The bridge to possible

The Blueprint to Building End-To-End Hybrid-Cloud AI Infrastructure

Nick Geyer, Cisco Systems Inc.
Eugene Minchenko, Cisco Systems Inc.
BRKCOM-1008

#CiscoLive

Cisco Webex App

Questions?

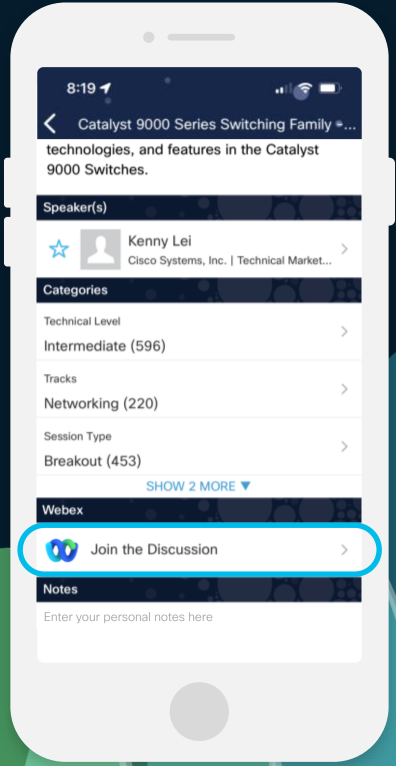
Use Cisco Webex App to chat
with the speaker after the session

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 June 7, 2024.

https://ciscolive.ciscoevents.com/
ciscolivebot/#BRKCOM-1008



Agenda

- Introduction
- AI Fundamentals & Impacts on Infrastructure Design Decisions
- Training Infrastructure & Network Considerations for AI Environments
- Inferencing, Fine-Tuning, & Compute Infrastructure
- Sizing for Inferencing
- AI Infrastructure Automation & Cisco Validated Designs
- Future Trends and Industry Impacts of AI Infrastructure Demands
- Summary

AI sets a new standard for Infrastructure

only 13% of Data Center management leaders say their network can accommodate AI computational needs.

AIOps

How can we harness all the data available to us to simplify data center operations?

Scale and Performance

Is our network AI-ready, with the ability to support data training and inferencing use cases?

Sustainability

How are we addressing corporate and regulatory sustainability requirements in our data center design?

What we know

Every organization's AI approach and needs are different

Build the Model | Training   **Optimize the Model** | Fine-tuning & RAG   **Use the Model** | Inferencing








What we’re hearing from IT infra and operations

Need consistency; avoid new islands of operations

Optimize for utilization and efficiency in many dimensions—support multiple projects, leverages GPUs wisely, power and cooling needs, lifecycle management

Comprehensive security protocols and measures

Support rapidly-evolving software ecosystem

Manage cloud vs. on-prem vs. hosted model

Straddle the training → fine tuning → inferencing → repeat model

Cisco's 2-Fold AI Strategy & Our Focus Today

Using AI to maximize YOUR experience with **Cisco products**

In

Develop AI tools across the Cisco portfolio that help manage networks more effectively

- *Delivering better results*
- *Providing intelligent guidance*
- *Providing better security*
- *Solving day-to-day challenges*

Enabling **YOUR infrastructure** to support adoption of AI applications

On

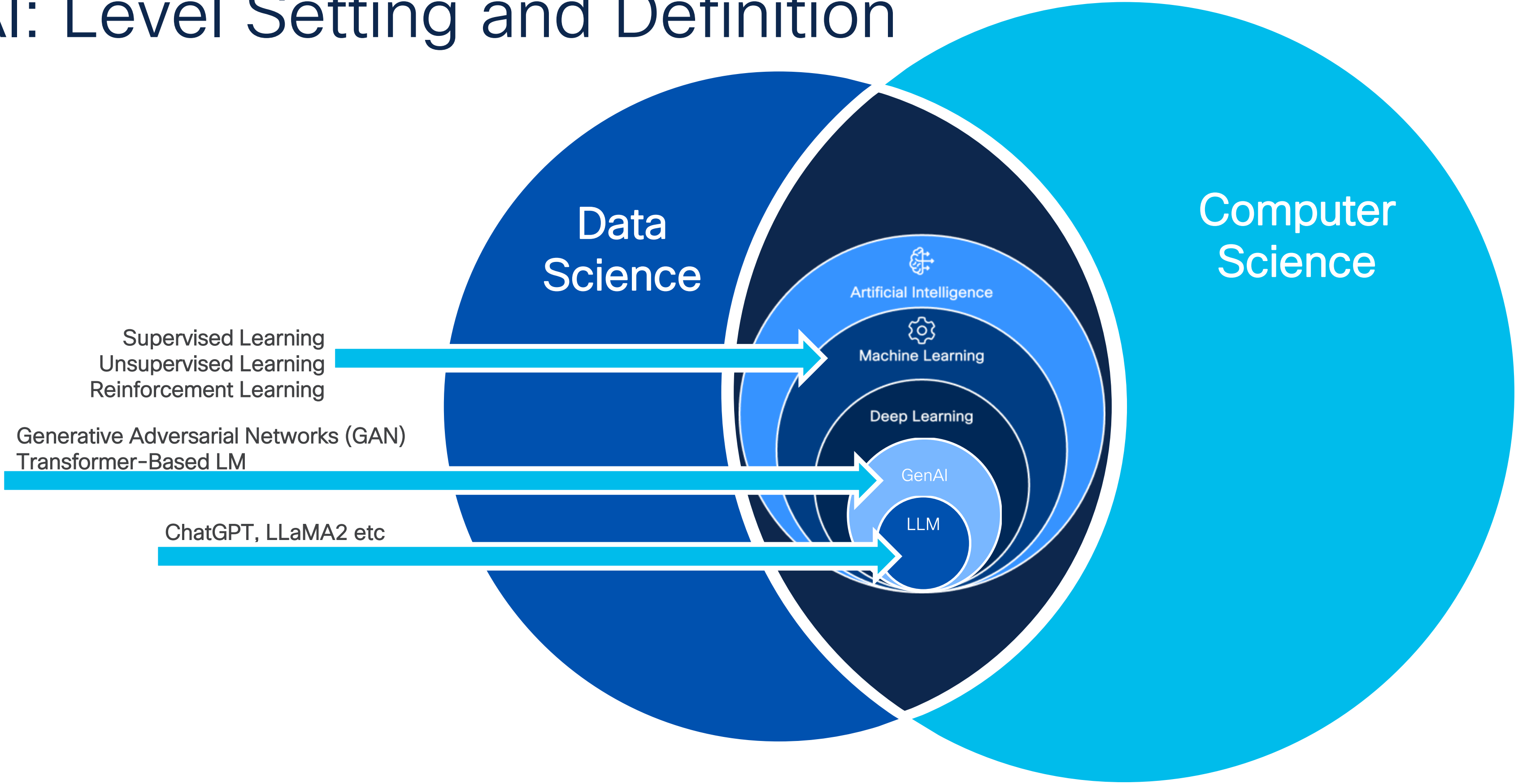
Develop products that help accelerate YOUR adoption of AI for your business solutions

- *High-speed networking for AI training and inference clusters*
- *Flexible compute building blocks to build AI compute clusters*

AI Fundamentals & Impacts on Infrastructure Design Decisions

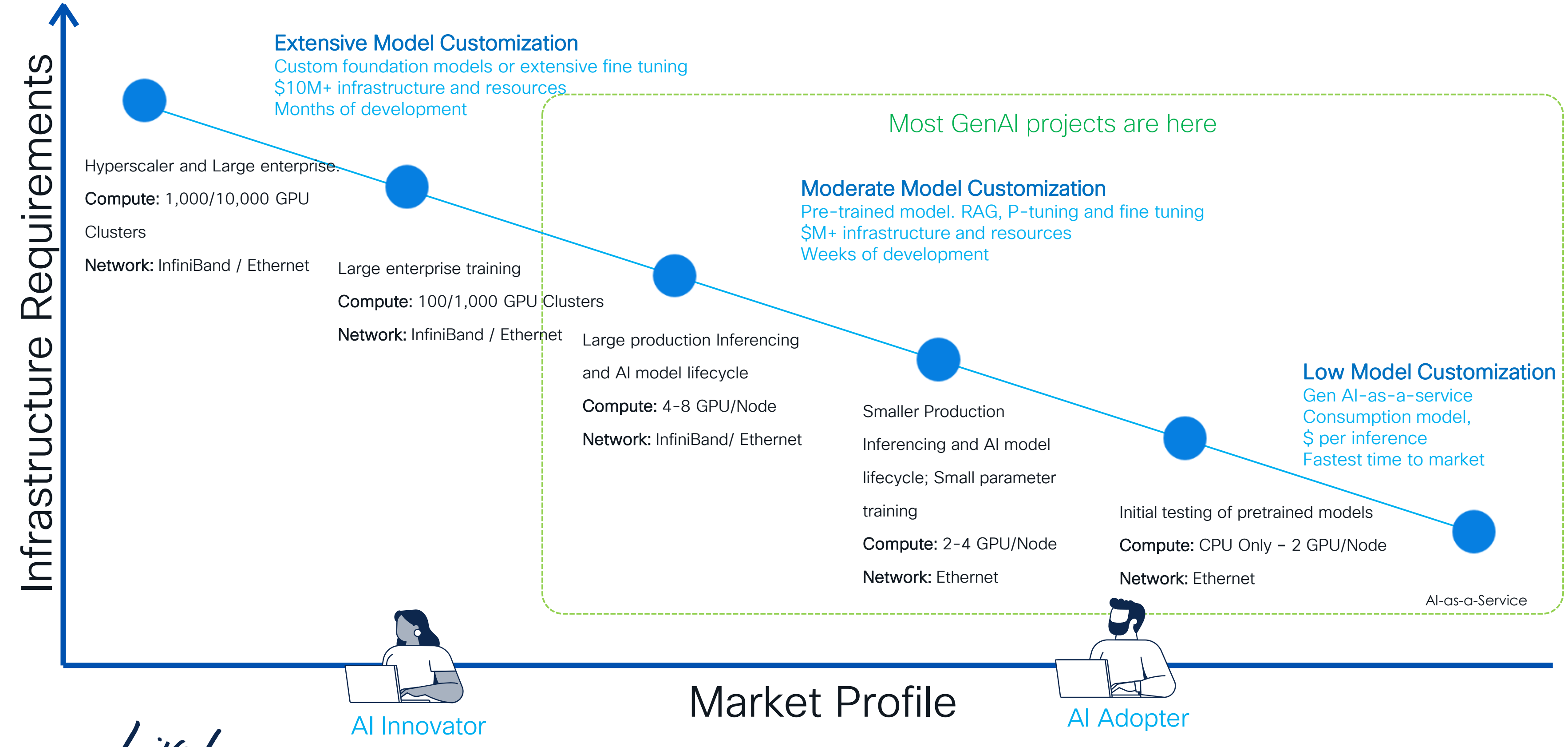
CISCO Live!

AI: Level Setting and Definition



AI Infrastructure Requirements

AI Infrastructure Requirements Spectrum



LLM Training vs. Fine Tuning vs. Inferencing

Model Stage	Relative Compute	Relative Utilization	Analogy
Inferencing	1X		Answering biology questions “What is the role of mitochondria?”
Fine Tuning	10X		Learning Biology: Words, terms, concepts
Training	100X		Learning the English language: Patterns of words

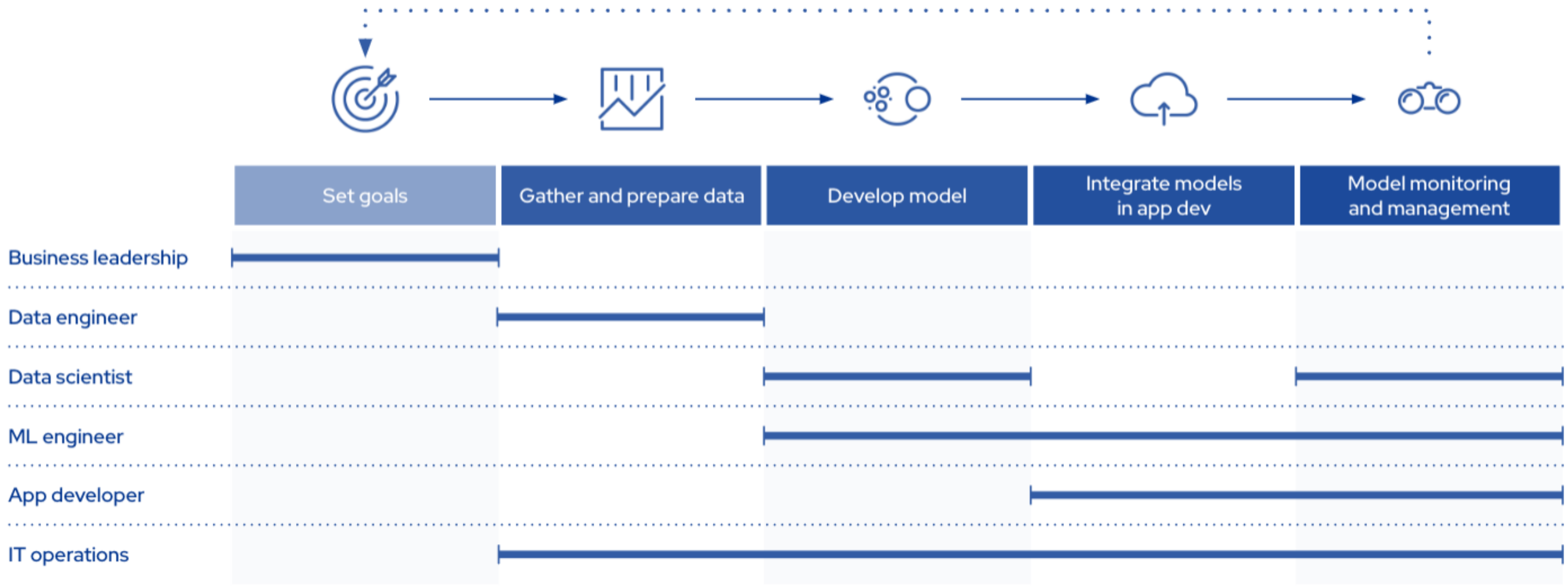
AI Maturity Model

Align customer capabilities to technology investment

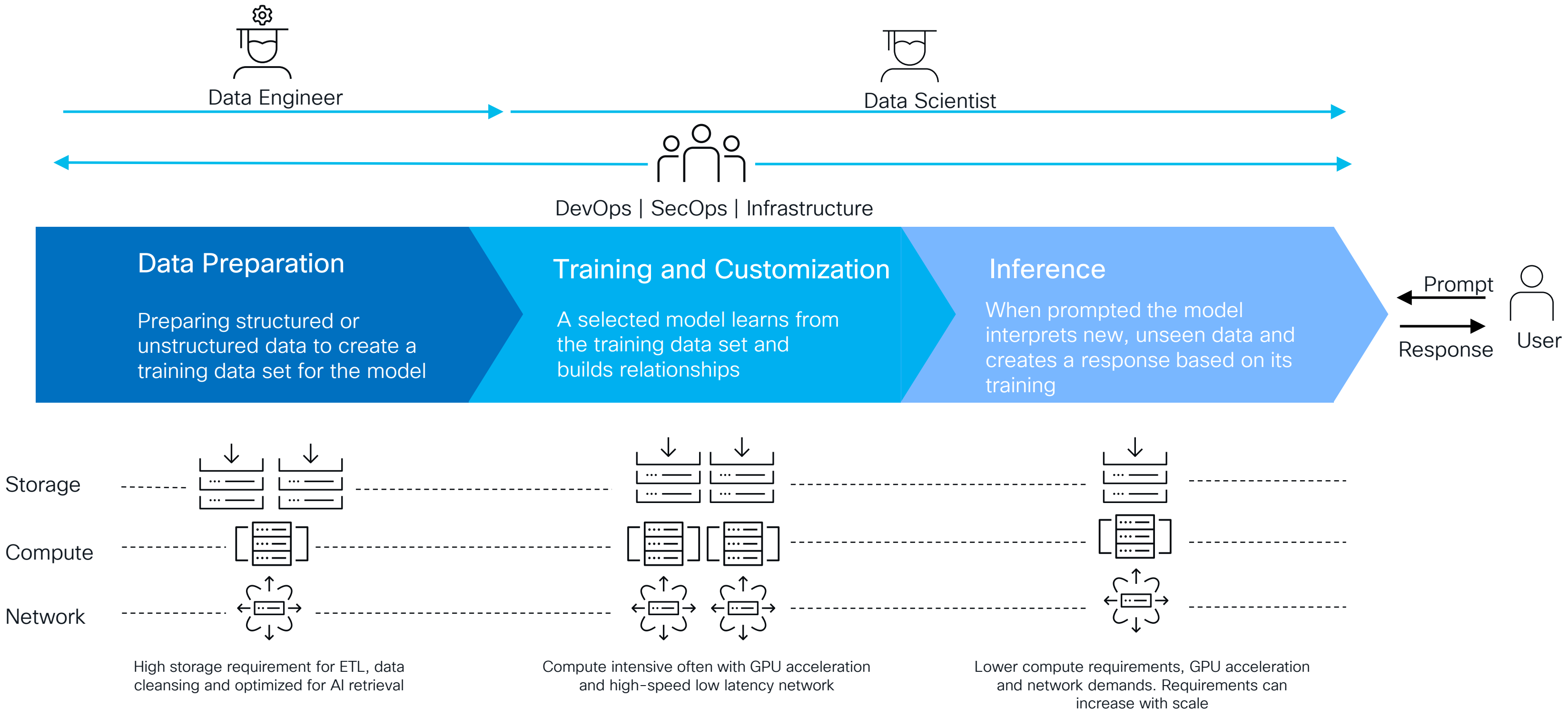
Exploratory	Experimental	Plan	Transform
Business use for AI not yet defined Data culture to support AI not established Exec agenda for AI not a priority No AI processes or technologies in place for implementation No investment in infrastructure to support AI workloads	Formulated short term AI strategy, proof of concept scenarios Exec, board support for AI, not across all lines of business. Small skillset of data science on staff Data advancement with policy and degree of governance using point solutions Trial AI adjacent technologies with future budget allocation	Defined AI standalone strategy, platform in place for quick wins, dedicated AI budget Decentralized support across staff, adequate resources for early stages Data gathering, analytics to centralized platform for variety of use cases AI used for internal processes – Billing automation, segment analysis	AI Strategy based on long term roadmap for new services, Framework defined to assure quality, format, ownership Data available in Realtime for predictive analysis A centralized platform model with pre-integrated AI capabilities.

Operationalizing AI/ML is not trivial

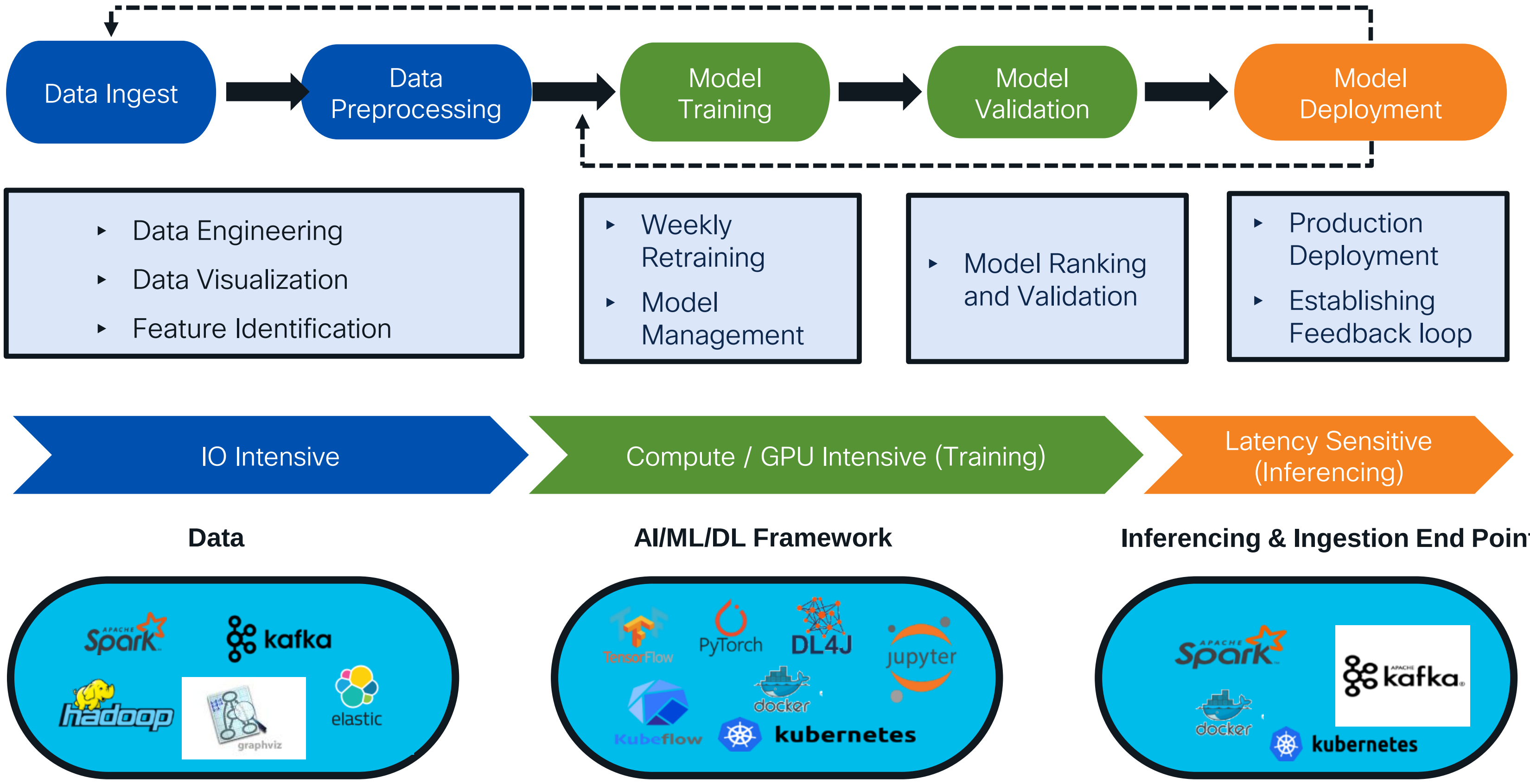
Everyone in your organization plays a critical role in a complex process



AI and Infrastructure Pipelines



Framework and Common Software



The need for flexible AI acceleration

1

For mixed workloads

Real-time video and audio streams	Group chat	Screen share	Recording	Generative AI for meeting summarization	Inference for real-time transcription or translation

Example: AI-enabled Video Conferencing App

General Purpose Compute (CPU) ⟷ AI Acceleration (GPU)

2

For the diversity of AI workloads

	Data Ingest and Preparation	Edge Inferencing	Data Center Inferencing	Fine Tuning	Large Foundational Training
GPUS		0 - 2	1 - 4	4 - 64	64 - 10K+

Network Considerations: Shared Fabric → Dedicated Fabrics

Revolutionizing AI workloads with 5th Gen Intel Xeon Scalable Processors

High Performance Features

- Intel AMX with built-in AI accelerator in each core
- Accelerated computations and reduced memory bandwidth pressure
- Significant memory reductions with BF16/INT8

Enhanced System Capabilities

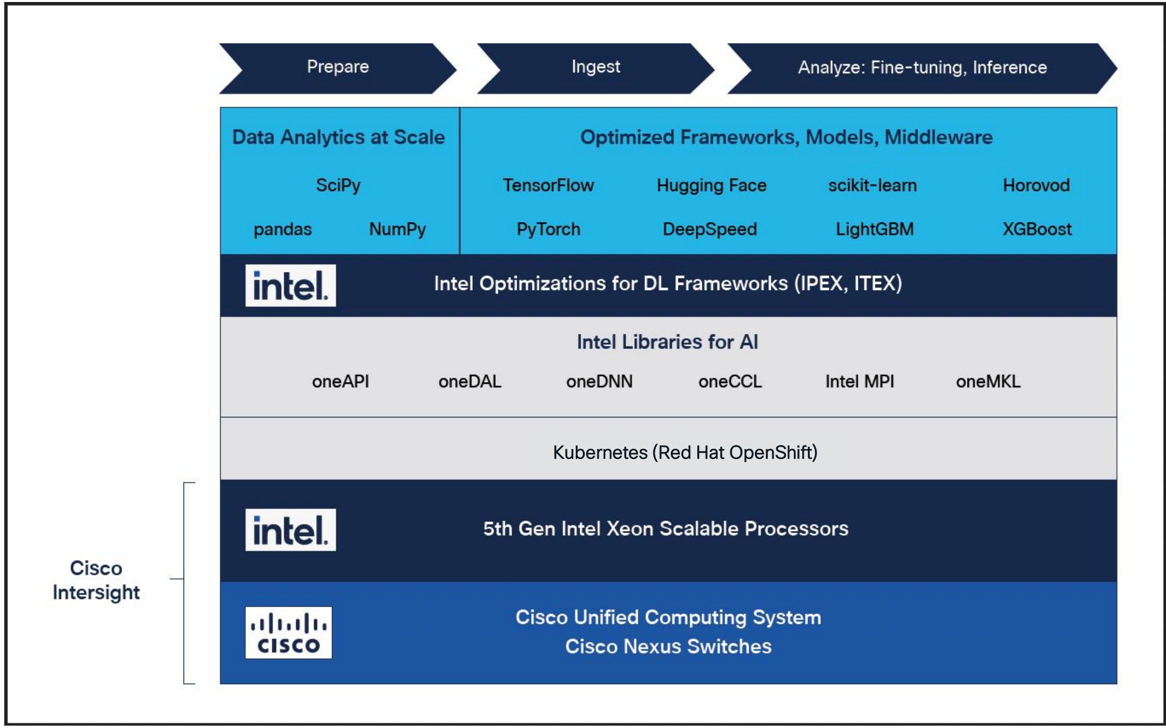
- Larger last-level cache for improved data locality
- Higher core frequency and faster memory with DDR5
- Intel AVX-512 for non-deep learning vector computations

Software Optimization

- Software suite of optimized open-source frameworks and tools
- Intel Xeon optimizations integrated into popular deep learning frameworks

TCO Benefits and Compatibility

- Lower operational costs and a smaller environmental footprint
- Available on UCS X-Series, C240, C220 platforms



Will Organizations Build Large Clusters with over 1000 GPUs?

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).

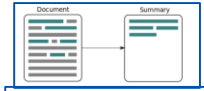
Similarly, fine-tuning an already trained model with domain-specific data sets requires fewer resources, often less than 100+ H100 scale GPUs. With these scales, both inference and fine-tuning do not need large GPU clusters on the same fabric.

https://blog.apnic.net/2023/08/10/large-language-models-the-hardware-connection/

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

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

Sample Large Language Model use Cases



Summarization

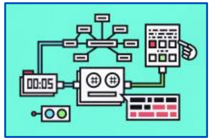
LLMs are highly effective in text summarization
tasks, in areas such as Academic Research, Business Report summary, Legal Analysis, Education materials, Emails, etc

Translation

Language translation is a key use-case for LLMs in areas such Travel & Tourism, Legal, Emergency Services, Education, Real-time translation.

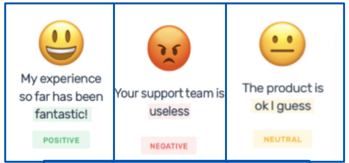

Dialog

Some examples of use cases for LLM chatbots include Customer Service, Personal Assistants, Tech Support, News and Information.



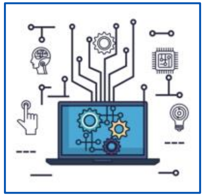
Text Generation

Use of LLMs for content creation, marketing, documentation, Business communication, product documentation, etc



Sentiment Analysis

Use LLMs to determine sentiment in areas such as comments, responses, content moderation, feedback, Market Research.


Code Generation

LLMs can be used to increase coding productivity, with tools such as co-pilot in areas web development, data analysis, Education tools, etc.

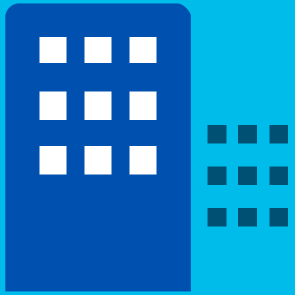
Enterprise Considerations to Define Requirements

- What is the use case?
- Am I Training? Fine Tuning? Inferencing? RAG?
- How much data am I training on?
- How many models am I training on?
- Am I using Private Data?
- Who is responsible for Management?

- Cost
- Accuracy
- Model Size
- User Experience (Response Time)
- Data Fidelity
- Concurrent User/Inputs

Where can this be run

Enterprises can choose where any model should be trained. Primarily there are two options:



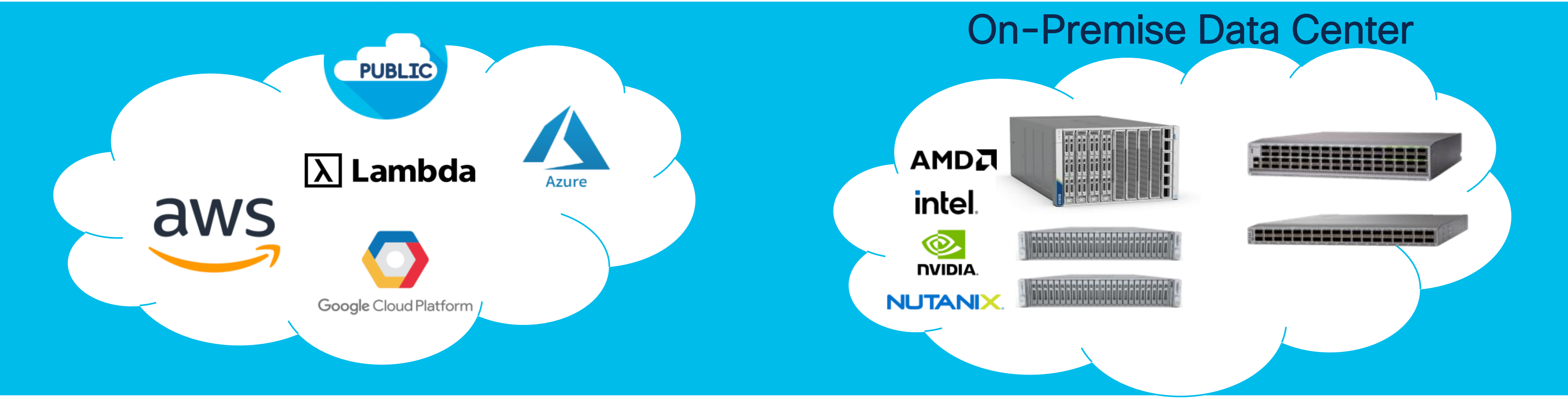
On Premises

- Always available for enterprise to use
- Flexibility for large enterprise to leverage same cluster for different functions
- Data is stored locally/ Data Sovereignty

Public Clouds

- Provides flexibility, pay for what you need
- Cost will grow with more data and training
- Challenge: Cost of egress data from the cloud, latency and lock in.

Smart Cloud, **not** Cloud First



Quantitative Trading Firm, London, UK (12,000 GPUs)

Example Hyperscaler Cost Model

Cloud Provider Lamba Labs @$1.99ph per (H100) GPU.
Potential Annual Cost: $210Million PA

Example On-Prem Cost Model

CoLo, Servers, Storage and NW
Potential Annual Cost: $130 Million PA (3 Years)

Bringing it all together

A helicopter view of an AI Deployment Journey

1. Deploy AI-ready infrastructure

Cisco validated designs

Cisco Intersight

ANSIBLE

HCI AI

FlexPod AI

FlashStack AI

2. Install common AI models from industry repositories

NVIDIA AI Enterprise

Red Hat OpenShift AI

NVIDIA NGC

3. Prep and inject data to fine tune the model

Customer data

010110
110010
001011

010110
110010
001011

010110
110010
001011

4. Periodic model updates and infrastructure scaling as required

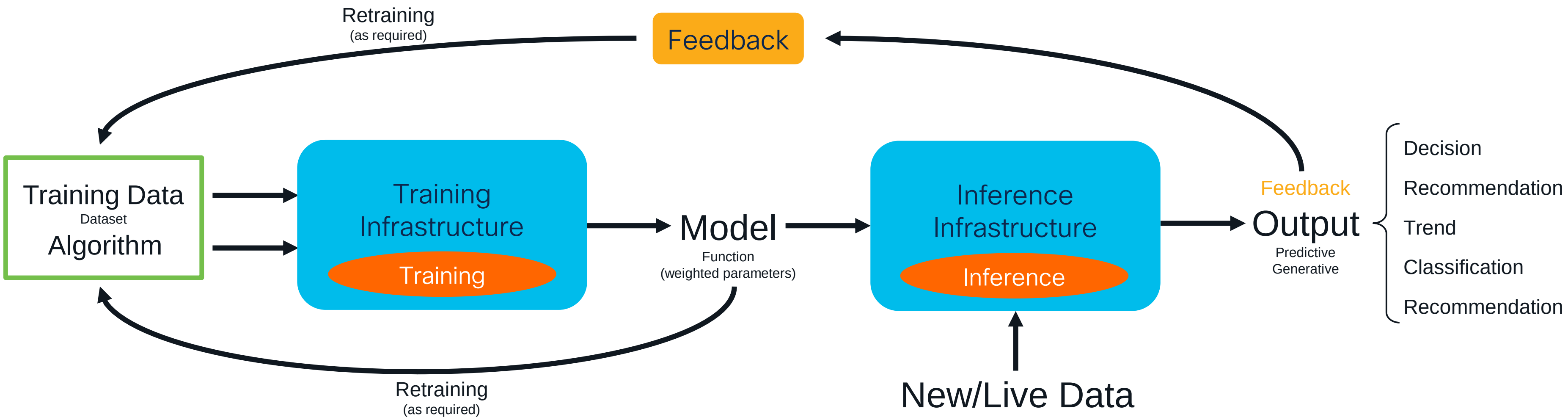
5. Deploy application for inferencing

Core

Edge

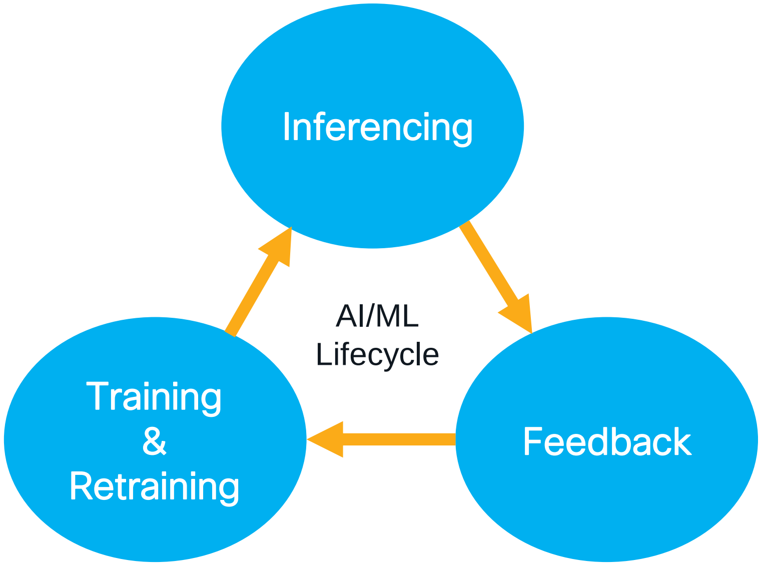
AI Training Infrastructure & Network Considerations for AI Environments

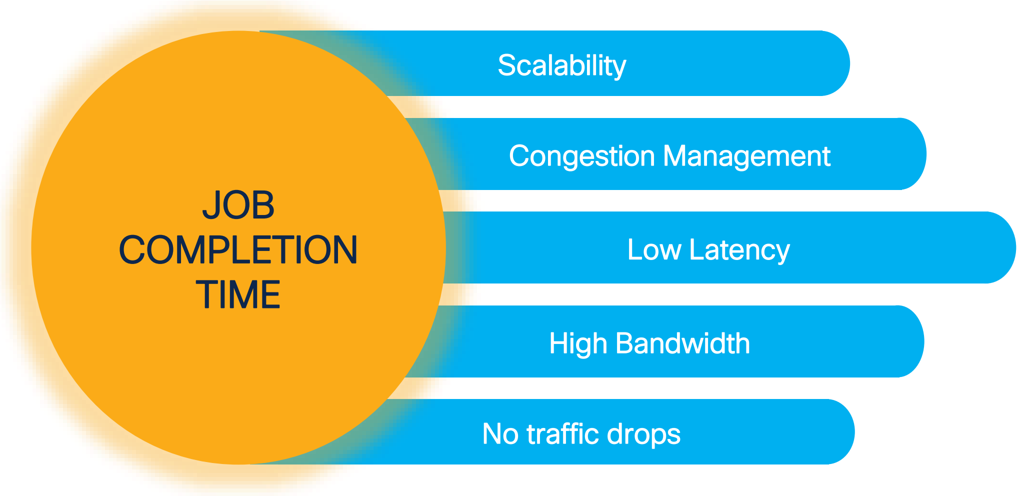
Breaking-down Machine Learning – The Process



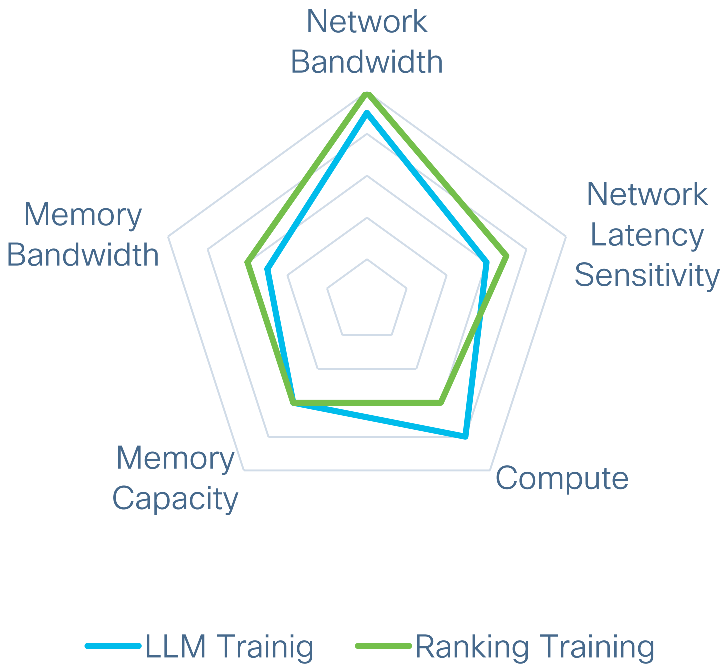
Architecting an AI/ML training cluster - Considerations

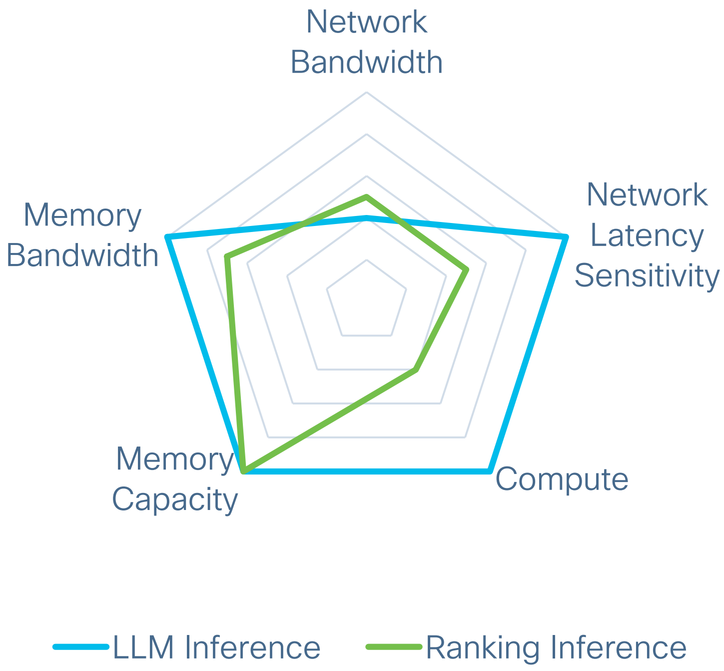
AI models and applications consume massive amounts of data,
and the data is constantly growing…
So, there are many challenges for the infrastructure to grow at the same scale as the data





Training and Inference Network Behaviors



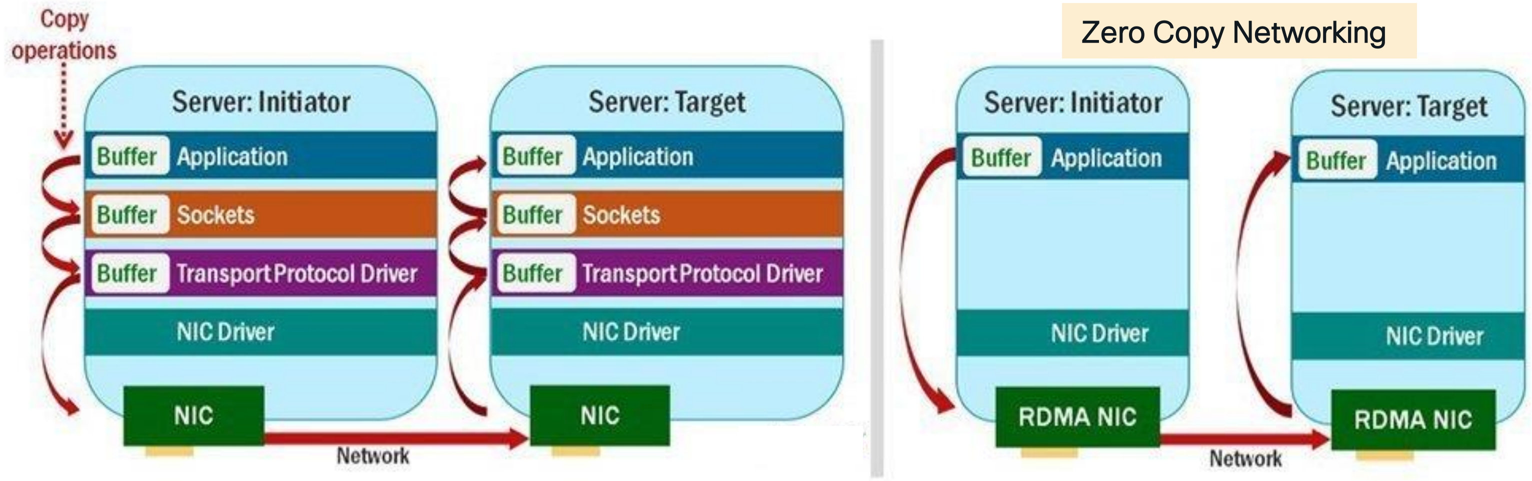


AI Networking: RDMA

Remote Direct Memory Access

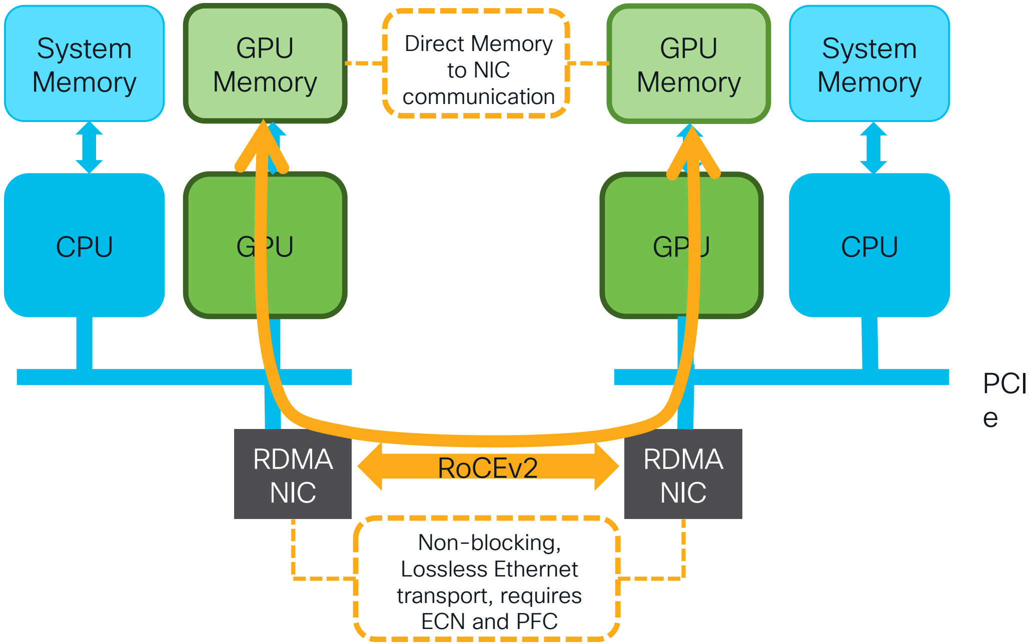
Benefits of RDMA

- Low latency and CPU overhead
- High network utilization
- Efficient data transfer
- Supported by all major operating systems



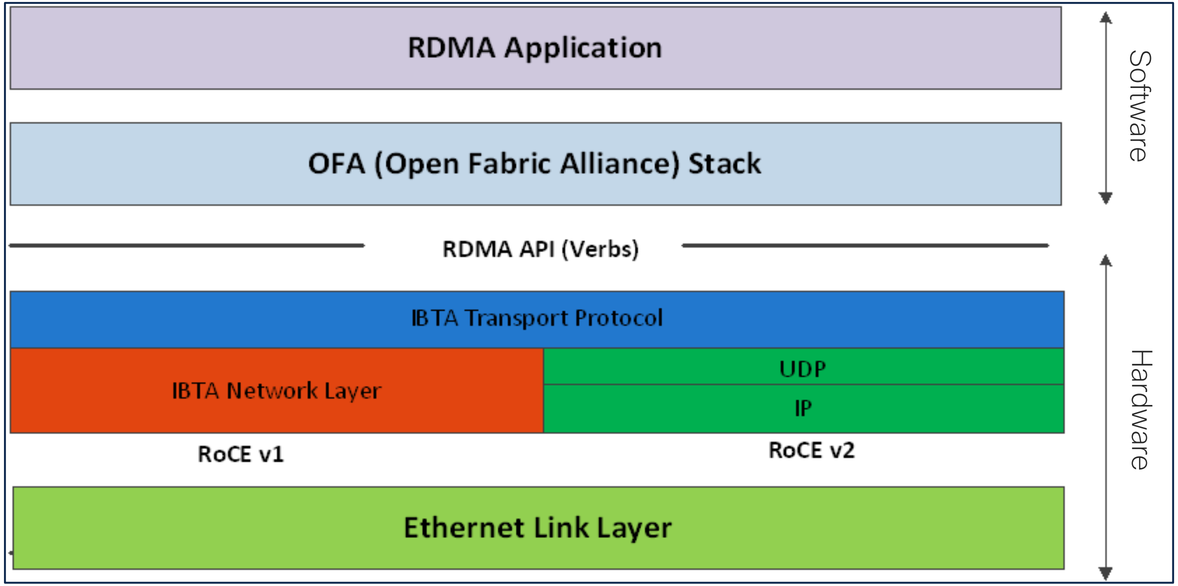
Remote Direct Memory Access (RDMA)....InfiniBand

- RDMA allows AI/ML nodes to exchange data over a network by accessing the bytes directly in the RAM
- Latency is very low as CPU and kernel can be bypassed
- RDMA data was natively exchanged over InfiniBand fabrics
- Later, RoCEv2 (*RDMA over Converged Ethernet*) protocol allowed the exchange over Ethernet fabrics



AI Networking: RoCE v1/RoCE v2 Protocol Stacks

RDMA Over Converged Ethernet



RoCE v1

- Ethernet link layer protocol
- Dedicated ether type (0x8915)
- Can be used with or without VLAN tag

RoCE v2

- Internet layer protocol – can be routed
- Dedicated UDP port (4791)
- UDP source port field is used to carry an opaque flow-identifier

RoCEv2: PFC and ECN Together for Lossless Transport

How does it work?

ECN is a layer 3 congestion avoidance protocol

ECN is an IP Layer Notification System allowing switches to indirectly inform the sources to slow down the throughput.

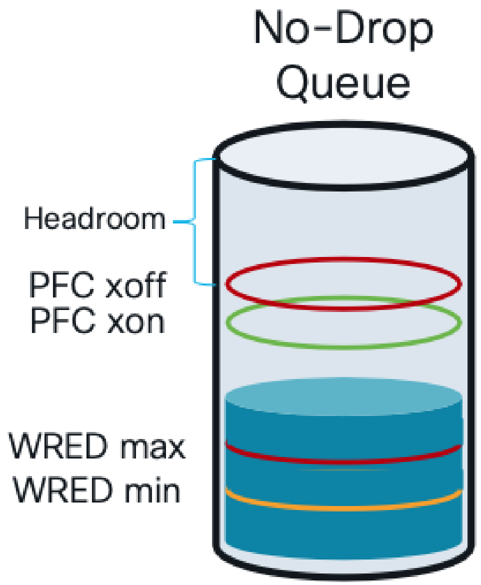
WRED thresholds are set low in no-drop queue.

- Signal early for congestion with CNP's, gives enough time for end points to react.

PFC is a layer 2 congestion avoidance protocol

PFC thresholds are set higher than ECN

- Oversubscription buffers can be filled quickly without giving time for ECN to react.
- PFC will react and mitigate congestion.

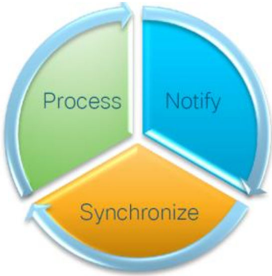


Data Center Quantized Congestion Notification

- IP ECN or PFC cannot alone provide a valid Congestion Management framework
- IP ECN signalling might take too long to relieve the congestion
- PFC can could introduce other problems like Head Of Line Blocking and unfairness for the flows
- The two of them together provide the desired result of having lossless RDMA communications across Ethernet networks (this is called **DCQCN**)
- The requirements are:
  - Ethernet devices compatible with both techniques
  - Proper configurations applied

AI/ML Flow Characteristics (Training Focused)

Execute instructions on GPU
High bandwidth compute can saturate network links



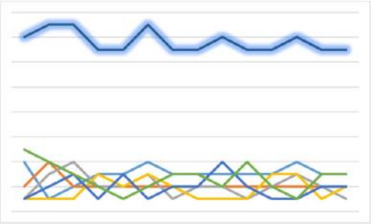
Send results of computation
Several methods, we'll focus just on one
All-to-All Collective (Everyone sends to everyone)

Wait for everyone to complete
Creates synchronization between GPUs
Computation stalls waiting for the slowest path
Job Completion Time (JCT) is based on the *worst-case* tail latency

Traditional DC Traffic Pattern

Cumulative Traffic

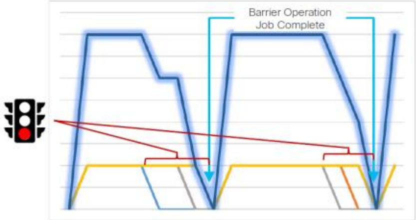
Individual Flows



Many asynchronous small BW flows
Chaotic pattern averages out to consistent load

AI (All-to-all Collective) Traffic Pattern

GPUs Stalled
Waiting for other GPUs to complete
(Due to network congestion from poor load balancing)



Cumulative Traffic

Individual Flows

Few synchronous high BW flows
Synchronization magnifies long tail latency & bad load balancing decisions

Bringing Visibility to AI workloads

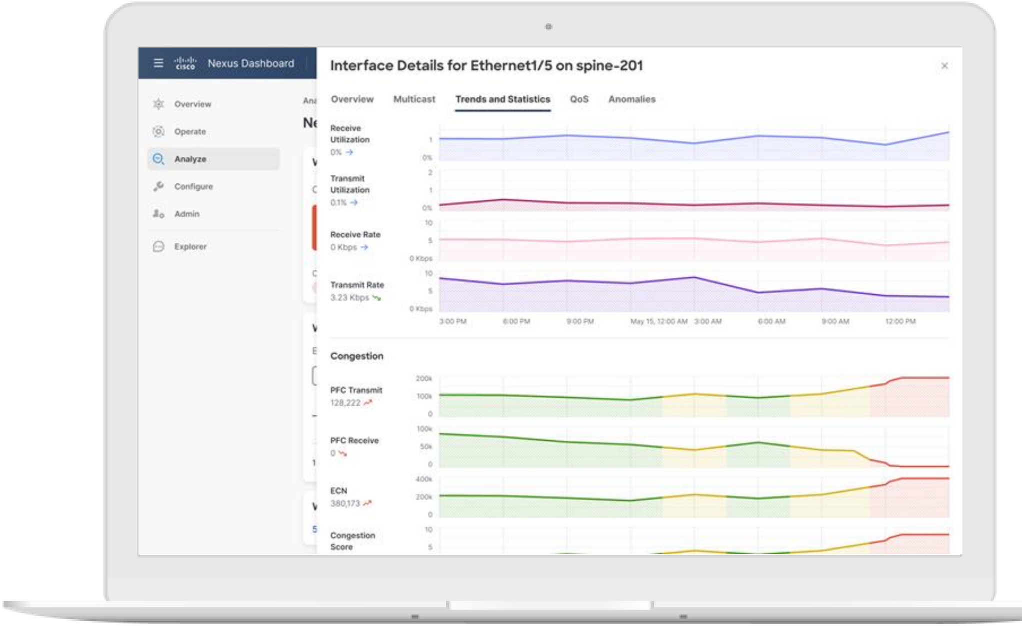
With the granular visibility provided by Cisco Nexus Dashboard Insights the network administrator can observe drops

Tune thresholds until congestion hot spots clear and packet drops stop in normal
traffic conditions

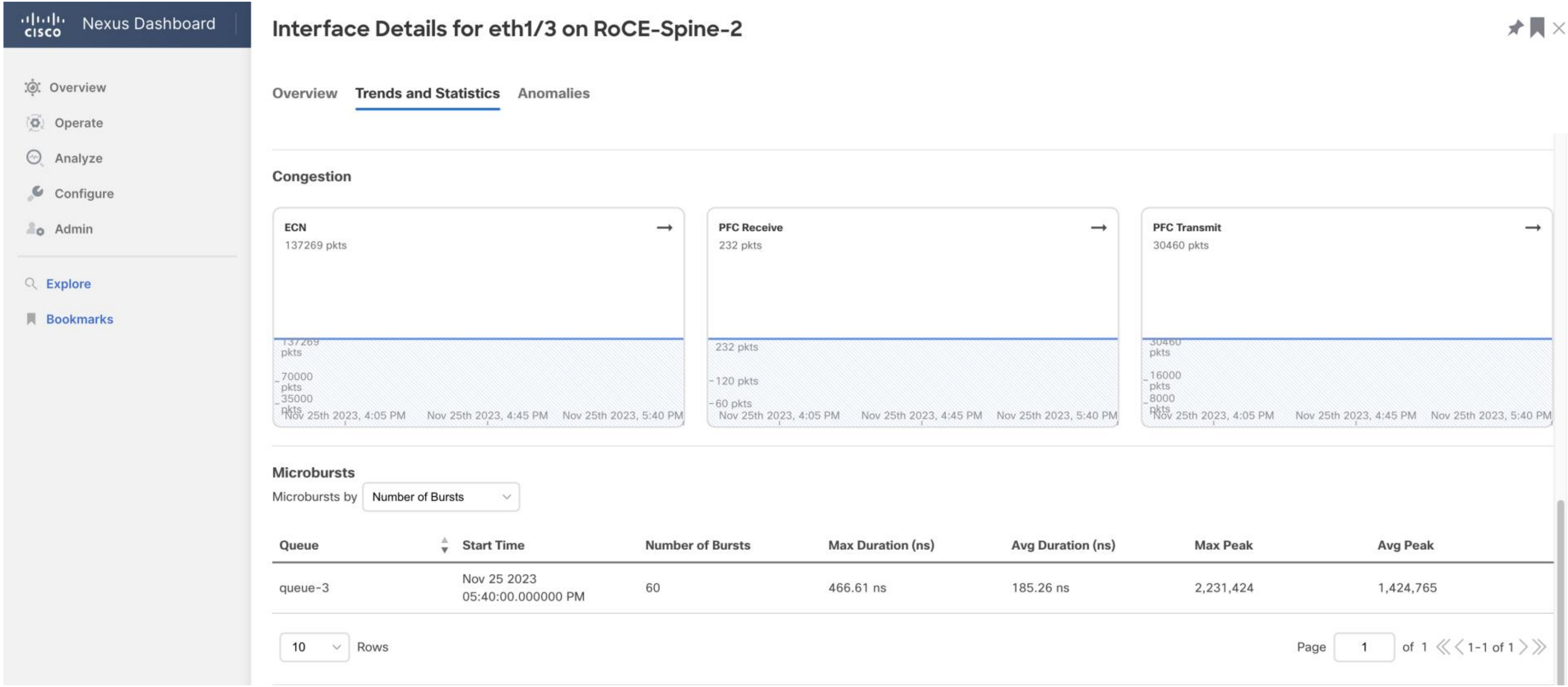
This is the first and most important step to ensure that the AI/ML network will cope with regular traffic congestion occurrences effectively

Monitoring These Events

- DCQCN leaves the fabric congestion management in a self healing status
- Still it is important to keep it under control:
  - Frequently congested links can be discovered
  - QoS policies can be tweaked with a direct feedback from the monitoring tools
- Nexus ASICs can stream these metrics directly to Nexus Dashboard Insights
- NDI will then collect, aggregate and visualize them all to provide insights to the operations team

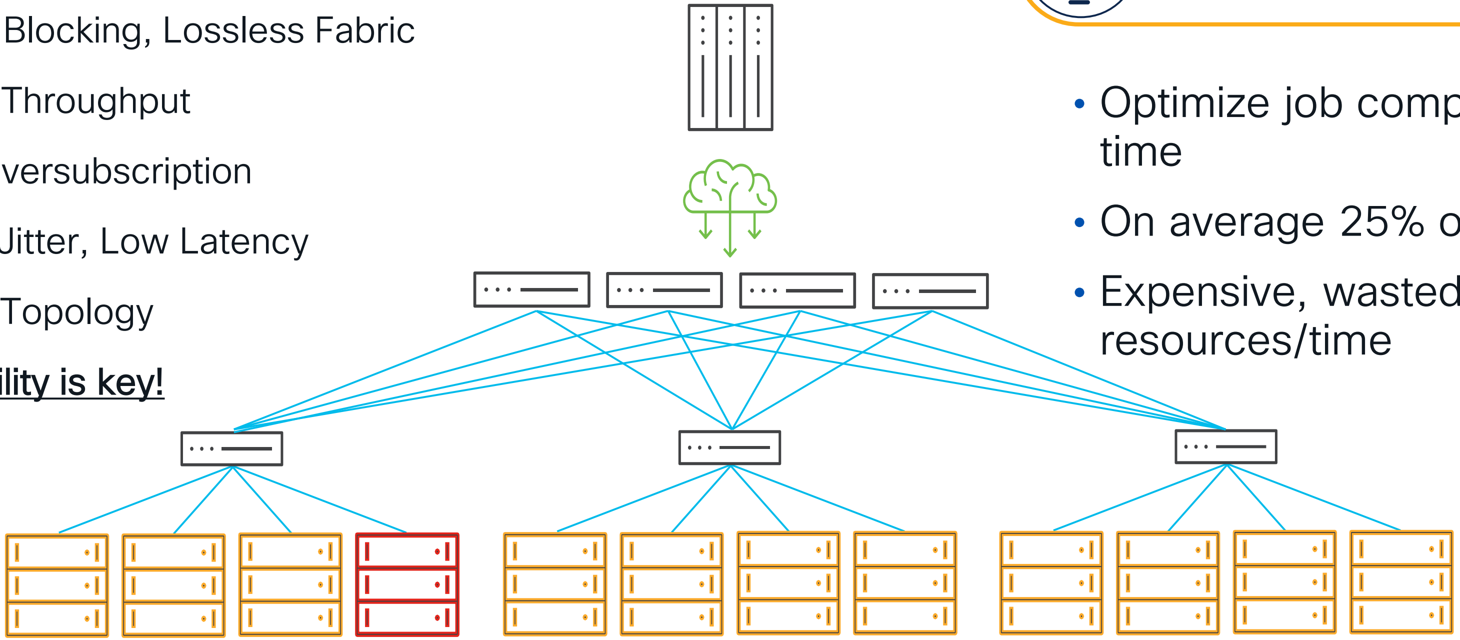


Nexus Dashboard Insights – Congestion Visibility



Designing a Network for AI success

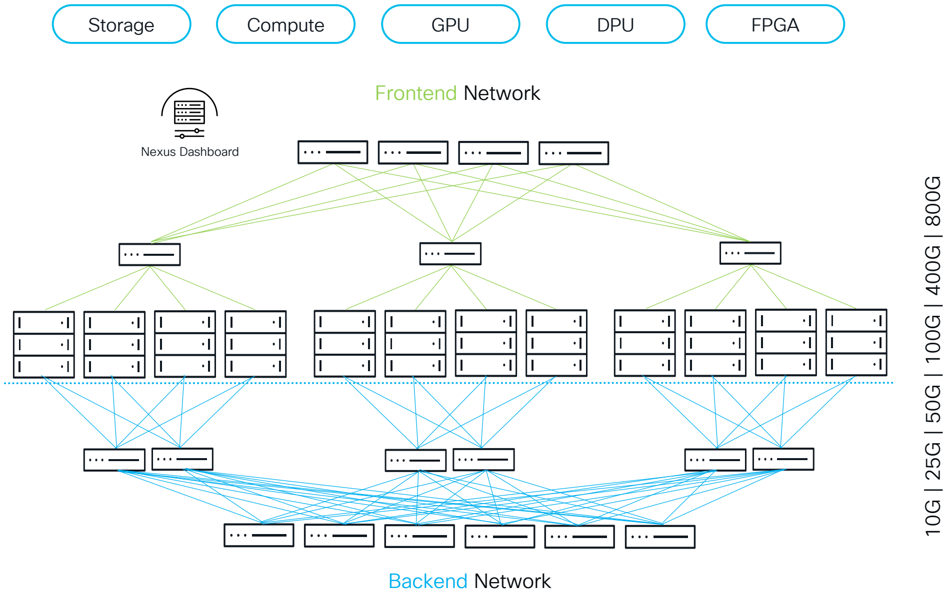
- Dedicated Network
- Non-Blocking, Lossless Fabric
- High Throughput
- No Oversubscription
- Low Jitter, Low Latency
- Clos Topology
- <u>Visibility is key!</u>



Stalled/Idle Job

- Optimize job completion time
- On average 25% of Jobs Fail
- Expensive, wasted resources/time

Do I need a backend network?



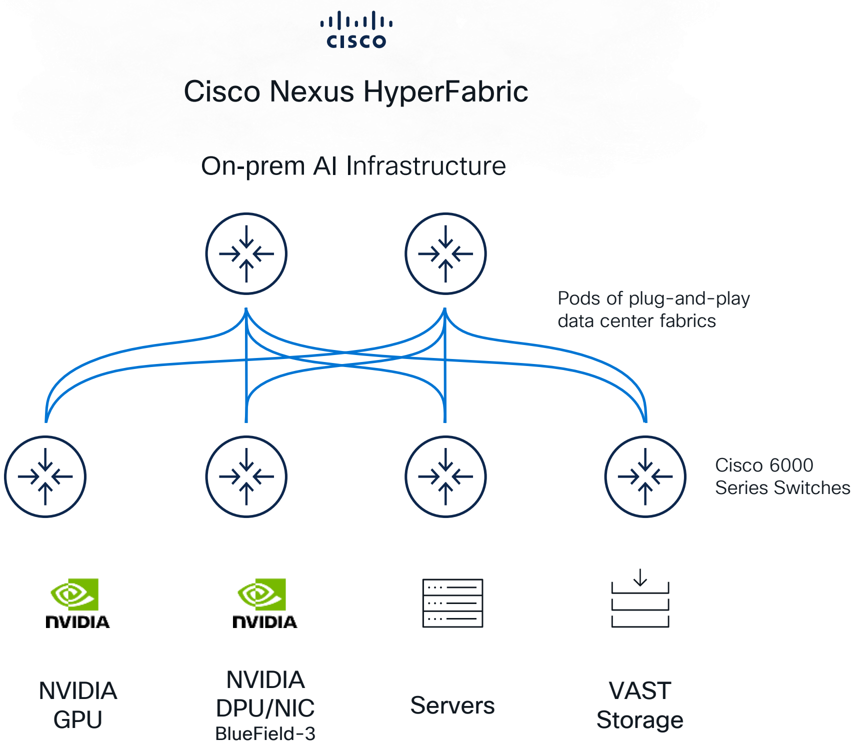
Lossless | High-Throughput | Low Jitter | Low-Latency

Cisco Nexus HyperFabric AI Cluster

in partnership with NVIDIA

Democratize AI Infrastructure	Visibility into full stack AI
Unified stack Including NVAIE	AI-native operational model
High-performance Ethernet	Cloud managed operations

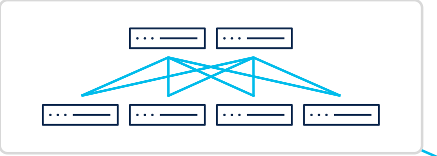
A solution that will enable you to spend time on AI innovation—not on IT.



Built on Cisco Silicon One and Optics innovations

A Simplified Backend Network for AI Environments

Cisco Nexus HyperFabric Use Cases

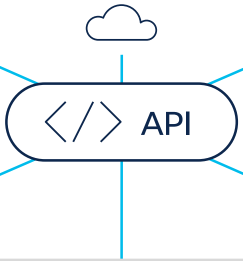


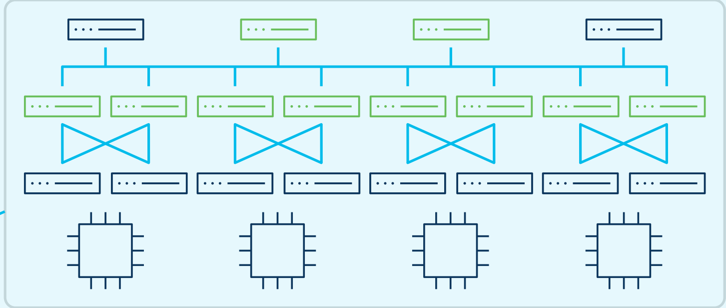
Build new data centers

- Ease-of-use for IT generalists
- Start small and grow fabrics (1+)
- Self-service for fabric tenants

Cloud SaaS Controller

- Single global UI for all owned fabrics
- Single global API endpoint
- Underlay and lifecycle automation

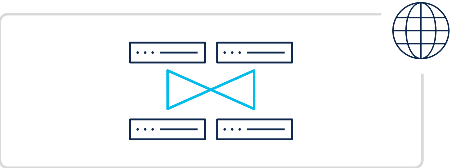




AI/ML/HPC fabrics

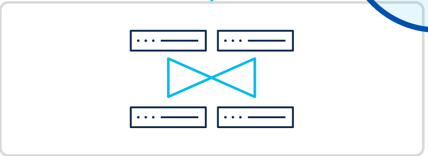
- Simple to deploy and manage
- Scalable AI-ready Ethernet fabric

Cisco Nexus HyperFabric AI cluster



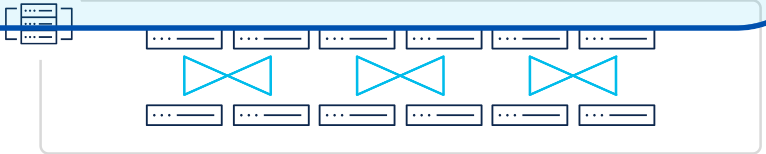
Extend data centers

- Plug-and-play deployment
- Easily expand to data center edge/colo
- Small fabrics of 1-2 switches



Downsize data center tooling footprint

- Data center anywhere with cloud controller
- Planning/design tools to help build rollout



Manage multiple customer data centers

- Managed from cloud
- Remote hands assistance

Building High-performance AI/ML Ethernet Fabrics

Maximizing customer choice and options

Cisco Nexus HyperFabric AI Cluster

Enterprise/ Public Sector/ Commercial

Cisco Cloud Managed as a Service, Full Stack

- Turnkey AI pod
- Nexus HyperFabric managed servers (BMC), NICs, and switches
- Converged ethernet infra
- Greenfield deployments only
- 400G -> 800G
  Cisco 6000 (Silicon One) switches

FCS Target CY25

Nexus 9000 with Nexus Dashboard

Enterprise/ Public Sector/ Commercial | Service Providers | Tier2 Web/ AI aaS

Private Cloud Managed, Interoperable

- General purpose AI multi-pod fabric
- Simplified network operations with Nexus Dashboard
- CVDs for converged ethernet infra
- Greenfield & brownfield deployments
- 100G -> 400G -> 800G
  Nexus (Cloud Scale & Silicon One) switches

Cisco 8000

Tier2 Web/ AI aaS | Hyperscalers

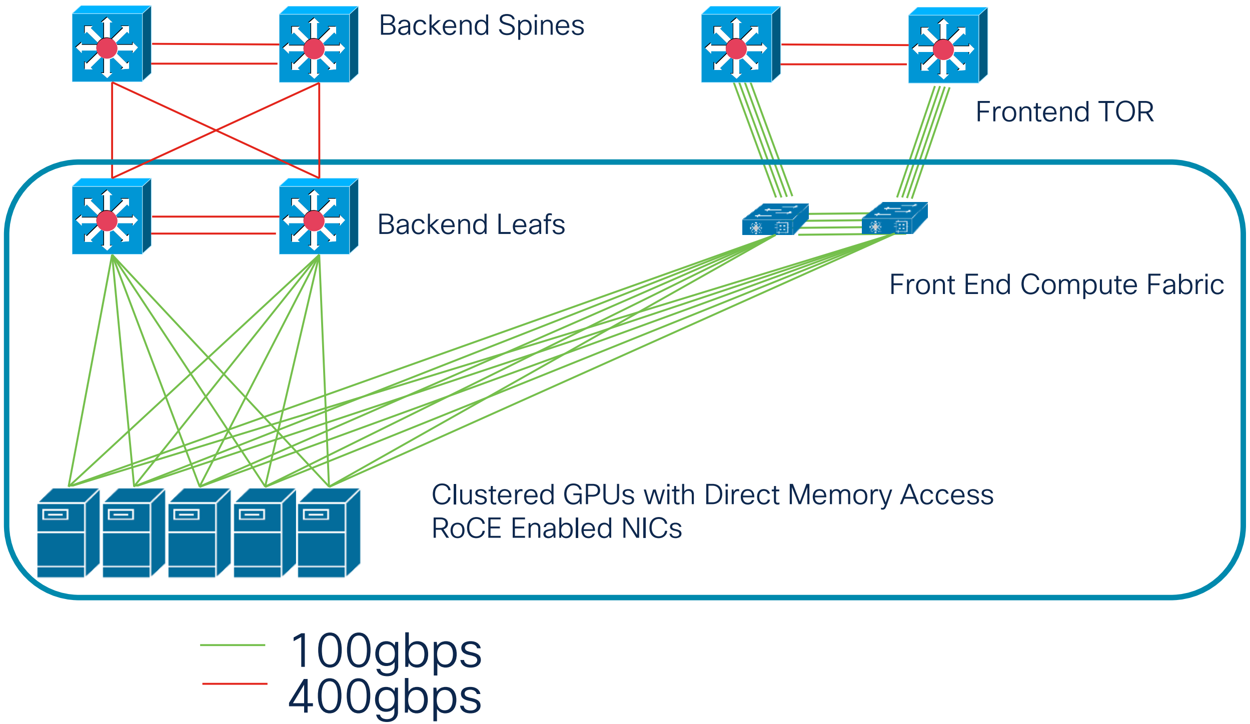
Customizable Solution BYO Management, SONiC / BYO-NOS

- Cisco validated SONiC or community sourced
- Customer assembled & operated
- ECMP* & Scheduled Ethernet** options
- Greenfield deployments
- 400G -> 800G
- Silicon One switches

FCS Target 2H 2024

Building an AI Workload Pod for Training

- Backend network for training
- 32 rack servers split across 2 racks
- Scale up to 30 pods per spine
  - 960 servers
  - 1920 GPUs
- Full RoCEv2 support on Compute



GPU Intensive Applications converged infrastructure example

Performance Testing

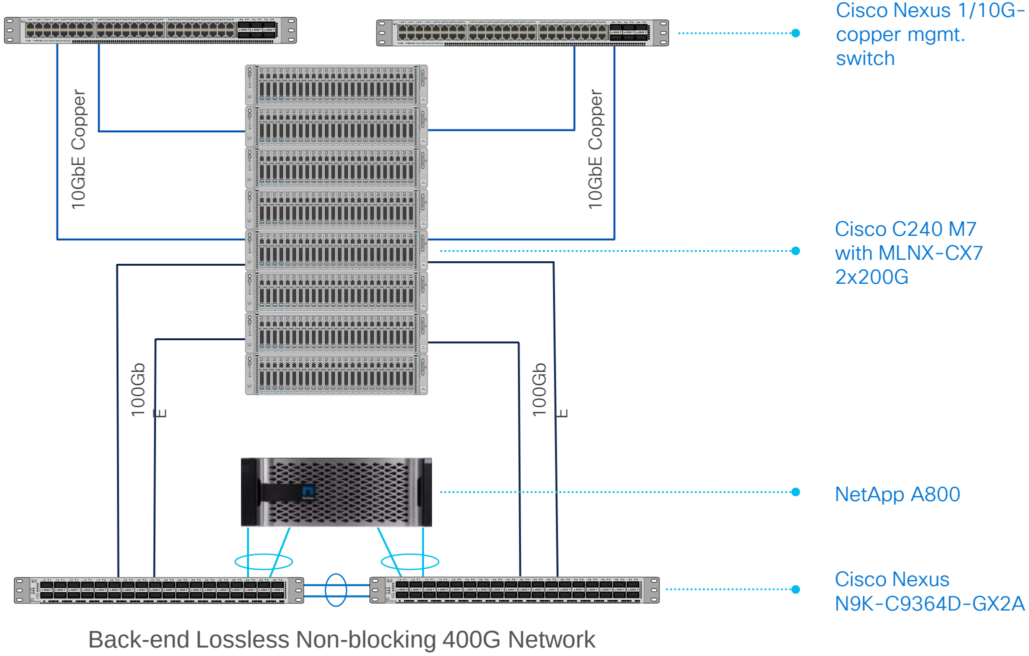
Linear Scalability demonstrated through benchmark tests on real life model simulation, showcasing consistent performance even with varying dataset sizes.

- Weather Simulation (MiniWeather)
- Nuclear Engineering (Minisweep)
- Cosmology (High Performance Geometric Multigrid)

Accelerated Deployment

- Centralized management and automation
- NVIDIA HPC-X Software Toolkit Setup & Configuration
- NetApp DataOps Toolkit to help developers, data scientists to perform numerous data management tasks

CVD Link



Cisco UCS C-Series Rack Server and NetApp AFF A400 storage array connected to Cisco Nexus 93600CD-GX leaf switch with layer 2 configuration for a single rack testing

The Blueprint For Today

Built to accommodate 1024 GPUs along with storage devices

Building a 400Gbps Back...

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... -

Building a Shared 100Gbps Network

Building a 400Gbps Back-End Netw...

Using Cisco Nexus Dashboard Fabr...

Conclusion

Related Materials

Building a 400Gbps Back-End Network

For higher performance use cases, where GPU to GPU traffic requires 400Gbps Ethernet connectivity, it may be preferred to build a separate network for these communications. These types of networks are often referred to as a "back-end network." Back-end networks generally are designed for GPUDirect (GPU to GPU) or GPUDirect Storage (GPU to storage) communications. The requirements for back-end networks are the same as any RoCEv2 network discussed in this document. The main difference for this type of network, assuming that it is connected at 400Gbps, is the recommended type of leaf and spine switches. To build a 400Gbps RoCEv2 network either the Cisco Nexus 9332D-GX2B or the Cisco Nexus 9364D-GX2A should be used as leaf switches. The Cisco Nexus 9364D-GX2A is the recommended spine switch. As with the network described earlier in this document a non-blocking fabric is required.

We will use the example of building a 512 GPU cluster. For this back-end network we will need to connect 512 400Gbps NICs. We will use the Cisco Nexus 9364D-GX2A as a leaf switch which will allow us to connect 32 GPUs/NICs to each leaf, leaving 32 ports for spine connectivity to build a non-blocking fabric. We will require 16 leaf switches to connect all 512 NICs.

To accommodate the amount of bandwidth coming from the leaf switches, 512 x 400Gbps ports are needed to the spines. We will use 8 Cisco Nexus 9364D-GX2A spine switches. Each leaf will connect to every spine using 4 x 400Gbps ports. The network is represented in the following diagram:

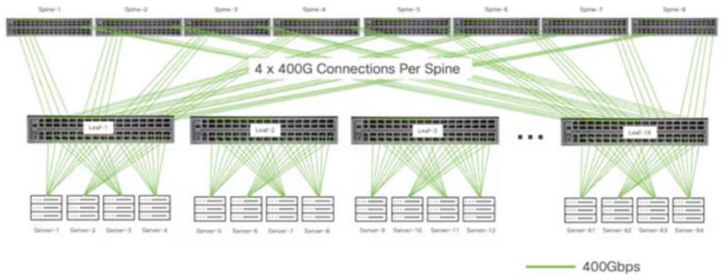


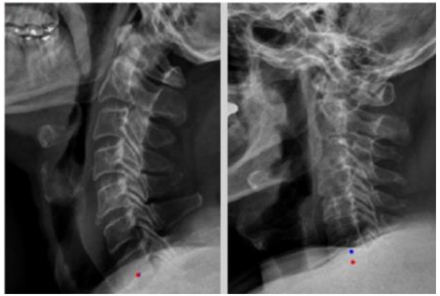
Figure 19. Spine-leaf network design, connecting 512 GPUs in a non-blocking network.

Inferencing, Fine-Tuning, & Compute Infrastructure

Model Inferencing Use Cases

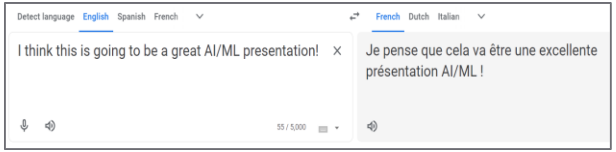
Productization Phase

Face recognition and computer vision

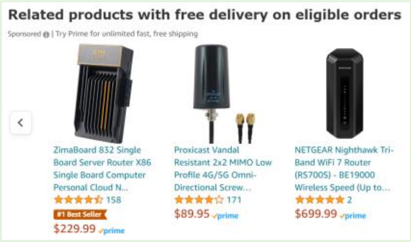


Analysis of medical images

Self-driving vehicles



Machine translation



Recommender systems

Conversational agents

Content generation
Images/Video/Voice

Large Language Models (LLMs)

Limitations for enterprise use

Hallucination	Can make stuff up, always has an answer
Sources	Where did the information come from ?
Outdated	Models maybe stale as quickly as it is released
Customize	Cannot personalize or use more current data
Update	Cannot edit the model to remove/change data

Training LLMs

Resource-Intensive and costly

Large Language Models are...

- Pre-trained on a large corpus of publicly available unlabeled data
- Training takes 1000s of GPUs over a span of months
- Requires periodic re-training to stay up to date

GPT-3 Large – 175B parameters
• Training Set Tokens: 300B
• Vocabulary Size: ~50k
• Number of GPUs: 10k x V100
• Training Time: One Month

Llama – 65B parameters
• Training Set Tokens: ~1-1.3T
• Vocabulary Size: ~32k
• Number of GPUs: 2048 x A100
• Training Time: 21 Days

Building LLMs from scratch is cost-prohibitive for the average Enterprise

Use Foundational Models

Starting point for most Enterprises

Foundational models (FM)

Download →

BERT
GPT
Llama
Mistral AI
Stable Diffusion
Cohere
Claude
BLOOM
...

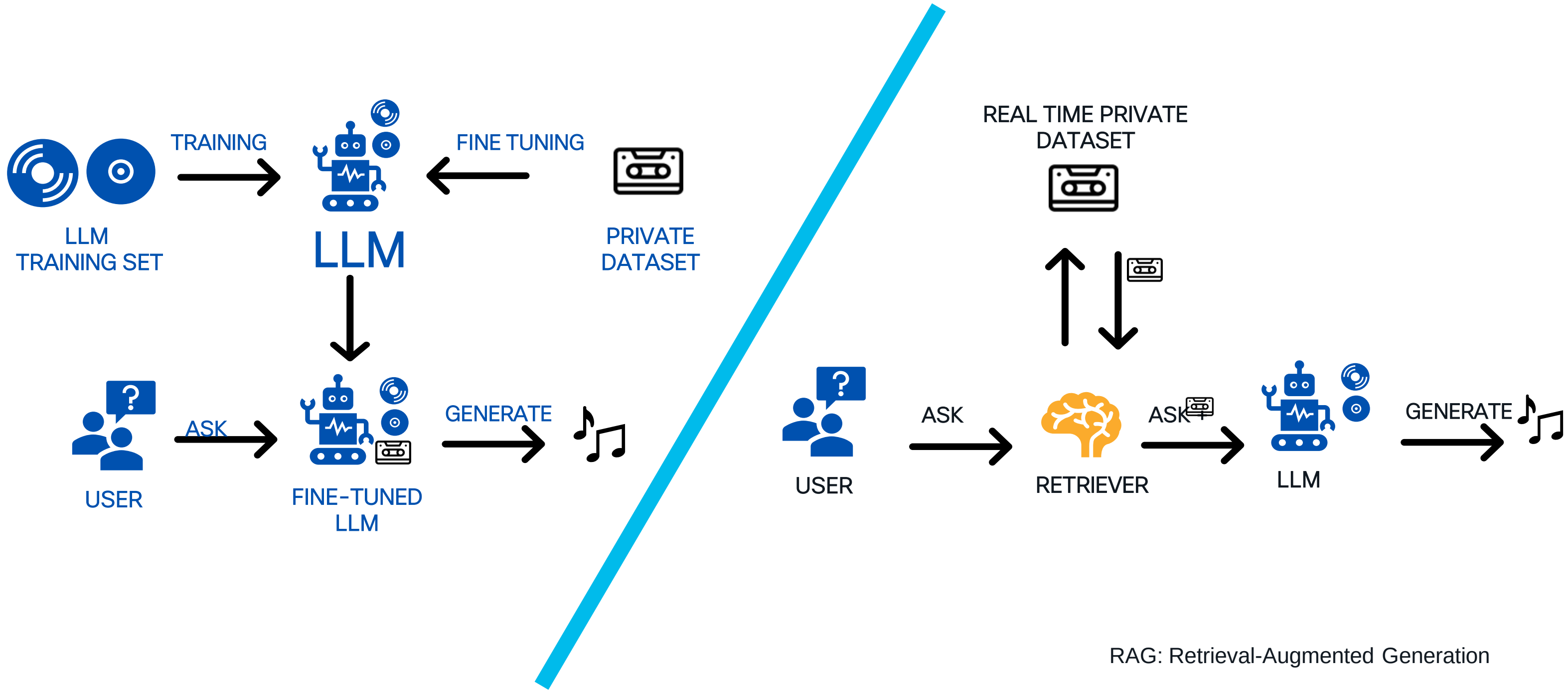
Pre-trained, general-purpose models

→ Customize or integrate directly for inferencing in enterprise applications

Model Size

- LLMs <100B
- Other Generative <1B
- Predictive <100M

LLM, Fine-Tuning and RAG?



RAG: Retrieval-Augmented Generation

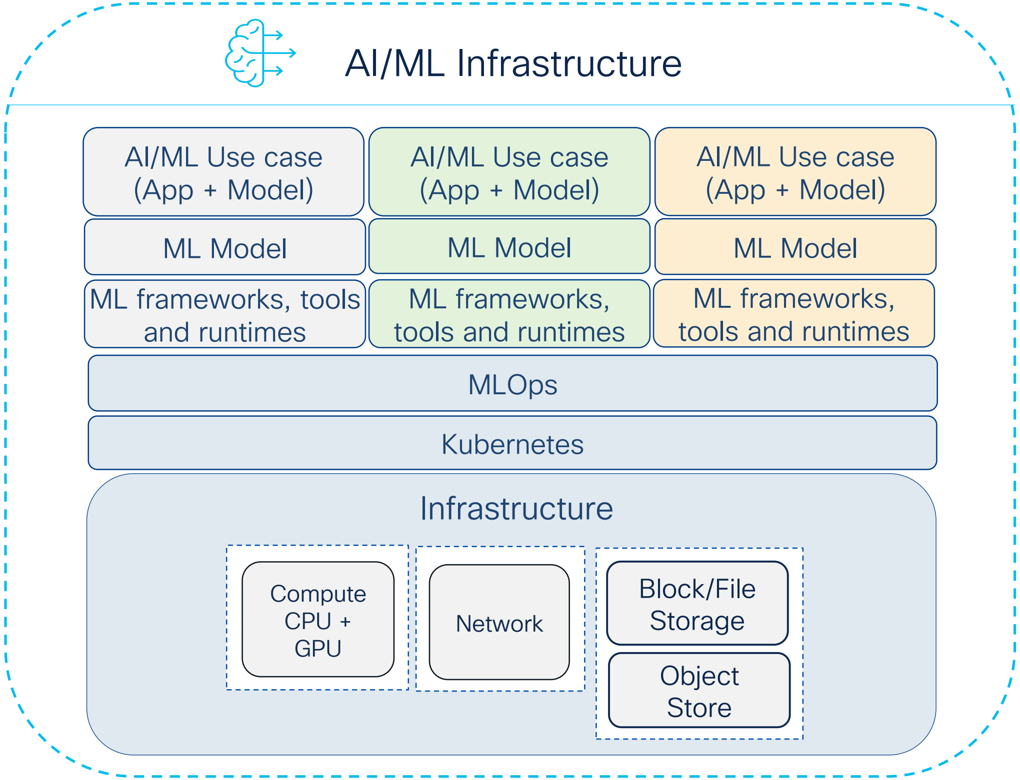
Business value of LLM + RAG

- RAG helps in mitigating hallucination or generation of incorrect or misleading information.
- Fine-tuning a pre-trained language model can be a resource-intensive process. RAG offers a cost-effective alternative.
- RAG generates context-aware responses by retrieving relevant data before crafting a response, this leads to clearer and more meaningful interactions with users.
- One of the major concerns with AI models is their "black box" nature, that is we are unsure of the source it has used to generate content. When RAG generates a response, it references the sources it used, enhancing transparency and instilling trust in the users.

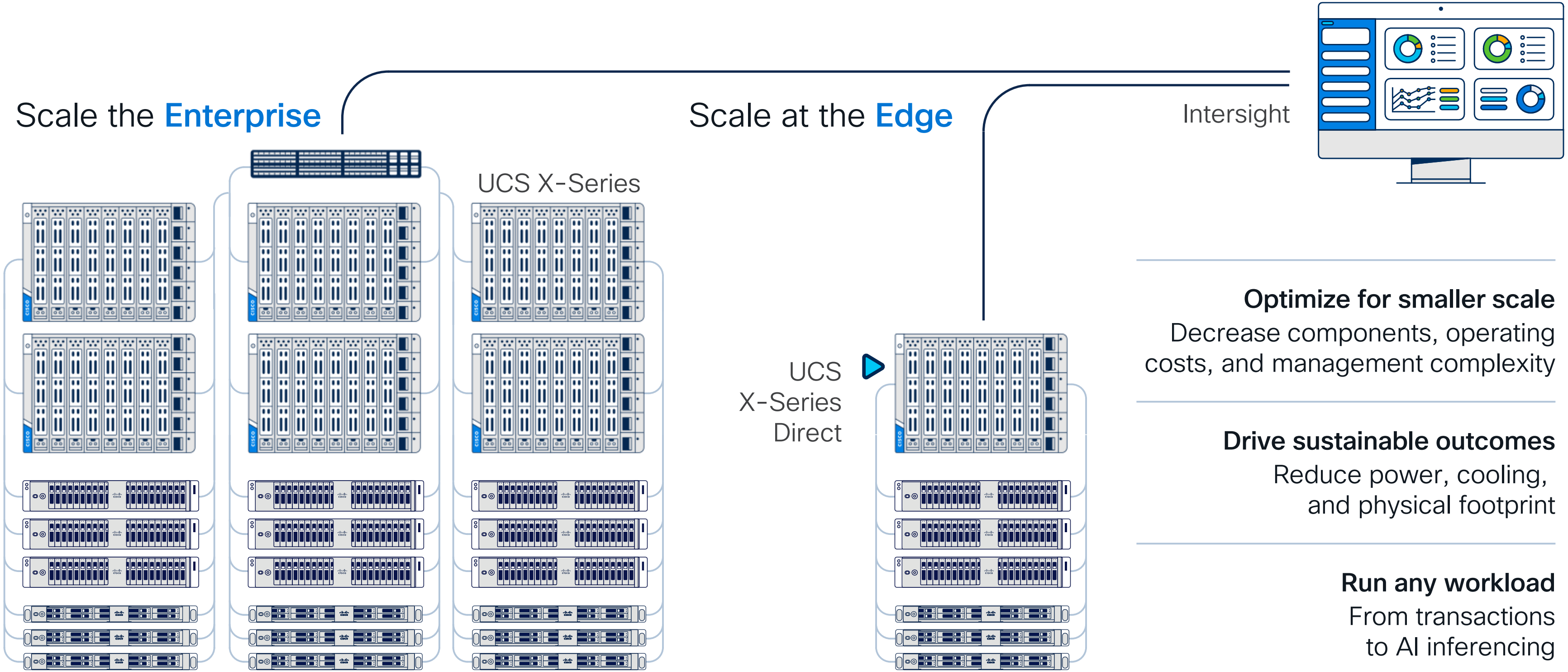
IT Infrastructure for Enterprise GenAI

High-level Architecture

Generative AI and Predictive AI Use Cases



Scale fine-tuning and inferencing compute from the data center to the edge



Optimize for smaller scale
Decrease components, operating costs, and management complexity

Drive sustainable outcomes
Reduce power, cooling, and physical footprint

Run any workload
From transactions to AI inferencing

Simpler, Smarter, More Agile

Fabric-Based adaptive Computing

Scale seamlessly to changing business needs

Faster deployment of applications

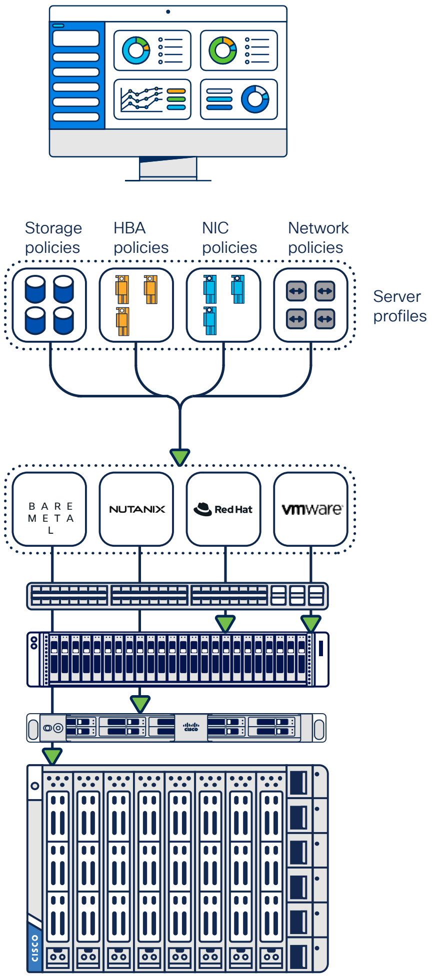
Greater control and flexibility

Better performance, resiliency, high availability

Less cost and complexity with fewer components

Innovative stateless server configuration

Infrastructure shapes to your specific workloads



Modular architecture

Ideal for AI component evolution

Investment preservation

- Convenience to upgrade or replace individual parts without overhauling the entire system
- Reduces cost and ensures that initial investments remain valuable over time

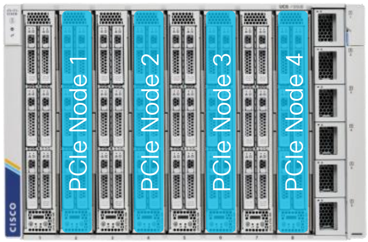
Multi-vendor support

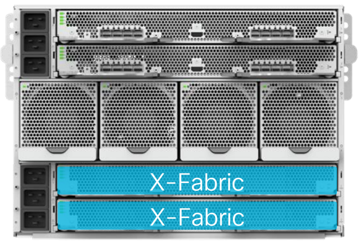
- Can select components from different vendors
- Best example is within CPU as you can move from AMD and Intel AMX to NVIDIA GPU A100 and then H100, or AMD in the future

Management & Upgradability

- Keep your technology stack current, adaptable, and competitive
- Cisco Intersight is a SaaS-based provides cloud-scale management from DC to edge

Modularity on X-Series





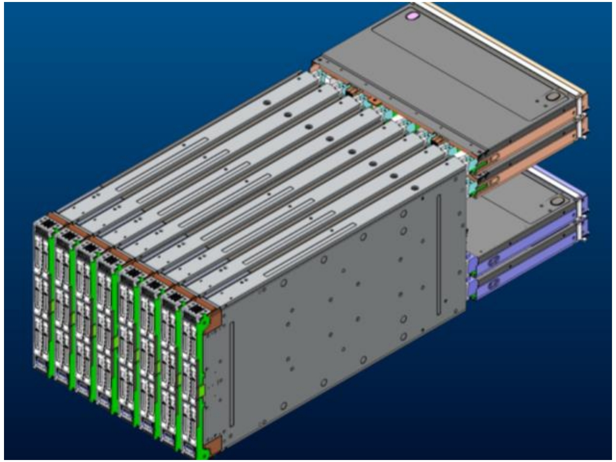
X-Series modular system decouples the lifecycles of CPU, GPU, memory, storage and fabrics - providing a perpetual architecture that efficiently brings you the latest innovations.

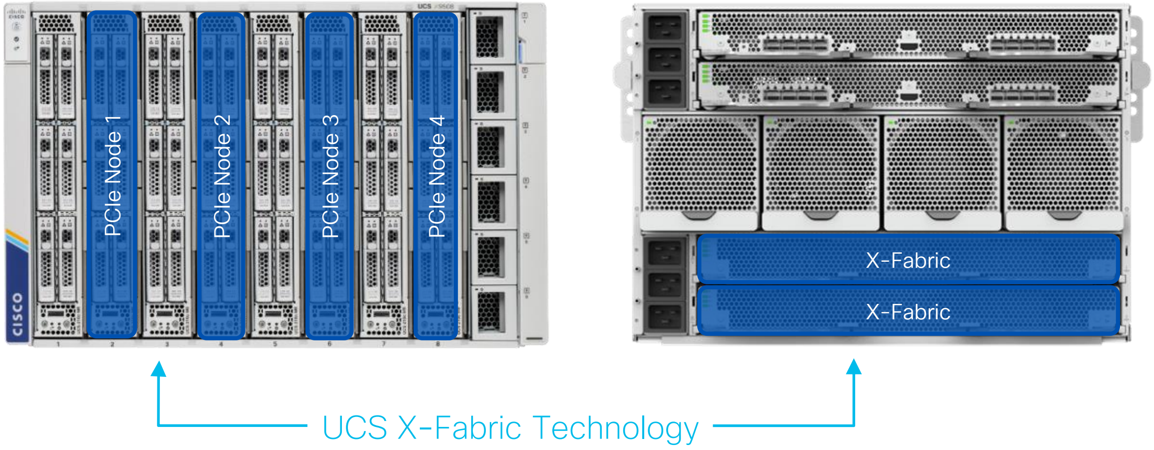
- Cloud-powered composability with Cisco Intersight
- Flexible GPU acceleration across server nodes
- No backplane or cables = easily upgrades

UCS X-Series for AI Workloads

Expandability and Flexibility

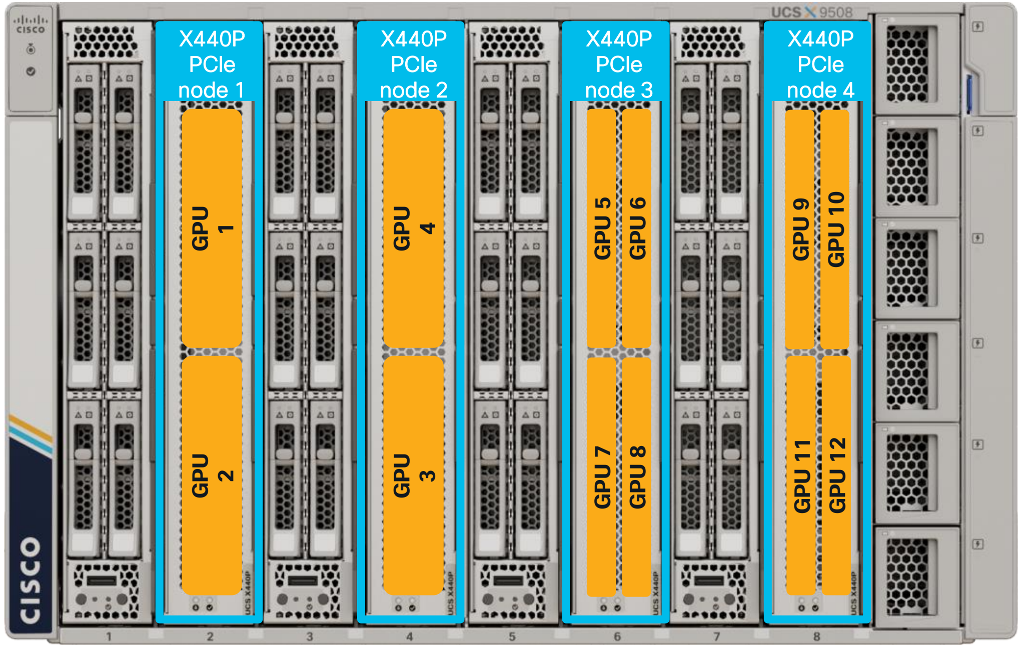
1. No backplane
2. X-Fabric
3. Server disaggregation (PCIe Node)





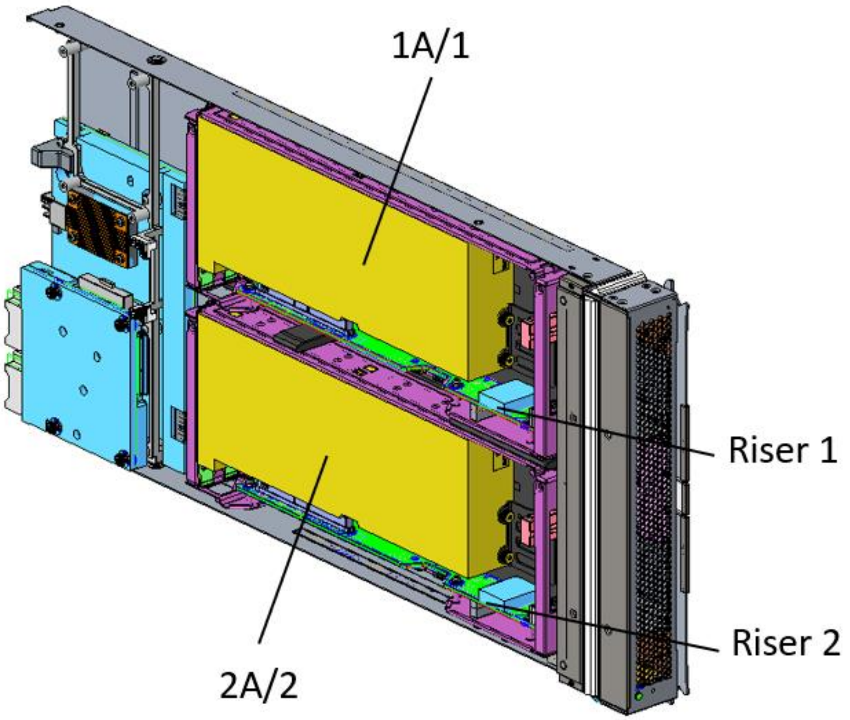
X440P PCIe Node

- Two different types
- Provides 2 or 4 PCIe slots per slot
- Connects via X-Fabric to adjacent compute node
- Dedicated power and cooling to GPU (no disks or CPUs blocking airflow)



Riser Style A

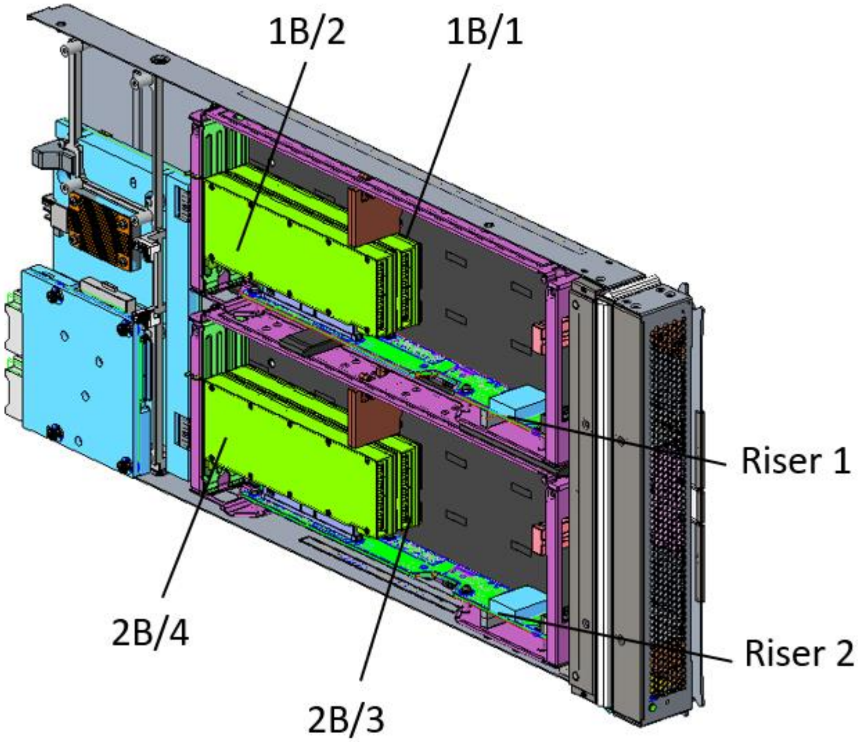
- Up to two dual width A16, A40, L40, L40S, A100 or H100 (NVL*), Flex170, MI210* GPUs
- One x16 per riser = 1 per CPU
- No mixing of GPUs



* planned

Riser Style B

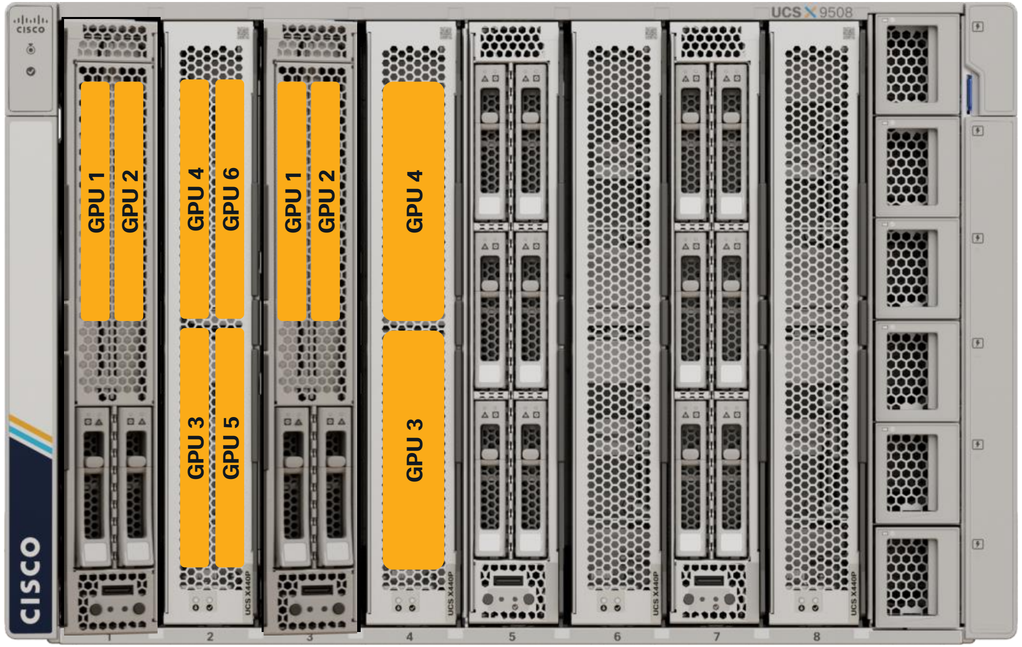
- Up to 4 single width T4/L4/Flex140 GPUs
- Two x8 per riser = 2 per CPU
- No mixing of GPU models



X210c/X215c Blade with GPU options

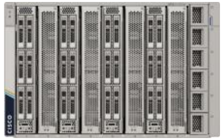
Additional Front Card GPU Options

- Up to six U.2 NVME drives
- Up to two GPUs
- Slides into front of X210C/X215C compute node
- Can be used with PCIe node to provide up to 6 GPUs per host
- Intel or AMD CPU



Cisco GPU-accelerated platforms offering

X-Series



Up to 24x HHHL GPUs or
8x FHFL GPUs per X9508 chassis

X210c M6/M7 2S Blades
2x NVIDIA T4 (MEZZ)

X210c M7 2S Blade
Intel Flex140 (MEZZ)

X210c M7 2S Blade
NVIDIA L4 (MEZZ)

X215c M8 2S Blade
NVIDIA L4 (MEZZ)

X440p + X210c M6/M7
4x NVIDIA T4 (M6 Only)
2x NVIDIA A16
2x NVIDIA A40
2 x NVIDIA A100-80

X440p + M7 (X210c & X410c)
2x NVIDIA H100-80
2x NVIDIA L40
4x NVIDIA L4
2x NVIDIA L40S

X440p + M7 (X210c & X410c)
4x Intel Flex140
2x Intel Flex170

Plan (Q3’24)

X440p + X210c M7
2x NVIDIA H100-NVL

X440p + X215c M8 AMD
2x NVIDIA H100-NVL
2x AMD MI210
4x NVIDIA L4
2x NVIDIA L40S
2x NVIDIA L40
2x NVIDIA A16

Plans are Subject to change

C-Series Rack Servers

C240 M6 INTEL
C245 M6 AMD

5x NVIDIA A10
3x NVIDIA A16
3x NVIDIA A30
3x NVIDIA A40
3x NVIDIA A100-80

8x NVIDIA L4
(C240 M6 only)

C240 M7 INTEL

3x NVIDIA A16
3x NVIDIA A30
3x NVIDIA A40
3x NVIDIA A100-80
2x NVIDIA H100-80
3x NVIDIA L40
8x NVIDIA L4
2x NVIDIA L40S
5x Intel Flex140
3x Intel Flex170

C220 M6 INTEL

3x NVIDIA T4
3x NVIDIA L4

C225 M6 AMD

3x NVIDIA T4

C220 M7 INTEL

3x NVIDIA L4
3x Intel FLex140

C245 M8 AMD

Plan (2H’24)

NVIDIA H100-80
NVIDIA L40S
NVIDIA L40
NVIDIA L4
NVIDIA H100-NVL
NVIDIA A16
AMD MI210

C225 M8 AMD

Plan (2H’24)

3x NVIDIA L4

Plans are Subject to change

Please Refer to the Server Specifications and HCL for detailed configuration support:
C-Series: https://www.cisco.com/c/en/us/support/servers-unified-computing/ucs-c-series-rack-servers/series.html#~tab-documents
X-Series: https://www.cisco.com/c/en/us/support/servers-unified-computing/ucs-x-series-modular-system/series.html#~tab-documents
UCS HCL: https://ucshcltool.cloudapps.cisco.com/public/

Sizing for Inferencing

LLM Inference Performance

How many GPUs do I need for inference?

Use Case	Model architecture	Context Length	GPU performance
▪ Determines model and minimum GPU ▪ CPU will also have an impact	▪ Impacts compute requirements per inference (TFLOPs)	▪ Will depend on the model ▪ Use average token size or vary token lengths in tests	▪ Will depend on its performance (TFLOPS) ▪ Use tests to verify performance

LLM Inferencing Performance

Objective and Subjective

Latency

- Time to first token
- Total Generation Time
- Time to second/next time

Throughput

- Requests per second dependent on concurrency and total generation time
- Tokens per second is the standard measure (> 30 per second)

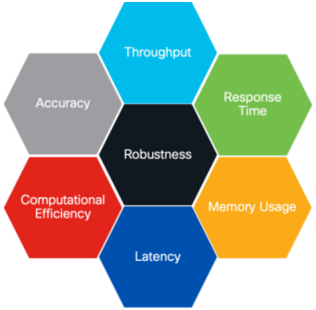
User experience – combination of low latency, throughput and accuracy

Prompt: What is Cisco UCS?

First Token

Cisco Unified Computing System (UCS) is a data center server computer product line composed of computing hardware, virtualization support, switching fabric, and management software. It was introduced by Cisco Systems in 2009.

43 Output Tokens



LLM Inference – Estimating Memory

How much memory does my model need?

For a given precision: FP32, FP16, TF16...

- Model Memory

  Precision in Bytes x # of parameters (P)

Example: Llama2 – 13B parameters

- Model Memory:

  13 billion x 2Bytes/parameter = 26GB

LLM Inference – Estimating Memory

How much memory does my model need?

For a given precision: FP32, FP16, TF16...

- Memory (Inference)

  Model Memory + ~20% overhead

Example: Llama2 – 13B parameters

- Memory (Inference):

  26GB + 20% overhead = 31.2GB

LLM Inference - GPU Estimation

Which GPU do I use?

Based on model memory, number of GPUs needed to load a 13B parameter model = any GPU with at least 32 GB

Similarly, a 70B parameter model, would require:
~2 A100-80 GPUs (168GB/80GB)

GPU Model	Memory (GB)	Memory Bandwidth (GB/s)	FP16 Tensor Core (TFLOP/s)
H100	80	2000	756
A100	80	1935	312
L40s	48	864	362
L4	24	300	121

LLM Inference - Methodology

How many GPUs do I need for inference?

For a given model and inferencing runtime, start with enough GPUs to load the model based on memory sizing

Vary concurrent inference requests and measure throughput and latency metrics for a given token length (context)

Vary batch sizes and measure throughput and latency - maximizes compute for non-RT use cases

Add a second GPU and repeat concurrent inference request and batch size tests (as needed)

Monitor GPU compute and memory utilization, along with inferencing performance, across all tests

Select a configuration that optimally balances latency, throughput and cost

Sample tool: https://github.com/openshift-psap/llm-load-test

Sample Performance Comparison with Nvidia A100

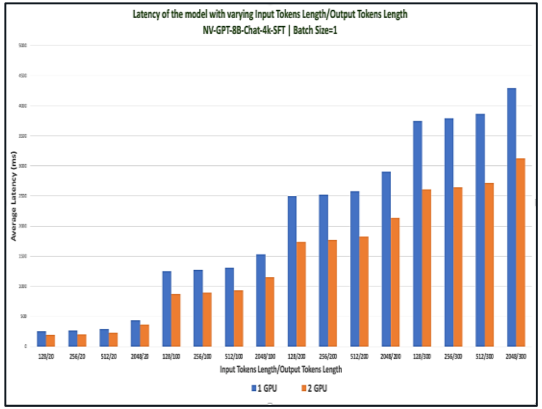
Llama 2 7B | NV-GPT-8B-Chat-4k-SFT | Llama2 13B

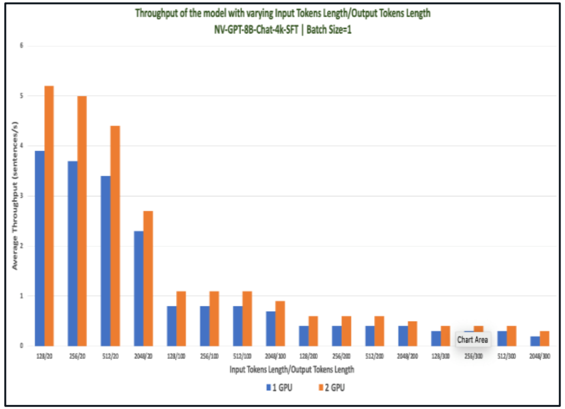
Llama 2 – 7B

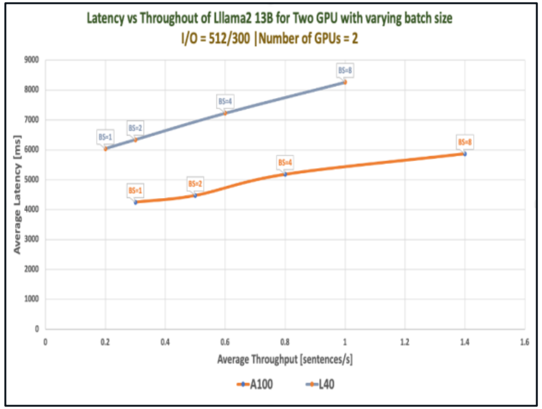
Input tokens Length: 128 and output Tokens Length: 20

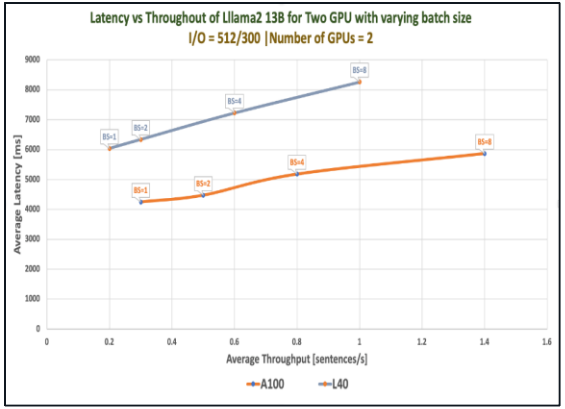
Batch Size	GPUs	Average Latency (ms)	Average Throughput (sentences/s)
1	1	241.1	4.1
2	1	249.9	8.0
4	1	280.2	14.3
8	1	336.4	23.8
1	2	197.1	5.1
2	2	204.1	9.8
4	2	230.2	17.4
8	2	312.6	25.5

Optimized price to performance ratio with FLASHSTACK AI



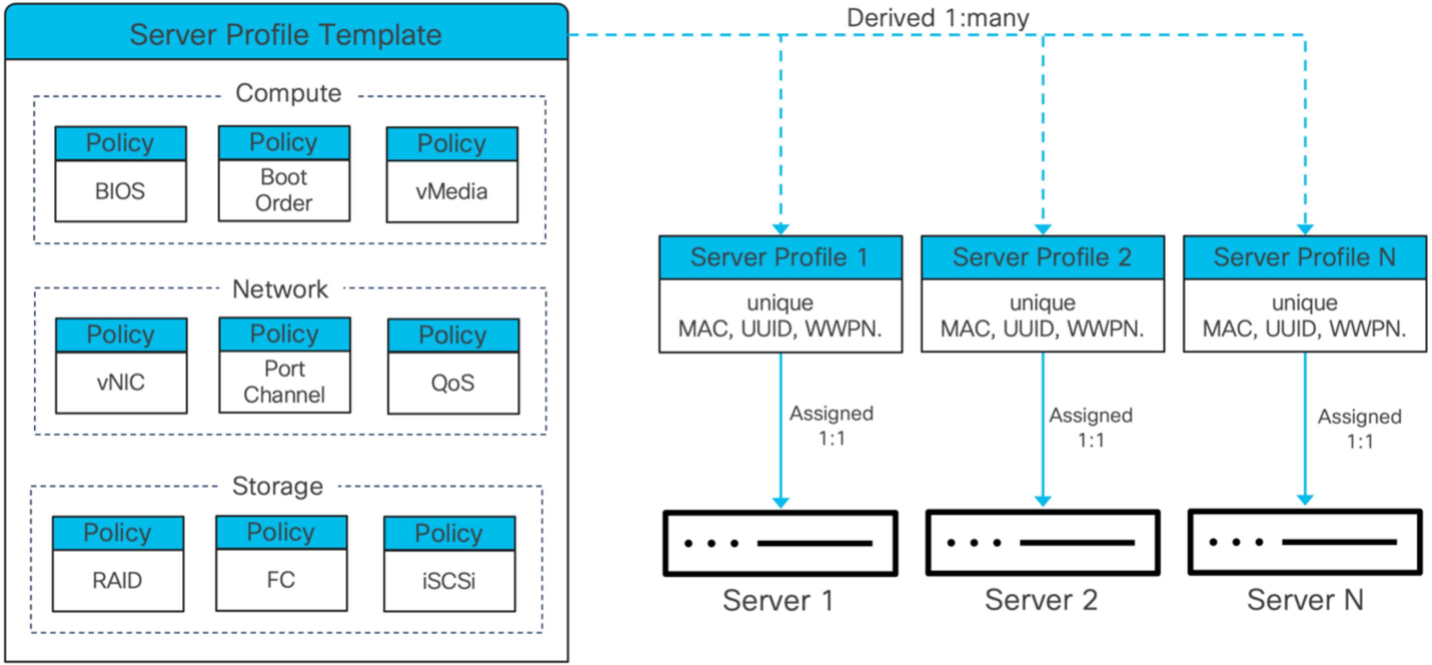




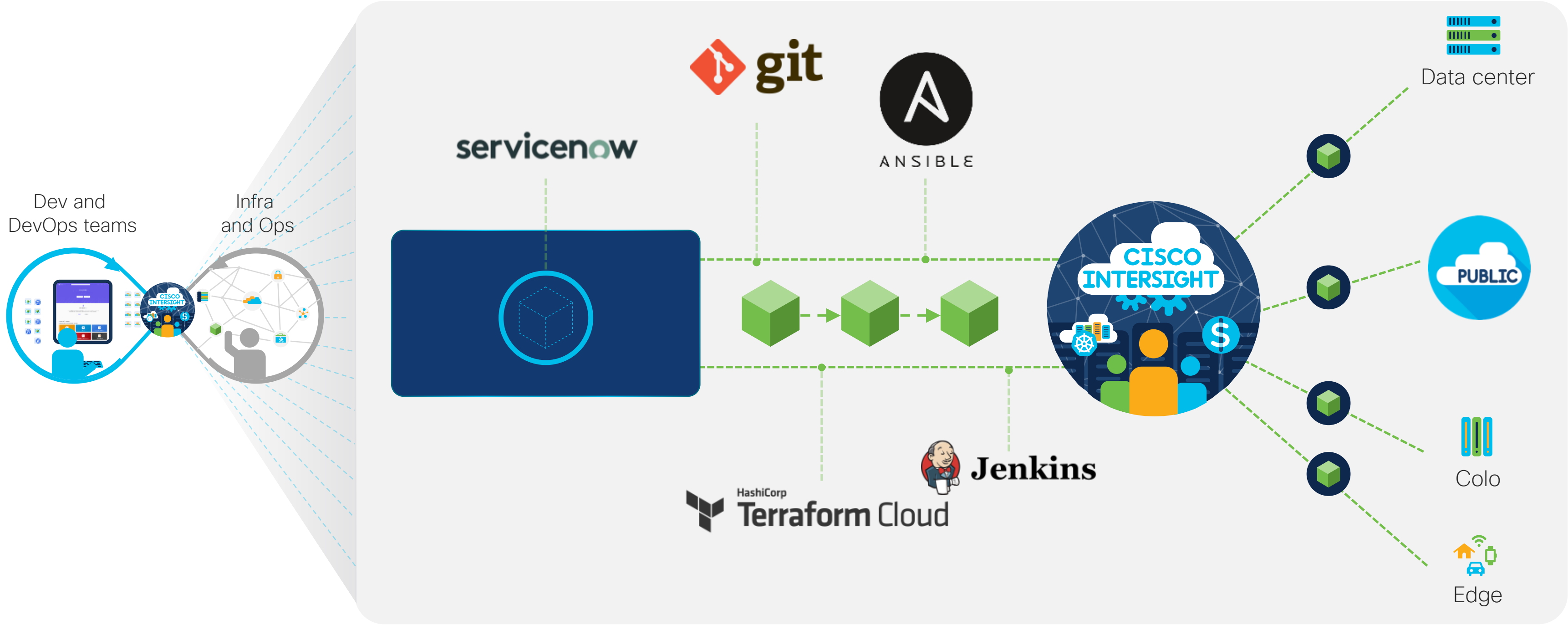


AI Infrastructure Automation

Policy based compute to scale operations



Integrate with DevOps to accelerate AI application delivery



Accelerate CI/CD processes and extend infrastructure as code (IaC) workflows by integrating Intersight into your DevOps toolchains

Simplify lifecycle management with integrated infrastructure and workload orchestration tools

Day 0/2: Operations (Full Stack Bare Metal)

Operational Challenges

- Lack of visibility across multiple infra and cluster deployments
- Difficulty gathering compliance and resource audits
- Capacity planning and inventory expansion

Optimization | Security | Supported | Alerts

Hybrid Cloud Admin

Red Hat
Hybrid Cloud Console

Telemetry – Infra Health, Alerts, Alarms, Security

Infra capacity management for expansion

CISCO INTERSIGHT

SaaS

On-Prem

K8s Admin

RED HAT OPENSHIFT Container Platform

Add/remove Bare Metal Nodes

Cluster Life-cycle

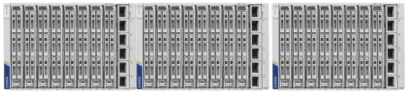
Cluster Upgrade/downgrade

Observability

Edge Site 1 — OPENSHIFT Cluster

Edge Site 2 — OPENSHIFT Cluster

Edge Site n — OPENSHIFT Cluster



UCS-X at the Edge sites

Intersight Private Appliance – Optional (Air-Gap Use Case)

Inventory (firmware, network, storage)

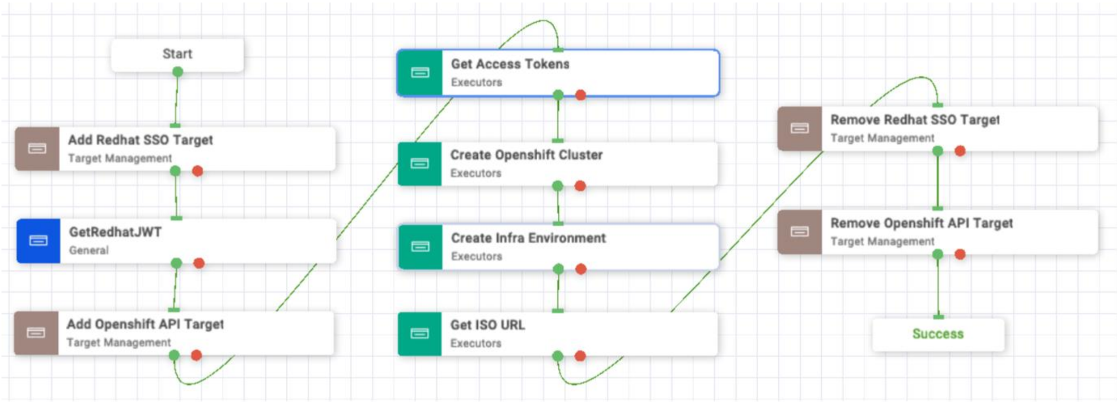
Field alerts & alarms, security advisories

Telemetry, metrics and actionable insights

Hardware Compatibility

RBAC, Multi-tenant

One-click Openshift cluster deployment



ⓘ Download and save your kubeconfig file in a safe place. This file will be automatically deleted from Assisted Installer's service in 10 days.

	Hostname	Role	Stat...	Discovered on	CPU...	Me...	Tota...	
›	00-25-b7-fa-01-01	Control plane node, Worker	Installed	4/26/2024, 9:41:34 PM	112	1.50 TiB	239.99 GB	⋮
›	00-25-b7-fa-01-02	Control plane node, Worker (bootstrap)	Installed	4/26/2024, 9:41:22 PM	112	1.50 TiB	239.99 GB	⋮
›	00-25-b7-fa-01-03	Control plane node, Worker	Installed	4/26/2024, 8:38:40 PM	112	1.50 TiB	239.99 GB	⋮

AI project deployment workflow example

1 Deploy AI-ready infrastructure

Cisco Intersight · ANSIBLE

NVIDIA AI Enterprise · Red Hat OpenShift

FLASHSTACK CI · FlexPod – A Cisco and NetApp Solution · NUTANIX

2 Deploy Red Hat OpenShift and other resources

portworx by Pure Storage · Red Hat OpenShift AI

- Image Registry
- Pipelines Artifacts Repo
- Model Repo

3 Deploy two projects* in Openshift AI

- Model delivery pipeline
- Application inferencing pipeline**

4 Load LLM from Hugging Face and explore/evaluate

5 Save and upload model to Model Repo

6 Deploy model serving runtime

vLLM

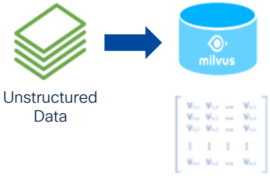
7 Deploy LLM for inferencing

8 Deploy Vector Database for RAG

milvus

Attu open-source GUI

9 Ingest Enterprise data to vector database



10 Deploy Q/A Chatbot App using Enterprise Knowledge Base

LangChain

11 Deploy GUI front-end for Q/A Chatbot

gradio

12 Deploy Enterprise Q/A Chatbot for inferencing

Core · Edge

* Workbenches/namespaces

** For demo purposes

Model Delivery Lifecycle

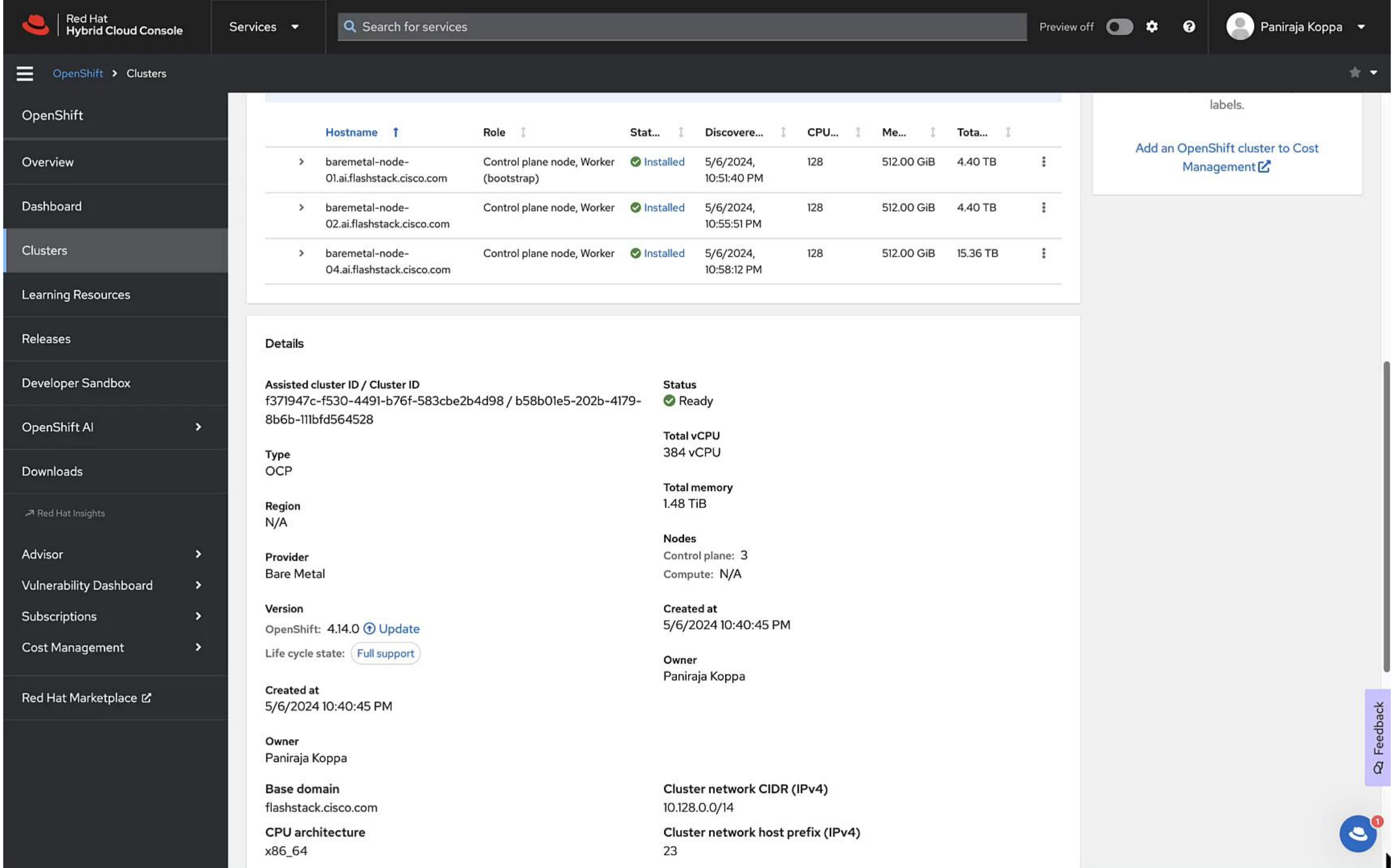
Streamline and scale using MLOps

- Scalability
- Governance
- Efficiency
- Reliability
- Adaptability

Iterate

Prepare Data	Experiment/Tune model	Serve and integrate with App	Monitor/Maintain model
Gathering & preparing data for AI	Apply scientific rigor to understand data and build/customize model	Model available for production inferencing	Track model quality, metrics and drift

Pace of AI/ML technology shifts require a strong foundation to adapt


Red Hat
Hybrid Cloud Console
Services
Search for services
Preview off
Paniraja Koppa
OpenShift
Clusters
OpenShift
Overview
Dashboard
Clusters
Learning Resources
Releases
Developer Sandbox
OpenShift AI
Downloads
Red Hat Insights
Advisor
Vulnerability Dashboard
Subscriptions
Cost Management
Red Hat Marketplace
Hostname
Role
Stat...
Discovere...
CPU...
Me...
Tota...
baremetal-node-01.ai.flashstack.cisco.com
Control plane node, Worker (bootstrap)
Installed
5/6/2024, 10:51:40 PM
128
512.00 GiB
4.40 TB
baremetal-node-02.ai.flashstack.cisco.com
Control plane node, Worker
Installed
5/6/2024, 10:55:51 PM
128
512.00 GiB
4.40 TB
baremetal-node-04.ai.flashstack.cisco.com
Control plane node, Worker
Installed
5/6/2024, 10:58:12 PM
128
512.00 GiB
15.36 TB
labels.
Add an OpenShift cluster to Cost Management
Details
Assisted cluster ID / Cluster ID
f371947c-f530-4491-b76f-583cbe2b4d98 / b58b01e5-202b-4179-8b6b-111bfd564528
Type
OCP
Region
N/A
Provider
Bare Metal
Version
OpenShift: 4.14.0 Update
Life cycle state: Full support
Created at
5/6/2024 10:40:45 PM
Owner
Paniraja Koppa
Status
Ready
Total vCPU
384 vCPU
Total memory
1.48 TiB
Nodes
Control plane: 3
Compute: N/A
Created at
5/6/2024 10:40:45 PM
Owner
Paniraja Koppa
Base domain
flashstack.cisco.com
CPU architecture
x86_64
Cluster network CIDR (IPv4)
10.128.0.0/14
Cluster network host prefix (IPv4)
23
Feedback

Cisco Validated Designs (CVD's) for AI

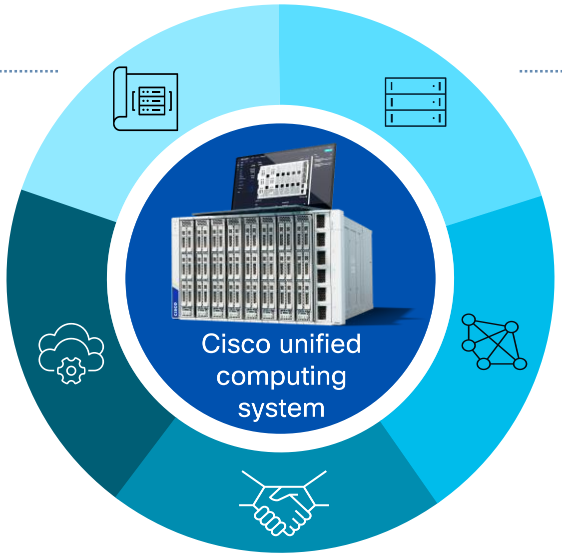
Cisco Validated Designs (CVD)

Accelerate

Ready to 'Go' solutions for faster time to value

Less risk

Reduce risk with tested architectures for standardized, repeatable deployments



Expert Guidance

CVDs provide everything from system designs to implementation guides, and ansible automation

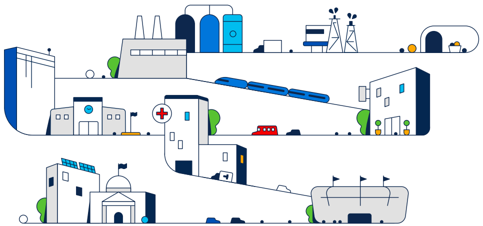
Cisco TAC support

Single point of contact for solution. Cisco will coordinate with partners as needed to resolve issues

Cisco Compute Coverage

UCS ONLY FLEXPOD FLASHSTACK NUTANIX

Explore Cisco validated AI demos showcasing a broad spectrum of AI technologies and practices ready to transform your business



Large Language Models (LLMs)

Discover the power of Large Language Model (LLM) inferencing as it seamlessly processes and generates human-like text in real-time.

Gen AI | NVIDIA AI Enterprise | Hugging Face | NVIDIA TRT-LLM | Text-to-Text

Retrieval Augmented Generation (RAG)

Experience an enterprise-grade Retrieval Augmented Generation (RAG) chatbot delivering responses tailored to your enterprise-specific content.

Gen AI | NVIDIA AI Enterprise | NVIDIA NIM | Vector Database | Text-to-Text

MLOps

Explore the cutting-edge of MLOps, where the efficiency of machine learning workflows meets the rigor of operational excellence.

Gen AI | Red Hat OpenShift AI | LangChain | Mistral | vLLM

Image Synthesis

Immerse yourself in the innovative world of text-to-image synthesis, where vivid images are conjured from descriptive language or existing photos.

Gen AI | NVIDIA AI Enterprise | Hugging Face | Diffusion Models | Text-to-Image

Image Analysis

Delve into the realm of Image Analysis, where advanced algorithms interpret and understand visual data with astonishing accuracy.

Predictive AI | Intel AMX | Kaggle | Keras Neural Network | Image-to-Text

FlexPod for Generative AI Inferencing

Optimized for AI

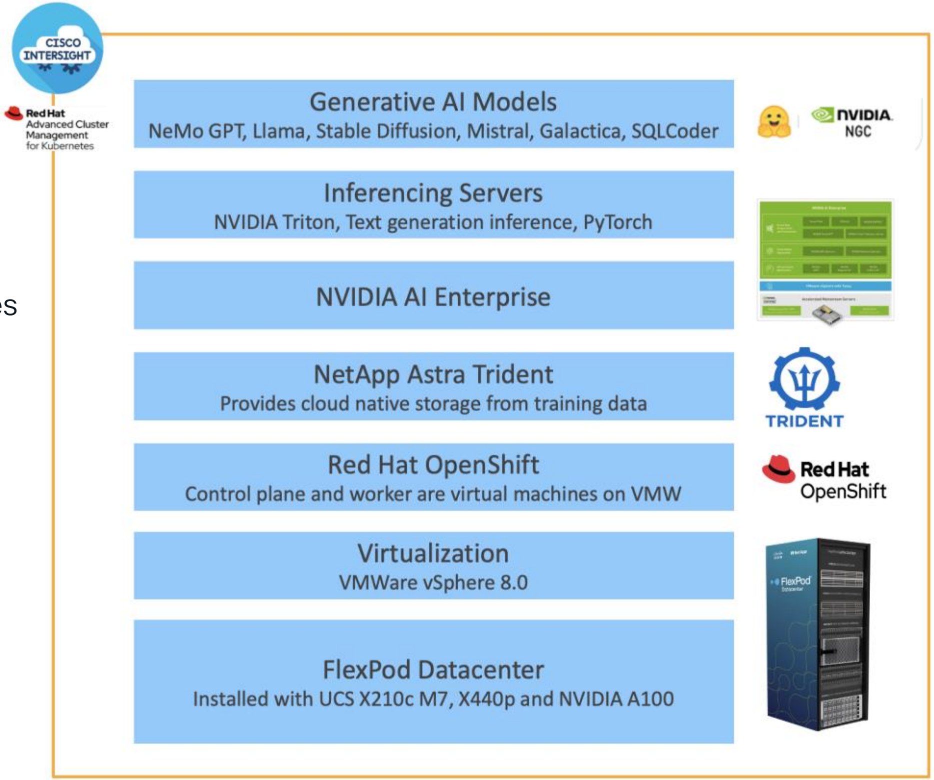
- Comprehensive suite of AI tools and frameworks with NVIDIA AI Enterprise that support optimization for NVIDIA GPU
- Validated NVIDIA NeMo with TRT-LLM that accelerates inference performance of LLMs on NVIDIA GPUs
- Metrics dashboard for insights into cluster and GPU performance and behavior

Accelerated Deployment

- Deployment validation of popular Inferencing Servers and AI models such as Stable Diffusion and Llama 2 LLMs with diverse model serving options
- Automated deployment with Ansible playbook

AI at Scale

- Scale discretely with future-ready and modular design



FlashStack for Generative AI | Inferencing with LLMs

Foundational Architecture for Gen AI

- Validated NVIDIA NeMo Inference with TensorRT-LLM that accelerates inference performance of LLMs on NVIDIA GPUs
- Validated models using Text Generation Inference server from Hugging Face
- Metrics dashboard for insights into infrastructure, cluster and GPU performance and behavior

Simplify and Accelerate Model Deployment

- Extensive breadth of validation of AI models such as GPT, Stable Diffusion and Llama 2 LLMs with diverse model serving options
- Automated deployment with Ansible playbook

Consistent Performance

- Consistent average latency and Throughput
- Better price to performance ratio

Cisco Intersight

Generative AI Models
Nemo GPT, Llama, Stable Diffusion

Inferencing Servers
NVIDIA Triton, Text Generation Inference, PyTorch

NVIDIA AI Enterprise
Advanced AI platform with advanced integration

Portworx Enterprise
Model repository and storage for applications

Red Hat OpenShift
Control plane and worker virtual machines

Virtualization
VMware vSphere

FlashStack Infrastructure
Cisco UCS X210 Compute nodes
Cisco UCS X440p PCIe nodes
Pure Storage FlashBlade or FlashArray
NVIDIA GPU accelerators

NVIDIA NGC

portworx enterprise

Red Hat OpenShift

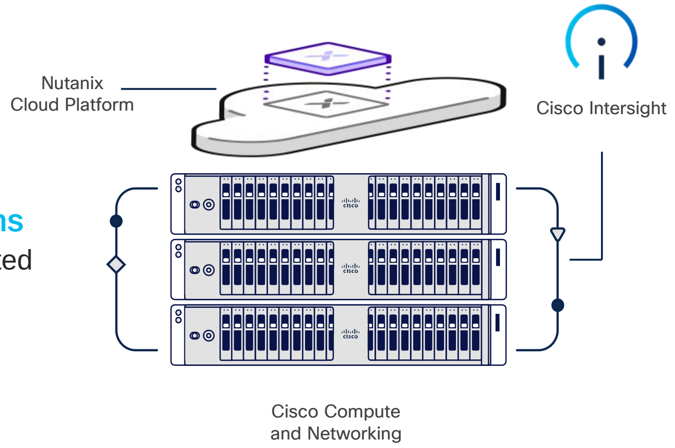
Cisco and Nutanix partner for AI: The Power of Two

Chat GPT-in-a-box

AI Everywhere
Existing apps and new experiences

Proven platforms
CVDs and automated playbooks

Secure foundation
End-to-end resiliency



Cisco Compute Hyperconverged GPT-in-a-Box

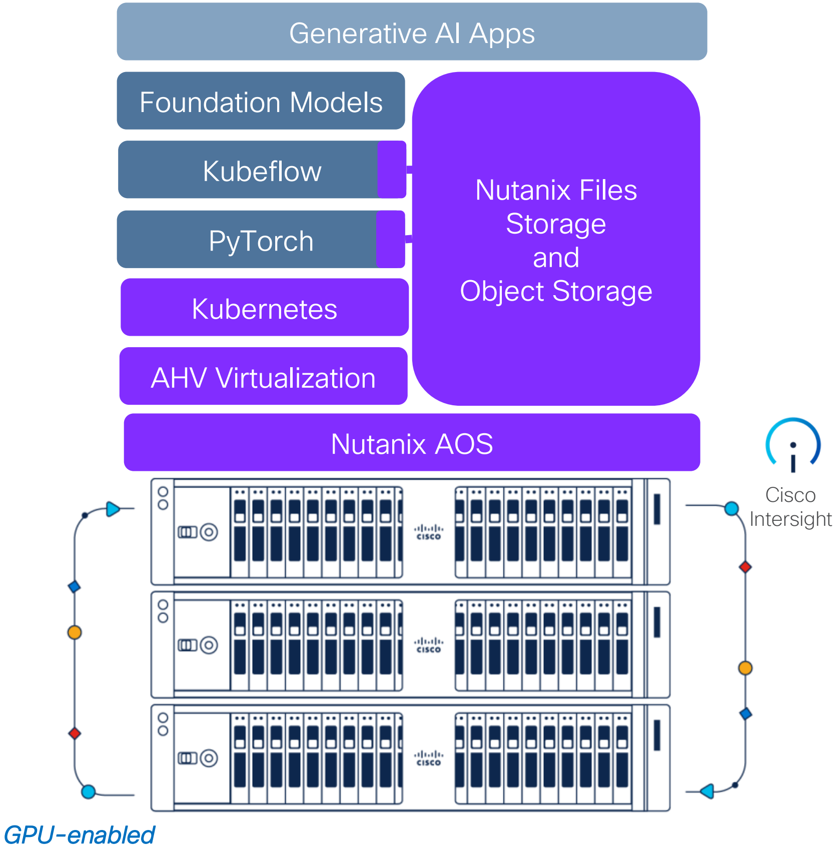
Deploy hybrid-cloud AI-ready clusters with Cisco Validated Designs (CVDs)

Business Challenges

- Optimized GenAI infrastructure
- Streamlined governance with enterprise software
- Sustainable energy use
- Hybrid cloud is complex

Benefits

- Risk reduction & fast time to market
- Streamline operations
- Proven performance
- Protect valuable data
- Simplified hybrid cloud operations



CVDs to simplify end-to-end AI infrastructure

1

CVD blueprint for AI networks

Best performing AI/ML networks, focus on application performance

Intelligent buffer, low latency, telemetry and RoCEv2

Dynamic congestion avoidance

One IP network for both front-end and back-end

Automation for day-2 operations

Validated designs for network and ecosystem partners

2 EXPANDED ROADMAP

CVDs for simplified AI-ready infrastructure

NVIDIA AI Enterprise

Red Hat OpenShift AI

GPT-in-a-box on Nutanix Hyperconverged

Gen-AI with Cloudera Data Platform

FlashStack

FlexPod

intel

3 NEW

CVD playbooks supporting common AI models

Large language models (GPT3, BERT, T5)

Computer vision models (ResNet, EfficientNet, YOLO)

Generative models (GANs, VAEs)

NVIDIA NGC

intel Developer Cloud

Future Trends and Industry Impacts of AI Infrastructure Demands

AI drives a better future

With a new kind of data center

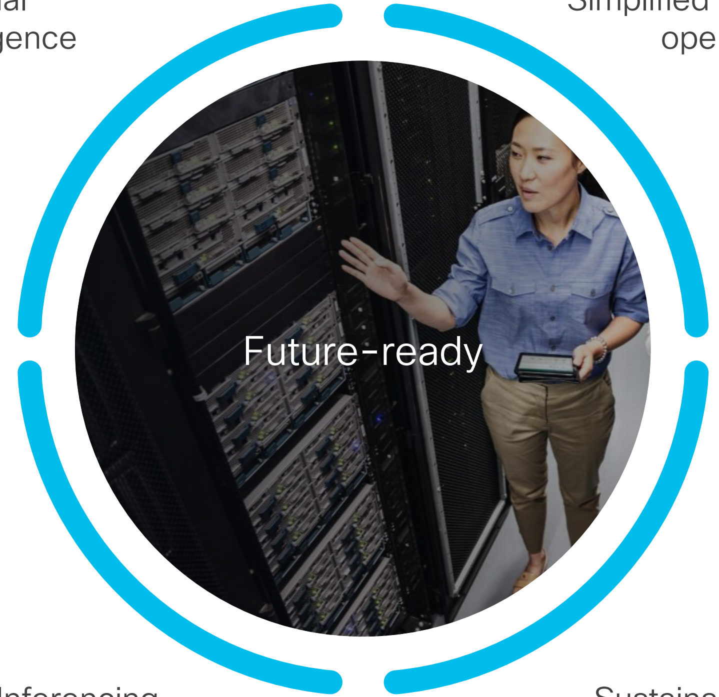
Simple, sustainable, future-ready

More programmability and control

Less operational complexity

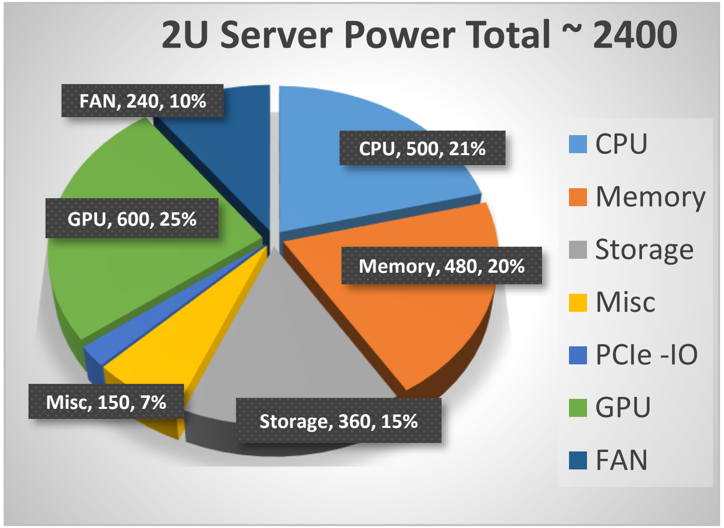
More efficient performance for new workloads

Less costly to build, deploy, and operate



Power & Cooling Trends

- CPU, GPU and Switch ASIC power requirements moving from ~350W TDP today to 400W+ and far beyond in the coming year(s)
- Traditional fan cooling consumes lot of power and less efficient as system power increases
- Passive cooling is approaching its limitation
- Liquid cooling technology to address future cooling requirement with significantly better cooling efficiency & reduced noise levels
- Closed loop liquid cooling provides a retrofit solution
- Future Data Center designs will need to provision for Rack level liquid cooling infrastructure (with external Cooling Distribution Unit - CDU)



Liquid Cooling Technologies

- PAO6: Zero GWP, cheaper, lower cooling capability
- FC-40: Better cooling, higher GWP
- Material compatibility

Single-Phase Immersion

- Better cooling, FC-3284
- Heatsink design is boiling enhancement coating
- Material compatibility
- High GWP

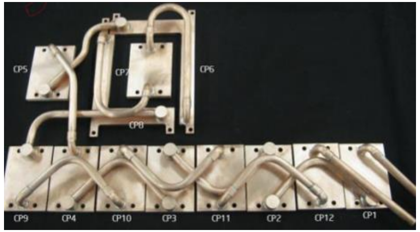
Two-Phase Immersion

- Better cooling, PG25
- Zero GWP
- Leaks can be catastrophic
- Requires parallel connections to avoid pre-heat

Single-Phase Cold Plate

- Better cooling, R134a, Novec7000 or other refrigerant
- Enables highly dense systems, series connections ok
- Leaks not catastrophic

Two-Phase Cold Plate



Compute Express Link (CXL)

Disaggregation Technologies

- Alternate protocol that runs across the standard PCIe physical layer
- Uses a flexible processor port that can auto-negotiate to either the standard PCI transaction protocol or alternate CXL transaction protocols
- First generation CXL aligns to 32 Gbps PCIe 5.0
- CXL usage expected to be a key driver for an aggressive timeline to PCIe 6.0
- Allows you to build fungible platforms

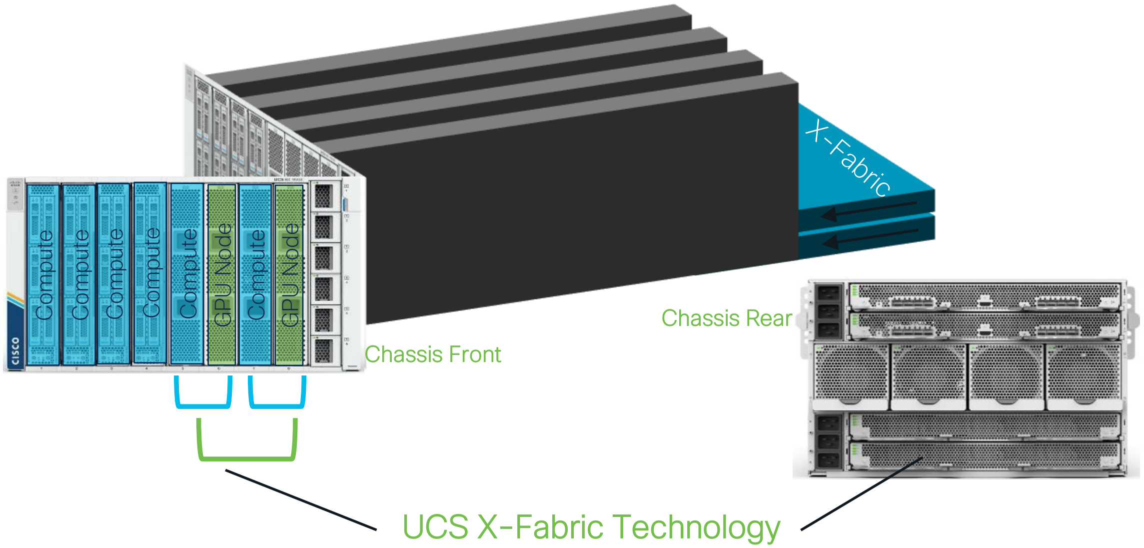
UCS X-Fabric Technology For Disaggregation

Open, modular design enables compute and accelerator node connectivity

Open standards: PCIe 4/5/6, CXL*
Not just another PCIe switch

No midplane nor cables = easy upgrades

Expandability to address new use cases in future (memory & storage nodes)



CXL will evolve out of PCIe for next generation speeds, cache coherency, shared-IO, memory

Expanding Ecosystem of Viable GPU Options

Available Now

GAUDI®

Native RoCE
Scaleup & out

Available via:
- HLS-1 Server (x8)
- SMC Server (x8)
- SDSC
- Public Cloud AWS: EC2

Available Now

GAUDI®2
(7nm)

Native RoCE
Scaleup & out

Available via:
- HLS-Gaudi2 Server (x8)
- SMC Server (x8)
- Aivres/IEI Server (x8)
- Intel Dev Cloud

1H CY 2024

GAUDI®3
(5nm)

Native RoCE
Scaleup & out

2024

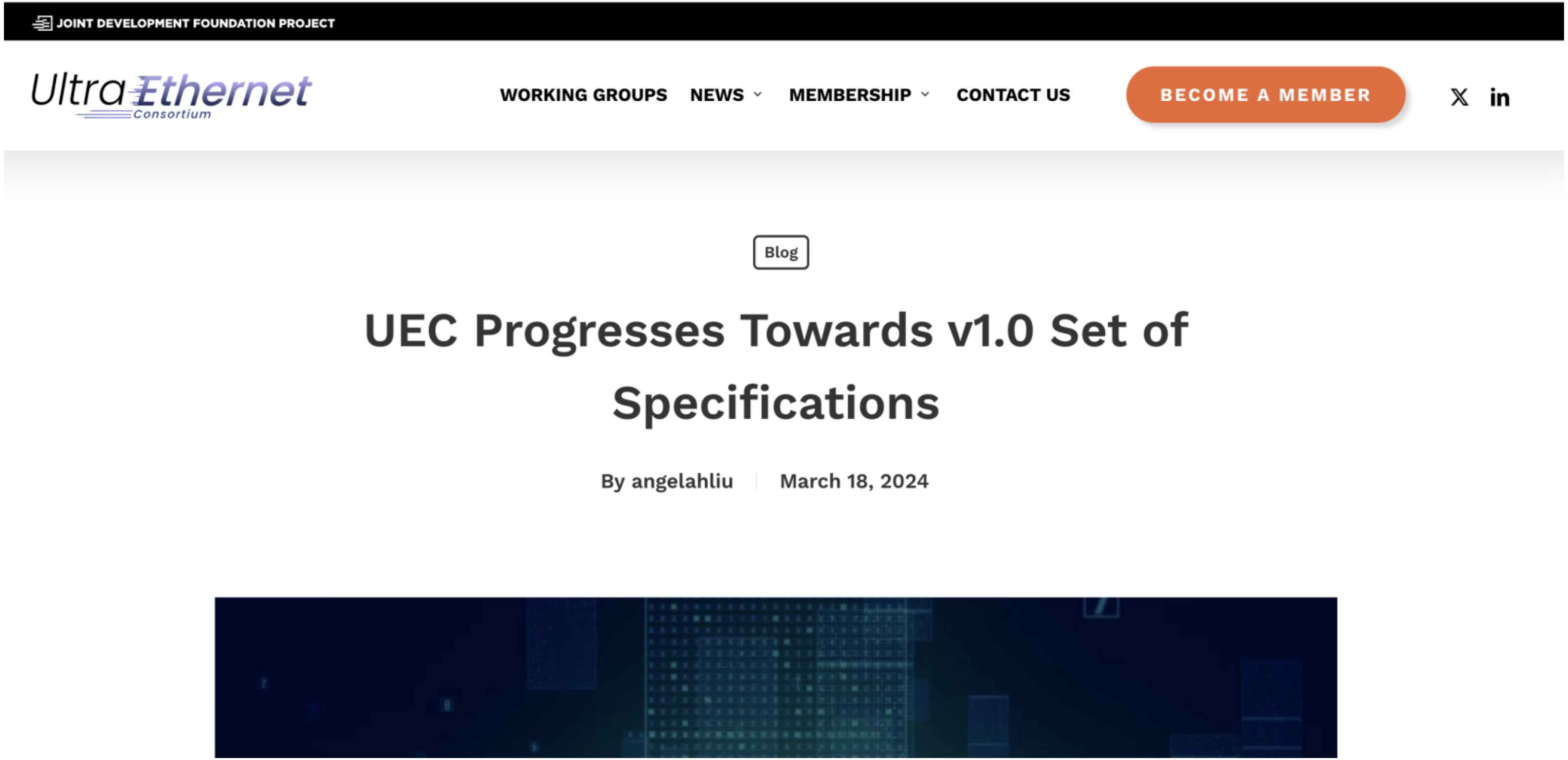
In Development

Next Generation
AI Accelerator:
Falcon Shores 1

Native RoCE
Scaleup & out

2025

Ultra Ethernet Consortium - UEC



https://ultraethernet.org/uec-progresses-towards-v1-0-set-of-specifications/

Ultra Ethernet Consortium - UEC

JOINT DEVELOPMENT FOUNDATION PROJECT

UltraEthernet Consortium

WORKING GROUPS NEWS MEMBERSHIP CONTACT US BECOME A MEMBER

Some of the key features described in the white paper are:

- Multi-path packet spraying
- Flexible ordering
- "State of the art", easily configured congestion control mechanisms
- End-to-end telemetry
- Multiple transport delivery services
- Switch offload (i.e., In-Network Collectives)
- Security as a first-class citizen co-designed with the transport
- Ethernet Link and Physical layer enhancements (optional)

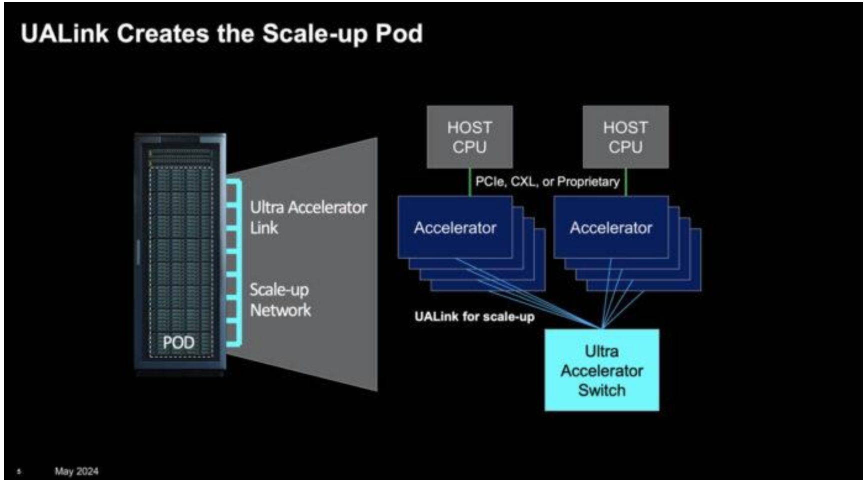
Open Standard NVLink Alternatives

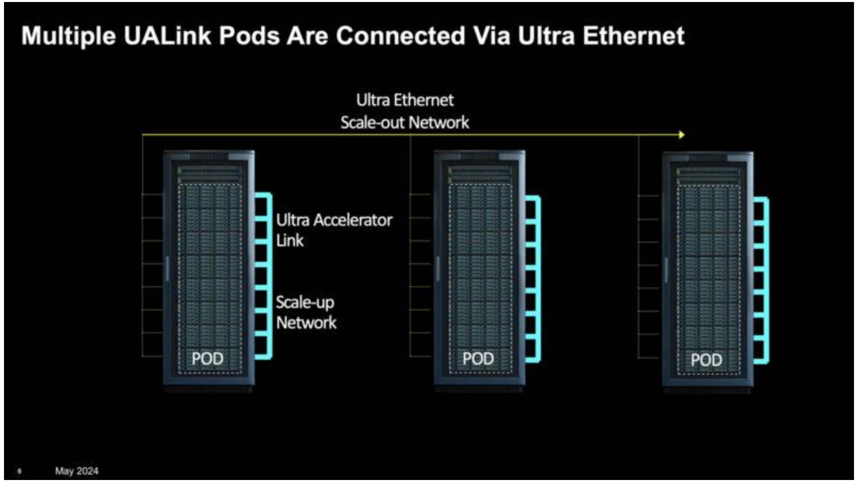
Introduction of Ultra Accelerator Link (UALink)

AMD, Broadcom, **Cisco**, Google, HPE, Intel, Meta, and Microsoft are announcing the formation of a group that will form a new industry standard, UALink, to create the ecosystem.

Low Latency, high bandwidth fabric for 100's of accelerators.

Interconnect for GPU<->GPU Communications





Silicon Photonics

Bringing Higher Data Rates, Lower Latency & Reduced Power Consumption

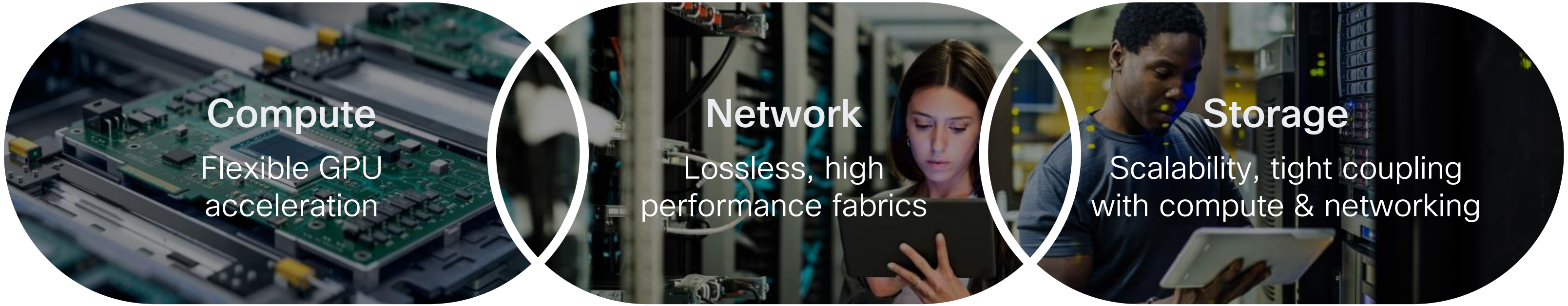
- Fiber Optic Photonics
  - Over length scales of hundreds or thousands of kilometers i.e undersea fiber optic links for internet
  - Majority of optical link involves light in fiber optic cable
  - Source Laser, Periodic Repeaters/amps and photodetector at receiver.
  - All components (lasers, amplifiers, photodetector optical modulators, splitters etc) are discrete and connected.

  == Very costly

- Silicon Photonics
  - Integrated Photonics Technology
  - All optical components directly created on same silicon-on-insulator (SOI) substrate i.e. compact photonics chips that can closely be integrated with CMOS logic.
  - All components are created on same substrate allowing optical components to be packed far denser than discreate optics can achieve.

Summary

Take Aways and Closing
- Cisco Makes AI Hybrid Cloud Possible



AI is pushing infrastructure requirements

Very few customers will train the largest models

Most will use pre-trained models with their own data and deploy associated inference models

The use cases must drive which AI models, methods, and techniques to utilize

AI consultants play a vital role in assessment, guidance, and adoption.

AI is driving the next push for modernized data center facilities, upgraded networks, compute, and storage and operational models

Major investments are not required to start. You can get started with CPU based acceleration and existing infrastructure

Complete Your Session Evaluations

Complete a minimum of 4 session surveys and the Overall Event Survey to be entered in a drawing to **win 1 of 5 full conference passes** to Cisco Live 2025.

Earn 100 points per survey completed and compete on the Cisco Live Challenge leaderboard.

Level up and earn **exclusive prizes!**

Complete your surveys in the **Cisco Live mobile app.**

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 www.CiscoLive.com/on-demand

Contact us at:
eminchen@cisco.com, nicgeyer@cisco.com

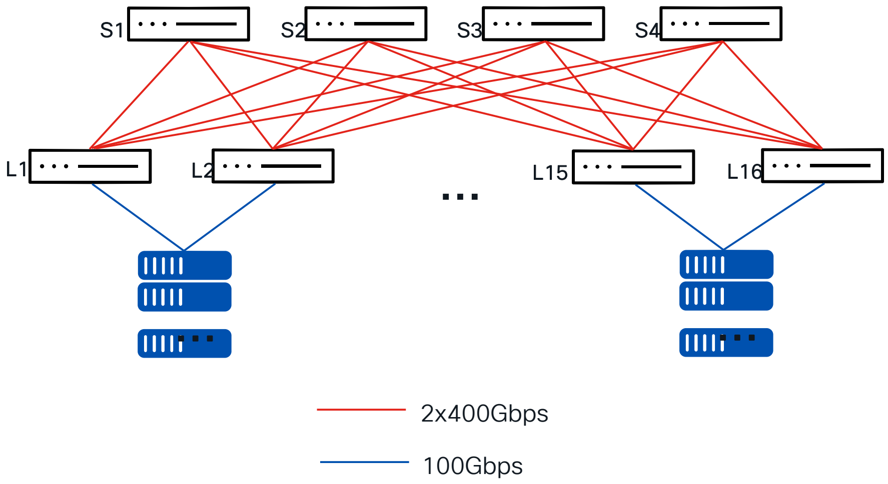
The bridge to possible

Thank you

Congestion in the fabric

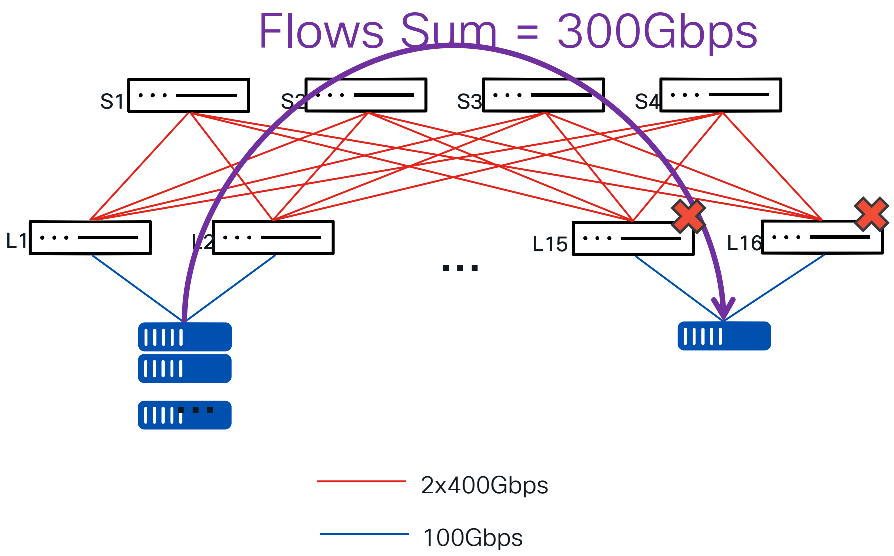
Congestion could always happen

- Congestion can always happen even in a non-blocking switch/fabric
- Let's consider the following example with some maths:
  - 16 ToR, each of them is dual-connected to every spine with 2x200Gbps links
  - Every ToR has 3.2Tbps of uplink capacity
  - Each ToR is attached to 26 dual-homed nodes via 100Gbps links
  - Every node could be firing up 200Gbps of traffic without affecting the uplinks capacity
- *But where is this traffic going?*



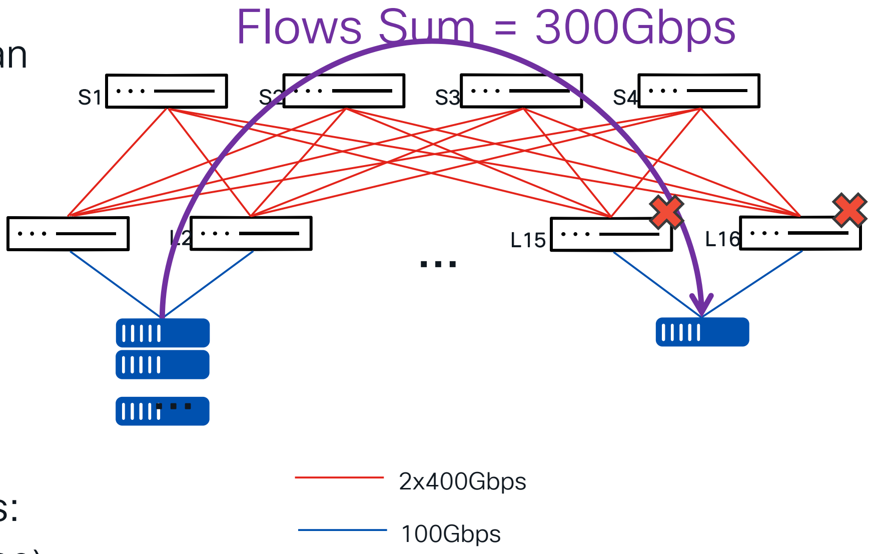
Congestion could always happen

- If traffic traffic aggregated in a node exceeds egress bandwidth capacity then we have congestion
- Impact depends on the data plane protocol.
- Protocols with congestion control capabilities, like TCP, can auto-adjust the flow throughput
- Other protocols, like UDP, have no concept about congestion control.



How RoCEv2 Solves This?

- RoCEv2 MUST run over a lossless network, retransmission must be avoided
- Ethernet networks are lossy by design, drops can happen
- RoCEv2 encapsulates data chunks over IP/UDP packets
- UDP doesn't have a native congestion control mechanism
- RoCEv2 uses the **Data Center Quantized Congestion Notification** scheme that relies primarily on two existing flow control techniques:
    - IP Explicit Congestion Notification (RFC 3168, 1999)
    - Priority Flow Control (802.1Qbb)



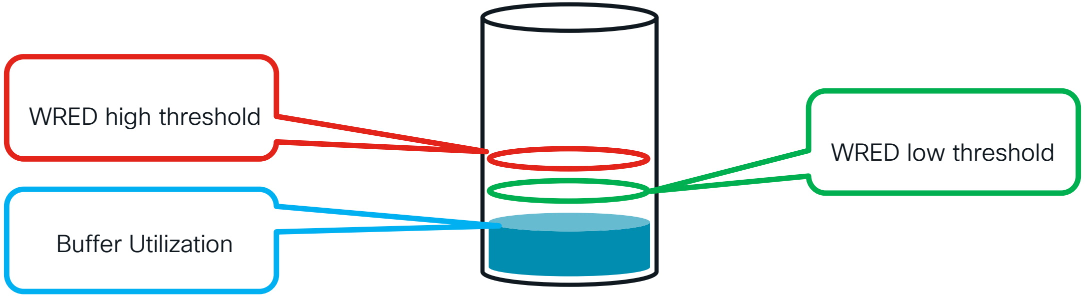
Data Center Quantized Congestion Notification

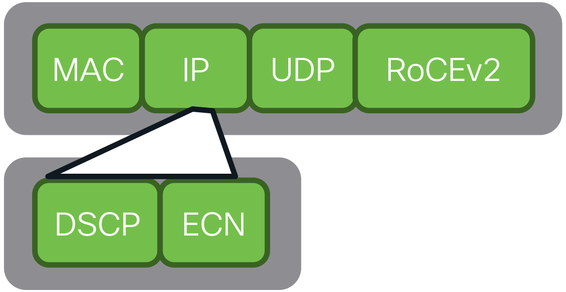
- IP ECN or PFC cannot alone provide a valid Congestion Management framework
- IP ECN signalling might take too long to relieve the congestion
- PFC can could introduce other problems like Head Of Line Blocking and unfairness for the flows
- The two of them together provide the desired result of having lossless RDMA communications across Ethernet networks (this is called **DCQCN**)
- The requirements are:
  - Ethernet devices compatible with both techniques
  - Proper configurations applied

Explicit Congestion Notification

Explicit Congestion Notification

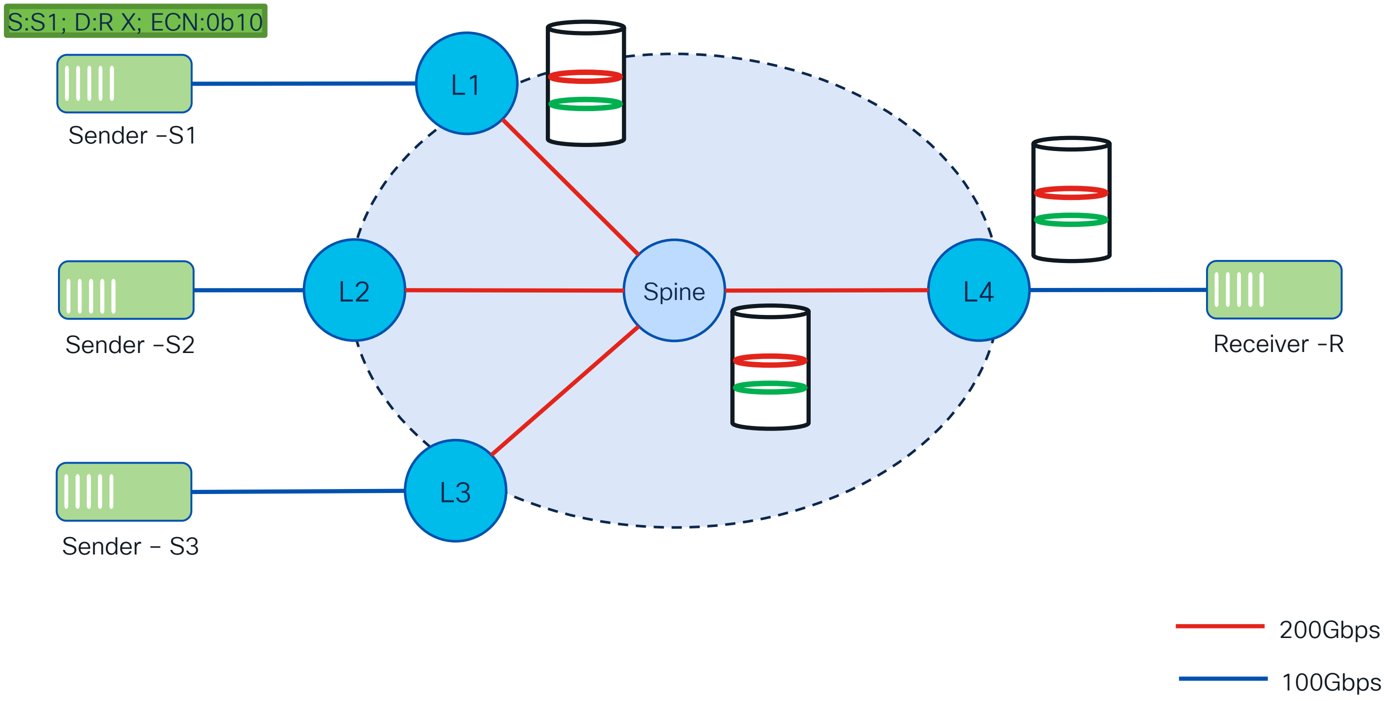
- ECN is implemented via QoS queuing policies leveraging WRED (Weighted Random Early Detection)
- Buffer utilization is constantly monitored, when the buffer goes above the low threshold then <u>some</u> packets get marked with the ECN bits to *0b11*. Only ECN capable packets are marked
- If it goes above the high threshold then <u>all</u> ECN capable packets are marked with 0b11



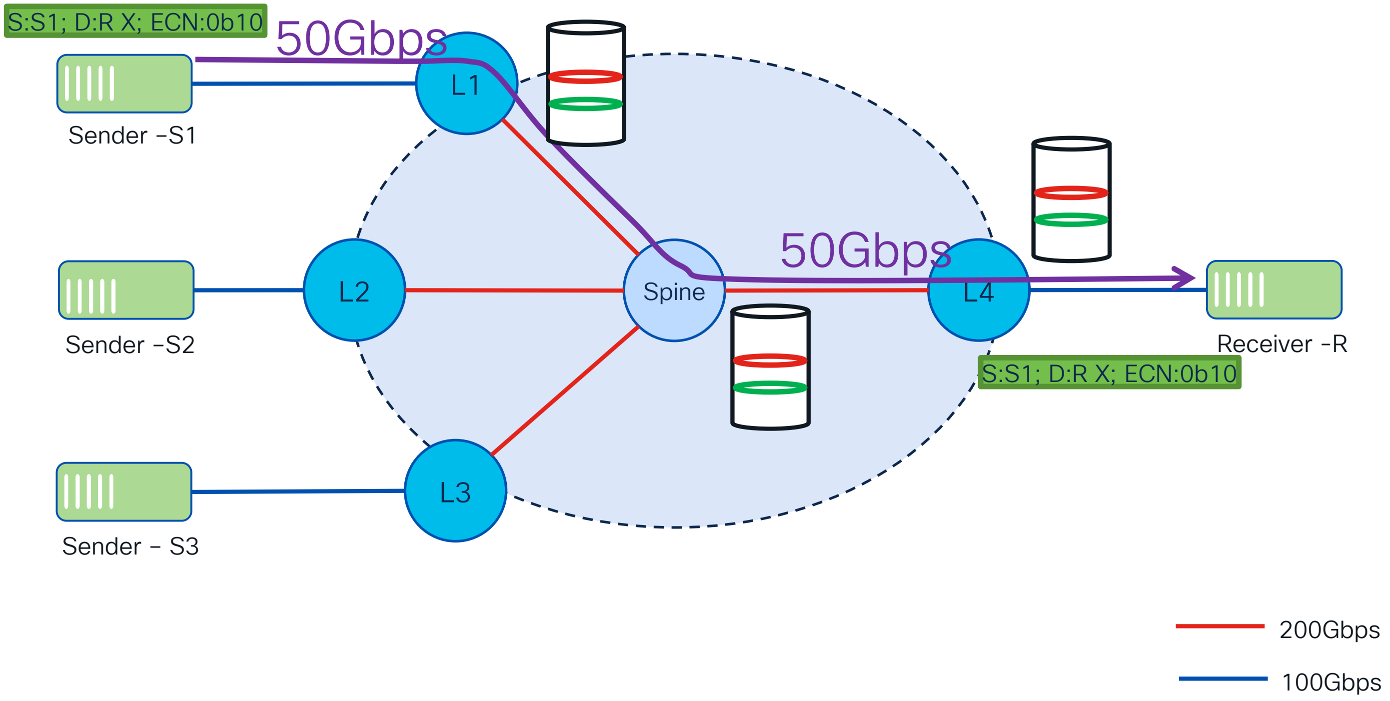


0b 00 --> Non ECN capable
0b 01 --> ECN capable
0b10 --> ECN capable
0b11 --> Congestion Experienced

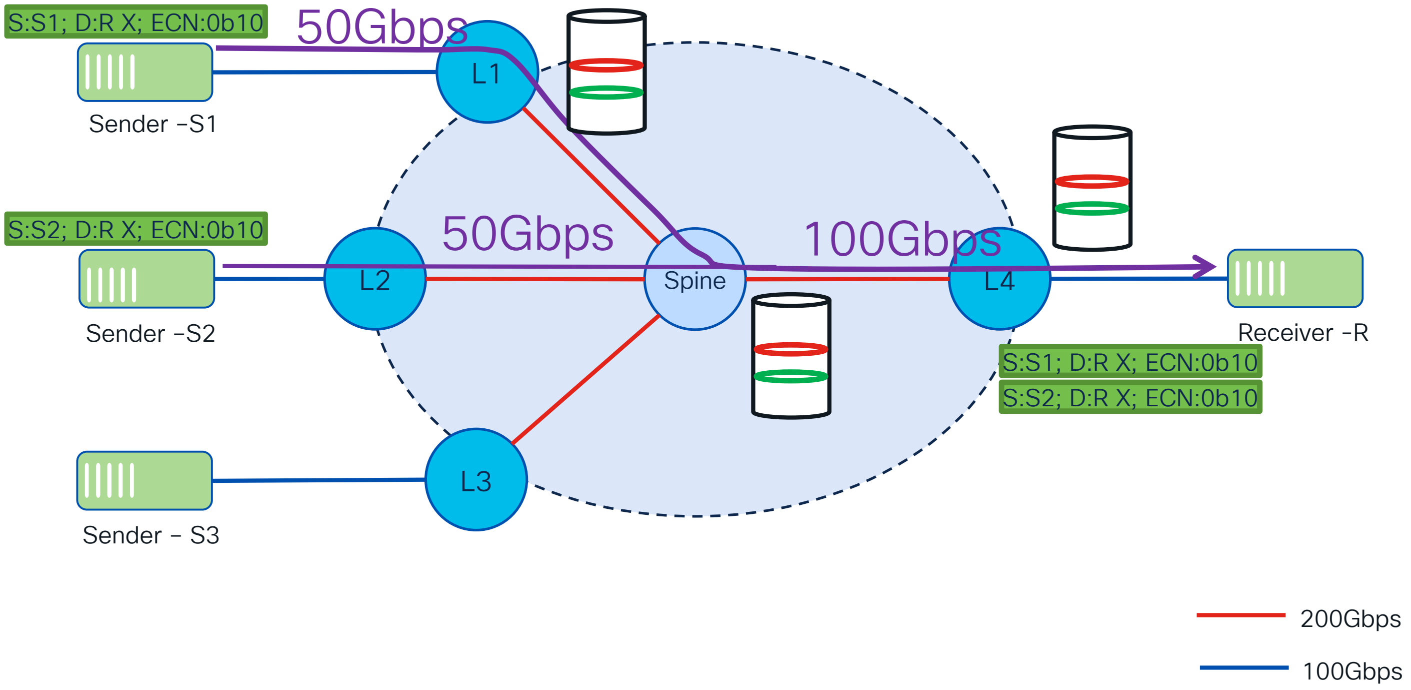
ECN In Action With RoCEv2



ECN In Action With RoCEv2

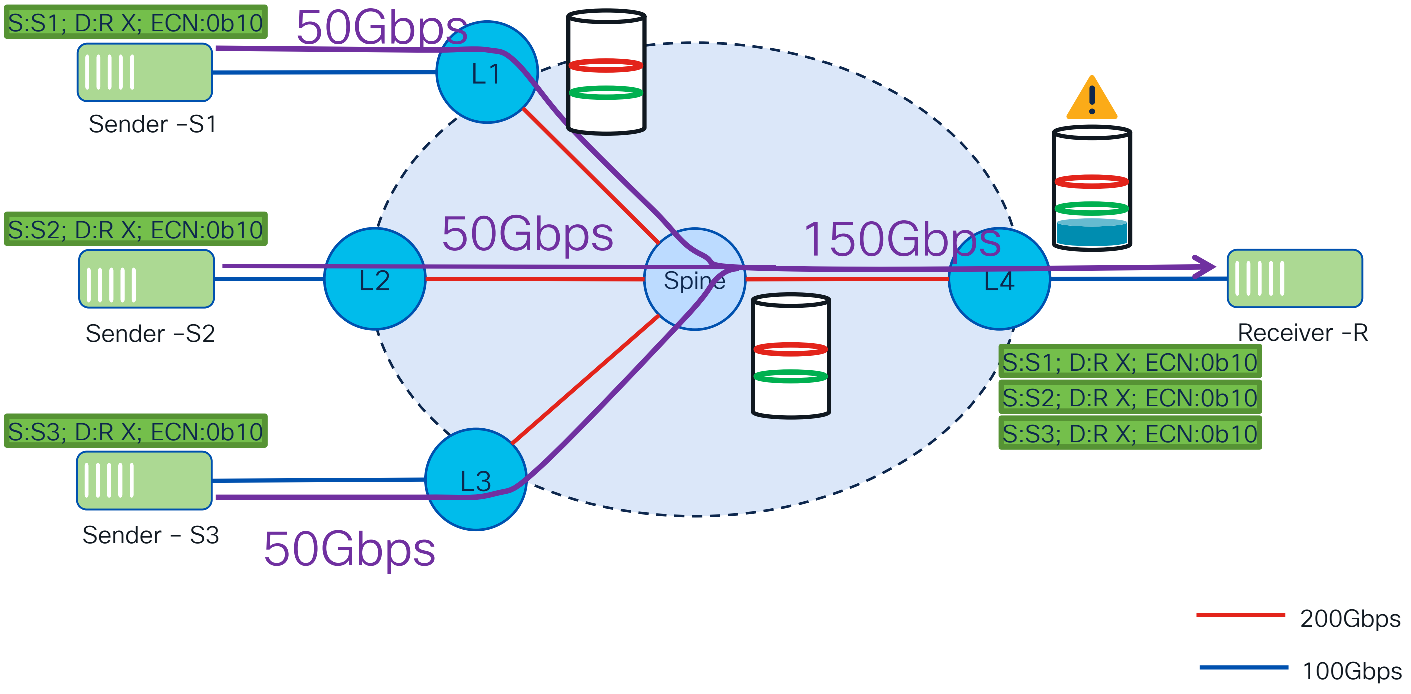


ECN In Action With RoCEv2

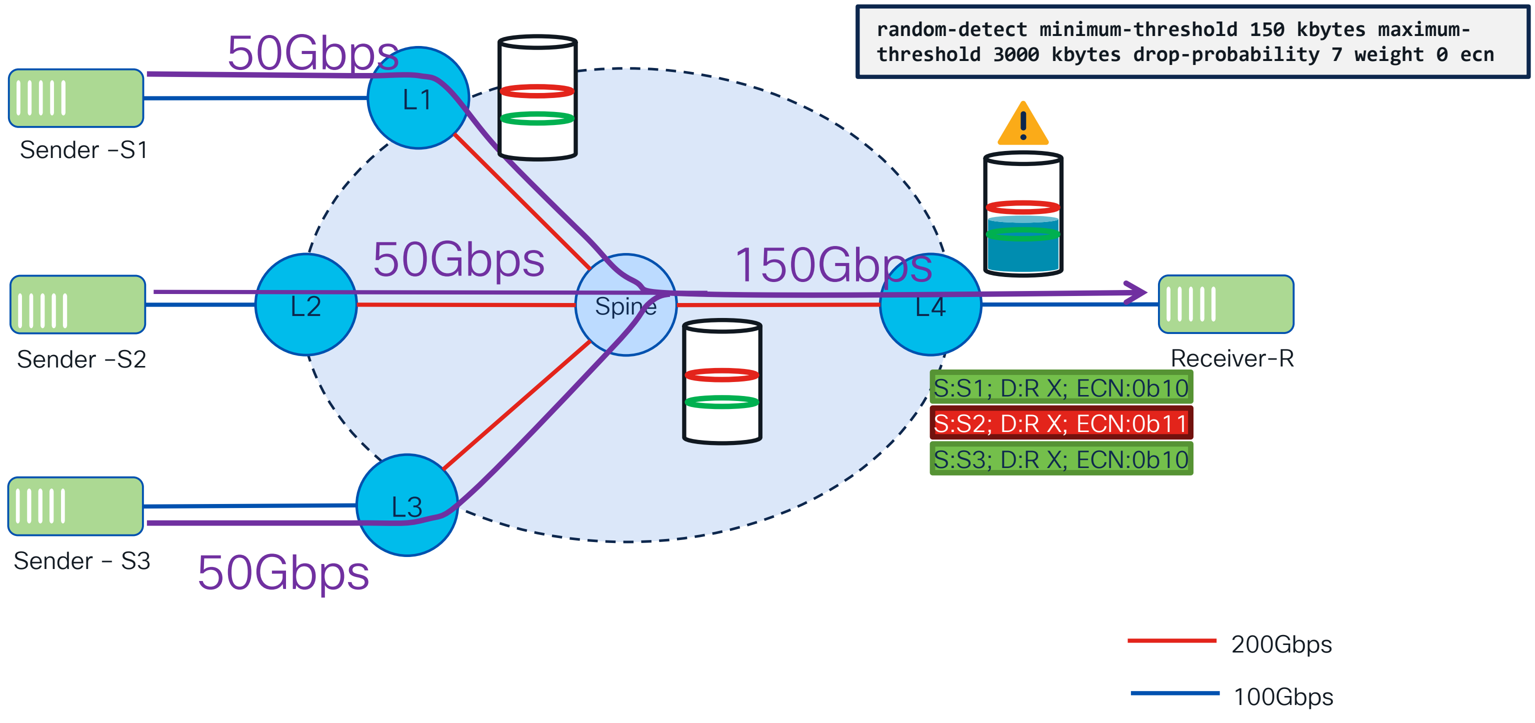


ECN In Action With RoCEv2

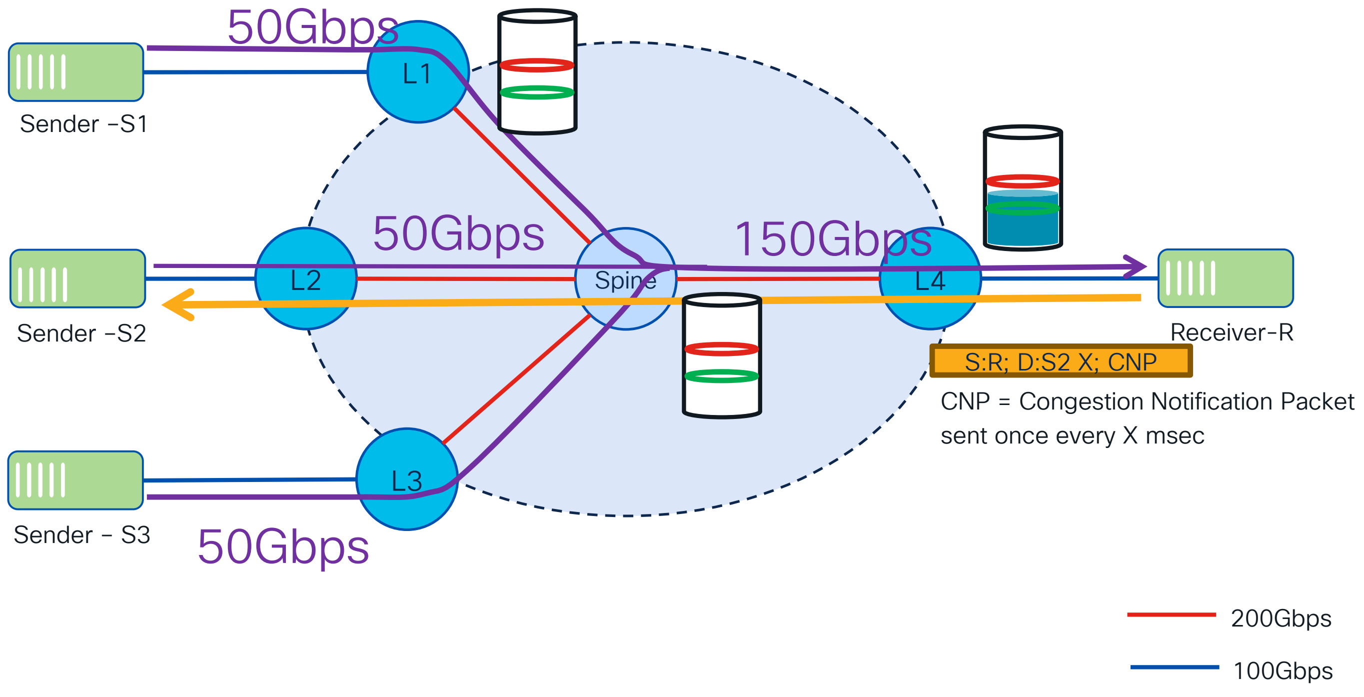
IMPORTANT: The next slides status changes and actions happen in <u>nanoseconds</u>



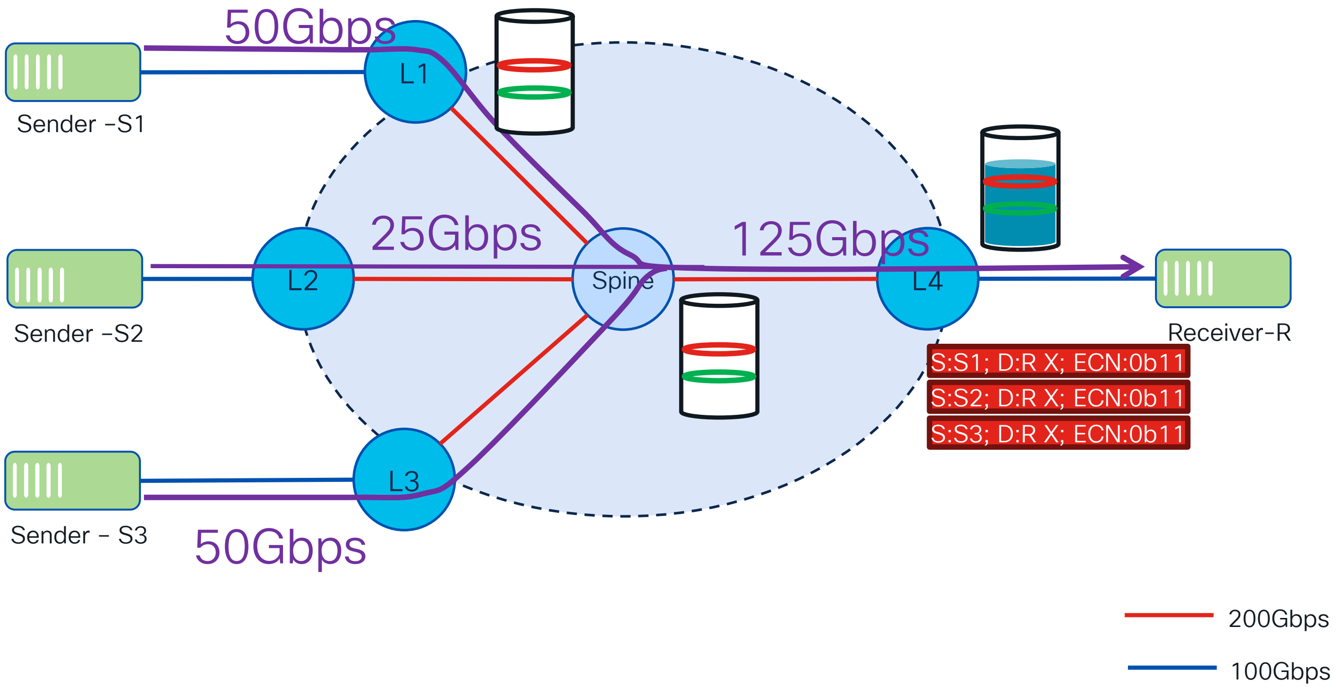
ECN In Action With RoCEv2



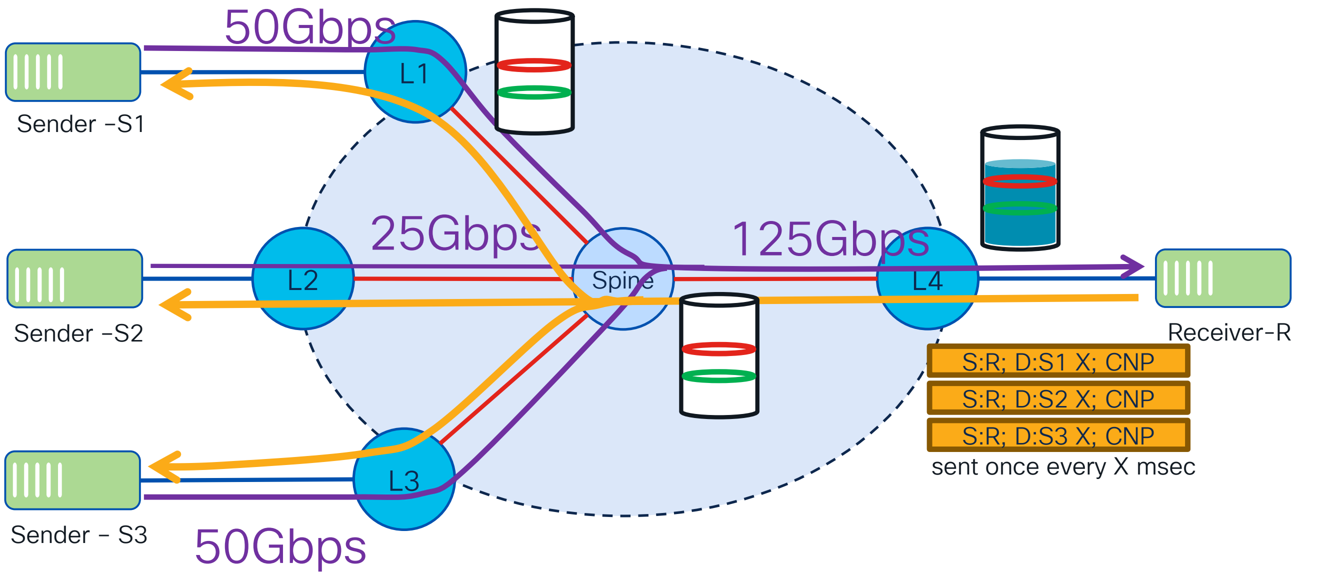
ECN In Action With RoCEv2



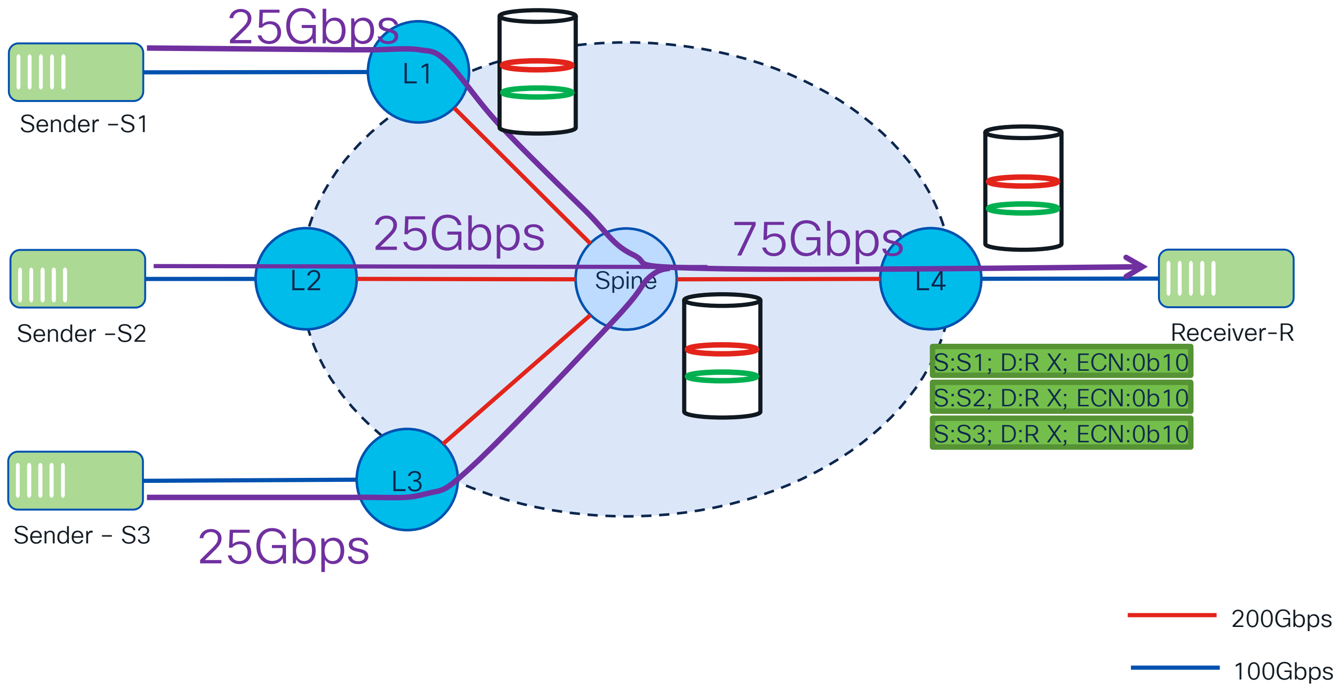
ECN In Action With RoCEv2



ECN In Action With RoCEv2

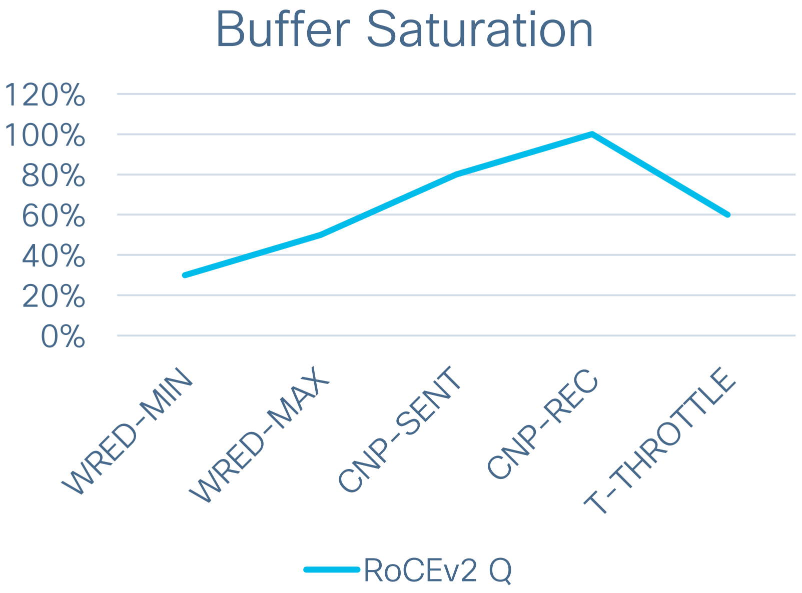


ECN In Action With RoCEv2



ECN In Action With RoCEv2

Considerations

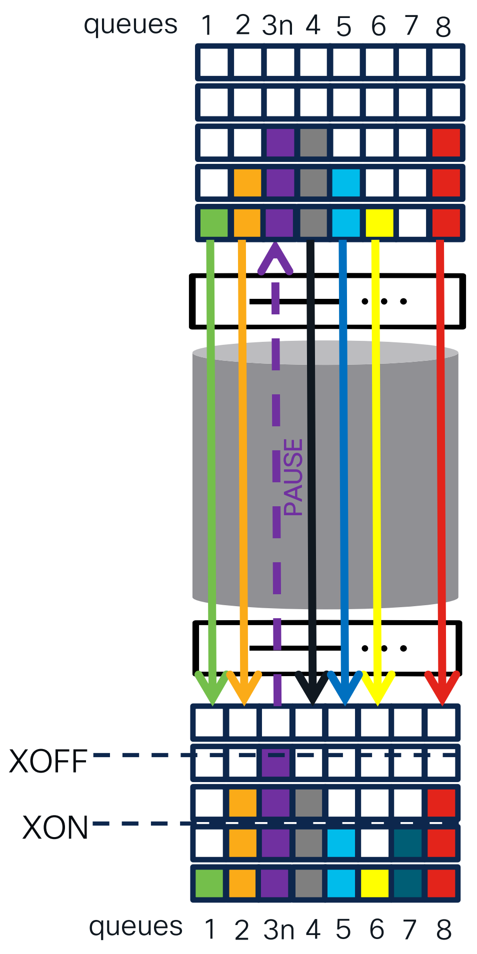


- Latency between ECN marking and subsequent throttling of the throughput rate could be significant
  - CNP packets must be prioritized!
- While notifications are running buffers might get fully saturated and this will cause a tail drop
- This is why DCQCN combines ECN with PFC

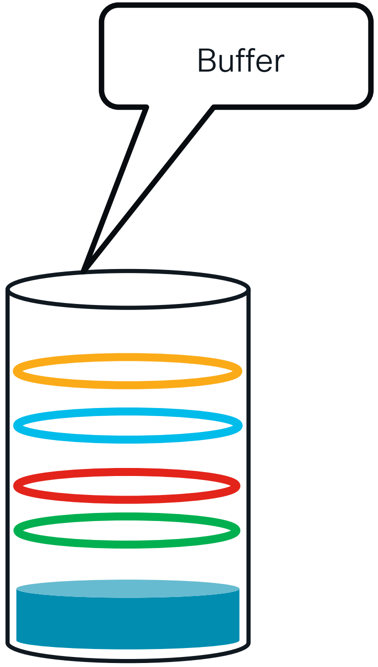
Priority Flow Control

Priority Flow Control

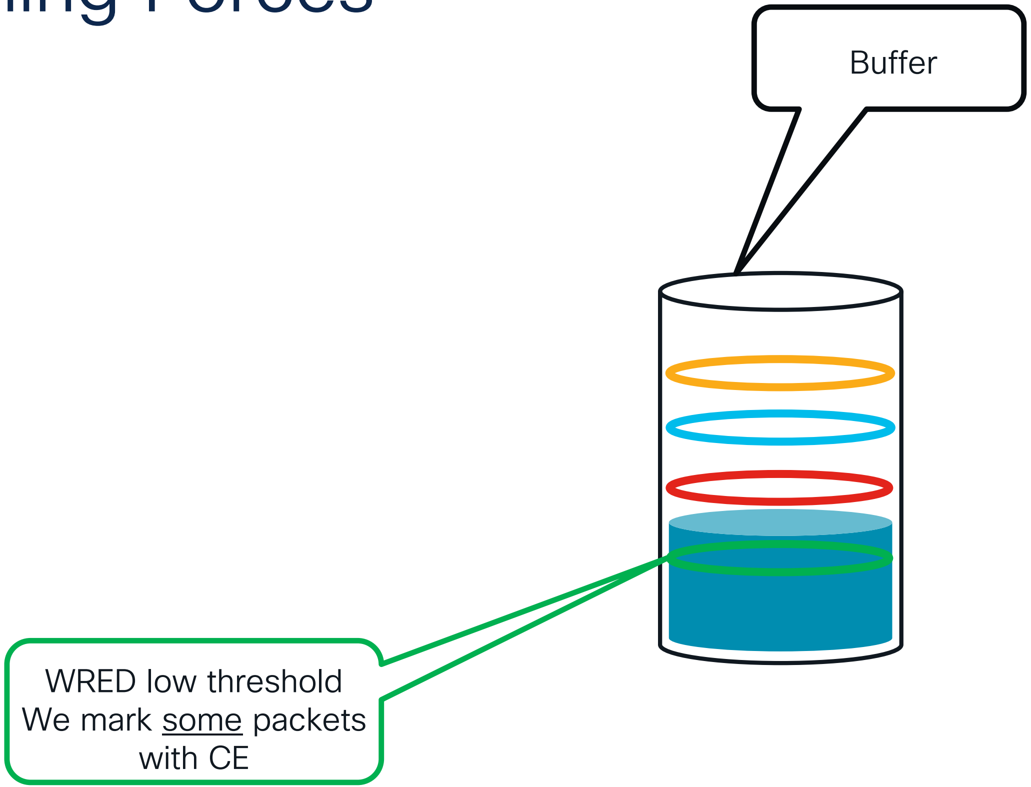
- With PFC we can define a no-drop queue
- Every time the queue reaches a defined threshold the almost saturated device sends pause frames to the devices causing that
- The device which receives it will stop forwarding packets classified for that queue and will place them into its buffer
- The process repeats from here until it reaches the original senders, at that point they will also stop temporarily sending packets
- By the time this happens all the buffers in the network should be flushed and forwarding can start again



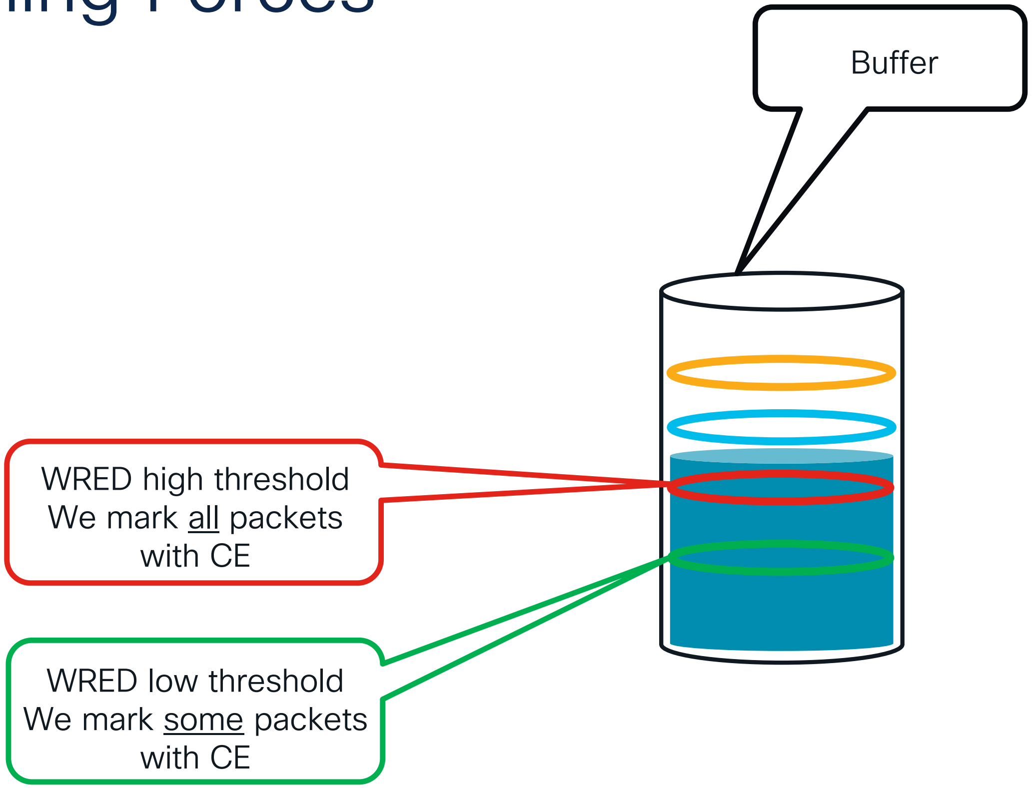
PFC and ECN Joining Forces



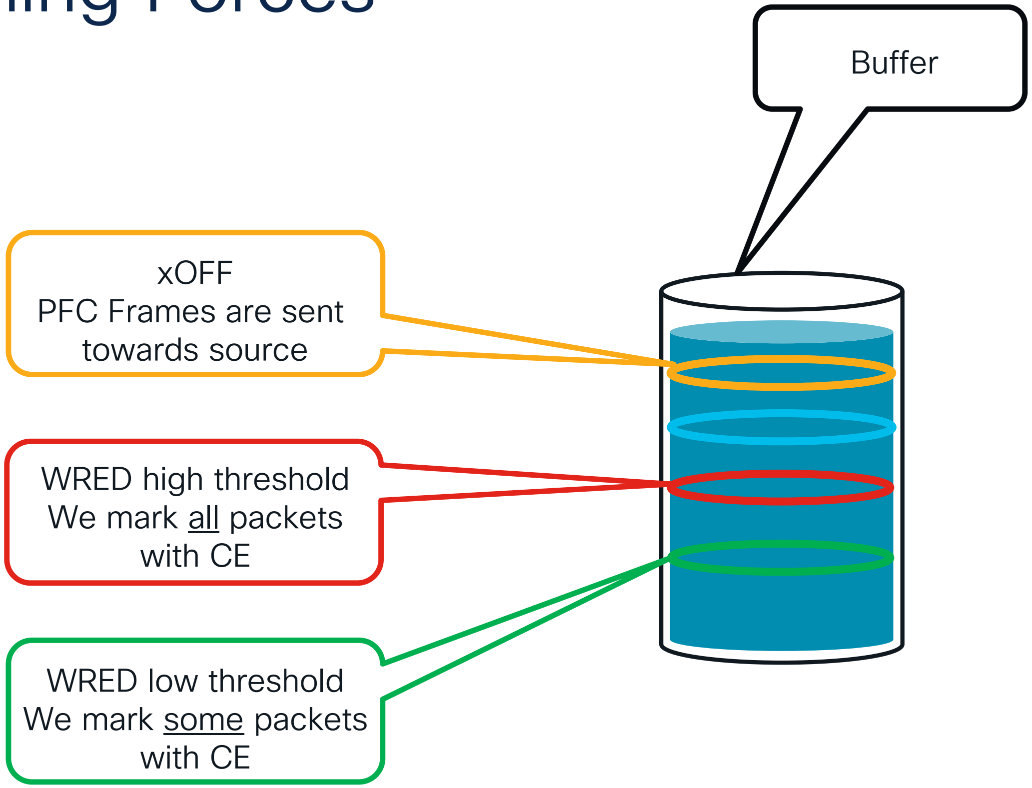
PFC and ECN
Joining Forces



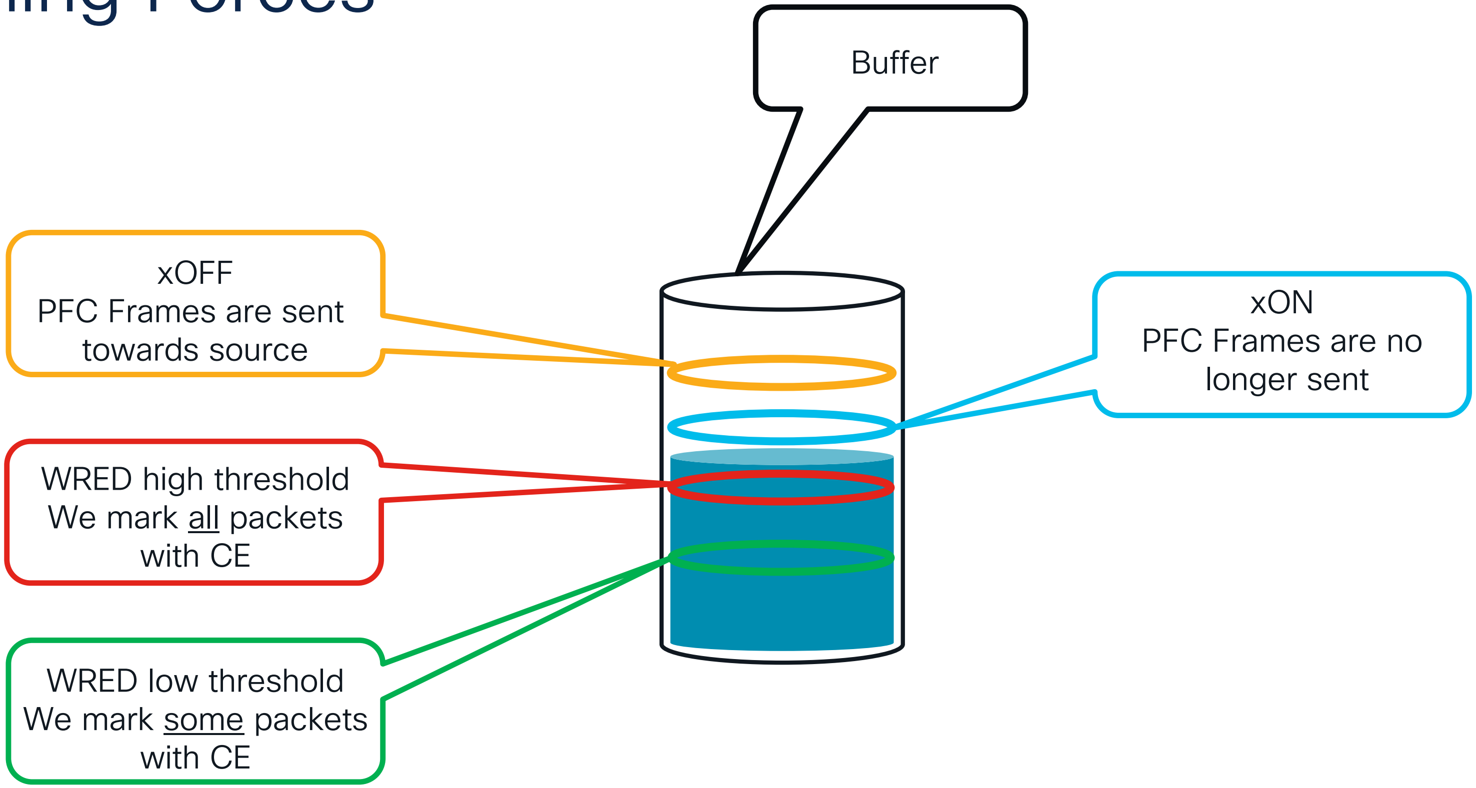
PFC and ECN
Joining Forces



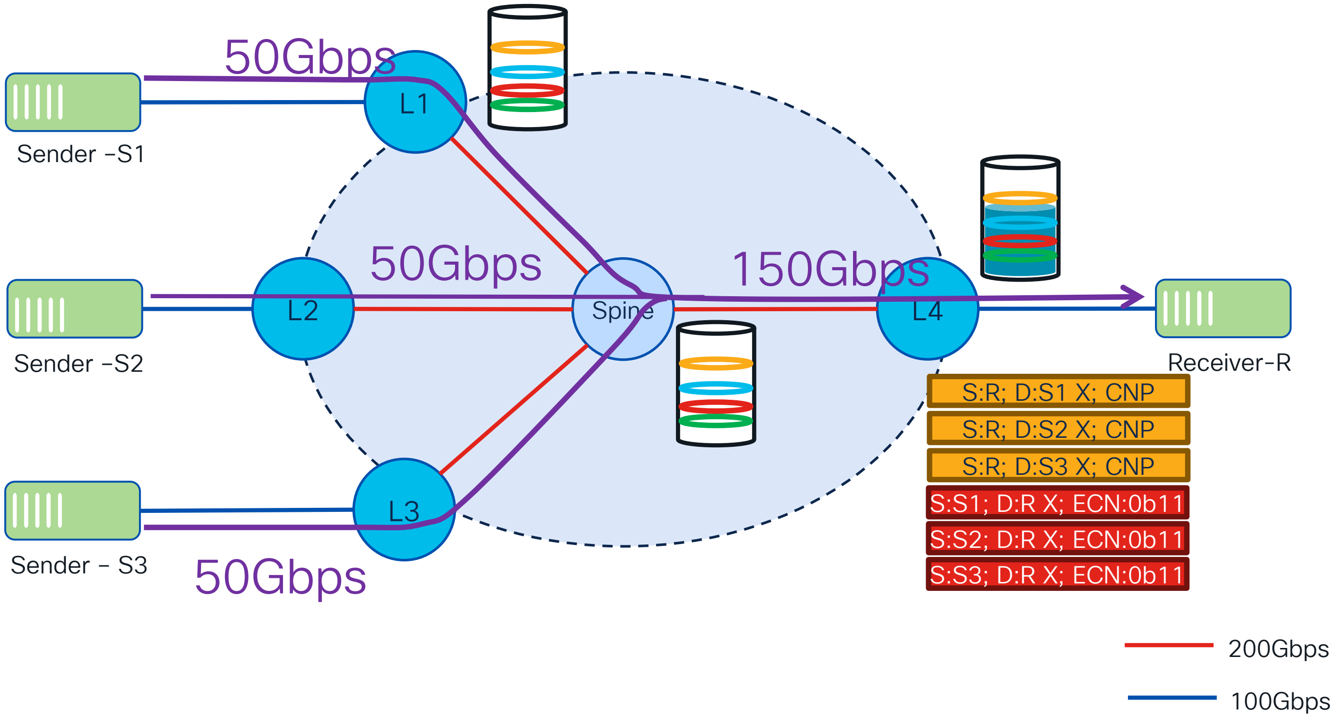
PFC and ECN Joining Forces



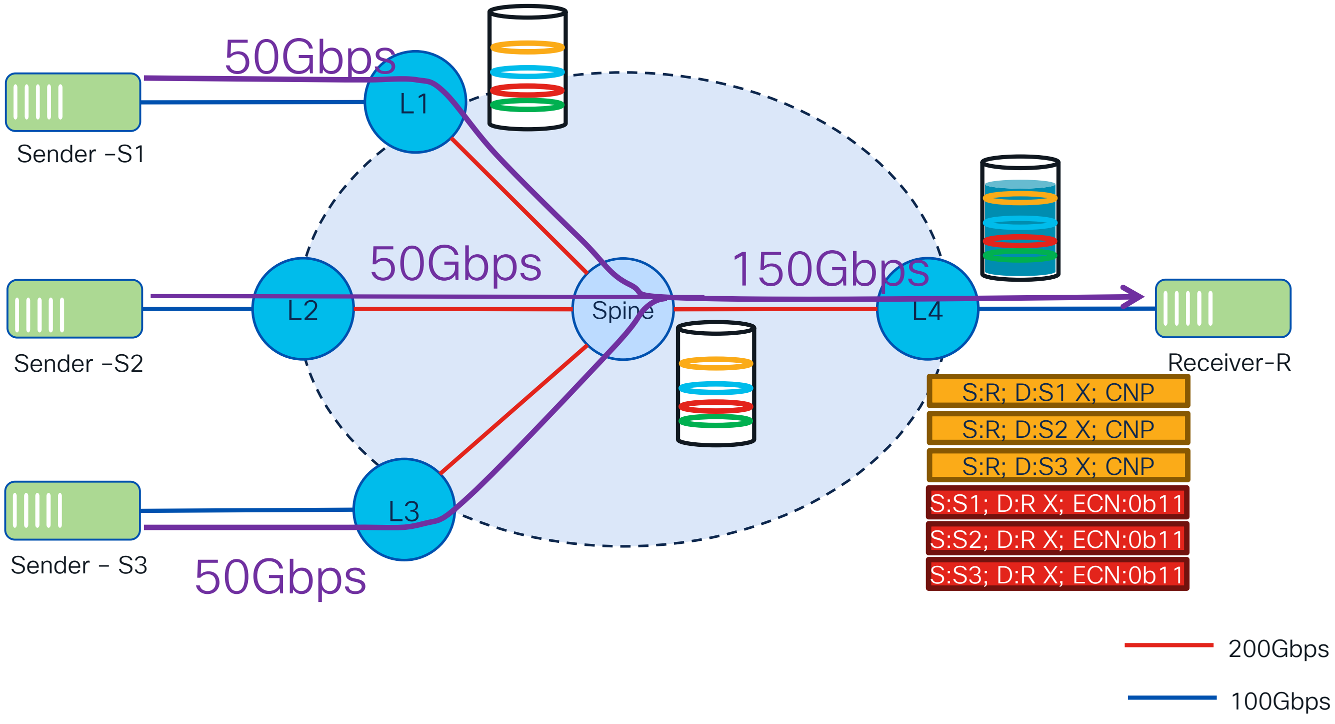
PFC and ECN Joining Forces



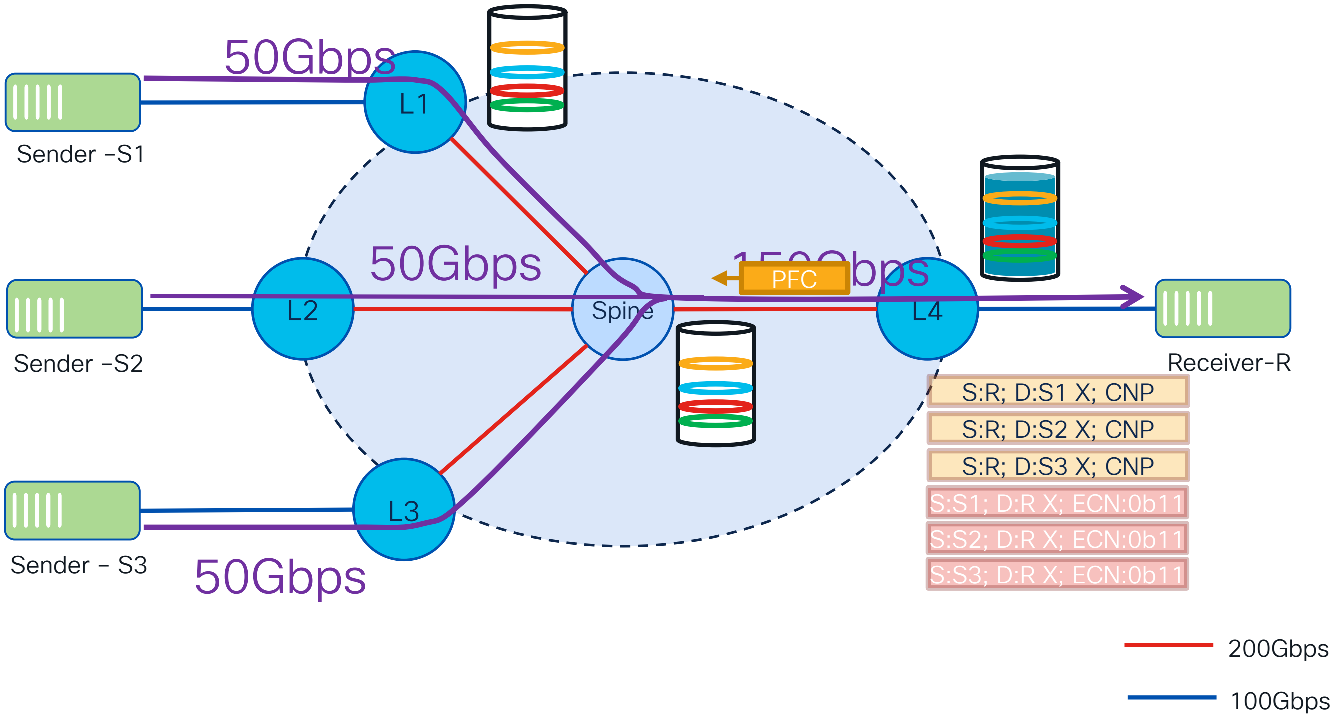
Priority Flow Control In Action With RoCEv2



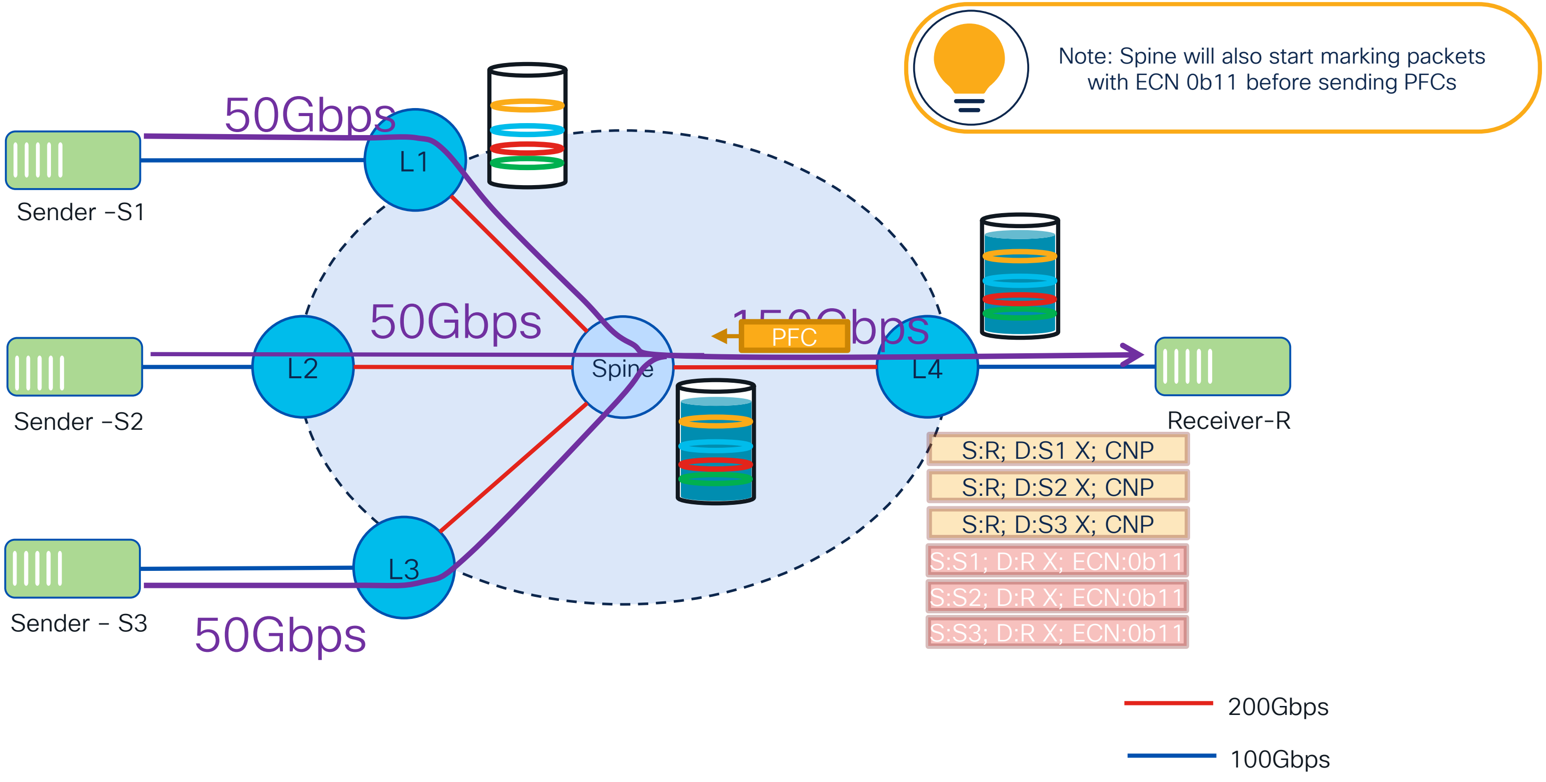
Priority Flow Control In Action With RoCEv2



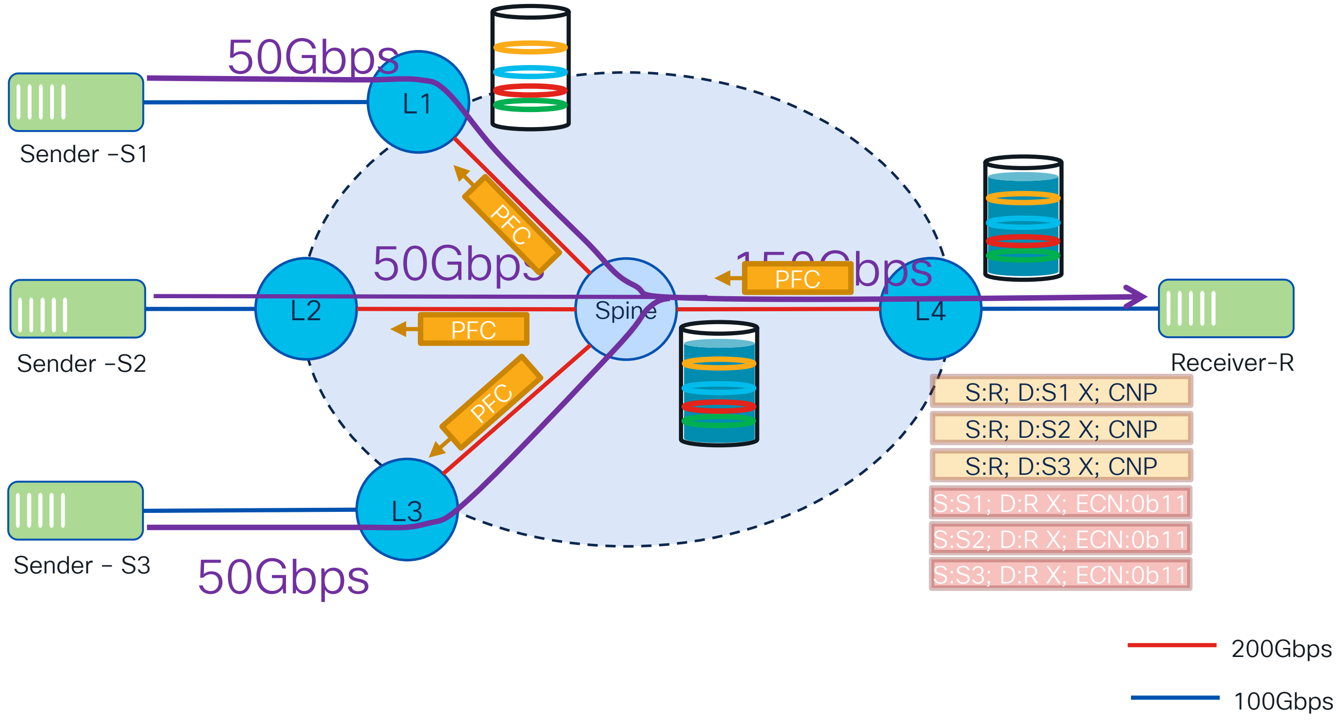
Priority Flow Control In Action With RoCEv2



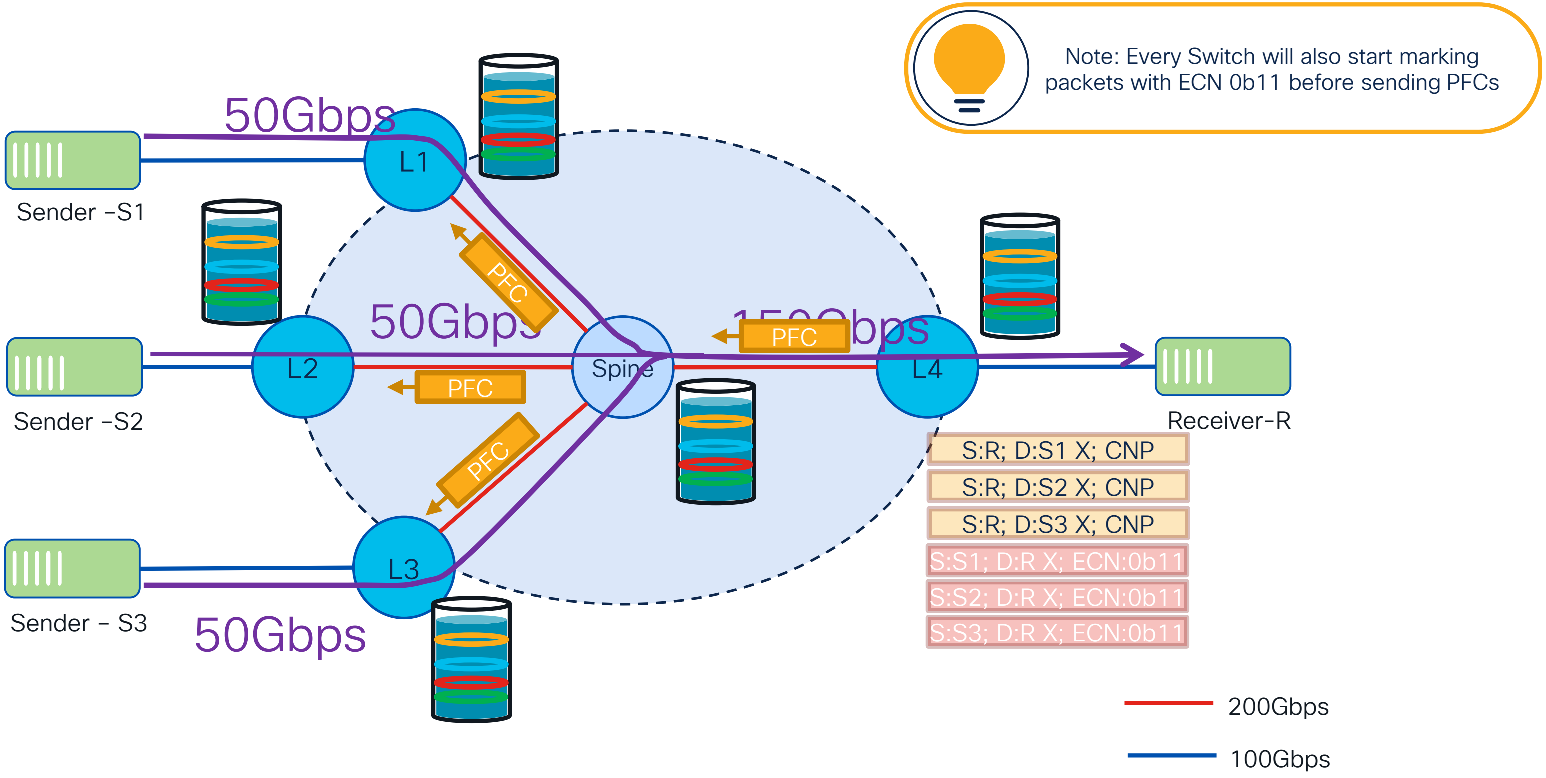
Priority Flow Control In Action With RoCEv2



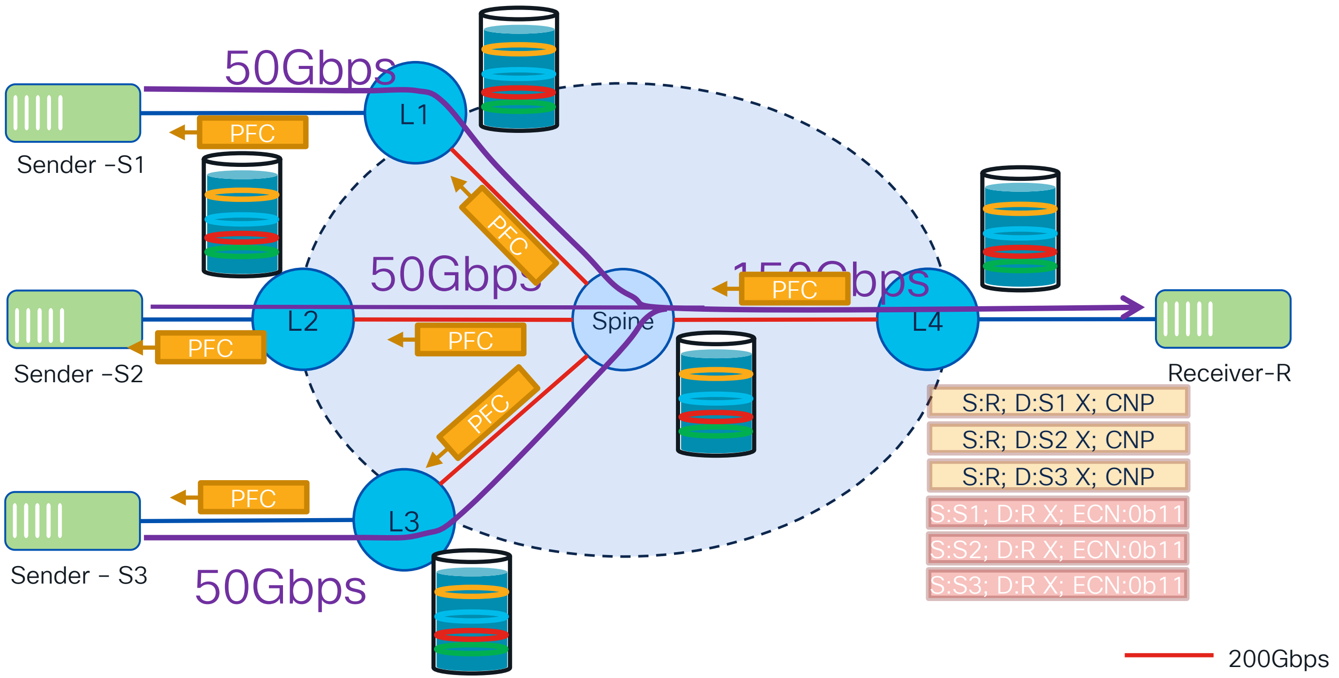
Priority Flow Control In Action With RoCEv2



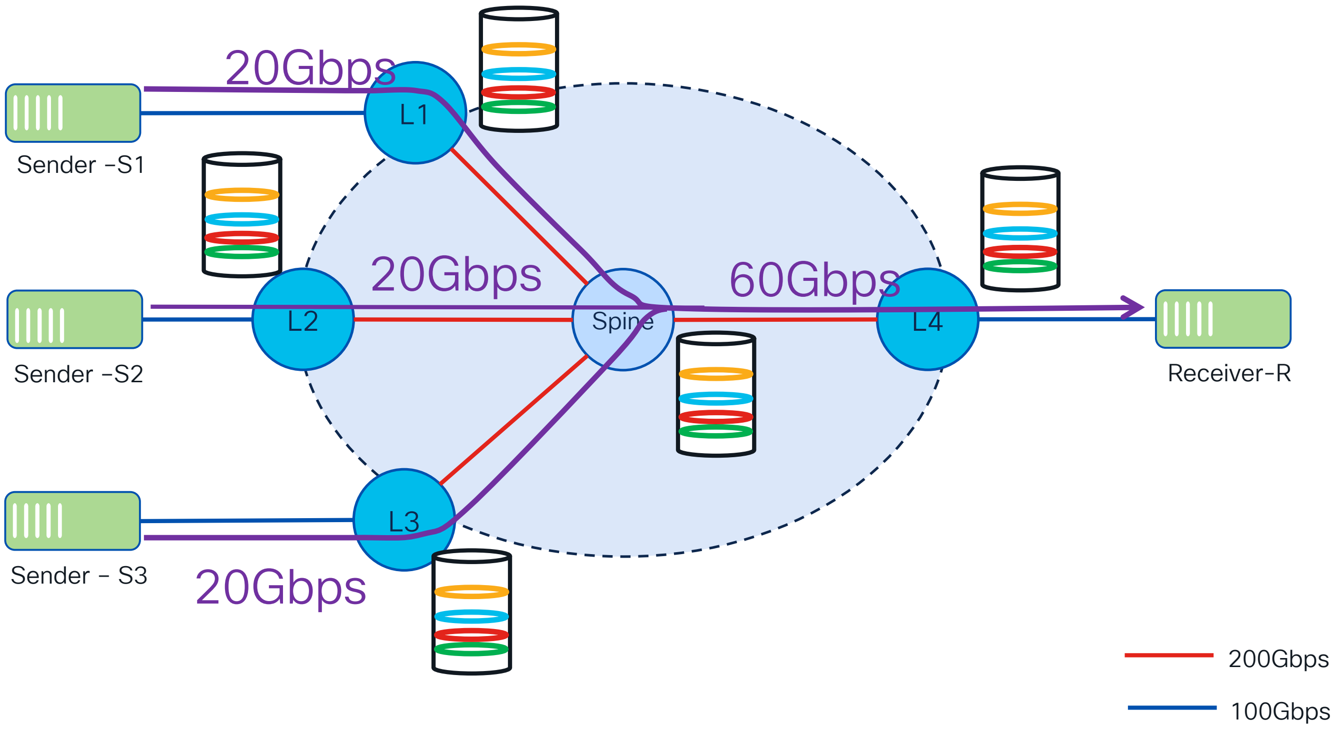
Priority Flow Control In Action With RoCEv2



Priority Flow Control In Action With RoCEv2



Priority Flow Control In Action With RoCEv2



ECN and PFC – What Each One Brings

RoCEv2 can leverage the use of both ECN and PFC to achieve its goals (i.e. lossless transport)

- ECN is an IP layer notification system. It allows the switches to indirectly inform the sources as soon as a threshold is reached and let them slow down the throughput
- PFC works at Layer 2 and serves as a way to use the buffer capacity of switches in the data path to temporarily ensure the no-drop queue is honoured. It effectively happens at each switch, hop-by-hop, back to the source, giving the source time to react without dropping packets
- ECN should react first, and PFC acts as a fail-safe if the reaction is not fast enough
- In any case the combo can help achieving a lossless outcome required by AI/ML traffic
- This collaboration of both is called *Data Center Quantized Congestion Notification* (DCQCN)
- All Nexus 9000 CloudScale ASICs support DCQCN

Alternatives to ECN with WRED

Approximate Fair Drop

- Nexus 9000 ASIC also implements advanced queuing algorithms that can avoid some non-optimized WRED results
- As an example WRED has no knowledge on which flows are consuming most of the bandwidth. ECN marking happens only based on probability
- AFD constantly tracks the amount of traffic exchanged and divides them in two categories:
  - Elephant Flows: long and heavy which will be penalized (ECN marked)
  - Mice Flows: short and light which will not be penalized(ECN marked)