortChannel0016 10.2.20.130/30": {},	21
<pre>"PortChannel0032": {},</pre>	22
"PortChannel0032 10.2.21.2/30": {},	23
<pre>"PortChannel0048": {},</pre>	24
"PortChannel0048 10.2.21.130/30": {},	26
<pre>"PortChannel0064": {},</pre>	27
"PortChannel0064 10.2.22.2/30": {},	28
<pre>"PortChannel0080": {},</pre>	30
"PortChannel0080 10.2.22.130/30": {},	31
<pre>"PortChannel0096": {},</pre>	32
"PortChannel0096 10.2.23.2/30": {},	
<pre>"PortChannel0112": {},</pre>	
"PortChannel0112 10.2.23.130/30": {},	

#### BGP\_NEIGHBOR.json

1

3

Δ

5

6

7

8

9

10

11

12

13

14

```
"BGP_NEIGHBOR": {
    "default|10.2.20.33": {
        "admin_status": "up",
        "asn": "65200".
        "holdtime": "30000"
        "keepalive": "10000".
        "local_addr": "10.2.20.34",
        "name": "mtv5-ag10-sfab1-sw3901:PortChannel0064".
        "nhopself": "0",
        "rrclient": "0"
   },
    "default|10.2.20.161": {
        "admin status": "up",
        "asn": "65200",
        "holdtime": "30000",
        "keepalive": "10000"
        "local_addr": "10.2.20.162",
        "name": "mtv5-ag11-sfab1-sw3902:PortChannel0064",
       "nhopself": "0",
        "rrclient": "0"
    }.
    "default|10.2.21.33": {
        "admin_status": "up",
        "asn": "65200",
        "holdtime": "30000".
        "keepalive": "10000",
       "local addr": "10.2.21.34",
        "name": "mtv5-aq13-sfab1-sw3903:PortChannel0064",
        "nhopself": "0",
        "rrclient": "0"
```

## **QoS** Configuration

PORT QOS MAP.json 1 1 2 2 "PORT QOS MAP": { 3 3 "Ethernet0": { "dscp\_to\_tc\_map": "CISCO", 4 4 "pfc\_enable": "3,4", 5 5 "pfc\_to\_queue\_map": "CISCO", 6 6 7 "pfcwd\_sw\_enable": "3,4", 7 8 "tc\_to\_pg\_map": "CISCO", 9 "tc\_to\_queue\_map": "CISCO" TC\_TO\_QUEUE\_MAP.json 10 }, 11 "Ethernet8": { 1 12 "dscp\_to\_tc\_map": "CISCO", "TC\_TO\_QUEUE\_MAP": { 2 13 "pfc\_enable": "3,4", "CISCO": { 3 "pfc\_to\_queue\_map": "CISCO", 14 4 "0": "0". 15 "pfcwd\_sw\_enable": "3,4", "1": "1". 5 16 "tc\_to\_pg\_map": "CISCO", "2": "2". 6 "tc\_to\_gueue\_map": "CISCO" 17 "3": "3". 7 18 }, "4": "4". 8 19 "Ethernet16": { "5": "5". "dscp\_to\_tc\_map": "CISCO", 9 20 21 "pfc\_enable": "3,4", "6": "6". 10 22 "pfc\_to\_queue\_map": "CISCO", "7": "7" 11 "pfcwd\_sw\_enable": "3,4", 23 12 24 "tc\_to\_pg\_map": "CISCO", 13 25 "tc\_to\_queue\_map": "CISCO" 14 26 },

WRED\_PROFILE.json



cisco / ila
#### **Generating Configurations**



## Configuring the Nexus Network

cisco ive!

#### Nexus 9300 - NDFC Fabric Setup

- Fabric Setup
  - 1. Run Fabric Creation Wizard
  - 2. Select Fabric Template
  - 3. Define Fabric Parameters

Fabric Name rtp5-prod-al.fe.yxlan				1	Create Fabric		
Pick Fabric				- I			
Data Center VXLAN EVPN >	Fabric Name rtp5-prod-ai_fe_vxlan					$\frown$	Select Type of Fabric ×
General Parameters Replication	Pick Fabric	Fabric Name	5		1	<u> </u>	Q Search Type of Fabric
BGP ASN*	Data Center VXLAN EVPN >	rtp5-prod-ai_fe_vxlan					
03303	General Parameters Re	Pick Fabric					Data Center VXLAN EVPN Fabric for a VXLAN EVPN deployment with Nexus 9000 and 3000 switches.
Enable IPv6 Underlay	VPC Deer Link VI AN Dance	Data Center VXLAN EVPN >					Enhanced Classic LAN
Enable IPv6 Link-Local Address	3600	General Parameters Replication vPC Protocols Sec	urity Advanced Resources Manageability Bootstrap Configu	ration Backup Flow Monitor			Fabric for a fully automated 3-tier Classic LAN deployment with Nexus 9000, 7000 and 3000 switches.
Fabric Interface Numbering*	Make vPC Peer Link VLAN	Manual Underlay IP Address Allocation	Checking this will disable Dynamic Underlay IP Address Allocations				Campus VXLAN EVPN Fabric for a VXLAN EVPN Campus deployment with Catalyst 9000 switches and Nexus 9000 switches.
Underlay Subnet IP Mask* 30	Ioopback	Underlay Routing Loopback IP Range* 10.2.0.0/22	Typically Loopback0 IP Address Range				BCP Fabric Fabric for an eBOP based deployment with Naxus 9000 and 3000 switches. Ontionally VXI AN EVPN can be enabled on top of the eBGP underlay.
Underlay Subnet IPv6 Mask Select an Option	vPC Auto Recovery Time () 360	Underlay VTEP Loopback IP Range* 10.3.0.0/22	Typically Loopback1 IP Address Range				Custom Network Fabric for flexible deployments with a mix of Nexus and Non-Nexus devices.
Underlay Routing Protocol* ospf	vPC Delay Restore Time (Ir 150	Underlay RP Loopback IP Range	Anycast or Phantom RP IP Address Range				Fabric Group Domain that can contain Enhanced Classic LAN, Classic LAN, IPFM, Classic
Route-Reflectors*	vPC Peer Link Port Channe 500	Underlay Subnet IP Range* 10.4.0.0/16	Address range to assign Numbered and Peer Link SVI IPs				urma, and external connectivity memory addics.
2020.0000.00aa	vPC IPv6 ND Synchronize	Underlay MPLS Loopback IP Range	Used for VXLAN to MPLS SB/LDP Handoff				
	vPC advertise-pip	Linderlay Routing Loophack IPv6 Range					
	vPC advertise-pip on Bord	Cindenay Routing Loopback in volkange	Typically Loopback0 IPv6 Address Range				
L		Underlay VTEP Loopback IPv6 Range					
	L		Typically Loopback1 and Anycast Loopback IPv6 Address Range				
		Underlay Subnet IPv6 Range					
	ا ر		IPv6 Address range to assign Numbered and Peer Link SVI IPs		J		

cisco Nexus Dashboard

≡ Fabric Controller

Overview
 Manage

lo Admin

🐵 Fabric Controlle

Manage > Fabrics

alle01-oc

Eabric Technoli

Multi-Fabric Domai

Fabric Type

Fabric Grou

ASN

NA

Fabrics

Fabrics

0 1

Fabric Healt

💙 Major

## Nexus 9300 Setup

- POAP and Configure Network •
  - 1. PreProvision Switches

Host N rtp5-6 rtp5rtp5rtp5-

rtp5rtp5-

rtp5rtp5-c rtp5-

rtp5rtp5-

- 2. Import with POAP
- 3. Deploy Network Configs

oy Configuration	- rtp8-prod-ai, fe , sxi	4.T						-
			Carefig Preview		Comp (	) And and a second s	3	
The lay of Figure .								Trees
which Harter	P Albert	Aire .	Serial Humber	Faliniu Ballus	Pending Carilia	Dates Description	Progress.	Respec Darkak
visia	HLIT CO	anter	PERSONAL INFORMATION	()****	# Gran	in spec		T Baryot
and the second s		toria are	ristenterw	(1110)	0.Linux	si-Spini		Respec
ri <b>e</b> whith		law de sanat	POSSEL AND V		Alives	$\leftarrow r_{0}m$		Bergere
	N2.44	tender spiller	1000000004	(1100	8 Silver	to Spec		Benere
-top-dati-	NUM	<i></i>	constants.	(1000	8.Lowery	$2-\Sigma_{\rm F} m_{\rm f}$		Beiges
- Carlor - Carlor	801 <b>0</b>	940 C	/000810816)	(1110)	#1.0mm	i-Spec	-	Respire
wheel-	sir 😋 i	900 C	PORTAL MADE	-	#1.0mm	e que		Auges
10 VIII		~	+incompany	(111)	@ Lines	a-tym		Breght
the which	12.41	-	*2420433043AA		Blines	e-Tper	-	Benymo
- Carlor			Concession and	-	These states	a disa	-	

ame	IP Address	Role		serialNu	mber	Model		Softwar	e
fefab1-sw	1042 10.115.	leaf		FDO2833	034E	N9K-C93600	CD-GX	10.3(6)	
-fefab1-sw	1041 10.115.	) leaf		FDO2833	037Q	N9K-C93600	CD-GX	10.3(6)	
-fefab1-sw	1032 10.115.	leaf		FDO2833	036J	N9K-C93600	CD-GX	10.3(6)	
-fefab1-sw	1072 10.115.	leaf		FDO2833	031Z	N9K-C93600	CD-GX	10.3(6)	
-fefab1-sw	3822 10.115.	bord	er	FDO2833	033E	N9K-C93600	CD-GX	10.3(6)	
-fefab1-sw	3904 10.115	bord	er spine	FDO2835	07P1	N9K-C93641	D-GX2A	10.3(6)	
-fefab1-sw	1031 10.115.	leaf		FDO2833	OVYS	N9K-C93600	CD-GX	10.3(6)	
fefab1-sw	1071 10.115.	leaf		FDO2830	OFZF	N9K-C93600	CD-GX	10.3(6)	
-fefab1-sw	3821 10.115.	bord	er	FDO2833	0313	N9K-C93600	CD-GX	10.3(6)	
-fefab1-sw	3903 10.115	bord	er spine	FDO2835	07NX	N9K-C9364	D-GX2A	10.3(6)	
-fefab1-sw	1022 10.115.	leaf		FDO2833	0335	N9K-C93600	CD-GX	10.3(6)	
-fefab1-sw	1062 10.115.	leaf		FDO2833	031M	N9K-C93600	CD-GX	10.3(6)	_
Piter by	Switches Links Interfac	es Interface	Groups Pela	Seciel Rendered	s VRFs 3	Confin Security	Event Analytic	s History Reso	Notes Metrics
	non	P Address	PLOTE	Gener Humber	-	Components	oper status	Ciscovery status	mootel
<b>S</b> f	6 fefant-sw1042	10.115	Leaf	FD02833034E	Normal	0 In-Sync	O Minor	0	NBK-CB3600CD-SK
)-f 🗆 👐	5 fefailt-sw1041	10,115	Leaf	FD028330370	Normal	ti-Syst	Theatthy		NBK-C93600CD-6X
	5 felab1-sw1032	10.115	Leaf	FD02833036J	Normal	Di-Sym	🗢 Minor		NEK-C93600CD-OX
- rtp1	5- fetab1-aw1072	10.115.	Leaf	FD02833031Z	(Normal)	🔵 in-Sync	🗢 Minor		N9K-C93600CD-GX
rtpl	5- felab1-aw3822	10.115.	Border	FD02833033E	Normal	in-dyne	🗢 Major	<b>O</b>	N9K-C93600CD-GX
i ripi	5 felsb1-ew3904	10.115.000	Border Spine	FD0283507P1	Normal	in-Syre	O Minor		NDK-C9364D-GX2A
i ripi	5- fefab1-aw1031	10.115	Leaf	FDQ28330VVS	Normal	in-Syre	V Healthy		NSK-C93600CD-GX
inter	5 fefabl-aw1071	10.115.	Leaf	FDC28300FZF	Normal	in-Syre	O Minor		NSK-C93500CD-GX
- rtpl	5 fefab1-sw3821	10,115.	Border	FD028330313	Normal	🔵 lin-Oyrai	🗢 Major	03	NSK-C93600CD-0X
in the	5 fefabl-sw3903	10.115	Border Spine	FDC283507NK	Normal	in-Syre	V Healthy		NSK-C9364D-GX2A
- rtp1	5- fofab1-sw1022	10.115	Leaf	FDC28330335	Normal	in-Syre	Thealthy		NBK-C03600CD-GX
- rtpl	5 fofab1-sw1082	10.115	Leaf	FD02833031M	Normal	in-dync	🗢 Major		NBK-C93600CD-GX
I ripi	5 tefab1-sw3812	10,115.	Border	FD028330338	Normal	() in dyna	Theatly		NSK-C93600CD-0X

#### Interface and QoS setup

- Define VPC Pairs
- Define Interface configurations (Interface Wizard)
- Define QoS Policy (standard templates)
- Deploy policy



cisco

## Why Ethernet?

cisco live!

#### Ethernet vs Infiniband



Key Takeaways

- 1. Ethernet trumps Infiniband due to its widespread use and cost advantages
- 2. Its flexibility allows much needed convergence in mixed use DCs
- 3. Infiniband will continue to prevail in some HPC and AI environments (Nvidia appliances)
- 4. Ultra Ethernet forum to deliver improvements to adapt Ethernet to Al

Note: Infiniband is not proprietary; open spec by IBTA

cisco /

#### Al Network Market Trends

cisco /

#### Al Back-End Networks - Ethernet Switch Port Growth

Worldwide Al Back-End Networks - Ethernet Switch Port Growth by Speed Source: Dell'Oro Group December 2023 - Al Networks for Al Workloads



NOTE: The graph above shows percentage of 2028 total ports.





cisco live!

## What is SONiC?



Democratize software components



cisco live

#### SONIC – The functional stack SONIC Feature Development



## SONiC - Feature Set

cisco ive!

SONiC Deployable - Technology Components

Management	Interfaces	Layer 3	Infrastructure
Telemetry (gRPC, gNMI) SNMP	L2 & L3 LAG	IPv4/IPv6 BGP	SWSS Unit Test Framework ConfigDB framework
Systog LLDP NTP TACACS+ WRED	SVI MTU setting Layer 2	BGP knobs (community, as-path- replace/relax, add-path, route aggregate, prefix- lists) FCMP	Overlay Static VxLAN Security
QoS PFC / ECN Everflow IPinIP ZTP	MAC / ARP Aging	BGP GR helper BGP MP ACL - IPv4 & IPv6 COPP RPL (route-policies) VRF	MACsec (8808) High Availability Warm Boot Fast Reload
Use Cases DC Fabric	Architectures		Consumers
DCI (L3/IP)	Overlay Fabric - VxLAN	Hyperscalers	Service Enterpris
CDN Fabric WAN LER/LSR*	Overlay - EVPN VxLAN* Core - MPLS/SR/SRv6*		Shippin

Under\* Development

## Congestion Management and Forwarding

cisco live!

## Explicit Congestion Notification (ECN)

- Purpose: Provides congestion notification in IP networks
- Function: Enables end-to-end congestion notification between two endpoints
- Mechanism:
  - Utilizes 2 LSB of the Type of Service field in the IP header
  - Upon congestion, it triggers the transmitting device to reduce the transmission rate using a Congestion Notification Packet (CNP) without halting traffic



ECN	ECN Behavior
00	Non ECN Capable
10	ECN Capable Transport (0)
01	ECN Capable Transport (1)
11	Congestion Encountered

## Priority Flow Control

#### Flow Control Mechanism - 802.1Qbb

- Also Known As: "Lossless Ethernet"
- Function: Enables Flow Control on a Per-Priority basis
- Alternate Name: Per-Priority-Pause
- · Key Benefits:
  - Allows both lossless and lossy priorities on the same wire.
  - Ensures traffic can operate over a lossless priority independently.
  - Other traffic on different priorities continues to transmit, relying on upper layer protocols for retransmission.





## **PFC Mechanism**

- **Thresholds**: Set in no-drop queue with headroom for "in-flight" packets.
- **Buffering**: Traffic buffered in no-drop queue during congestion.
- **PFC Frames**: Sent to sender when queue utilization exceeds xoff threshold.
- Queue Management: Stops sending PFC frames when utilization drops below xon threshold.



#### Fabric Forwarding Operation with ECMP



cisco live!

## Day 2 Ops: SONiC

cisco live!



#### SONiC Day 2 Ops – Grafana Dashboards



cisco/ill

#### SONiC Day 2 Ops - Operational Runbooks



cisco / ile

#### SONiC Day 2 Ops - Monitoring and Backup

#### • Zabbix

← → C	수 은 뿐 단
🗅 Cloud-ACI 🔯 DCH-SDX-Micros. 🖸 Anable 🔯 APICs 🔯 Apple 🔯 Ciscolin/Cancun 🖸 Python 🔯 Facul 🖓 Tetration 😍 111. Defining Func. 🤤 About   Benline 🖈 Boolmark Manager	🖀 Free Online XML F 🔶 Git - Book 🛛 🔅 🗅 Al B
Q Search or jump to (2) cnd+k	O
🚍 Home > Dashboards > EVENTS DETAIL 🕁 🥰	🕑 Last 24 hours 👻 🔾
فرطان بصبيها والألا بصاليهما وبالبرجا وبتالي منتهد وبصبي والمائع	
0 17200 18200 19200 20200 2100 22200 2300 00:00 01:00 02:00 03:00 04:00 05:00 05:00 05:00 08:00 09:00 10:00 11:00 12:00 13:00 14:00 15:00 18:00	
ENDPOINT LATENCY TREND	1
	Name
1ms	- mtv5-ag11-sfab1-sw1008.cisco.com - ZABB0C_WEST
900 με	i i i i i i i i i i i i i i i i i i i
800 µs	
700 µs	i i i i i i i i i i i i i i i i i i i
800 µ2	
400M	
05/28 18:00 05/28 20:00 05/28 22:00 05/29 00:00 05/29 02:00 05/29 04:00 05/29 06:00 05/29 08:00 05/29 10:00 05/29 14:00 05/29 16:00	
ENDPOINT SSH AVAILABILITY TREND	
2	Name
18	<ul> <li>mtv5-ag11-sfab1-sw1006.cisco.com - ZABBIX_WEST</li> </ul>
15	i i i i i i i i i i i i i i i i i i i
14	
12	
1	
0.8	
0.6	

#### • Bitbucket

←	→ C ( gitscm.cisco.com/projects/SONIC/repos	/sonic-backups/brov	wse/mtv5-ag14-s	fab1-sw3904-	econ
	Training 🗎 Cloud 🗎 Infra 🗎 Upgrade 🗎 Training	ACI-Everywhere	🗎 X-Functional	🗎 Diagrams	
	Bitbucket Your work Projects Repositories -				
0	sonic / sonic-backups				
(±)	Source				
۱ţ	থি master ৺ ··· sonic-backups / mtv5-ag14-sfab1-s	w3904-econ /			
-C\$	Source	Description			
0					
_	mtv5-ag14-sfab1-sw3904-econ-config_db_v1.json	Backup Complete			
$\diamond$	twtv5-ag14-sfab1-sw3904-econ-config_db_v2.json	Backup Complete			
¢	mtv5-ag14-sfab1-sw3904-econ-config_db_v3.json	Backup Complete			

cisco ile

## Day 2 Ops: Nexus

cisco live!

#### Nexus Day 2 Ops – NDFC Inventory and Events

cisco Nexus Dashboard	🛞 Febric Controller -						0
E Fabric Controller * Overview Manage O, Analyze	Manage > Patric Software Fabric Software Learn More Overview Images Ima	age Policies Devices I	listory				Refer
3 <sub>0</sub> Adren	Inages	NGS Integes		Pakies	<b>()</b> • NEXE	C	• 052.mod4
	Fabrio	Current Version(6)	Policy	Status	Fabric Type	Switches	0
	aw01-cob	9.3(10)	None	None Prepare	Classic LAN	4	
	aar01-prod	7.9(12)N1(1) and 2 more	None	C Name Prepare	Classic LAN	16	
	alin01-oob	7.3(12)N1(1)	None	C None Prepare	Classic LAN	12	
	43 items found						Rowsperpage 50



E Fabric Controller	Nonage 2 Inventory Inventory									
e Heren	-									
a manage	Switches Interfaces									
C Analyze										
le Admin	1 Biordey additions									Autors -
	Switz	P Address	Relation	Serial Number	Faderic Hume	Mode	Config Status	Oper-Status	Noorvery Status	man i
	C CLAR BE PROPERTIES	N3 154.	Last	PD01834005V	CL-AR-FROM: VIE AN- FUTN	(Terrar	( the last	😨 Minar	•	NEW COMMO CATA
		10.186	Level	FUCUED400UX	CLAN Fabric VIILAN- LINN		() to ly m	C Marrie	œ	NIN CEDERD GREA
		NY 1996.	Last	PD01834007U	C-APPEND VEAN-	<b></b>	( the last	() Institut	•	NR CENHO CATA
		10.194	Leef	FOCOBD400TC	CLAN Fabric-VILAN DIPN		(Teller)	(† 1000)	•	NEW-CEOBAD-GXGA
	CL-N-66-FM2RC-SPINET	10.108	Spine	F001834007X	CL-M-Fabric-VILAM- FaPN		(The last	()	•	NIN-CE1840-CX28
	CLANE MARCHINES	NO 1886	Spine	10038030782	CLAFEDRO-VILAN- DIPN	••••	(The last	(9 martin)	œ	NOH-CEOBAD-GIGR

Fabric Controller	Switches Interfaces								
Manage									
Anatyze	Device Nerse contains be $\times$							ER CHARA	tions -
Adrein	Fabric Name	Device Name	Interface	Admin Status	Oper. Status	Reason	Policies	Overlay Network	Sync
	CL-AI-Fabric-VXLAN-EVPN	CL-AI-BE-FABRIC- LEAF01	mgmt0	1 Up	1 Up	ok	int_mgmt	NA	
	CL-AI-Fabric-VXLAN-EVPN	CL-AI-BE-FABRIC- LEAF01	Vian1	🕹 Down	🕹 Down	Administratively down	NA	NA	@ N
	CL-AI-Fabric-VXLAN-EVPN	CL/AI-BE-FABRIC- LEAF01	Vien2000	1 Up	1 VP	ok	NA.	MyV88_50000	<b>8</b> N
	CL-AI-Fabric-VXLAN-EVPN	CL-AI-BE-FABRIC- LEAF01	Loopback0	1 Up	T UP	ok	int_fabric_loopback_11_1	NA	
	CL-AI-Fabric-VXLAN-EVPN	CL-AI-BE-FABRIC- LEAF01	Loopback1	1 Up	1 Up	ok	int_fabric_loopback_11_1	NA	
	CL-AI-Fabric-VXLAN-EVPN	CL-AI-BE-FABRIC- LEAF01	Ethernet1/1	1 Up	1 Up	ok	int_trunk_host	NA	
	CL-AI-Fabric-VXLAN-EVPN	CL-AI-BE-FABRIC- LEAF01	Ethernet1/2	1 Up	1 up	ok	int_trank_hest	NA	

Feb	rie O	verview - Ci	L-Al-Fabric-VI	LAN-EYPN						8,3061 -
Ove	nie	Switches	Links Into	rlaces interface (	droups Policies Networks VRFs	Services Socurit	y Event Analytics His	tory Researces	Matrica	
-	larre	Gleared	Alarma Even	ts Recent Tasks						
1	File	by attributes								
	0	•	Severity	Source	Hame	Category	Creedion Time	Policy	Messge	AckUser
	0	140410	- Minis	10.100	CL-N-BD-FNBRIC-LEAFEI	INTERFACE STATUS	Jun 24 2025, 10:14:10	discovery	Interface environment/Mit in uncleasing state up observatives not connected	
		140400	MICH	10.898.	CL-N-80-MBNIC-LEARCE	INTERVACE STATUS	Jun 28 2025, 1014-04	discovery.	Interface attempt (VE in undesired state up down Unit not connected	
	-	100770		CL-40-F187G	Patric (Patric, Newpare-Satricest Co-	CVTTERIAL	Inc. 24 2005 10-54-54	be fable error	VTP Like Subnet IP Hange duplicate with Sparic:	

cisco ile

#### Nexus Day 2 Ops – NDI Performance Monitor



5-	-befab	2-sw390	5												-	
ten Ca	mactivity .	formation (	Novinceine .													
infaces.	L3 Helghbo	rs Endpoint	a d'C Dana	ice Molice	n flouin											
i in																
increasily i	anal	-	inis liates	Operational	Station.	Fype										
C	5	Ter 1	i mant	(† 1945) († 1945)		Read H										
	•															
in the second	Amongly Land	- Derectional Taxeet	-													
				Beighburn	Links.	Dates	Repairs Obligation	Receive Byles	Paceiro Packata	Receive Natio	Instantic Spins	Instanti Padeda	Farmed Rela	Instant Official Inst	t	0
	(Const	ett Dav	Dene	Augustan International International International	inter The	Tana (Tan)	Receive Unitative	Territor Biglion	Pacoles Packets	Receive Rate 28.12 Maps	Laterated Radios	Transmit Padhole	Faccord Role 28.1875ger	Tenend Writedow	1111-1	0
	(1111) (1111)	ett Class	Prysinal Physical	Applaam Applaam Infatte Infatte Infatte Infatte Infatte Infatte	1000 (100) (100)	(7.96) (7.96)	100000 (Misailan 1	Nacira Byles 14678 18677 M	Hannine Flackata 2004-0142 9109-0142	Nooine Nam 28-12 Maps 24-36 Maps	Transmit Ratur 1.28 TB 1.85 TS	Transmit Producter 2014/02/08/1 Transmit Transmit	Saranai Balo 28.1876pe 21.16785pe	Tenand Etherine	11114 11144	0
	(111) (111)	#11 Days #11 Days #11 Days	Protect Protect	Register Register Indat- Indat		(7.96) (7.96) (7.96) (7.96)	Rocie: (Master 6 6 6	Noore Bytes 14278 0001140 12678	Hooine Packets 2004-01-02 500-000 2004-000	Record Halls 2012 Maps 21.00 Maps 20.20 Maps	Lanned Bytes L20130 L20130 L20130	Lansenti Paskula 20400887 798815645 1010088	Lancent Hale 28-William 28-William		1000 1000 1	0
6110 6110		*** Cape *** Supe *** Cape	Presal Presal Presal	Applaten Begleten Islatz Islat	100 (190 (190 (190 (190	(7.96) (7.96) (7.96) (7.96) (7.96)	Rocie Maater	Norm Nor 1873 Nor 11 No 12678 1875	Nooise Italais 2004-112 2004-1	Киссин Кайл 28 12 Маря 29 28 Маря 29 28 Маря 29 28 Маря	lassed liples 128 TB 188 TS 488 IT CB 158 TS	Insense Pachole 20402087 Insenses Insenses Insenses	Exerced Bale 28 William 29 William 28 William 29 William	Toosed Dilatos 9 9 9 9 9 9	1000 1000 1 1 1 1	0
arias arias arias arias	(111) (111) (111) (111)	*** Days *** Days *** Days *** Days	Prysial Prysial Prysial	Appleader Appleader India I	100 (7.00) (7.00) (7.00) (7.00) (7.00)	(19) (29) (29) (29) (29) (29)	Rocie Madee	Harris Dela 14878 18878 1897 1897 1897 1897 1897	House Holds 2004 House 2004 House 2004 House 2004 House 2004 House 2004 House	Record Data 20.12 Maps 20.20 Maps 20.20 Maps 20.20 Maps 20.20 Maps	Insense Bytes 128 TB 108 TS 400 TO 108 TS 108 TS	Insense Parkon 204020887 Insenses Insenses 20802084 20802084	Execution 28 Ether 29 - Other 29 - Other 21 - Other 21 - Other 21 - Other	1 menet Minutes 4 4 5 6 6 7 8	111-1 111-1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	0







#### Nexus Day 2 Ops – NDI Advanced Capabilities

CECO	Contraction of the second s				
Coervee F Mariege	Angjar y Kangkan Ma <b>Sangjar Ma</b> Sang and Sandard series and de alarsa f anglas taking direct for yon ing an induce metger software and had it you network.				
Andree	Complexes (at the) Dataset your failers y pergelaxes with subject arrange	Conformance Keep track of your hardware and syffware life system	Policy CAM (AXEMy) Ventus par referred polices	Sourcestivity	
leokasta	Leg Colector Dates and and yet ingo from your enormation	Traffic Analytics Memory real services between programmer, and drops	D Suntainability Explore your fabrics energy usage, cord, and emissions	Conjuste configurations and differences is pair April 1 between two points in time	
	Pre-Change Add One Versi the prevent of configuration changes	Reg Scan Later stops and potential bugs affecting year returns			

St. Overview	View Connectivity Analysis Nov	ı				
,∲ Masage	processing sector	T				
O Analyza	face Device Internet	L				
la Adrin	1 0 0 6 and a Line_vides   🖗 2 all ( 🌒 6 C28/23 563.3.58	L				
Smarth and Explore	Lige filmandes ~	l				
		L				
	Kaagish Sylaw 🖞 🖉 RAW 🗋 RAW 🔄 Considered Declare					
	Centre Tree Del Tree Bar Tree Sie Sours NGC Dedenter NGC Sours VLIII Filer Type Are 37 M2N, SARATABER MR	l				
		L				
	🛓 Analysis General	l				
	• ITTE PROFILE VIEW INTO THE O	l				

0 10000 A 1000p	Advisories (					
C anagen Ag annen	all Addrefen () autorites -					
2 Barris and Explore	14 C				The Arrest Op	••
1 Instants	Aktive (and Campy ************************************					·
		Addressly Lovell	Callegray	faire .	R-01	
	C Statist Preservating	-	Best Facilities	ware on the	niget and the field of a set for fight ground at the poster free at the set	·
	Configure 40 with automations	-	Bell Talles		April Carlado and NY April 2010 April 2010 These all (2010)	•
	Configure 1989/14 consuming and apply Allia	-	Ber Parties	1014101-0.00.0-	nyi ang katala ang Milanda ja jaban Kanada ja jaban	•
	C The second sequel benear protocol for (2) access (201)	-	Bell Pallor	-	April - California and Arr Alfranci - A., Jan, Johns Trans all (14 April)	-
	Trails may present during	-	But Particular	101003-00.00.0-	nati 🌰 schiel auf 10 Adramatica (sa., chen Tana al (21 ana)	-
		-	Bol Failler		And a second second and a second second second and a second secon	- 1

C Overview	Create Pre-Change Ar	alysis	
Manage			
Analyze		Oeneral	
Admin	Catolica Catolica		
	2) Change Simulation	Pre-Change Analysis Name -	
Search and Explore		Enter Name	
Bookmarks	(a) furning	Description (Optional)	
		Enter Description	
		Fabric *	
		Bellect Fabric	
		Snapshot -	



## Use Cases



## A Sample of Current Usage

Team	Use Case	Business Outcomes
Webex Audio (Babblelabs)	Improving codec development for noise cancellation and lower bandwidth data prediction	Improved Webex Teams feature performance
Webex Video	The team trains video AI models for Cisco, including background replacement, gesture recognition, and face landmarks.	Improved Webex Teams feature performance
Cisco SBG	Specialized code generation LLMs to reason over Cisco Security devices; Custom pretrained LLMs built for the cybersecurity domain.	Increase Cisco's security posture trough AI-driven service & application log analysis
I2C & Revenue (Core Finance)	Order Collection Optimization, Payment predictors, Al Ops predictors, Payment fraud analysis	Drive a reduction in late- and non- paying clients through the use of Al
Observability (SNOW)	Avoid business impacting incidents caused from IT changes	>80% accuracy in high-risk change detection leading to corrective action to avoid incidents Future Vision: GAI workflows to add approval paths, and suggestions to reduce Change Request scopes

cisco live!

## Lessons Learned

cisco live!

#### Al presents challenges for IT teams



#### INFRASTRUCTURE DEMANDS

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

cisco / ile

## Key Takeaways

#### Power Efficiency:

Prioritize power-efficient solutions.

#### Scalable CLOS Fabric Design:

- Focus on target CLOS Fabric Design.
- Ensure a scalable two-tier design.

#### • Fixed Switches:

- Utilize fixed switches for optimal performance.
- Achieve lower latency with single ASIC designs.

#### NoS Choices:

- SONIC is targeted for Hyperscalers and SPs.
- Nexus is targeted for Enterprises.
- Congestion Management:
  - Implement effective QoS is very important.

#### BGP for Control Plane:

• Utilize BGP for an efficient control plane.

#### Simplicity and Reusability:

• Aim for simplicity and reusability in designs.

#### Planning and Expectations:

- Cross-Team Planning and Double-check the details.
- Set Clear expectations.

#### • Al Evolution:

- Rapid advancements in AI technology.
- Embrace a mindset of continuous learning.
- Iteration and Evolution:
  - Iterate and evolve with ongoing improvements.
  - Follow a Crawl/Walk/Run methodology for implementation.





## Webex App

#### Questions?

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

#### How

- 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.

cisco / illa



### 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'





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



## Thank you

cisco Live!



# GO BEYOND