

The background features a vibrant, abstract design with a color gradient from dark blue on the left to bright yellow and white on the right. The design consists of overlapping, wavy horizontal bands and a series of radiating lines that create a sense of motion and energy, resembling a stylized sunburst or a dynamic wave pattern.

cisco *Live!*

Let's go

#CiscoLiveAPJC



The bridge to possible

Building a High-Performance AI DC

Using Nexus, UCS, and NVIDIA GPUs

Andy Sholomon
Director of Technical Marketing
@asholomon
BRKAPP-2698

CISCO *Live!*

#CiscoLiveAPJC

Cisco Webex App

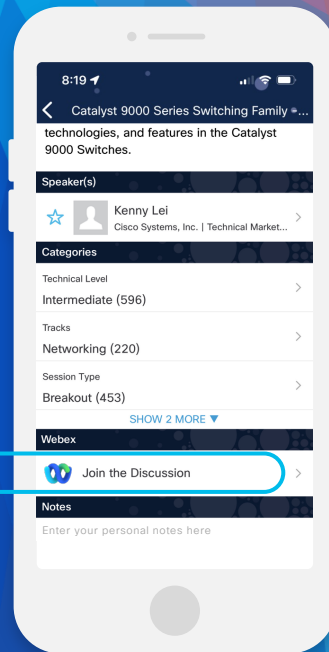
Questions?

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 December 22, 2023.

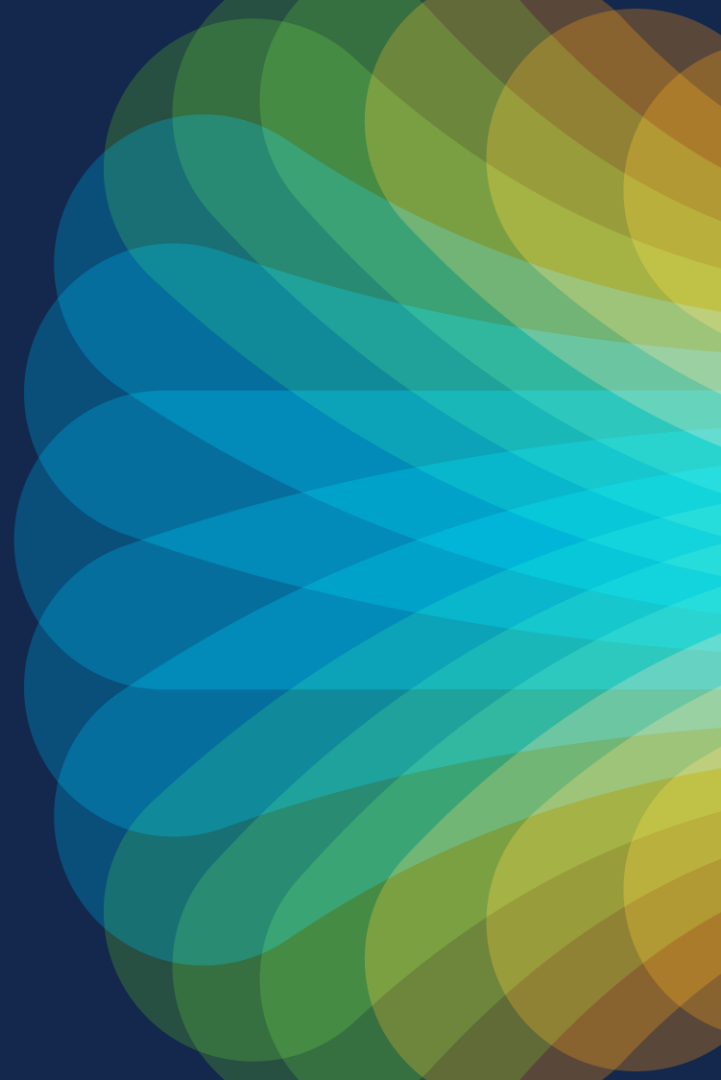


<https://ciscolive.ciscoevents.com/ciscolivebot/#BRKAPP-2698>

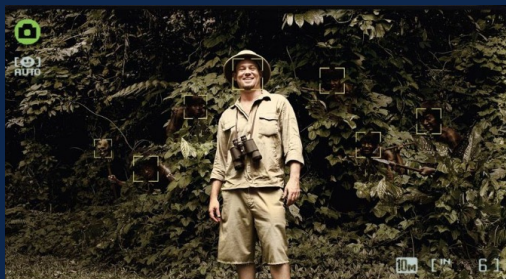
Agenda

- AI/ML Applications
- What Are Their Requirements?
- AI/ML Blueprint for DC Networks
- Industry Talk
- Conclusion

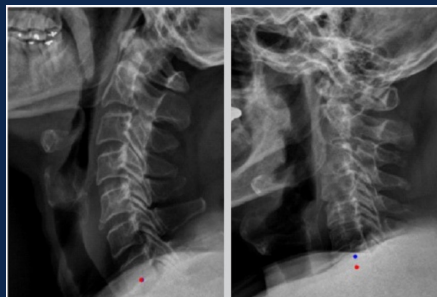
AI/ML Applications



AI is everywhere!



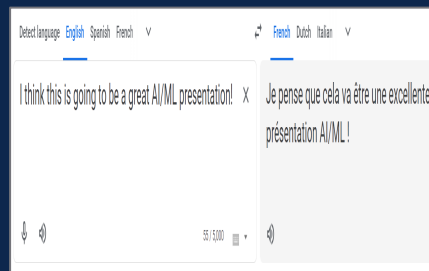
Face recognition



Analysis of medical images



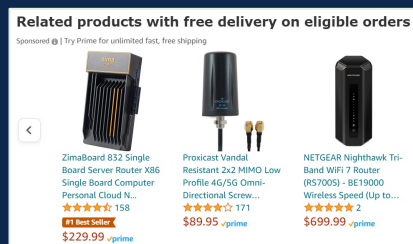
Self-driving vehicles



Machine translation



Conversational agents



Recommender systems



Image generation

Main Types of AI/ML Workloads

Training: Training refers to the process of using a machine learning algorithm to build a model. Training involves the use of a deep-learning framework and training dataset.

Inference: Inference refers to the process of using the trained machine dataset to make a prediction.

Dog or Muffin?



LLMs

LLMs stands for Large Language Models. These are machine learning models that have been trained on massive amounts of text data, such as books, articles, and web pages, to understand and generate human language. OpenAI's GPT-4 has 1.5 Trillion parameters.

There are now hundreds of open-source LLMs available. Currently, the most powerful is Llama-2, released by Meta in July 2023, with 70 billion parameters. Open-source LLMs have the advantage over proprietary models that they can be fine-tuned for task-specific goals.

What It Takes to Train an LLM

	GPT-3 large	LLaMa
Vocabulary size	50,257	32,000
Sequence length	2048	2048
Parameters in the largest model trained	175B	65B
Tokens in the training dataset	300B	1 - 3T
Number of GPUs	10,000 V100 GPUs	2048 A100GPUs
Training time	One month	21 days

Table 1 — Comparison of GPT-3 and LLaMA models.

With 8 GPUs per server it takes 1250 servers running for one month to train an LLM like GPT-3

The estimate is it would take ~6000 H100 GPUs or 750 servers today

256 servers for LLaMa

Power

DATASHEET



NVIDIA DGX A100

The universal system for AI infrastructure.



The Challenge of Scaling Enterprise AI

Every business needs to transform using artificial intelligence (AI), not only to survive, but to thrive in challenging times. However, the enterprise requires a platform for AI infrastructure that improves upon traditional approaches, which historically involved slow compute architectures that were siloed by analytics, training, and inference workloads. The old approach created complexity, drove up costs, constrained speed of scale, and was not ready for modern AI. Enterprises, developers, data scientists, and researchers need a new platform that unifies all AI workloads, simplifying infrastructure and accelerating ROI.

The Universal System for Every AI Workload

Part of the NVIDIA DGX™ platform, NVIDIA DGX A100 is the universal system for all AI workloads—from analytics to training to inference. DGX A100 sets a new bar for compute density, packing 5 petaFLOPS of AI performance into a 6U form factor, replacing legacy compute infrastructure with a single, unified system. DGX A100 also offers the unprecedented ability to deliver a fine-grained allocation of computing power, using the Multi-Instance GPU (MIG) capability in the NVIDIA A100 Tensor Core GPU. This enables administrators to assign resources that are right-sized for specific workloads.

Available with up to 640 gigabytes (GB) of total GPU memory, which increases performance in large-scale training jobs up to 3X and doubles the size of MIG instances, DGX A100 can tackle the largest and most complex jobs, along with the simplest and smallest. This combination of dense compute power and complete workload flexibility makes DGX A100 an ideal choice for both single and scaled DGX deployments. Every DGX system is powered by **NVIDIA Base Command™** for managing single node as well as large-scale Slurm or Kubernetes clusters, and includes the NVIDIA AI Enterprise suite of optimized AI software containers.

Unmatched Level of Support and Expertise

SYSTEM SPECIFICATIONS

NVIDIA DGX A100 640GB		
GPUs	8x NVIDIA A100 80GB Tensor Core GPUs	
GPU Memory	640GB total	
Performance	5 petaFLOPS AI 10 petaOPS INT8	
NVIDIA NVSwitches	6	
System Power Usage	6.5 kW max	
CPU	Dual AMD Rome 7742, 128 cores total, 2.25 GHz (base), 3.4 GHz (max boost)	
System Memory	2TB	
Networking	Up to 8x Single-Port NVIDIA ConnectX-7 200 Gb/s InfiniBand Up to 2x Dual-Port NVIDIA ConnectX-7 VPI 10/25/50/100/200 Gb/s Ethernet	Up to 8x Single-Port NVIDIA ConnectX-6 VPI 200 Gb/s InfiniBand Up to 2x Dual-Port NVIDIA ConnectX-6 VPI 10/25/50/100/200 Gb/s Ethernet
Storage	OS: 2x 1.92TB M.2 NVME drives Internal Storage: 30TB (8x 3.84 TB) U.2 NVMe drives	
Software	DGX OS / Ubuntu / Red Hat Enterprise Linux / Rocky – Operating System NVIDIA Base Command – Orchestration, scheduling, and cluster management NVIDIA AI Enterprise – Optimized AI software	
Support	Comes with 3-year business-standard hardware and software support	

System Power Usage 6.5 kW max

Power

Datasheet



NVIDIA DGX H100

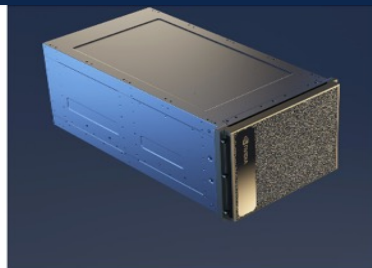
The gold standard for AI infrastructure.

Artificial intelligence has become the go-to approach for solving difficult business challenges. Whether improving customer service, optimizing supply chains, extracting business intelligence, or designing leading-edge products and services with generative AI and other transformer models, AI gives organizations across nearly every industry the mechanism to realize innovation. And as a pioneer in AI infrastructure, NVIDIA DGX™ provides the most powerful and complete AI platform for bringing these essential ideas to fruition.

NVIDIA DGX H100 powers business innovation and optimization. Part of the DGX platform and the latest iteration of NVIDIA's legendary DGX systems, DGX H100 is the AI powerhouse that's the foundation of NVIDIA DGX SuperPOD™, accelerated by the groundbreaking performance of the NVIDIA H100 Tensor Core GPU. The system is designed to maximize AI throughput, providing enterprises with a highly refined, systemized, and scalable platform to help them achieve breakthroughs in natural language processing, recommender systems, data analytics, and much more. Available on-premises and through a wide variety of access and deployment options, DGX H100 delivers the performance needed for enterprises to solve the biggest challenges with AI.

The Cornerstone of Your AI Center of Excellence

AI has bridged the gap between science and business. No longer the domain of experimentation, AI is used day in and day out by companies large and small to fuel their innovation and optimize



Specifications

GPU	8x NVIDIA H100 Tensor Core GPUs
GPU memory	640GB total
Performance	32 petaFLOPS FP8
NVIDIA® NVSwitch™	4x
System power usage	10.2kW max
CPU	Dual Intel® Xeon® Platinum 8480C Processors 112 Cores total, 2.00 GHz (Base), 3.80 GHz (Max Boost)
System memory	2TB
Networking	4x QSFP ports serving 8x single-port NVIDIA ConnectX-7 VPI ➤ Up to 400Gb/s InfiniBand/Ethernet 2x dual-port QSFP112 NVIDIA ConnectX-7 VPI ➤ Up to 400Gb/s InfiniBand/Ethernet
Management network	10Gb/s onboard NIC with RJ45 100Gb/s Ethernet NIC Host baseboard management controller (BMC) with RJ45
Storage	OS: 2x 1.92TB NVMe M.2
Internal storage:	8x 3.84TB NVMe U.2
Software	NVIDIA AI Enterprise – Optimized AI software NVIDIA Base Command – Orchestration, scheduling and infrastructure management

System power
usage

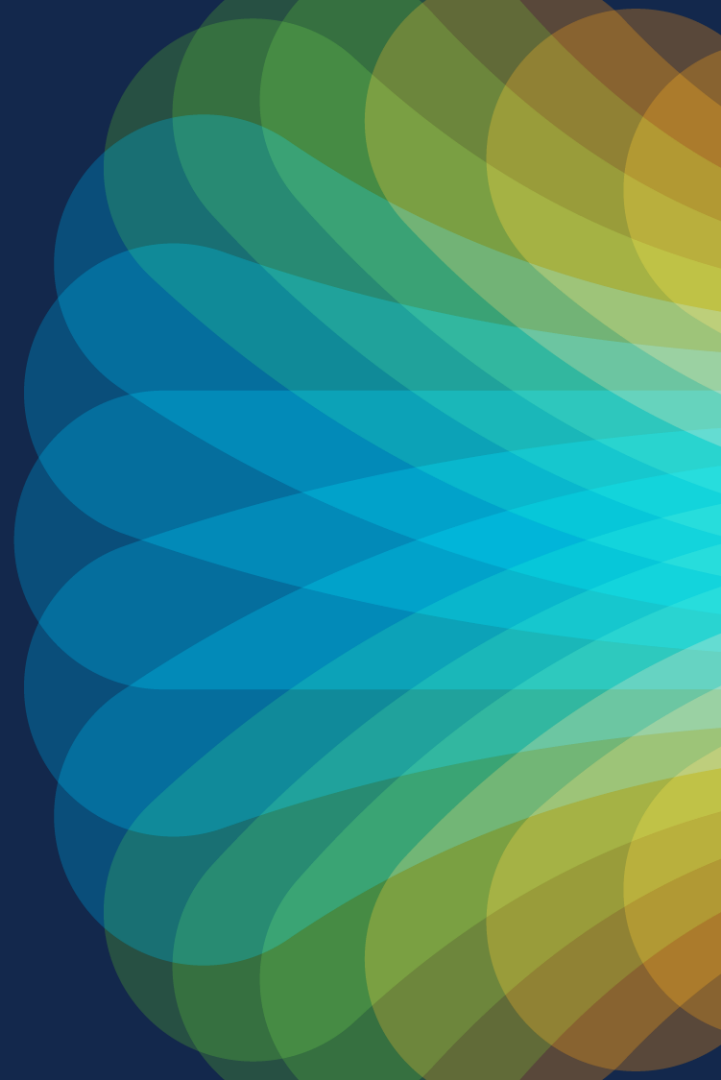
10.2kW max

LLMs

Feature/Model Parameters	GPT-4 1.5 Trillion	Bard 1.6 Trillion	LLaMA 1.2 Trillion	Flan-UL2 20 Billion	BLOOM 176 Billion
Training Data	WebText-like corpus	WebText-like corpus	WebText-like corpus	Publicly available data	Multilingual web corpus
Training Objectives	Language modeling	Language modeling	Language modeling	Mixture-of-Denoisers (MoD)	Not specified
Special Features	Improved prompt design	Improved prompt design	Improved prompt design	Universally effective across NLP tasks	Open-access, multilingual
How to Access	Via OpenAI API	Via Google Workspace	Application required	Not specified	Open-source
Released By	OpenAI	Google	Meta AI	Google Research	BigScience Workshop
Dataset Used for Training	WebText-like corpus	WebText-like corpus	WebText-like corpus	Publicly available data	Multilingual web corpus
Multimodal Capabilities	No	No	No	No	Yes
Multilingual Support	Yes	Limited	Yes	Yes	Yes

<https://deepchecks.com/llm-models-comparison/>

99% of customers
will not be building
infrastructure to train
their own LLMs



Inference and Fine Tuning

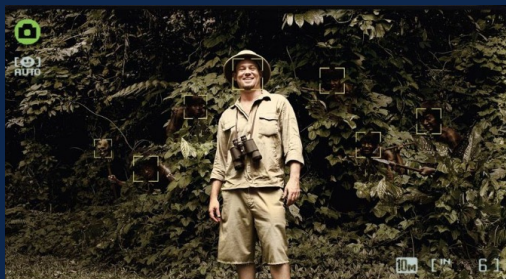
A note on inference/fine-tuning workloads

Inference in LLM is the process of using a trained model to generate responses to the user prompts, usually through an API or web service. For example, when we type in a question in a ChatGPT session, an inference process is run on a copy of the trained GPT-3.5 model hosted somewhere in the cloud to get us the response back. Inference needs a lot less GPU resources than training. But, given the billions of parameters in the trained LLM model, inference still needs multiple GPUs (to spread parameters and the computation). For example, Meta's LLaMA model typically needs 16 A100 GPUs for inference (as opposed to 2,000 used for training).

Similarly, fine-tuning an already trained model with domain-specific data sets requires fewer resources, often less than 100+ H100 scale GPUs. With these scales, both inference and fine-tuning do not need large GPU clusters on the same fabric.

Many customers will build GPU clusters in their existing DCs for training “smaller” models, for fine tuning existing models, and to do inference or generative AI.

LLMs everywhere?



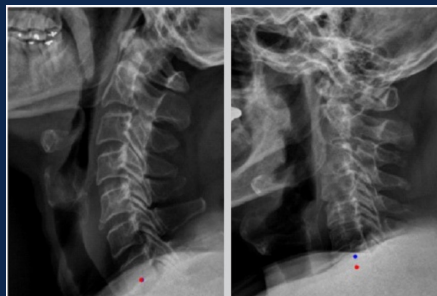
Face recognition



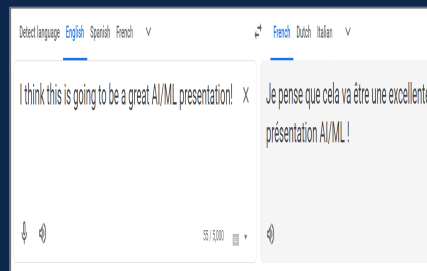
Self-driving vehicles



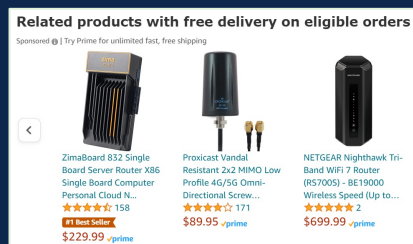
Conversational agents



Analysis of medical images



Machine translation



Recommender systems



Image generation

Industry Talk

Ultra Ethernet Consortium – UEC

The screenshot shows the Ultra Ethernet Consortium website. The top navigation bar includes links for WORKING GROUPS, NEWS, MEMBERSHIP, and CONTACT US, along with a prominent orange button labeled 'BECOME A MEMBER'. Below the navigation bar is a list of frequently asked questions. The question 'When and where can I find the technical specifications and roadmap of UEC?' is highlighted with a red rectangular box. The answer to this question is 'Stay tuned: UEC will announce when technical documents and roadmap are available. These documents will be published on the UEC website.'

https://ultraethernet.org

Ultra Ethernet Consortium

WORKING GROUPS NEWS MEMBERSHIP CONTACT US BECOME A MEMBER

Who should join UEC and what is the process and cost?

Are there different membership levels?

How is UEC organized and operated, and does it have presence outside of the United States?

When and where can I find the technical specifications and roadmap of UEC?

Stay tuned: UEC will announce when technical documents and roadmap are available. These documents will be published on the UEC website.

Where can I subscribe to news of the UEC?

When can we expect products based on UEC specifications to be available in the market?

What are the planned routes to market once products are available and where can users buy them?

Ultra Ethernet Consortium – UEC

UEC Areas of Innovation

Feature	UEC Focus
Multi-Pathing	Packet spraying
Flexible Ordering	Out-of-order packet delivery with in-order message delivery
AI & HPC Congestion Control	Configuration-free, low latency, programmable
E2E Telemetry	To enable Sender or receiver based congestion control
Simplified RDMA	Replaces RoCE, streamlines APIs
Security	Scalable, 1 st class citizen
High Scale	1,000,000 endpoints

ECMP, Telemetry Assisted Ethernet and Fully Scheduled Fabrics



Evolved Networking: the AI/ML Challenge Cisco Silicon One

Rakesh Chopra
Cisco Fellow
October 18, 2022, 3:30pm. Room 210 B/F



OCP
GLOBAL
SUMMIT

OCTOBER 18-20, 2022
SAN JOSE, CA

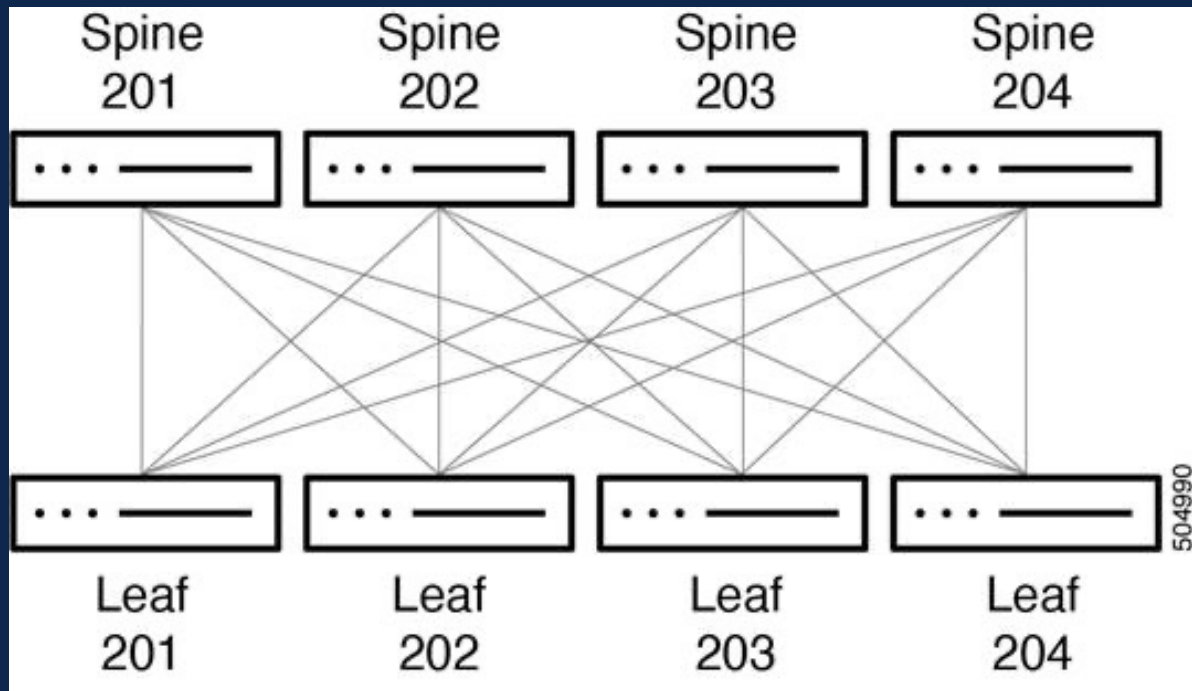
 rakchopr@cisco.com

 www.linkedin.com/in/rakesh-chopra/

 [@Rakesh_Chopra1](https://twitter.com/Rakesh_Chopra1)

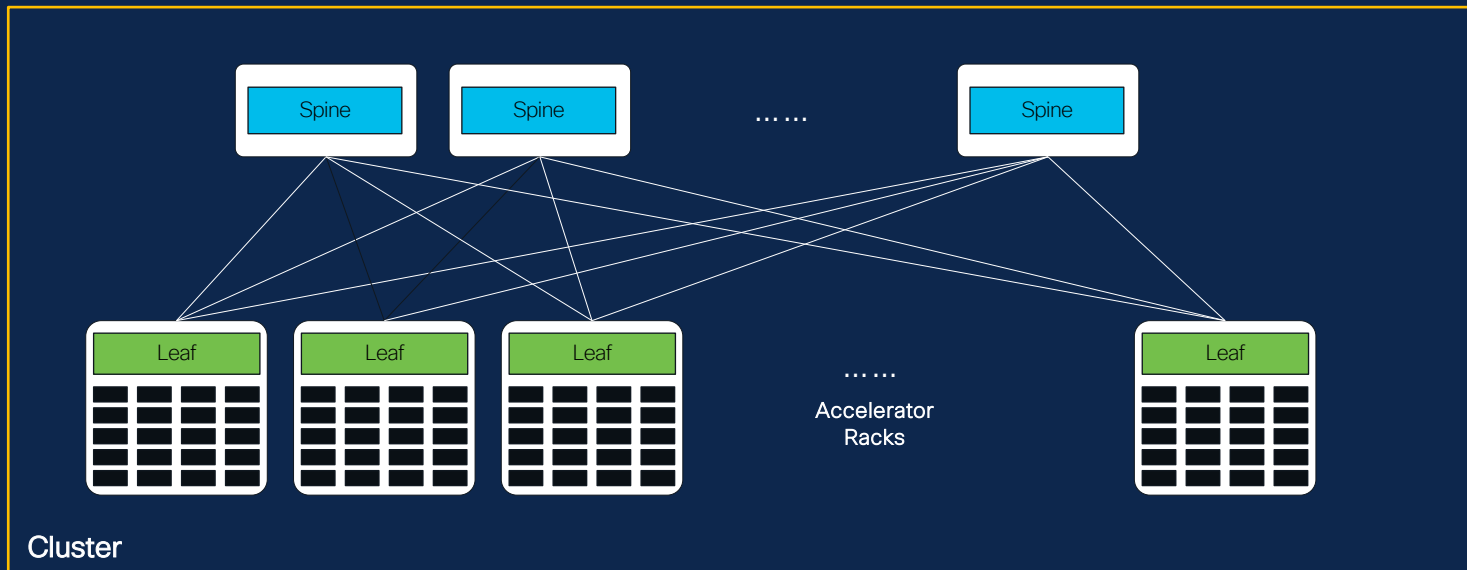
Special thanks to Nadav Chachmon & Ofer Iny for their incredible contributions to this presentation

ECMP, Telemetry Assisted Ethernet and Fully Scheduled Fabrics



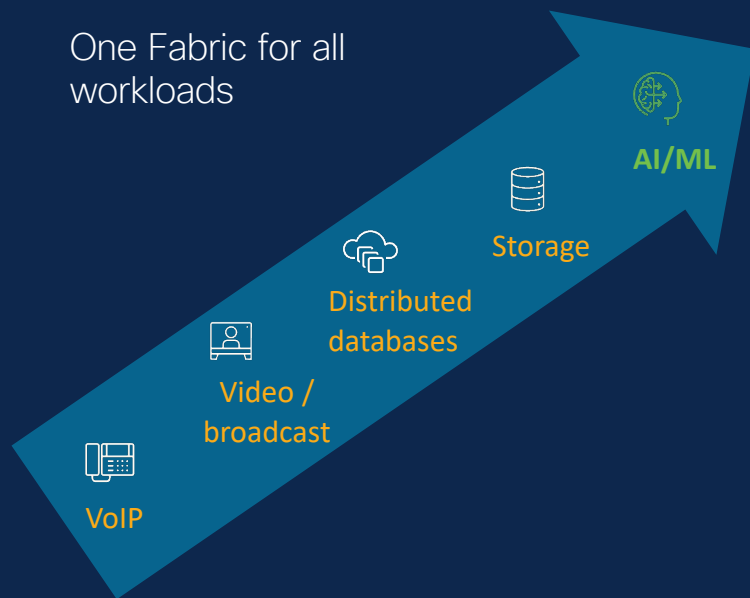
Distributed Scheduled Fabric

32K+ GPUs/cluster
Leaf to spine packet spray
VoQ & Request Grant
PFC & ECN Leaf to Host



AI/ML: High Performance Ethernet Fabrics

One Fabric for all workloads



Systems & Solutions

Primary target

Enterprise/
Public Sector/
Commercial

Service
Providers

Tier 2
Web

Nexus 9K with
Nexus Dashboard

Disaggregated or
Components

Primary target



Hyperscalers

8000 with SONIC

What About Infiniband?

AI Fabric POC test procedure and results

GPU to GPU TEST

NCCL (Nvidia Collective Communication Library) Test, ALL GPU to ALL GPU

Verification

- GPU server 1 <-> GPU server 2

```
# mpirun --allow-run-as-root -np 16 -H dgx01:8,dgx02:8 -mca pml ucx  
-bind-to numa -map-by numa -x LD_LIBRARY_PATH -x  
NCCL_DEBUG=INFO -x NCCL_SOCKET_IFNAME=enp12s0,^enp226s0 -x  
UCX_NET_DEVICES=enp12s0,enp18s0,enp75s0,enp84s0,enp141s0,enp14  
8s0,enp186s0,enp204s0 -x NCCL_IGNORE_CPU_AFFINITY=1 -x  
NCCL_UCX_TLS=rc,cuda,cuda_copy -x NCCL_UCX_MAX_RNDV_RAILS=4 -x  
NCCL_MAX_EAGER_RAILS=4 -x NCCL_UCX_RNDV_THRESH=0 -x  
UCX_MEMTYPE_CACHE=n -x NCCL_PLUGIN_P2P=ucx -x  
NCCL_COLLNET_ENABLE=1 -x NCCL_MAX_P2P_NCHANNEL=1 -x  
NCCL_ALGO=Tree,CollNet -x NCCL_DEBUG_SUBSYS=graph,tuning,init,p2p  
-x NCCL_NET_GDR_READ=1 -x NCCL_NET_GDR_LEVEL=5 -x  
NCCL_IB_PCI_RELAXED_ORDERING=1 -x NCCL_NET_P2P_LEVEL=5 -x  
CUDA_DEVICE_ORDER=PCI_BUS_ID /bmt/mkis/workspace/nccl-  
tests/build/all_reduce_perf -b 8 -e 8G -f 2 -g 1 -n 10 -w 100 -p 0 -z 0 -  
m 1 -t 1 -c 1 | tee result/2node_1set_date +%Y%m%d_%H%M$.log
```

```
root@dgx01:/bmt/mkis/workspace/nccl_result# ./run_nccl.sh 2node.distro  
nThread 1 mPops 1 mBytes 0 mMaxBytes 5889934592 step: 2(factor) warmup iters: 100 iters: 10 agg iters: 1 validation: 1 graph: 0  
Using devices  
Rank 0 Pid 1181172 on dgx01 device 0 [0x07] NVIDIA A100-SXM4-80GB  
Rank 1 Pid 1181174 on dgx01 device 1 [0x08] NVIDIA A100-SXM4-80GB  
Rank 2 Pid 1181174 on dgx01 device 2 [0x07] NVIDIA A100-SXM4-80GB  
Rank 3 Pid 1181175 on dgx01 device 3 [0x08] NVIDIA A100-SXM4-80GB  
Rank 4 Pid 1181176 on dgx01 device 4 [0x07] NVIDIA A100-SXM4-80GB
```

2 x NVIDIA DGX A100 6U server with 8 x NVIDIA A100 Tensor Core GPUs
8 x Mellanox ConnectX-6
2TB of Mem

```
size count type redop root time algw busbw #wrong time algw busbw #wrong  
(B) (elements) (us) (GB/s) (GB/s) (us) (GB/s) (GB/s)  
8 2 Float sum -1 46.46 0.00 0.00 0 46.36 0.00 0.00 0  
16 4 Float sum -1 46.52 0.00 0.00 0 46.73 0.00 0.00 0  
32 8 Float sum -1 46.76 0.00 0.00 0 46.85 0.00 0.00 0  
64 16 Float sum -1 44.88 0.00 0.00 0 45.04 0.00 0.00 0  
128 32 Float sum -1 45.92 0.00 0.01 0 45.27 0.00 0.01 0  
256 64 Float sum -1 47.07 0.01 0.01 0 46.39 0.01 0.01 0  
512 128 Float sum -1 48.28 0.01 0.02 0 47.25 0.01 0.02 0  
1024 256 Float sum -1 49.04 0.02 0.04 0 48.99 0.02 0.04 0  
2048 512 Float sum -1 50.41 0.04 0.08 0 49.67 0.04 0.08 0  
4096 1024 Float sum -1 54.07 0.08 0.14 0 54.00 0.08 0.14 0  
8192 2048 Float sum -1 56.70 0.14 0.27 0 54.86 0.15 0.28 0  
16384 4096 Float sum -1 60.50 0.27 0.51 0 65.72 0.25 0.47 0  
32768 8192 Float sum -1 58.57 0.34 1.05 0 59.13 0.55 1.04 0  
65536 16384 Float sum -1 61.64 1.06 1.99 0 63.90 1.03 1.92 0  
131072 32768 Float sum -1 78.02 1.68 3.15 0 74.62 1.76 3.29 0  
262144 65536 Float sum -1 81.41 3.22 6.04 0 77.95 3.37 6.31 0  
524288 131072 Float sum -1 86.47 6.06 11.97 0 86.40 6.07 11.98 0  
1048576 262144 Float sum -1 102.7 10.21 19.14 0 101.8 10.30 19.31 0  
2097152 524288 Float sum -1 159.2 13.17 24.69 0 155.6 13.48 25.27 0  
4194304 1048576 Float sum -1 178.3 23.52 44.10 0 168.5 24.90 46.48 0  
8388608 2097152 Float sum -1 211.8 39.40 74.25 0 205.6 40.76 76.43 0  
16777216 4194304 Float sum -1 329.3 50.96 95.54 0 334.2 50.20 94.13 0  
33554432 8388608 Float sum -1 554.1 60.56 113.53 0 554.8 60.48 113.41 0  
67108864 16777216 Float sum -1 999.4 67.15 123.90 0 986.9 68.00 127.50 0  
134217728 33554432 Float sum -1 1591.1 70.16 133.55 0 1502.8 70.54 132.24 0  
268435456 67108864 Float sum -1 3678.0 72.98 136.85 0 3734.1 71.89 134.79 0  
536870912 134217728 Float sum -1 6914.1 77.65 145.59 0 6892.0 77.50 146.06 0  
1073741824 268435456 Float sum -1 14583.8 74.59 139.85 0 12813 83.79 157.10 0  
2147483648 536870912 Float sum -1 23464 84.47 158.75 0 23459 84.43 158.35 0  
4294967296 1073741824 Float sum -1 50812 84.53 158.49 0 42890 86.29 128.05 0  
8589934592 2147483648 Float sum -1 99908 85.99 161.21 0 123801  
Out of bounds values : 0 OK  
# Avg base bandwidth : 49.4928
```

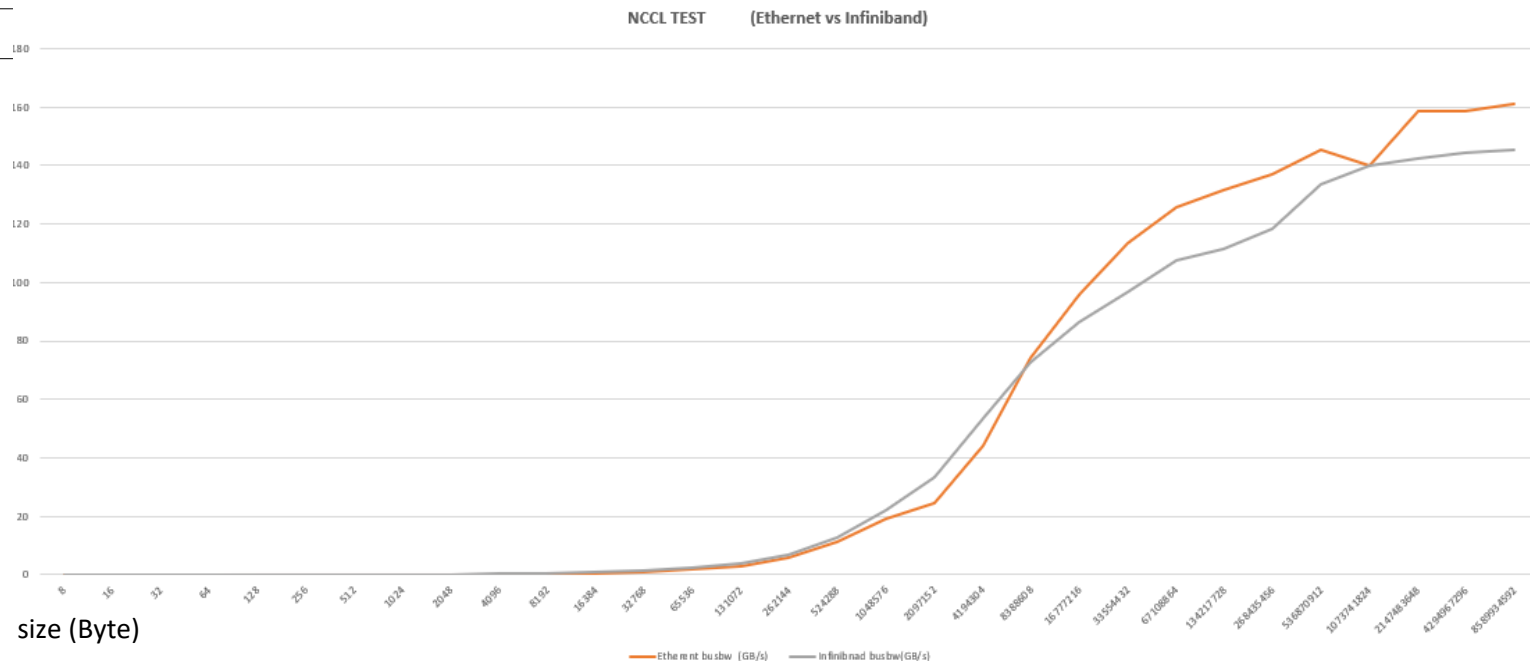
NVIDIA-SMI 470.161.03 Driver Version: 470.161.03 CUDA Version: 11.4									
GPU Name	Persistence-M	Bus-Id	Disp.A	Volatile Uncorr. ECC	Fan Temp	Perf	Pwr:Usage/Cap	Memory-Usage	GPU-Util Compute M.
MIG M.									
0 NVIDIA A100-SXM...	On	00000000:07:00.0 Off							
3 NVIDIA A100-SXM...	On	00000000:1E:00.0 Off							
N/A 35C	P0 140W / 400W	25724MiB / 40536MiB	100%	Default Disabled					
4 NVIDIA A100-SXM...	On	00000000:87:00.0 Off							
N/A 45C	P0 158W / 400W	25804MiB / 40536MiB	100%	Default Disabled					
5 NVIDIA A100-SXM...	On	00000000:90:00.0 Off							
N/A 42C	P0 167W / 400W	25724MiB / 40536MiB	100%	Default Disabled					
6 NVIDIA A100-SXM...	On	00000000:B7:00.0 Off							
N/A 45C	P0 177W / 400W	25804MiB / 40536MiB	100%	Default Disabled					
7 NVIDIA A100-SXM...	On	00000000:BD:00.0 Off							
N/A 44C	P0 133W / 400W	25724MiB / 40536MiB	100%	Default Disabled					

Max 161 GB/s

AI Fabric POC test procedure and results

GPU to GPU TEST

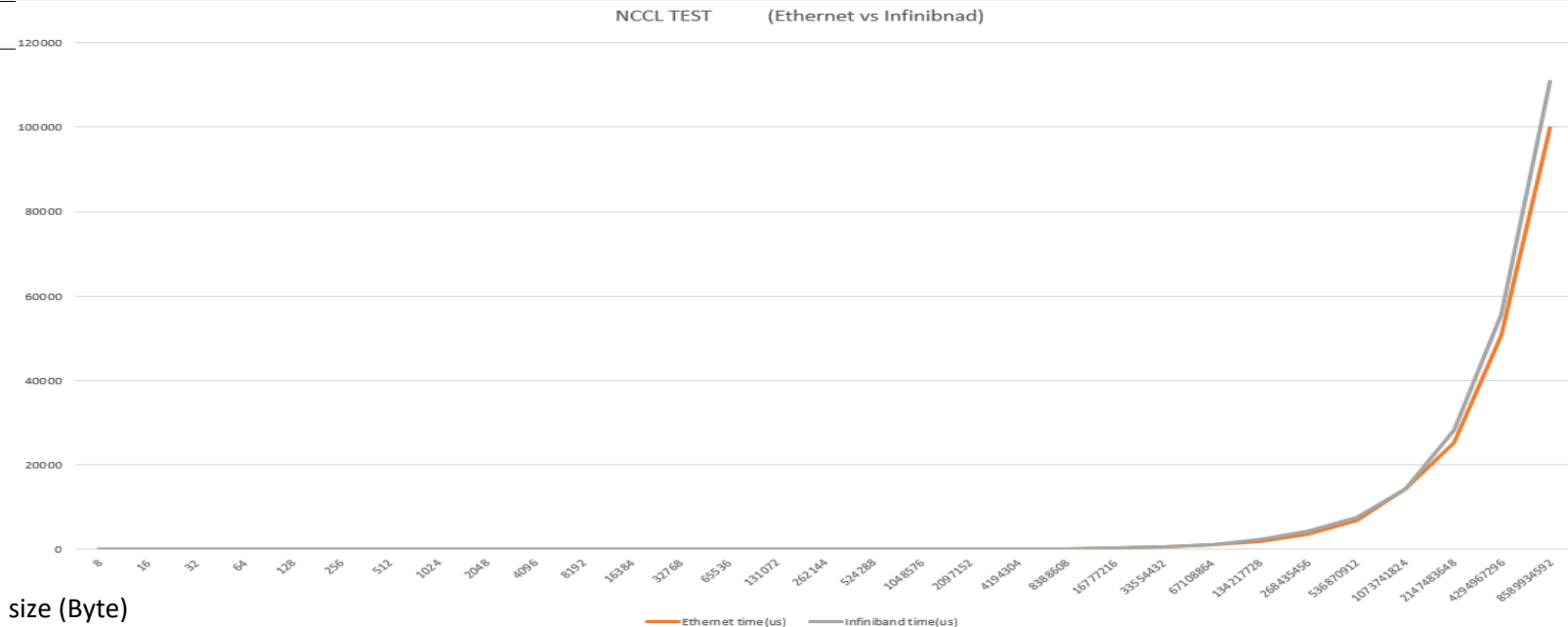
NCCL test result comparison chart (Ethernet vs Infiniband)



AI Fabric POC test procedure and results

GPU to GPU TEST

NCCL test result comparison chart (Ethernet vs Infiniband)

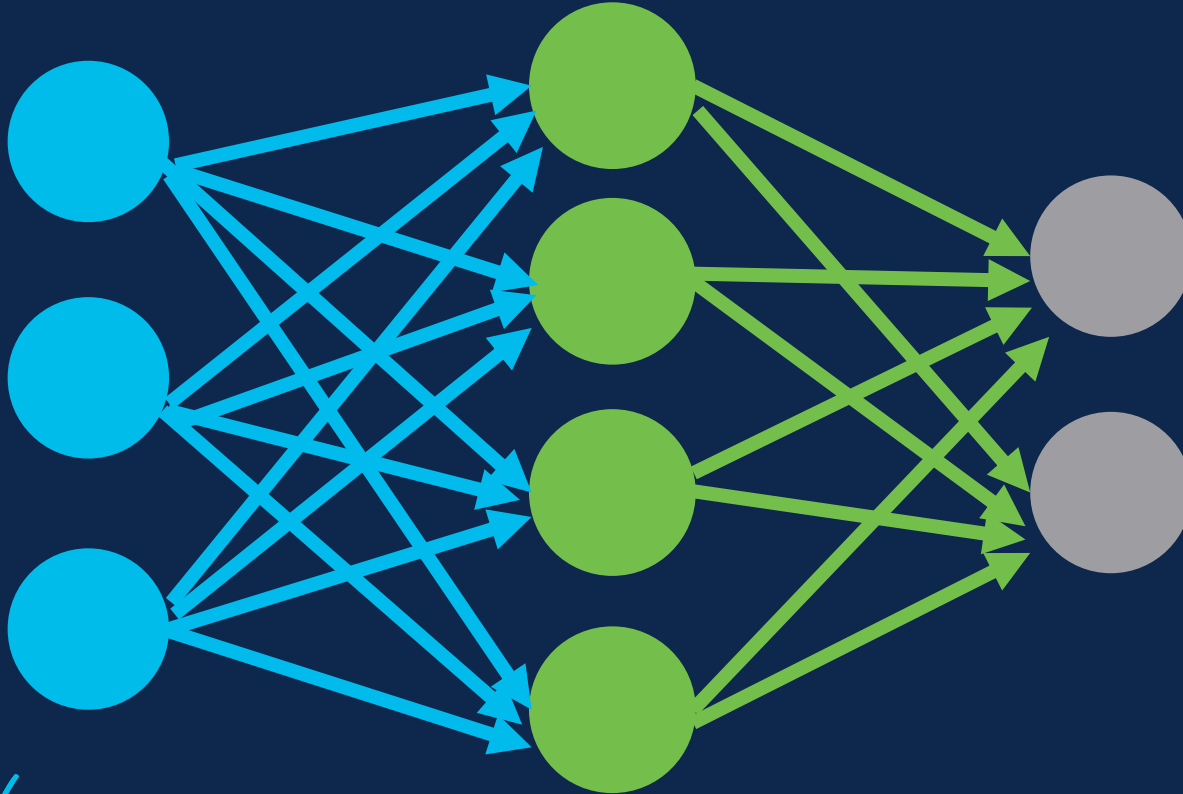


What Are AI/ML Application Network Requirements

What is a Neural Network?

A neural network is a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain. It is a type of machine learning process, called deep learning, that uses interconnected nodes or neurons in a layered structure that resembles the human brain.

What is a Neural Network?



Neural Network Example (Very Oversimplified)

Should I go surfing?

1. Are the waves good? (Yes: 1, No: 0)
2. Is the line-up empty? (Yes: 1, No: 0)
3. Has there been a recent shark attack? (Yes: 0, No: 1)

Then, let's assume the following, giving us the following inputs:

$X_1 = 1$, since the waves are pumping

$X_2 = 0$, since the crowds are out

$X_3 = 1$, since there hasn't been a recent shark attack

Now, we need to assign some weights to determine importance. Larger weights signify that particular variables are of greater importance to the decision or outcome.

$W_1 = 5$, since large swells don't come around often

$W_2 = 2$, since you're used to the crowds

$W_3 = 4$, since you have a fear of sharks

<https://www.ibm.com/topics/neural-networks>

Neural Network Example (Very Oversimplified)

Should I go surfing?

Finally, we'll also assume a threshold value of 3, which would translate to a bias value of -3. With all the various inputs, we can start to plug in values into the formula to get the desired output.

$$\hat{Y} = (1*5) + (0*2) + (1*4) - 3 = 6$$

If we use the activation function, we can determine that the output of this node would be 1, since 6 is greater than 0. In this instance, you would go surfing;

<https://www.ibm.com/topics/neural-networks>

Parallelism Mechanisms (collectives)

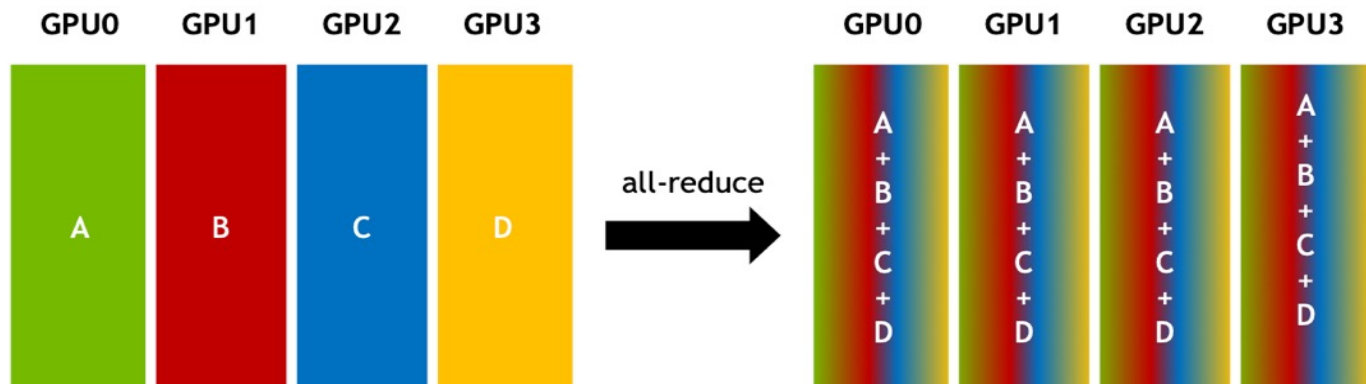
- NVIDIA Collective Communications Library (NCCL) supports several data/job distribution methods called collectives, namely:
 - AllReduce
 - Broadcast
 - Reduce
 - AllGather
 - ReduceScatter
 - RingReduce
- Each method has its pros and cons; choice is left to the developer
 - Python's PyTorch (top library for deep learning) defaults to AllReduce*
 - Informal research (scanning Git repos, reading blogs) shows virtually every developer leaves this unchanged

https://pytorch.org/docs/stable/ddp_comm_hooks.html

Different Traffic Patterns and Job Sizes

ALL-REDUCE

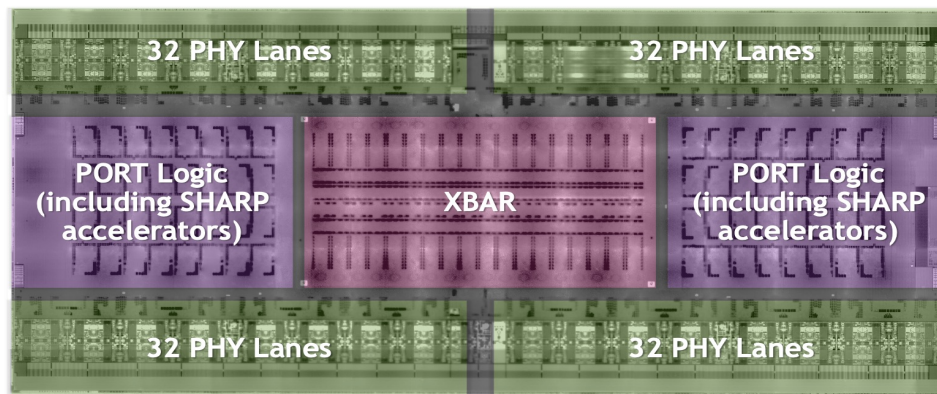
Combine data from all senders; deliver the result to all participants



GPU Communications “Inside” the Server

NVLINK4 NVSWITCH

Chip Characteristics



Largest NVSwitch Ever

- TSMC 4N process
- 25.1B transistors
- 294mm²
- 50mmX50mm package (2645 balls)

Highest Bandwidth Ever

- 64 NVLink4 ports (x2 per NVLink)
- 3.2TB/s full-duplex bandwidth
- 50Gbaud PAM4 diff-pair signaling
- All ports NVLink Network capable

New Capabilities

- 400GFLOPS of FP32 SHARP (other number formats are supported)
- NVLink Network management, security and telemetry engines



Main Infrastructure Requirements for Training

In a DL training infrastructure, it is crucial to get as much raw compute power and as many nodes as you can afford. Think multi-core processors and GPUs.

The more nodes and the more mathematical accuracy you can build into your cluster, the faster and more accurate your training will be.

Huge training datasets require massive networking and storage capabilities to hold and transfer the data, especially if your data is image-based or heterogeneous. Plan ahead for adequate networking and storage capacity, not just for strong computing.

Main Infrastructure Requirements for Inference

Inference clusters should be optimized for performance. Think simpler hardware with less power than the training cluster but **with the lowest latency possible**.

Throughput is critical to inference. **The process requires high I/O bandwidth** and enough memory to hold both the required training model(s) and the input data without having to make calls back to the storage components of the cluster.

Data center resource requirements for inference are typically not as great for a single instance compared to training needs.

Backend and Frontend Connectivity

The “**backend network**” is used only for GPU-to-GPU connectivity

The “**frontend network**” is used for all other communications and can be used for IP storage access as well

For very large clusters an IP **storage network** can also be used

Multiple Connections May Be Required For AI Workloads

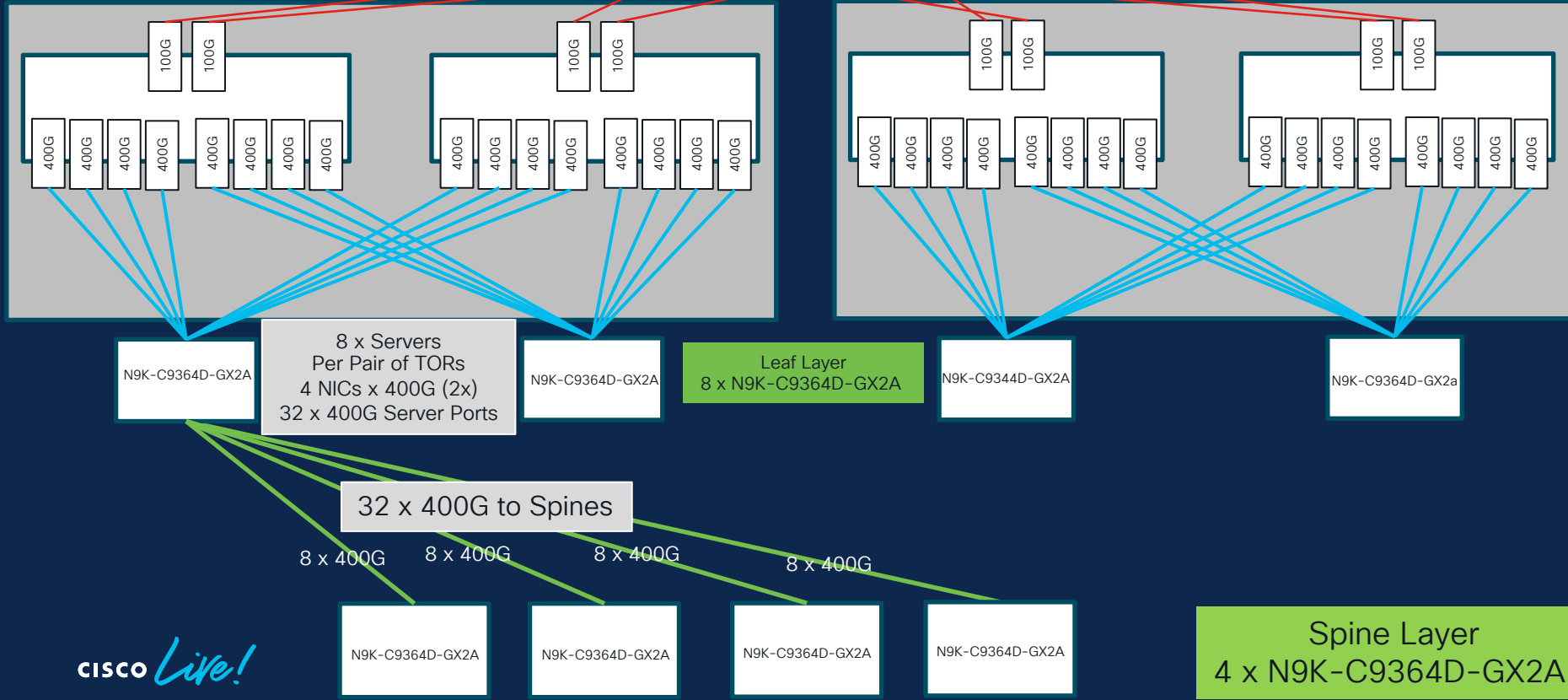
Nexus design

All 32 Servers connect
2 x 100G NICs = 64 x 100G
Across 2 x N9K-C9364C-GX

N9K-C9364C-GX

N9K-C9364C-GX

Front End / Storage & Outside World
2 x N9K-C9364C-GX



What is Needed to Build The Best AI/ML DC Network?

- Non-Blocking Fabric

- Low Latency

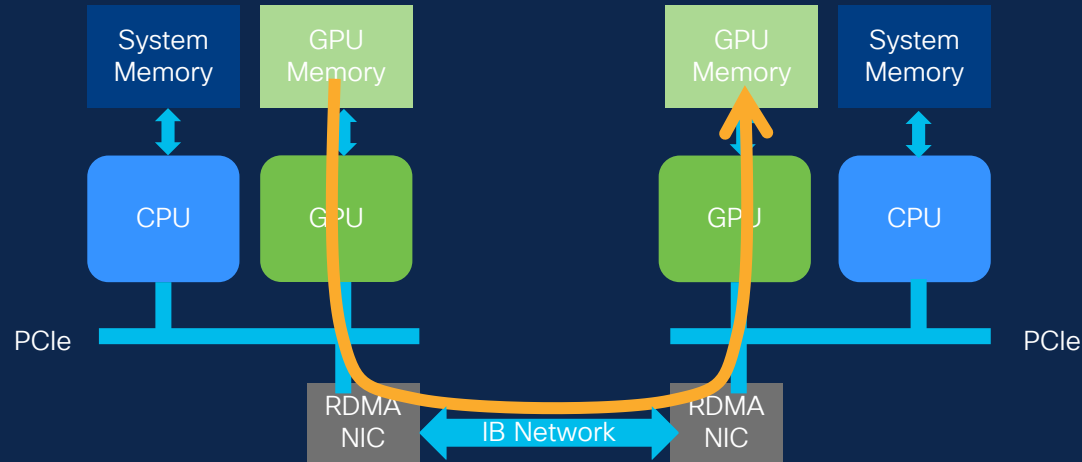
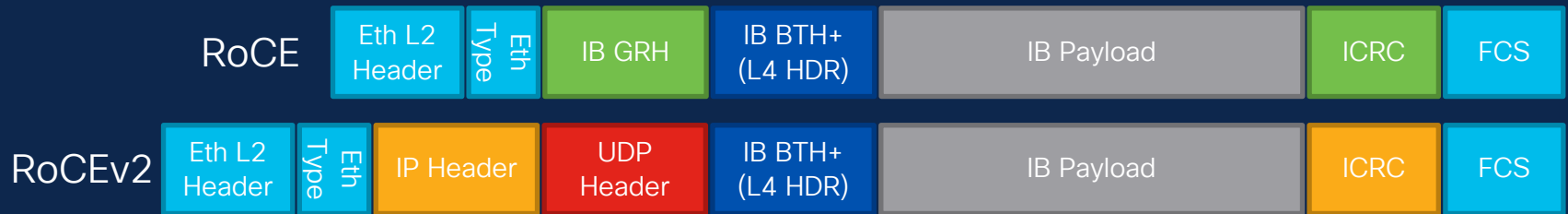
- Lossless



RoCE

- Visibility

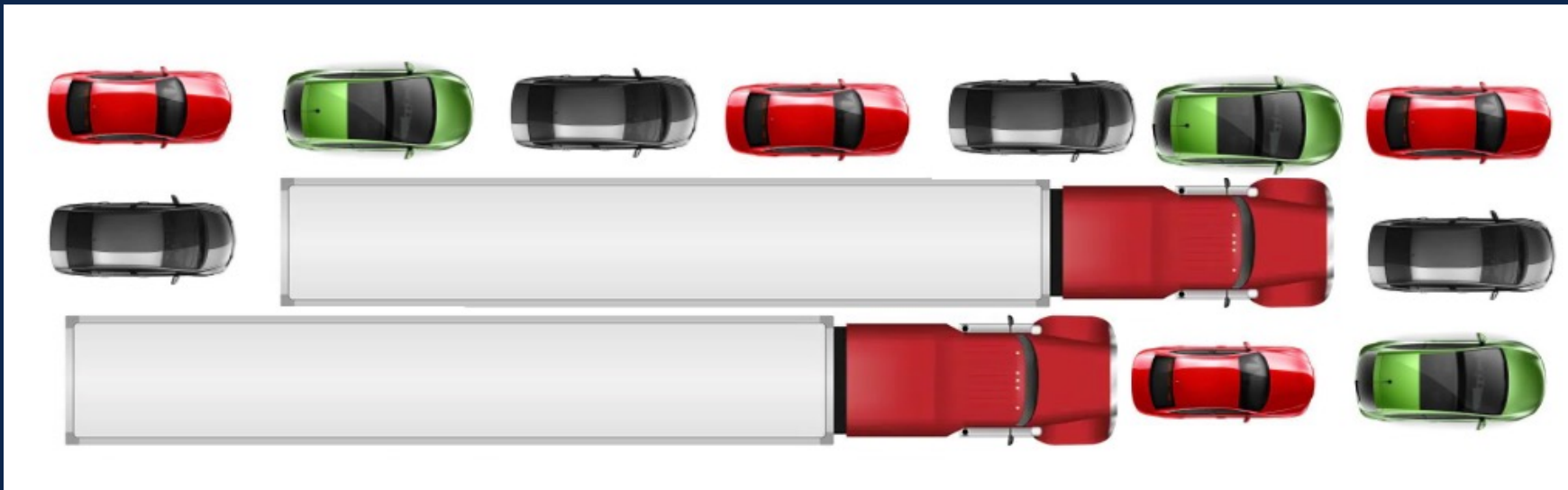
RDMA Over Converged Ethernet – RoCE



Lossless Ethernet

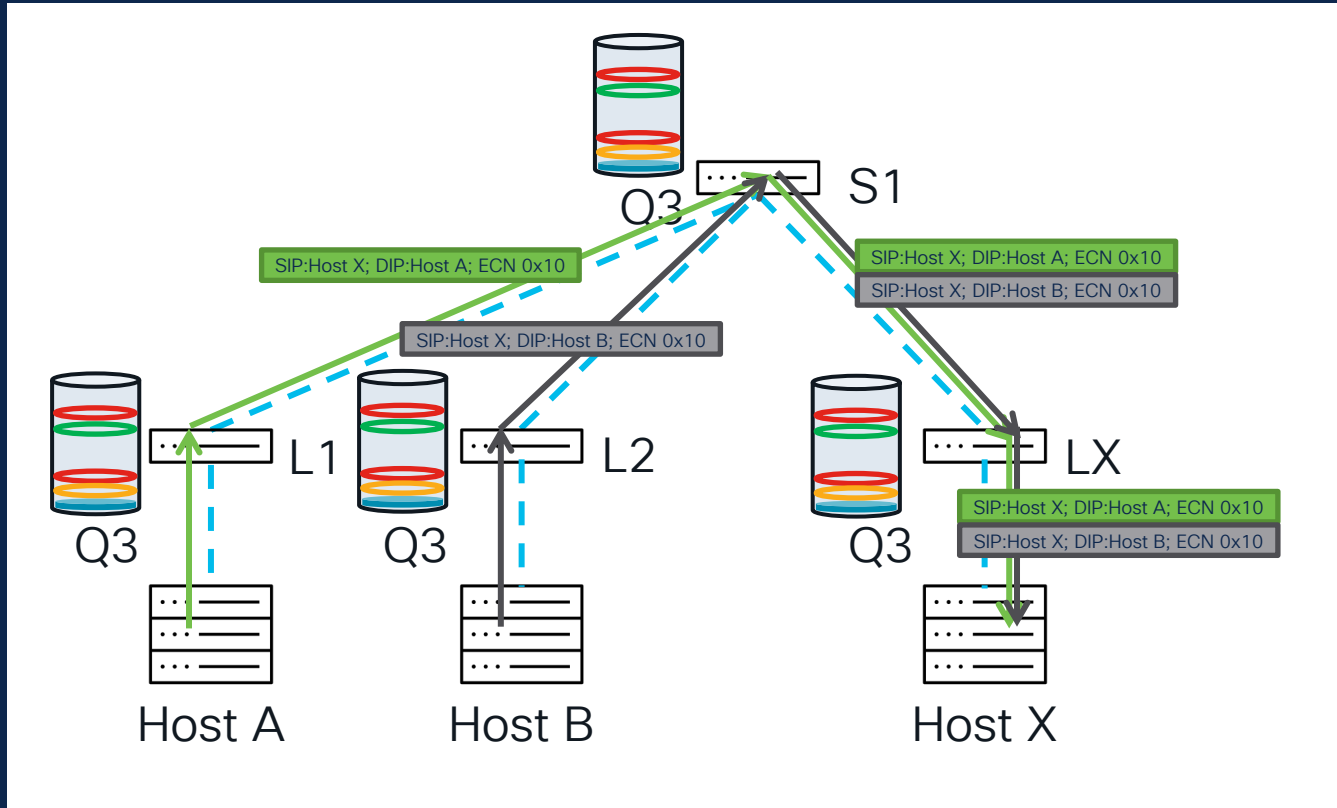
- Lossless Ethernet for RoCEv2
 - End-to-End
 - ECN (Explicit congestion Notification) + WRED or AFD
 - PFC (Priority Flow Control)
- Buffers (smart and the right amount)
- Tuning

Intelligent Buffers

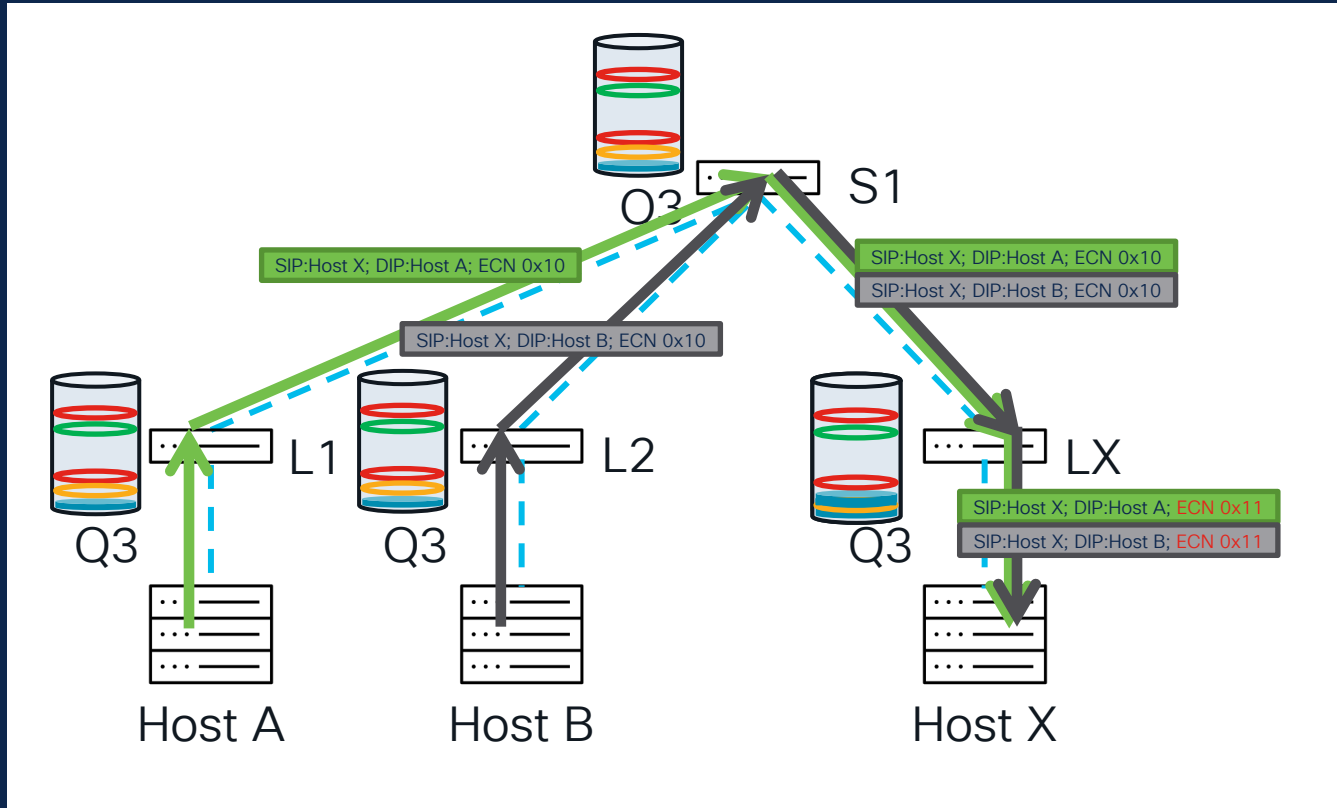


Congestion Management Needs to Work As a System

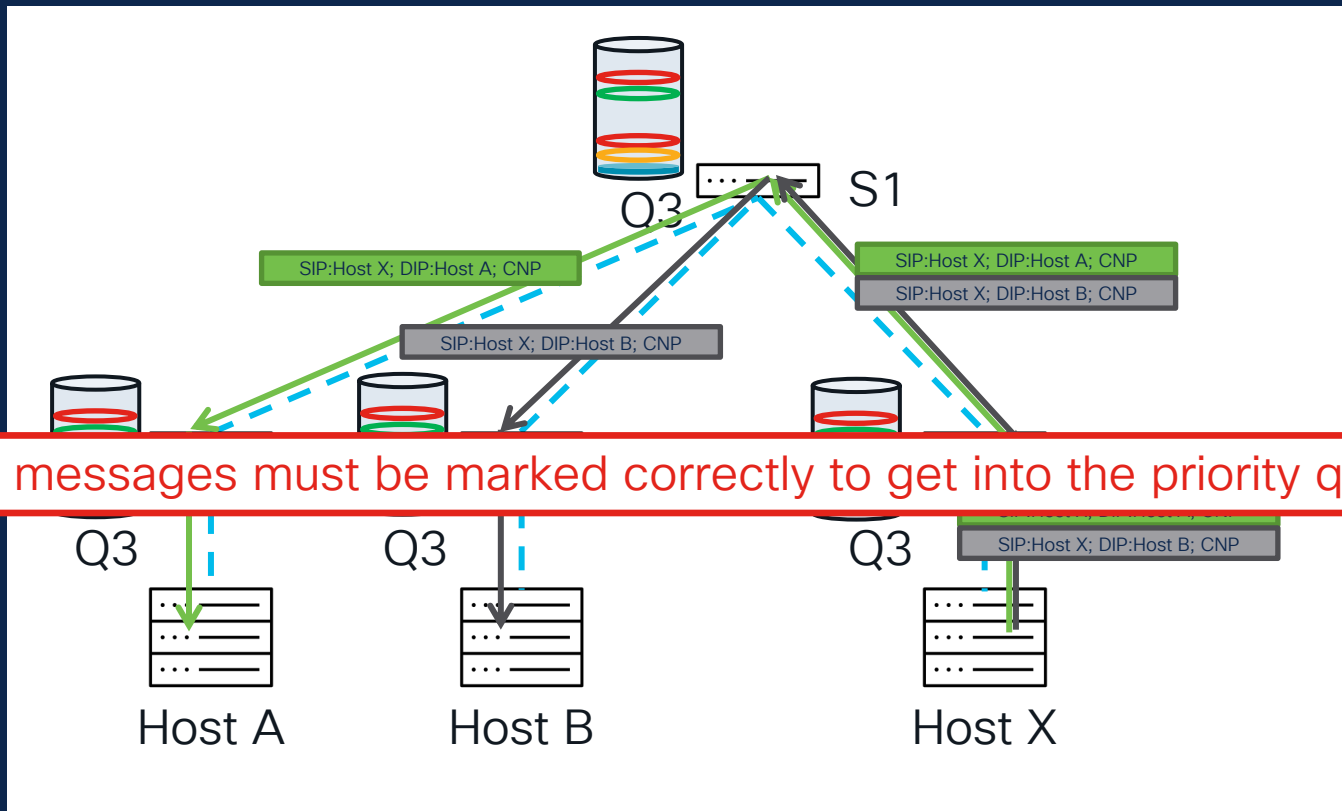
No Congestion Example



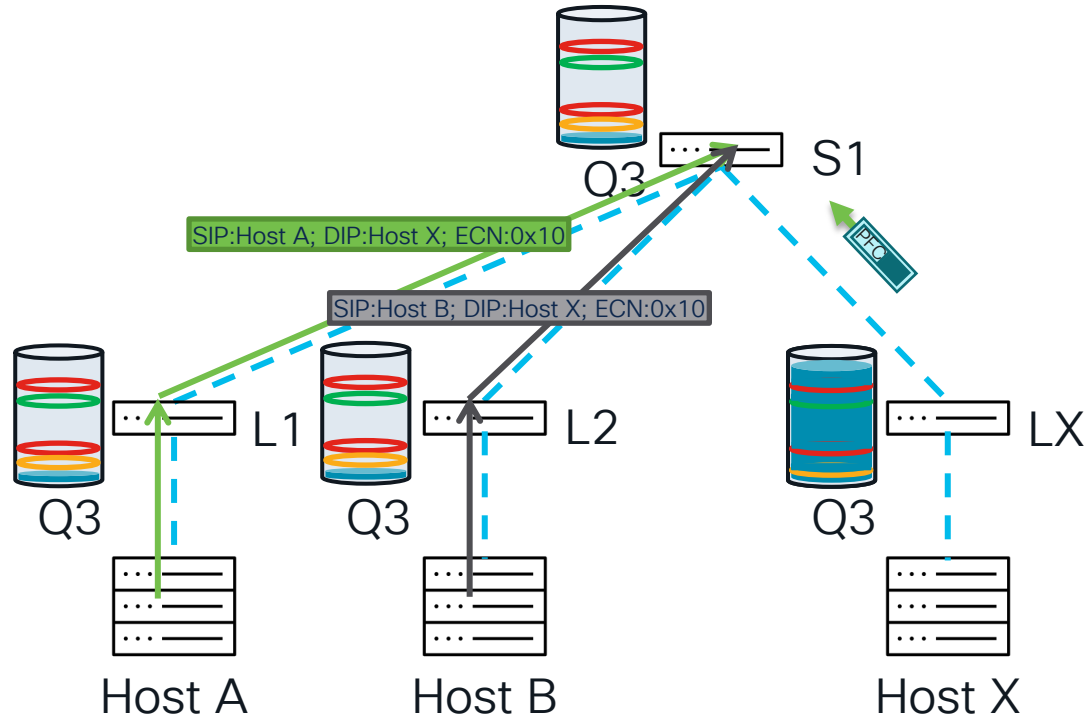
Light Congestion Example



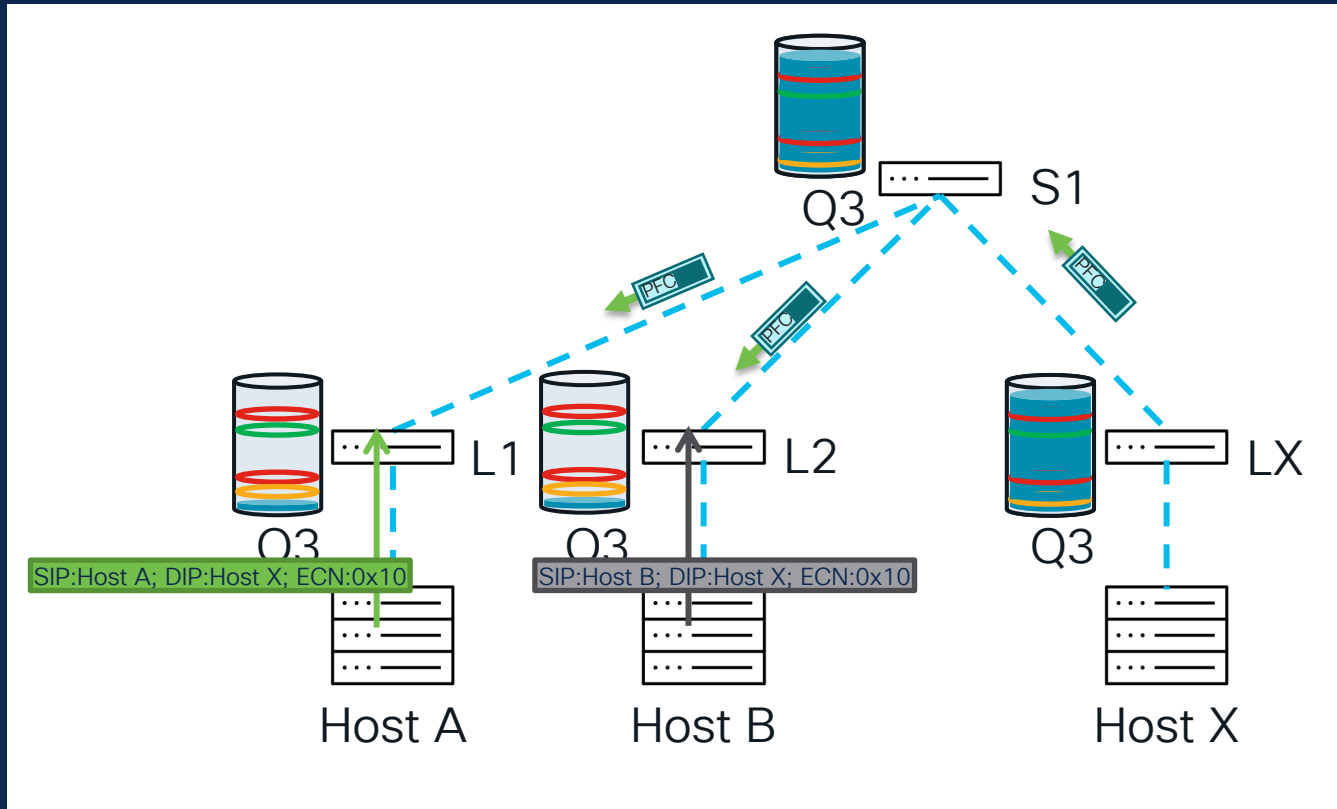
Light Congestion Example –CNP Message



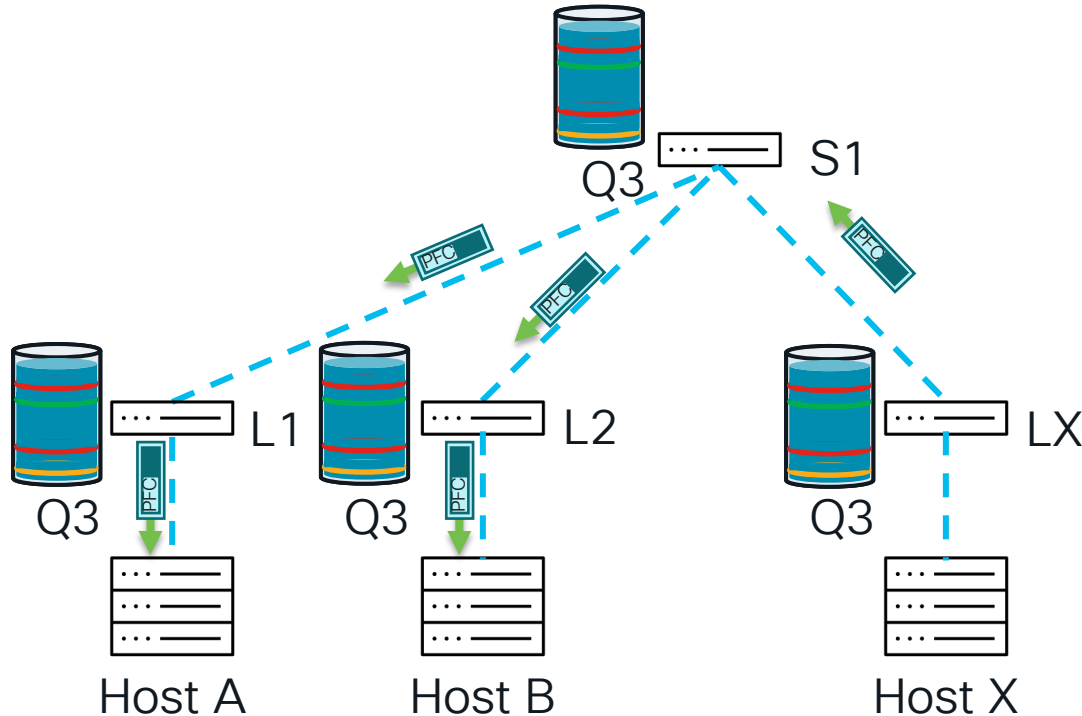
PFC Congestion Leaf



PFC Congestion Leaf + Spine



PFC Congestion Leaf + Spine + Leaf



To get RoCEv2 to work
as expected it is
critical to configure the
NICs and OS correctly

Configure Your Ethernet NICs

Mellanox ConnectX-6 NIC
2 Port 100G Ethernet NIC

cat /etc/netplan/00-installer-config.yaml

<CLIP>

ens1f0np0:

dhcp4: false

ens1f1np1:

dhcp4: false

vlan:

ens1f0np0.101:

id: 101

link: ens1f0np0

addresses:

- 192.168.101.11/24

routes:

- to: 192.168.0.0/16

via: 192.168.101.1

metric: 100

ens1f1np1.101:

id: 101

link: ens1f1np1

addresses:

- 172.16.101.11/24

routes:

- to: 172.16.0.0/16

via: 172.16.101.1

metric: 100

VLAN Config Required for PFC

VLAN Encap is 101

Configure the “marking” Ubuntu Example

```
sudo apt-get install vlan
```

```
sudo bash -c 'echo 1 > /sys/class/net/ens1f0np0/ecn/roce_np/enable/3'
sudo bash -c 'echo 1 > /sys/class/net/ens1f1np1/ecn/roce_np/enable/3'
```

← Enable ECN for class 3

```
sudo bash -c 'echo 6 > /sys/class/net/ens1f0np0/ecn/roce_np/cnp_802p_prio'
sudo bash -c 'echo 6 > /sys/class/net/ens1f1np1/ecn/roce_np/cnp_802p_prio'
```

← Enable CNP L2 egress priority to 6, “enable CNP”

```
sudo bash -c 'echo 48 > /sys/class/net/ens1f0np0/ecn/roce_np/cnp_dscp'
sudo bash -c 'echo 48 > /sys/class/net/ens1f1np1/ecn/roce_np/cnp_dscp'
```

← Map CNP to DSCP 48

```
sudo mlx_qos -i ens1f0np0 --trust=dscp --pfc 0,0,0,1,0,0,0,0
sudo mlx_qos -i ens1f1np1 --trust=dscp --pfc 0,0,0,1,0,0,0,0
```

← Map PFC to priority 3

```
sudo sysctl -w net.ipv4.tcp_ecn=1
```

← Enable ECN for traffic

```
sudo cma_roce_mode -d mlx5_0 -p 1 -m 2
sudo cma_roce_mode -d mlx5_1 -p 1 -m 2
```

← Set mode to RoCEv2 on the NIC

```
sudo cma_roce_tos -d mlx5_0 -t 96
sudo cma_roce_tos -d mlx5_1 -t 96
```

← Set TOS = DSCP CS3 for RoCEv2 traffic

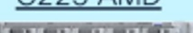
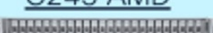
```
sudo vconfig set_egress_map ens1f0np0.101 4 3
sudo vconfig set_egress_map ens1f1np1.101 4 3
```

← Egress map RoCE traffic to priority 3

Recommendations for Server Tuning*

*These recommendations were done in
conjunction with [Vijay Durairaj](#)

Cisco GPU-Accelerated Platform Offerings

X-Series	C-Series Rack Servers				
 <p>Up to 24x T4 GPUs or 8x A-Series GPUs per X9508 Chassis</p>  <p><u>X210c M6/M7 Blade</u> 2x NVIDIA T4 (MEZZ)</p> <p><u>X210c M7 Blade (2H23)</u> Intel Flex140 (MEZZ)</p> <p><u>X210c M7 Blade (1H24)</u> NVIDIA L4 (MEZZ)</p> <p><u>X440p M6/M7 Node Only</u> 4x NVIDIA T4 (M6 Only) 2x NVIDIA A16 2x NVIDIA A40 2 x NVIDIA A100-80</p> <p><u>X440p M7</u> 2x NVIDIA H100 2x NVIDIA L40 4x NVIDIA L4</p> <p><u>X440p M7 (Q3'23)</u> 4x Intel Flex140 2x Intel Flex170</p>	<p><u>C220 M5</u></p>  <p>2x NVIDIA T4</p>	<p><u>C240 M5</u></p>  <p>2x NVIDIA M10 6x NVIDIA T4 2x NVIDIA A100-80</p> <p>2x NVIDIA V100 (EoS 6/21) 2x NVIDIA RTX6000 (EoS 5/22) 2x NVIDIA RTX8000 (EoS 5/22) 2x NVIDIA A100 40GB (EoS 3/22)</p>	<p><u>C480 M5</u></p>  <p>3x NVIDIA M10 9x NVIDIA T4 6x NVIDIA A40 6x NVIDIA A100-80</p> <p>6x NVIDIA V100 (EoS 6/21) 6x NVIDIA A100 40GB (EoS 3/22)</p>	<p><u>C220 INTEL C225 AMD</u></p>  <p><u>C220/C225 M6</u> 3x NVIDIA T4</p> <p><u>C220 M7 (Q3'23)</u> 3x NVIDIA L4 3x Intel FLex140</p>	<p><u>C240 INTEL C245 AMD</u></p>  <p><u>C240/C245 M6</u> 5x NVIDIA A10 3x NVIDIA A16 3x NVIDIA A30 3x NVIDIA A40 3x NVIDIA A100-80</p> <p><u>C240 M7 (Q1'23)</u> 3x NVIDIA A16 3x NVIDIA A30 3x NVIDIA A40 3x NVIDIA A100-80</p> <p><u>C240 M7 (Q3'23)</u> 2x NVIDIA H100 3x NVIDIA L40 8x NVIDIA L4 5x Intel Flex140 3x Intel Flex170</p>

System Tunning Recommendations

- BIOS tunings:

- **Processor Configuration:**

- Intel Virtualization Technology* : <Enabled>

- Intel Hyper-Threading Technology* : <Enabled>

- CPU Performance* : <HPC>

- **Power & Performance Configuration:**

- Processor C1E* : <Disabled>

- Processor C6* : <Enabled>

- Package C-State control* : <C0/C1 State>

- Uncore Frequency Scaling*: <Enabled>

- Thermal : Fan Control Policy : <Maximum Power>
- PCIe : Width, Speed, Max Payload Size, Max Read Request
- OS tunings : CPU Power, TCP Parameters & Memory Settings

C1E is one of the automatic power-saving features which is triggered when the system is idle. **Best to disable it when overclocking**

C6 BIOS will automatically disable CPU core and cache for power saving

In C0/C1 state all of the cores are locked at maximum performance and will cause a higher consumption of power

Uncore Frequency Scaling enabled stops the processor from dynamically changing frequencies based on workload to save power



BIOS Tuning – CPU Performance (HPC)

The screenshot displays the Cisco Intersight web interface for BIOS tuning. The top navigation bar includes the Cisco logo, 'Intersight', 'Infrastructure Service', and a search bar. The breadcrumb trail indicates the path: Policies > BIOS > ucs-m7-bios_NVCERT. The main section is titled 'Edit' and contains a sidebar with tabs for 'General' and 'Policy Details'. The 'Policy Details' tab is active, showing a list of BIOS settings. The 'CPU Performance' setting is highlighted with a dashed red box and set to 'hpc'. The 'Intel HyperThreading Tech' setting is also highlighted with a dashed red box and set to 'platform-default'. The 'Intel(R) VT' setting is highlighted with a dashed red box and set to 'enabled', with the 'enabled' option further emphasized by a solid red rectangle. Other settings visible include 'Intel Dynamic Speed Select' (platform-default), 'Intel Speed Select' (platform-default), 'Intel Turbo Boost Tech' (platform-default), 'I/O Error Enable' (platform-default), 'KT1 Prefetch' (platform-default), and 'Power Technology' (platform-default).

Overview

Policies > BIOS > ucs-m7-bios_NVCERT

Edit

General

2 Policy Details

Frequency Floor Override
platform-default

CPU Performance
hpc

platform-default

custom

Power Technology

General

2 Policy Details

Intel Dynamic Speed Select
platform-default

Intel HyperThreading Tech
platform-default

Intel Speed Select
platform-default

Intel Turbo Boost Tech
platform-default

Intel(R) VT
disabled

platform-default

enabled

disabled

I/O Error Enable
platform-default

KT1 Prefetch
platform-default

BIOS Tuning – Processor Power States

The screenshot displays the Cisco Intersight web interface for BIOS tuning. The top navigation bar includes the Cisco logo, 'Intersight', 'Infrastructure Service', and a search bar. The breadcrumb trail shows 'Policies > BIOS > ucs-m7-bios_NVCERT'. The main section is titled 'Edit' and contains a sidebar with navigation links: Overview, Operate, Servers, Chassis, Fabric Interconnects, HyperFlex Clusters, Integrated Systems, Analyze, Explorer (marked 'New'), and Configure. The 'Policy Details' tab is active, showing a list of settings. Two settings are highlighted with red boxes: 'C0 C1 State' under 'Package C State Limit' and 'disabled' under 'Processor C1E'. Both settings are currently set to 'platform-default'.

Package C State Limit

Setting	Value
C0 C1 State	platform-default
Auto	Auto
C0 C1 State	C0 C1 State

Processor C1E

Setting	Value
disabled	disabled
platform-default	platform-default
enabled	enabled
disabled	disabled

Thermal – Fan control Policy

The screenshot displays the Cisco Intersight interface for editing a Thermal policy. The breadcrumb trail is Policies > Thermal > ucsx-thermal. The left sidebar contains navigation links: Overview, Operate, Servers, Chassis, Fabric Interconnects, HyperFlex Clusters, Integrated Systems, Analyze, Explorer (with a 'New' button), Configure, Profiles, and Templates. The main content area is titled 'Edit' and has two tabs: 'General' and 'Policy Details' (which is active). Under 'Policy Details', there is a section for 'Add policy details' with filters for 'All Platforms', 'UCS Server (Standalone)', 'UCS Server (FI-Attached)', and 'UCS Chassis'. The 'Fan Control' section is visible, showing a 'Fan Control Mode' dropdown menu. The dropdown is open, showing options: 'Balanced', 'Low Power', 'High Power', 'Maximum Power' (highlighted with a red box), and 'Acoustic'. The 'Maximum Power' option is also highlighted with a dashed purple box.

Intersight | Infrastructure Service

Search

Policies > Thermal > ucsx-thermal

Overview

Edit

General

2 Policy Details

Add policy details

All Platforms | UCS Server (Standalone) | UCS Server (FI-Attached) | UCS Chassis

Fan Control

Fan Control Mode

Maximum Power

Balanced

Low Power

High Power

Maximum Power

Acoustic

OS Tuning – 1

- *Mellanox Automatic tuning for Performance*

```
# mlx_tune -p HIGH_THROUGHPUT
```

- *Disable the TCP Timestamps for better CPU Utilization*

```
# sysctl -w net.ipv4.tcp_timestamps=0
```

- *Enable the TCP selective acks option for better throughput:*

```
# sysctl -w net.ipv4.tcp_sack=1
```

- *Increase the maximum length of processor input queues:*

```
# sysctl -w net.core.netdev_max_backlog=300000
```

- *Increase the TCP maximum and default buffer sizes using setsockopt():*

```
# sysctl -w net.core.rmem_max=4194304
```

```
# sysctl -w net.core.wmem_max=4194304
```

```
# sysctl -w net.core.rmem_default=4194304
```

```
# sysctl -w net.core.wmem_default=4194304
```

```
# sysctl -w net.core.optmem_max=4194304
```

Part of the `mlx_tune` utility. This command does all the NIC optimization recommended by Mellanox



Linux

The rest of these are recommendations by Mellanox for “Tuning the Network Adapter for Improved IPv4 Traffic Performance”

<https://mellanox.my.site.com/mellanoxcommunity/s/article/linux-sysctl-tuning>

OS Tuning - 2

- *Increase memory thresholds to prevent packet dropping:*

```
# sysctl -w net.ipv4.tcp_rmem="4096 87380 4194304"
```

```
# sysctl -w net.ipv4.tcp_wmem="4096 65536 4194304"
```

- *Enable low latency mode for TCP:*

```
# sysctl -w net.ipv4.tcp_low_latency=1
```

```
# sysctl -w net.ipv4.tcp_adv_win_scale=1
```

- *Disable OS watchdog & Numa balancing:*

```
# echo '0' > '/proc/sys/kernel/nmi_watchdog'
```

```
# echo 0 > /proc/sys/kernel/numa_balancing
```

- *Fine tuning Kernel task scheduler :*

```
# echo 100000000 > /proc/sys/kernel/sched_min_granularity_ns
```

```
# echo 50000000 > /proc/sys/kernel/sched_migration_cost_ns
```

These are recommendations by Mellanox for “Tuning the Network Adapter for Improved IPv4 Traffic Performance”

<https://mellanox.my.site.com/mellanoxcommunity/s/article/linux-sysctl-tuning>



Linux

Set to maximum acceptable values in nanoseconds (1 sec and 1/2 sec)



PCIE Tuning for the Best Performance

Default PCIE Attributes : Width | Speed | Max Payload Size | Max Read Request

```
[root@ucs19 ~]# lspci | grep Mell
5e:00.0 Ethernet controller: Mellanox Technologies MT28908 Family [ConnectX-6]
5e:00.1 Ethernet controller: Mellanox Technologies MT28908 Family [ConnectX-6]
[root@ucs19 ~]# lspci -s 5e:00.0 -vvv | grep PCIeGen
                [V0] Vendor specific: PCIeGen4 x16
[root@ucs19 ~]# lspci -s 5e:00.0 -vvv | grep Speed
LnkCap: Port #0, Speed 16GT/s, Width x16, ASPM not supported, B
, L1 unlimited
LnkSta: Speed 8GT/s, Width x16, TrErr- Train- SlotClk+ DLActive
LnkCtl2: Target Link Speed: 16GT/s, EnterCompliance- SpeedDis-
[root@ucs19 ~]# lspci -s 5e:00.0 -vvv | grep MaxReadReq
                MaxPayload 256 bytes, MaxReadReq 512 bytes
[root@ucs19 ~]# lspci -s 5e:00.0 -vvv | grep DevCtl: -C 2
DevCap: MaxPayload 512 bytes, PhantFunc 0, Latency L0s unlimite
ExtTag+ AttnBttn- AttnInd- PwrInd- RBE+ FLReset+ SlotPow
DevCtl: Report errors: Correctable- Non-Fatal- Fatal- Unsupport
RlxdOrd+ ExtTag+ PhantFunc- AuxPwr- NoSnoop+ FLReset-
MaxPayload 256 bytes, MaxReadReq 512 bytes
[root@ucs19 ~]#
```

List PCI command to find out where the Mellanox NICs are connected

List PCI command to show location 5e:00.0 settings. Speed determines the number of PCIe transactions possible. The speed is measured in GT/s which stands for "billion transactions per second"

The PCIe Max Payload Size determines the maximum size of a PCIe packet, or PCIe MTU (similar to networking protocols). This parameter is set only by the system and depends on the chipset architecture (e.g. x86_64, Power8, ARM, etc).

Ubuntu/Debian: `sudo apt install pciutils`

PCIe Tuning for the Best Performance

Tuned PCIe Attributes : Width | Speed | Max Payload Size | Max Read Request

```
[root@ucs19 ~]# lspci -s 5e:00.0 -vvv | grep DevCtl: -C 2
DevCap: MaxPayload 512 bytes, PhantFunc 0, Latency L0s unlimited, L1 unlimited
        ExtTag+ AttnBtn- AttnInd- PwrInd- RBE+ FLReset+ SlotPowerLimit 0.000W
DevCtl: Report errors: Correctable- Non-Fatal- Fatal- Unsupported-
        RlxdOrd+ ExtTag+ PhantFunc- AuxPwr- NoSnoop+ FLReset-
        MaxPayload 256 bytes, MaxReadReq 512 bytes

[root@ucs19 ~]# setpci -s 5e:00.0 68.W=5036
[root@ucs19 ~]# lspci -s 5e:00.0 -vvv | grep DevCtl: -C 2
DevCap: MaxPayload 512 bytes, PhantFunc 0, Latency L0s unlimited, L1 unlimited
        ExtTag+ AttnBtn- AttnInd- PwrInd- RBE+ FLReset+ SlotPowerLimit 0.000W
DevCtl: Report errors: Correctable- Non-Fatal+ Fatal- Unsupported-
        RlxdOrd+ ExtTag+ PhantFunc- AuxPwr- NoSnoop+ FLReset-
        MaxPayload 256 bytes, MaxReadReq 4096 bytes
```

The first digit is the PCIe Max Read Request size selector.

The acceptable values are:
0 - 128B, 1 - 256B, 2 - 512B,
3 - 1024B, 4 - 2048B
and 5 - 4096B.

Ubuntu/Debian: `sudo apt install pciutils`

What Does The Network Configuration Look Like?

QOS Configs Classification

CLASS MAP

```
class-map type qos match-all class-q3
  match dscp 24
class-map type qos match-all class-q7
  match dscp 48
```



Put packets with DSCP 24 into
“class-q3”

Put Packets with DSCP 48 into
“class-q7”

POLICY MAP

```
policy-map type qos QOS_classification_policy
  class class-q3
    set qos-group 3
  class class-q7
    set qos-group 7
  class class-default
    set qos-group 0
```



Put packets from class-q3 into
queue 3

Put packets from class-q7 into
queue 7

Everything else goes into default or
queue 0

QOS Configs Queuing and Scheduling

POLICY MAP “type queuing”

```
policy-map type queuing custom-8q-out-policy
```

```
class type queuing c-out-8q-q7
```

```
priority level 1
```

```
class type queuing c-out-8q-q6
```

```
bandwidth remaining percent 0
```

```
class type queuing c-out-8q-q5
```

```
bandwidth remaining percent 0
```

```
class type queuing c-out-8q-q4
```

```
bandwidth remaining percent 0
```

```
class type queuing c-out-8q-q3
```

```
bandwidth remaining percent 60
```

```
random-detect minimum-threshold 150 kbytes maximum-threshold 3000 kbytes drop-probability 7 weight 0 ecn
```



Queues 7 is set up as a priority queue



Queues 3 received 60% of the bandwidth and ECN marking enabled starting at 150KB of buffer utilization, drop probability of 7%

QOS Configs Queuing and Scheduling

POLICY MAP cont.

```
class type queuing c-out-8q-q2  
  bandwidth remaining percent 0  
class type queuing c-out-8q-q1  
  bandwidth remaining percent 0  
class type queuing c-out-8q-q-default  
  bandwidth remaining percent 40
```



Queues 1 and 2 receive 0%
guaranteed bandwidth

Default queue receives 40%

QOS Configs cont.

POLICY MAP “type queuing” cont.

```
policy-map type network-qos custom-8q-nq-policy
<snip>
class type network-qos c-8q-nq3
    mtu 9216
    pause pfc-cos 3
<snip>
```



Use a policy-map of type network-qos to add PFC to queue 3

MTU is used as a way to calculate headroom for the non-drop queue.

SYSTEM QOS

```
system qos
service-policy type network-qos custom-8q-nq-policy
service-policy type queuing output custom-8q-out-policy
```



Attach policy maps of "type queuing" and "type network-qos" system wide.

This will trigger ECN marking and that ports configured with PFC will receive and honor those frames, as well as generate pause frames as needed

QOS Configs Interfaces

```
interface Ethernet1/1
service-policy type qos input QOS_classification_policy
priority-flow-control mode on
priority-flow-control watch-dog-interval on
```

Attach the service policy to the interface.

Enable PFC

Enable PFC Watch Dog using a default interval of 100 milliseconds.



AI/ML Blueprint for DC Networks

The Blueprint For Today

Built to accommodate 1024 GPUs along with storage devices

The screenshot shows the Cisco website's product page for the 'Cisco Data Center Networking Blueprint for AI/ML Applications'. The page features a dark blue header with the Cisco logo and navigation links. The main content area has a white background with a large title and a breadcrumb trail. A left sidebar contains a 'Table of Contents' with links to various sections. The main content area lists the sections: Introduction, RoCEv2 as Transport for AI Clusters, AI Clusters Require Lossless Networks, How to Manage Congestion Efficiently in AI/ML Cluster Networks, and Conclusion. A right sidebar contains links to 'Bias-Free Language' and 'Contact Cisco', and a download section with 'Save', 'Download', and 'Print' icons.

Cisco Products and Services Solutions Support Learn Partners Explore Cisco Search

Products & Services / Cloud and Systems Management / Cisco Nexus Dashboard Fabric Controller / White Papers /

Cisco Data Center Networking Blueprint for AI/ML Applications

Updated: May 24, 2023

Bias-Free Language Contact Cisco

Introduction Table of Contents Introduction RoCEv2 as Transport for AI Clu... AI Clusters Require Lossless N... + How to Manage Congestion E... + How Visibility into Network Be... + Network Design to Accommod... + Conclusion Related Materials

Introduction RoCEv2 as Transport for AI Clusters AI Clusters Require Lossless Networks Explicit Congestion Notification (ECN) Priority Flow Control (PFC) How to Manage Congestion Efficiently in AI/ML Cluster Networks How ECN Works How PFC Works Using ECN and PFC Together to Build Lossless Ethernet Networks

Save Download Print

The Blueprint and the CVD

Leaf 201

Table of Contents

Introduction

Building a Non-Blocking Network...

Building a Lossless Ethernet Fabric...

Using NDFC to Build Your AI/ML...

Configuration Examples

Leaf 201

Spine 201

Versions of Software Used in T...

```
class-map type qos match-all class-q3
match dscp 24
class-map type qos match-all class-q7
match dscp 48
policy-map type qos QOS_classification_policy
class class-q3
set qos-group 3
class class-q7
set qos-group 7
class class-default
set qos-group 0
policy-map type queuing custom-8q-out-policy
class type queuing c-out-8q-q7
priority level 1
class type queuing c-out-8q-q6
bandwidth remaining percent 0
class type queuing c-out-8q-q5
bandwidth remaining percent 0
class type queuing c-out-8q-q4
bandwidth remaining percent 0
class type queuing c-out-8q-q3
bandwidth remaining percent 60
random-detect minimum-threshold 150 bytes maximum-threshold 3000 bytes drop-probability 7 weight 0 ecn
class type queuing c-out-8q-q2
bandwidth remaining percent 0
class type queuing c-out-8q-q1
bandwidth remaining percent 0
class type queuing c-out-8q-q-default
bandwidth remaining percent 40
system qos
```

Using NDFC to Build Your...

Table of Contents

Introduction

Building a Non-Blocking Network...

Building a Lossless Ethernet Fabric...

Using NDFC to Build Your AI/ML...

Configuration Examples

Versions of Software Used in T...

Add Switches - Fabric: ROCEv2

Switch Addition Mechanism?

Seed Switch Details

Fabric	Switch	Authentication Protocol	Username
ROCEv2	172.22.131.216	MD5	admin
Password	Max Hops	Presence config	
384	2	Disabled	

Discovery Results

Switch Name	Serial Number	IP Address	Model	Version	Status	Progress
switch	FOC1722910E	172.29.165.182	N9K-C3047P-10E	7.0(302T)	Not Reachable	
switch	FOC1722918D	172.29.165.232	N9K-C3047P-10E	7.0(302T)	Not Reachable	
Leaf205	FOC03445J8M	172.22.131.238	N9K-C9300C2-GB	15.2(5)	Manageable	
Leaf206	FOC03389408	172.22.131.239	N9K-C9300C2-GB	15.2(5)	Manageable	

After you add the switches, NDFC deploys the fabric-wide configuration from the template. In this example, the BGP AS number is the same for all leaf switches, and you configure the BGP AS number through a policy template. All leaf switches are selected to receive this configuration.

Create Policy

Select Switches

Search Switches

Select All

Show Selected

leaf-201

leaf-202

leaf-203

leaf-204

leaf-205

leaf-206

spine-201

Nexus Dashboard Fabric Controller Automation

Create Fabric

N9K Cloud Scale Platform Queuing Policy

Select an Option

Queuing Policy for all 92xx, -EX, -FX, -FX2, -FX3, -GX series switches in the fabric

N9K R-Series Platform Queuing Policy

Select an Option

Queuing Policy for all R-Series switches in the fabric

Other N9K Platform Queuing Policy

Select an Option

Queuing Policy for all other switches in the fabric

Enable AI / ML QoS and Queuing Policies

☒

Configures QoS and Queuing Policies specific to N9K Cloud Scale switch fabric for AI / ML network loads

AI / ML QoS & Queuing Policy*

AI_Fabric_QOS_100G

Queuing Policy based on predominant fabric link speed: 400G / 100G / 25G

AI_Fabric_QOS_400G

AI_Fabric_QOS_100G

AI_Fabric_QOS_25G

Enable MACsec in the fabric

Cisco Type 7 Encrypted Octet String

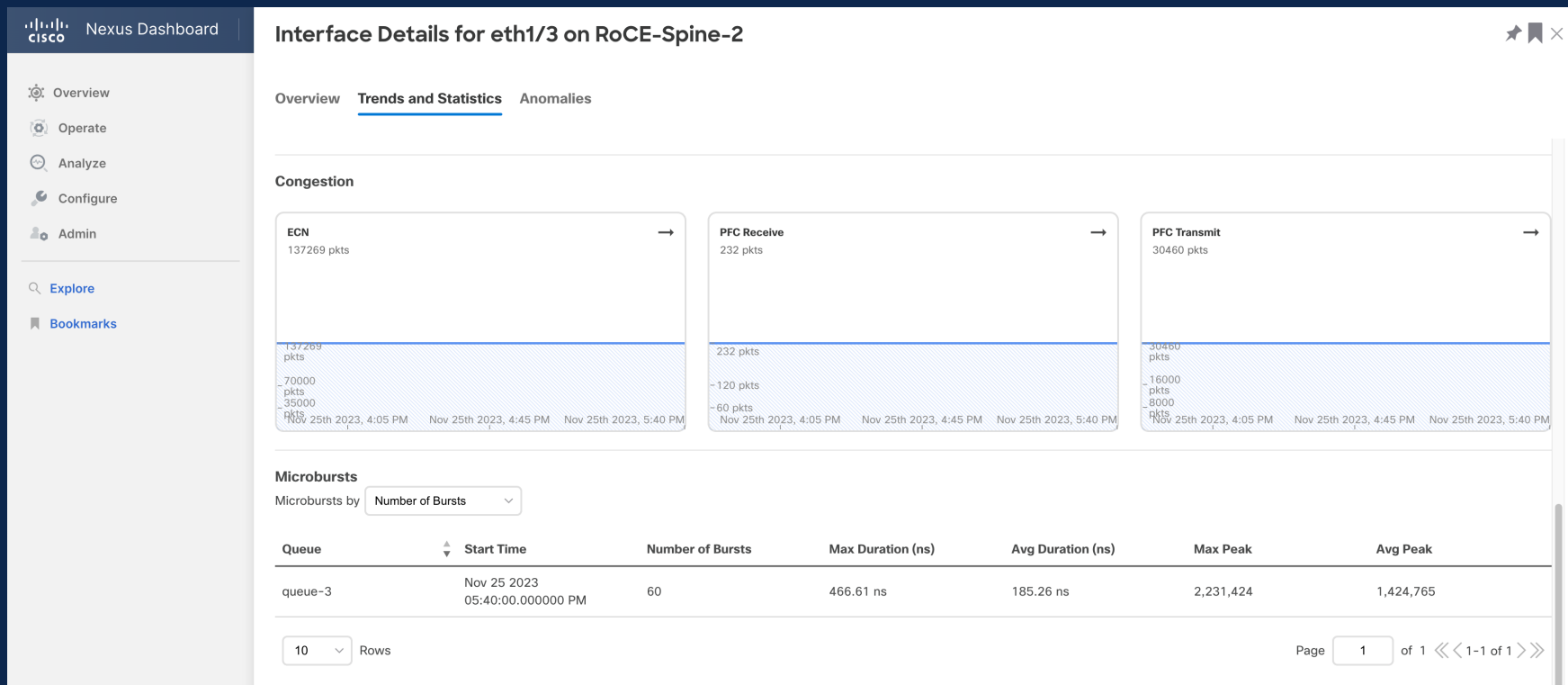
The Blueprint For Today

The screenshot shows the GitHub interface for the repository 'allenrobel / NDFC-AIML-Fabric'. The repository is public and has 0 stars, 0 forks, and 2 watchers. The main branch is 'main'. The repository description is 'Ansible playbook to create an NDFC fabric which supports AI/ML workloads'. The file list includes:

File	Commit Message	Time Ago
doc	Update with current 'show running-config ipqos'	1 hour ago
inventory	Initial commit	5 days ago
.gitignore	Ignore .graffle files	47 minutes ago
AIML_Fabric.yml	Template name will probably change, so let's not hardcode it	2 hours ago
AI_Cluster_QOS_template.template	Adding new QOS template	2 hours ago
README.md	Add topology diagram	5 days ago
ansible.cfg	Initial commit	5 days ago

The README.md file is visible, showing the title 'NDFC-AIML-Fabric'.

Nexus Dashboard Insights - Visibility



Conclusion

Conclusion

- The AI/ML market and technology is moving quite quickly
- If you are building a very large training cluster please speak with your sales team
- Use the best practices in the presentation to get started
- We have a blueprint and best practices to get you started on your journey today
- The blueprint will continue evolving as standards and equipment change



Did you know?

You can have a
one-on-one session with
a technical expert!

Visit Meet the Expert in The HUB
to meet, greet, whiteboard & gain
insights about your unique questions
with the best of the best.



Meet the Expert Opening Hours:

Tuesday	3:00pm – 7:00pm
Wednesday	11:15am – 7:00pm
Thursday	9:30am – 4:00pm
Friday	10:30am – 1:30pm

Session Surveys

We would love to know your feedback on this session!

- Complete a minimum of four session surveys and the overall event surveys to claim a Cisco Live T-Shirt



Continue your education



- Visit the Cisco Showcase for related demos
- Book your one-on-one Meet the Expert 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



The bridge to possible

Thank you

CISCO *Live!*

#CiscoLiveAPJC

The background is a vibrant, abstract graphic featuring a spectrum of colors (red, orange, yellow, green, blue) arranged in a radial pattern that resembles a stylized sunburst or a rainbow. The colors transition from red on the left to blue on the right, with yellow and green in the center. The overall effect is bright and energetic.

cisco *Live!*

Let's go

#CiscoLiveAPJC