

Building a High-Performance AI Data Center

cisco Live !

Janith Mantriratna
AI Infrastructure Specialist

Session ID: BRKDCN-2613

Cisco Webex App

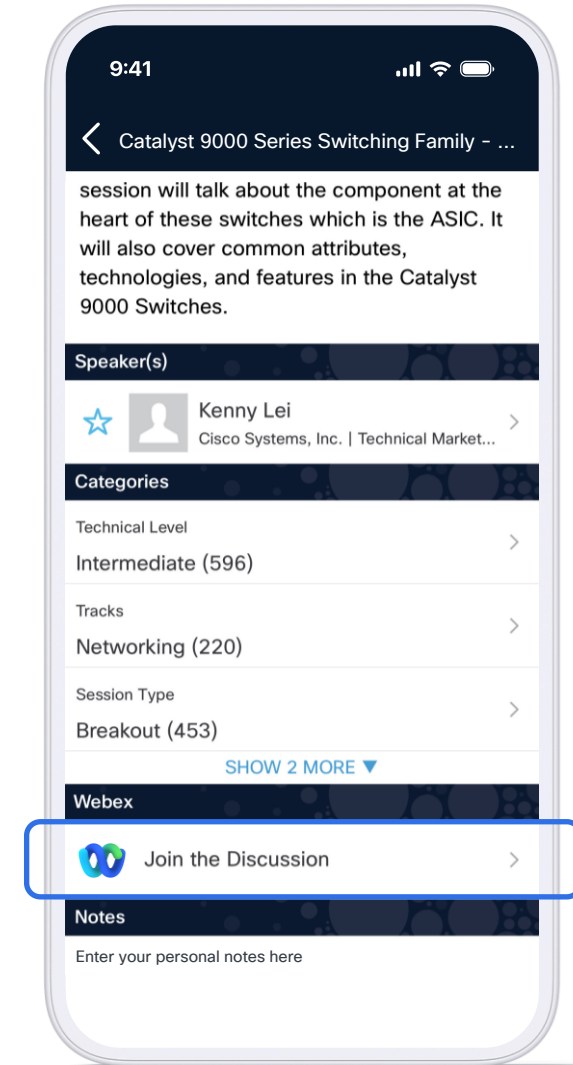
Questions?

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 14 November 2025.



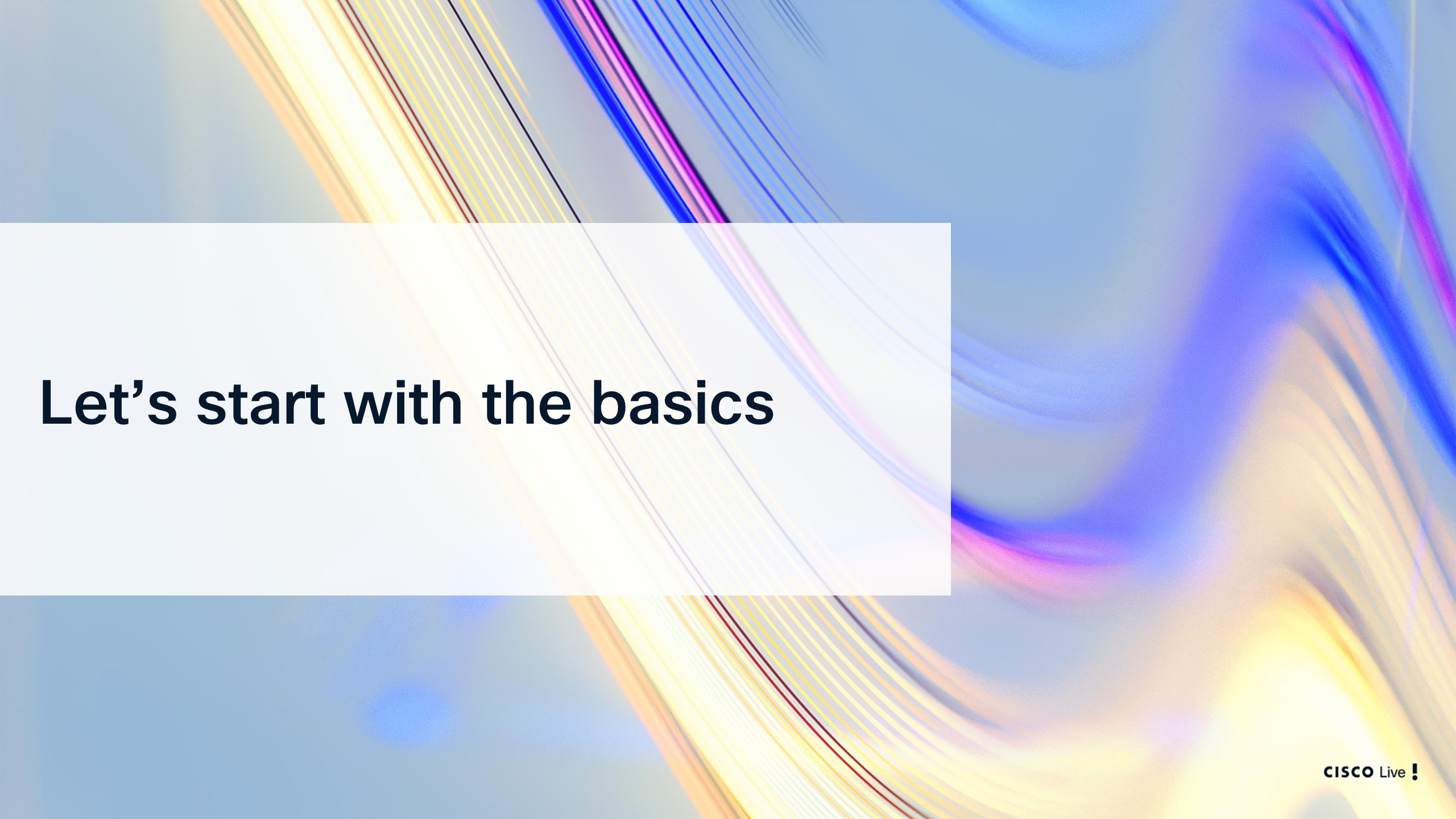
<https://cislive.ciscoevents.com/cislivebot/#BRKDCN-2613>



Image credit: <https://ampd.com.au/our-work/dc-rivals-hypercoaster/>

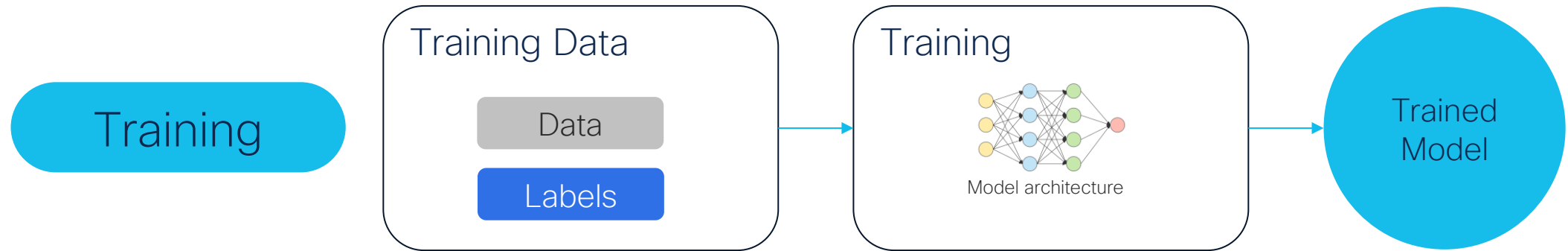
Agenda

- 01 Reviewing some basics**
- 02 Compute Infrastructure**
- 03 AI Backend Networks–
The building blocks**
- 04 AI Backend Networks–
Finetuning for optimal performance**
- 05 Why Security matters**
- 06 Bringing everything together**
- 07 Wrap up**

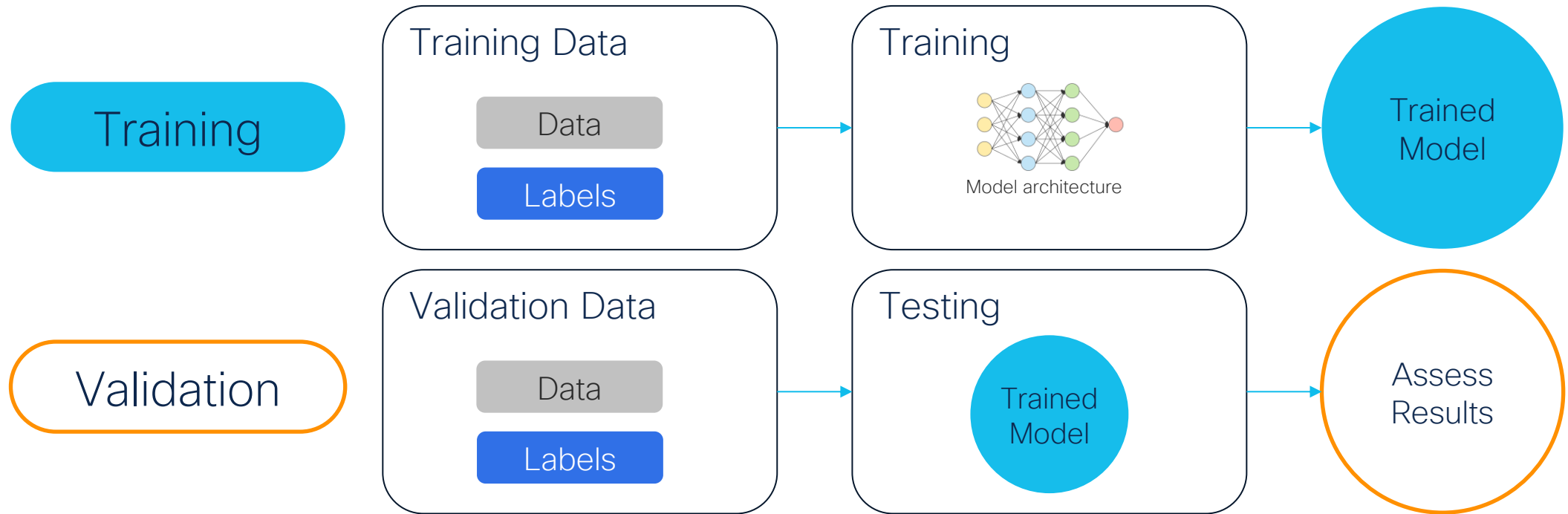


Let's start with the basics

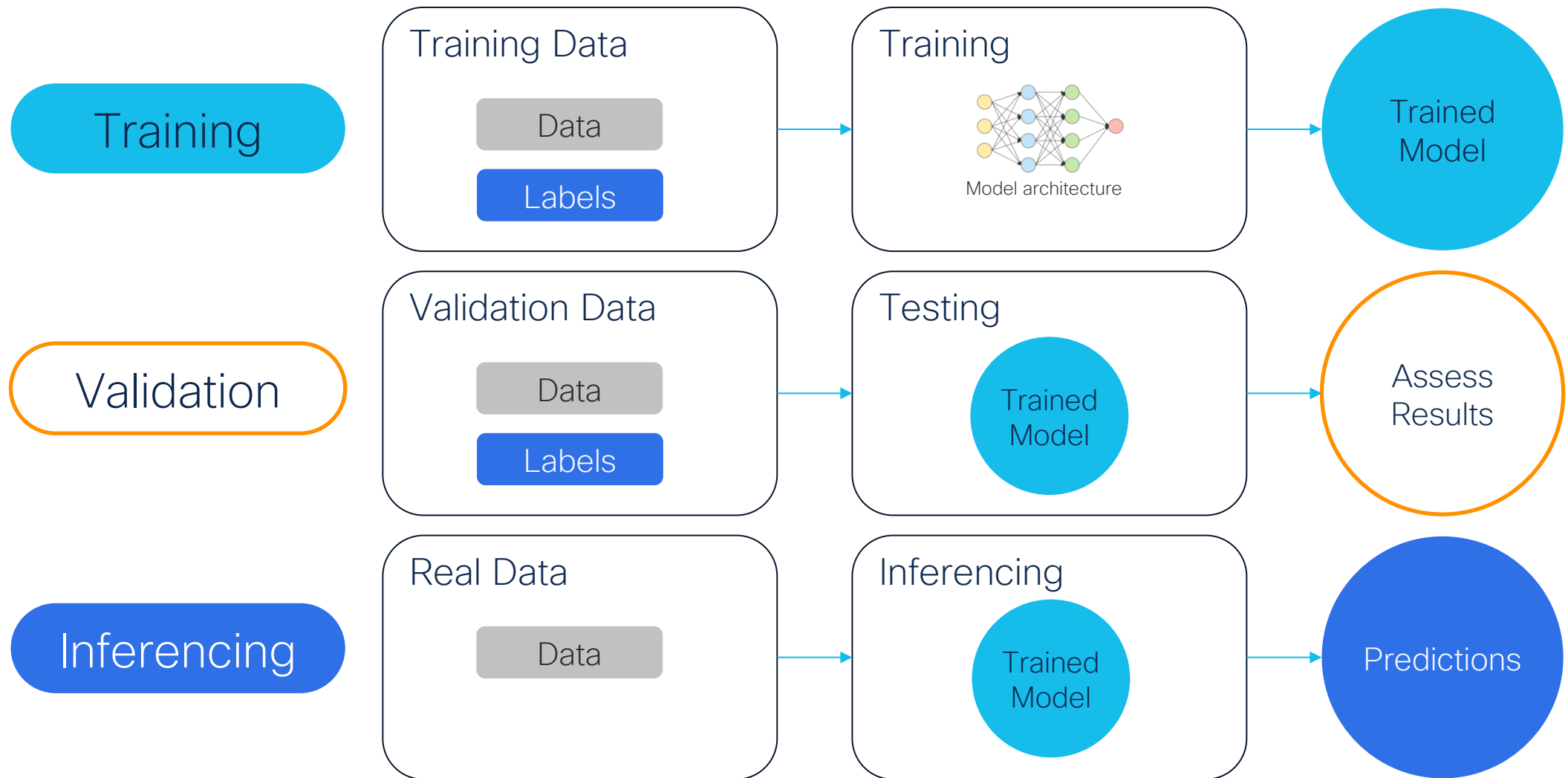
Machine Learning Overview



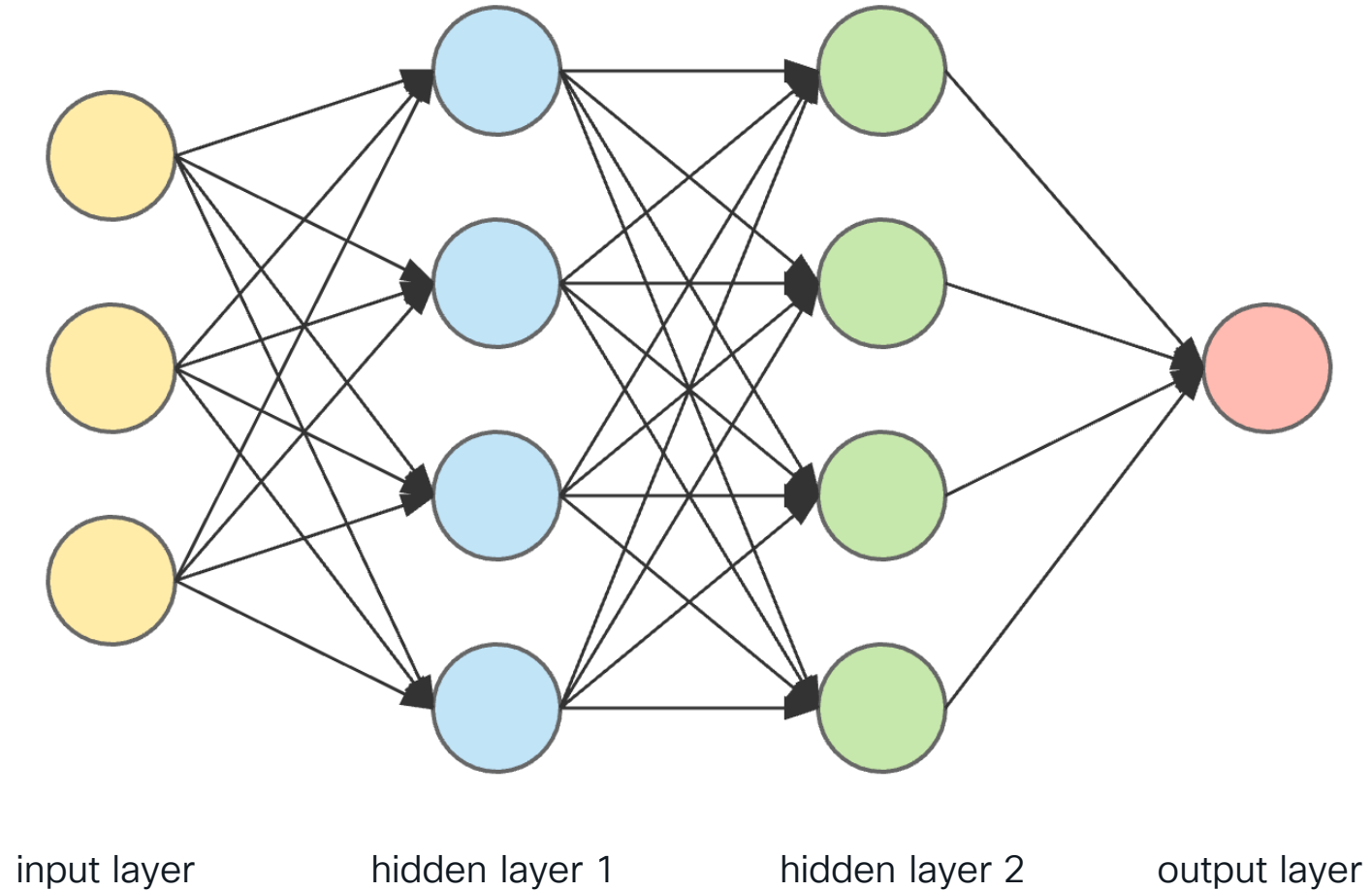
Machine Learning Overview



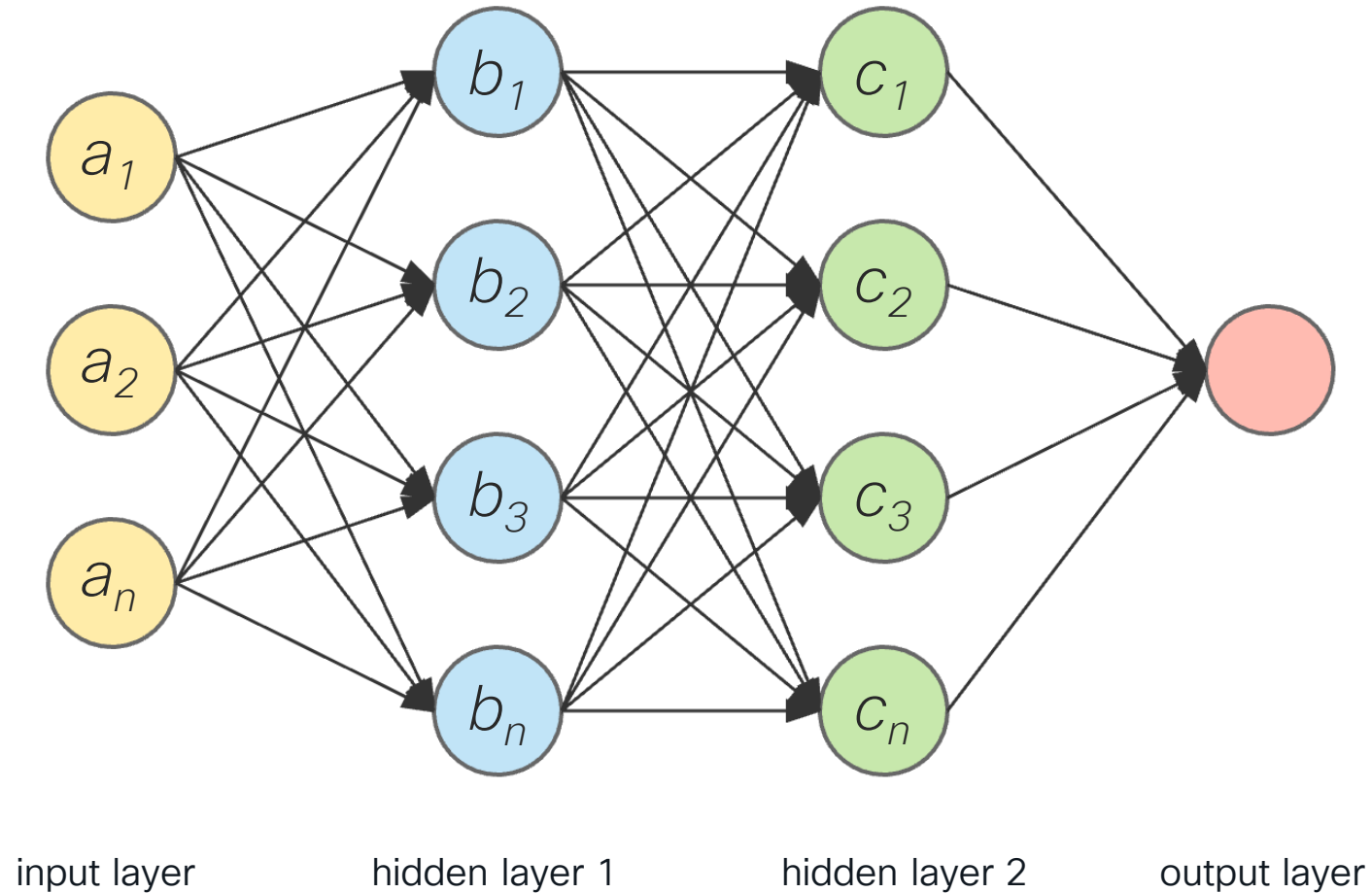
Machine Learning Overview



Model Architecture

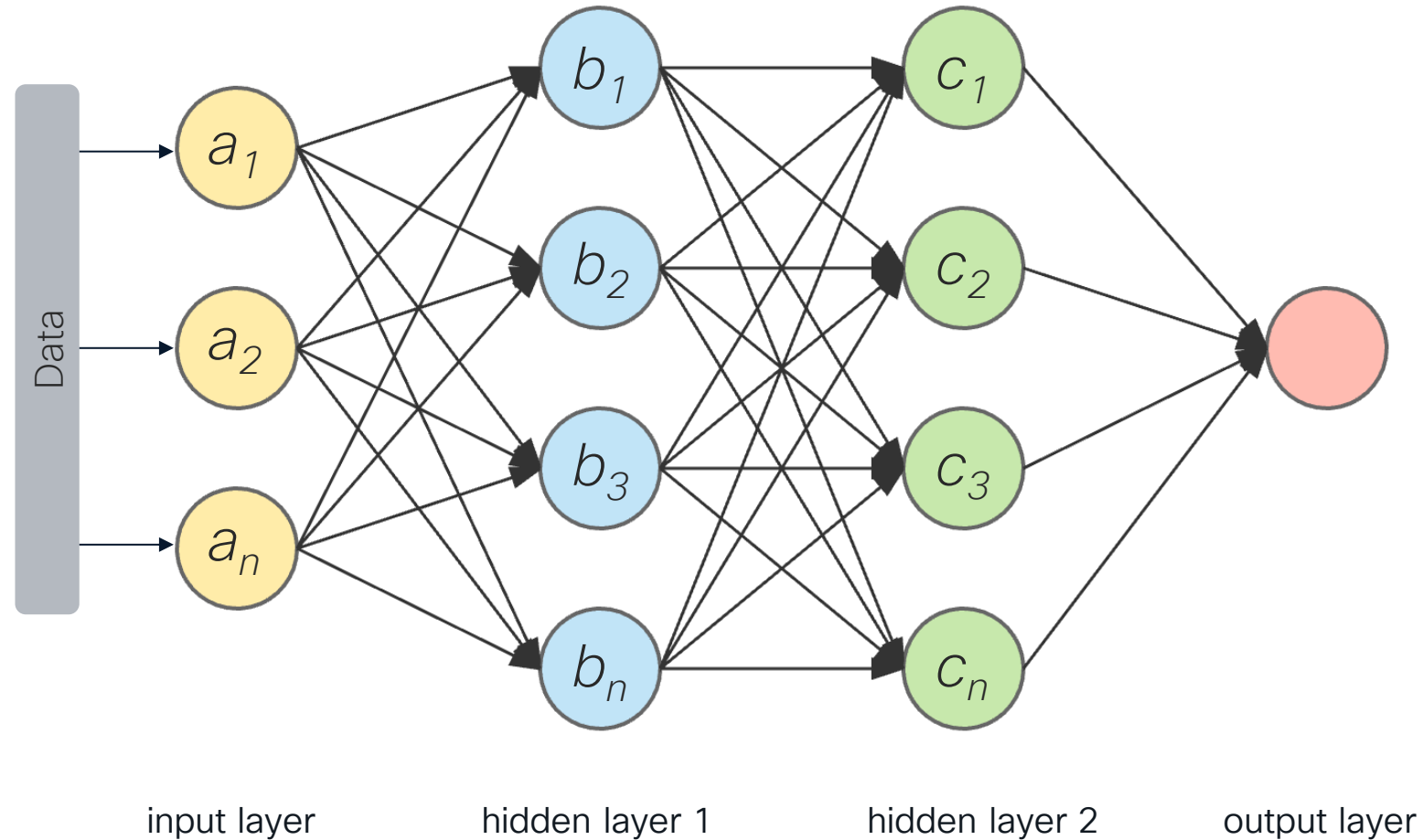


Model Training



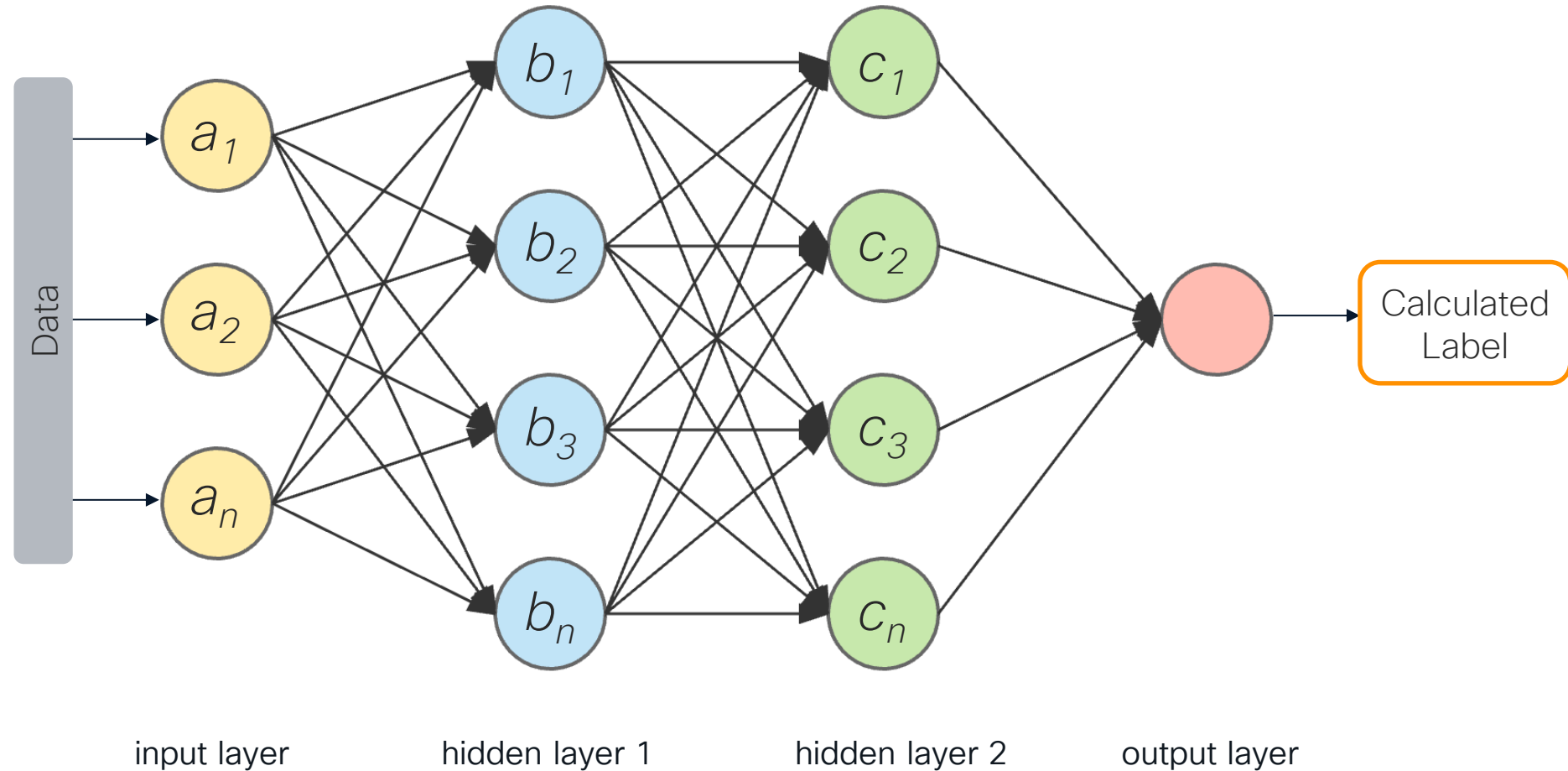
Feed forward phase

Model Training

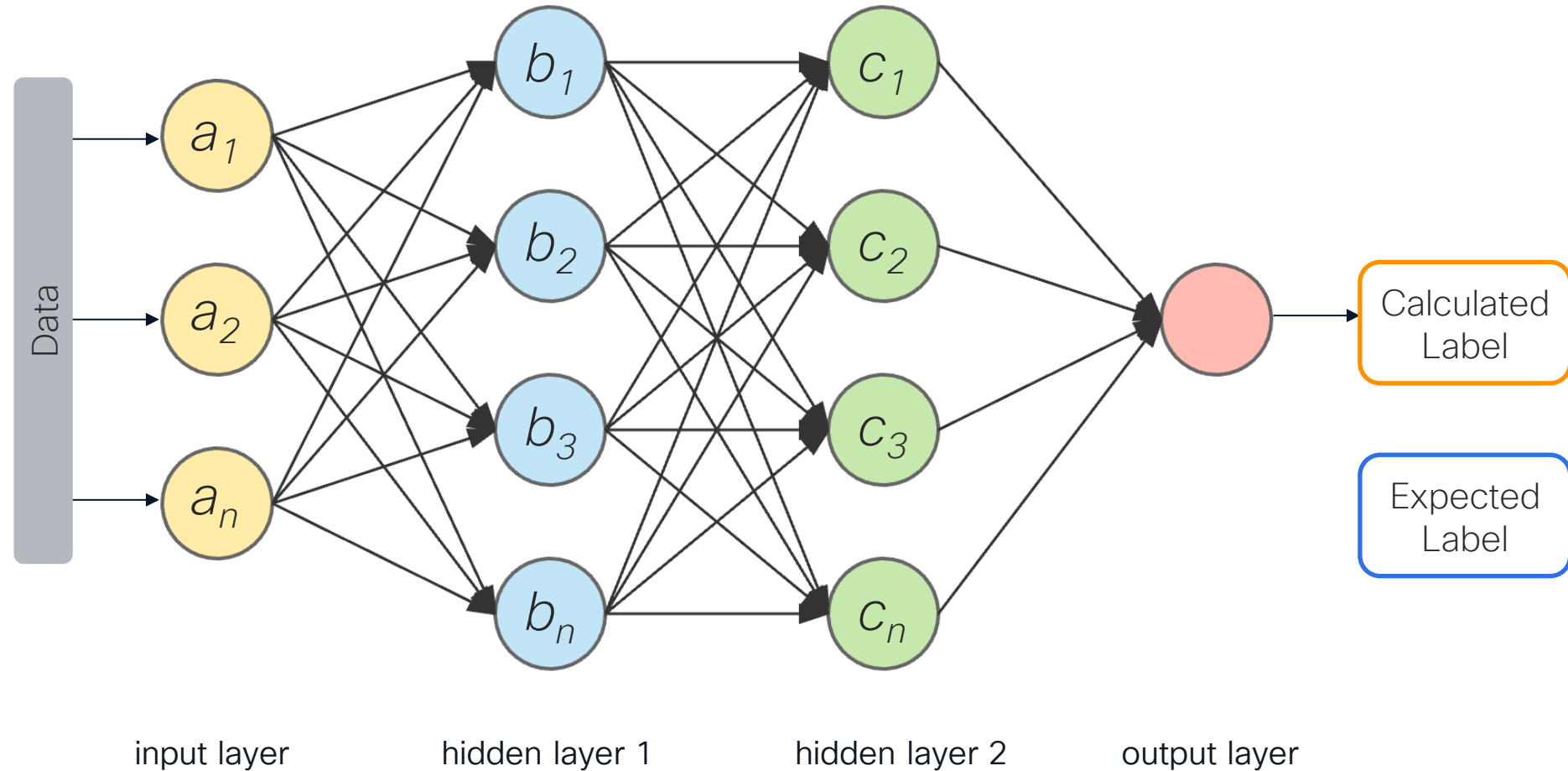


Feed forward phase

Model Training

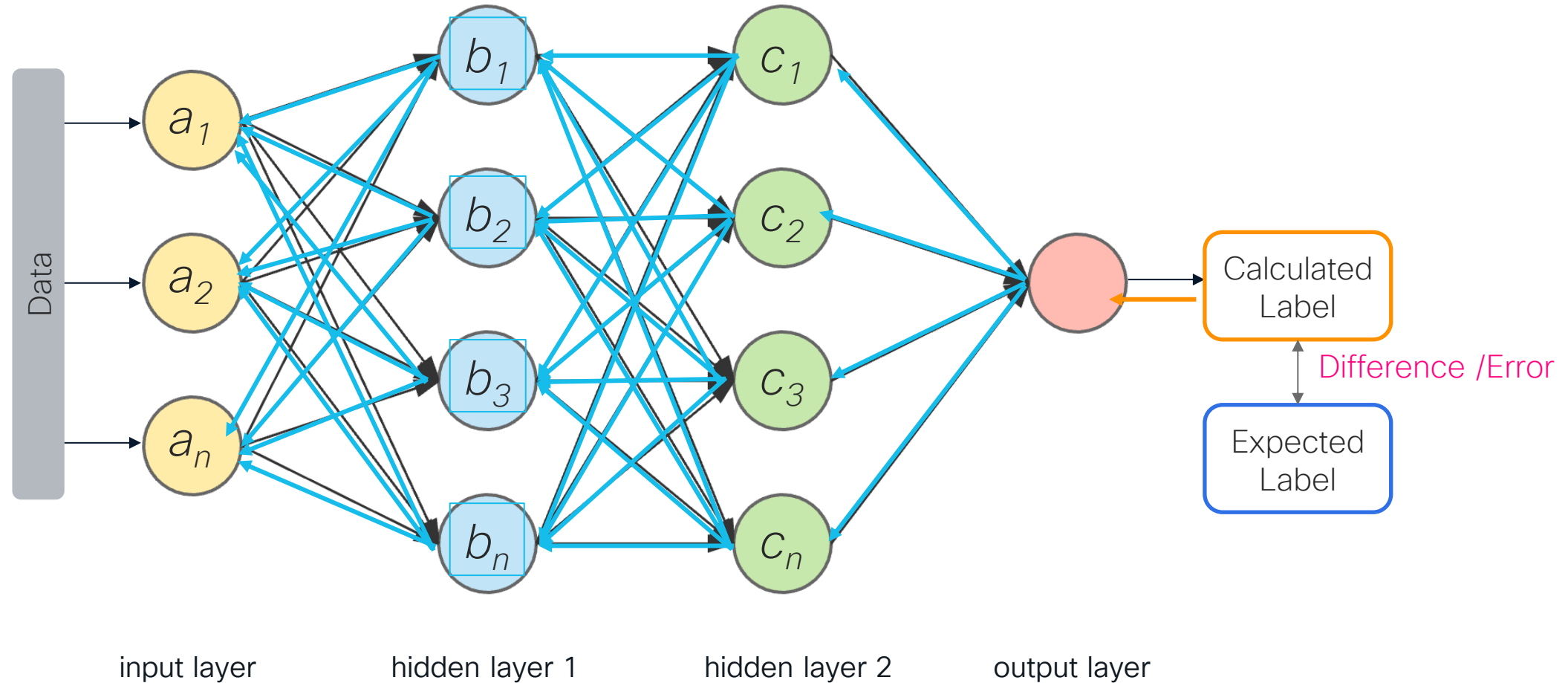


Model Training

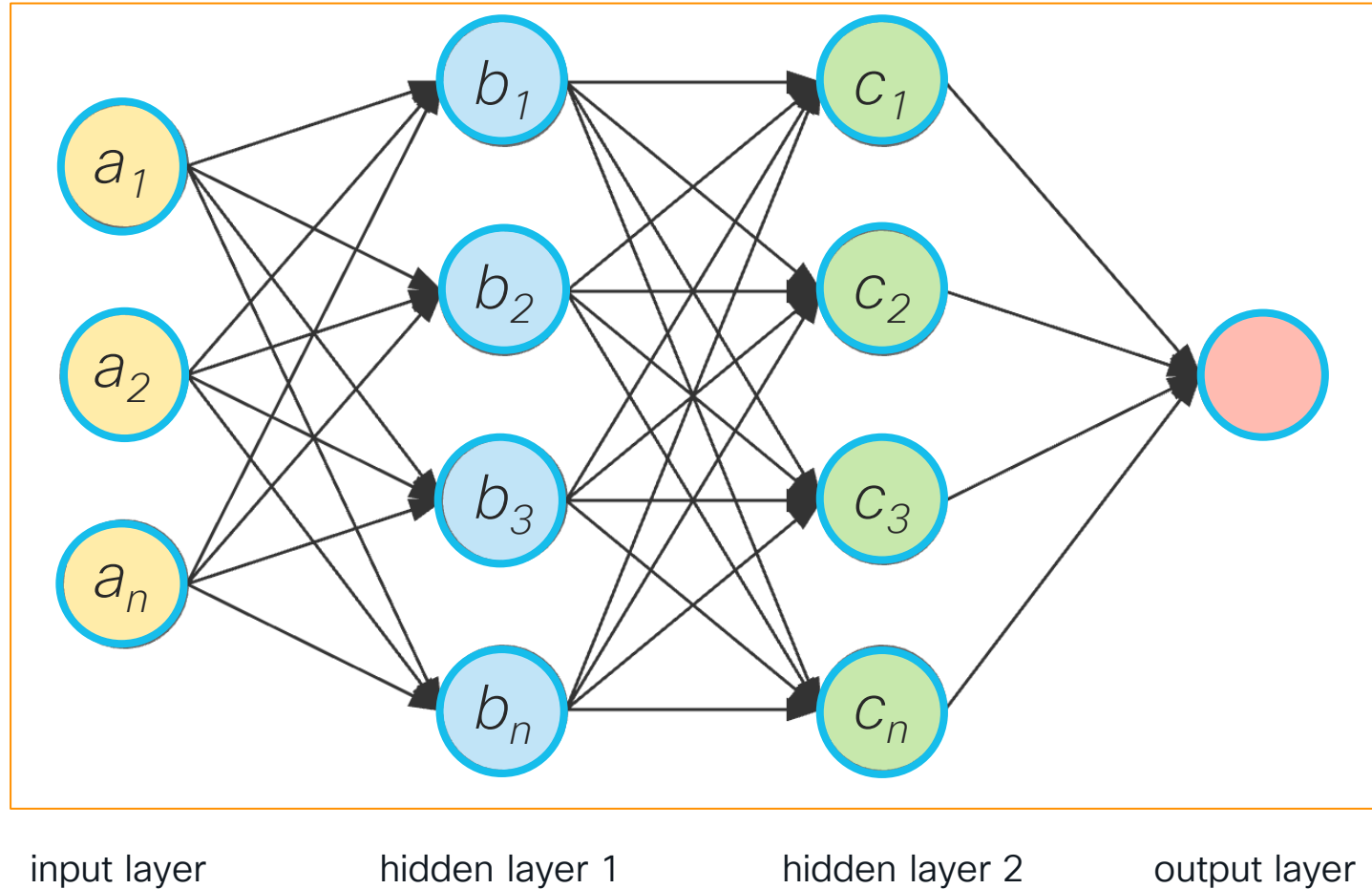


Feed forward phase

Model Training

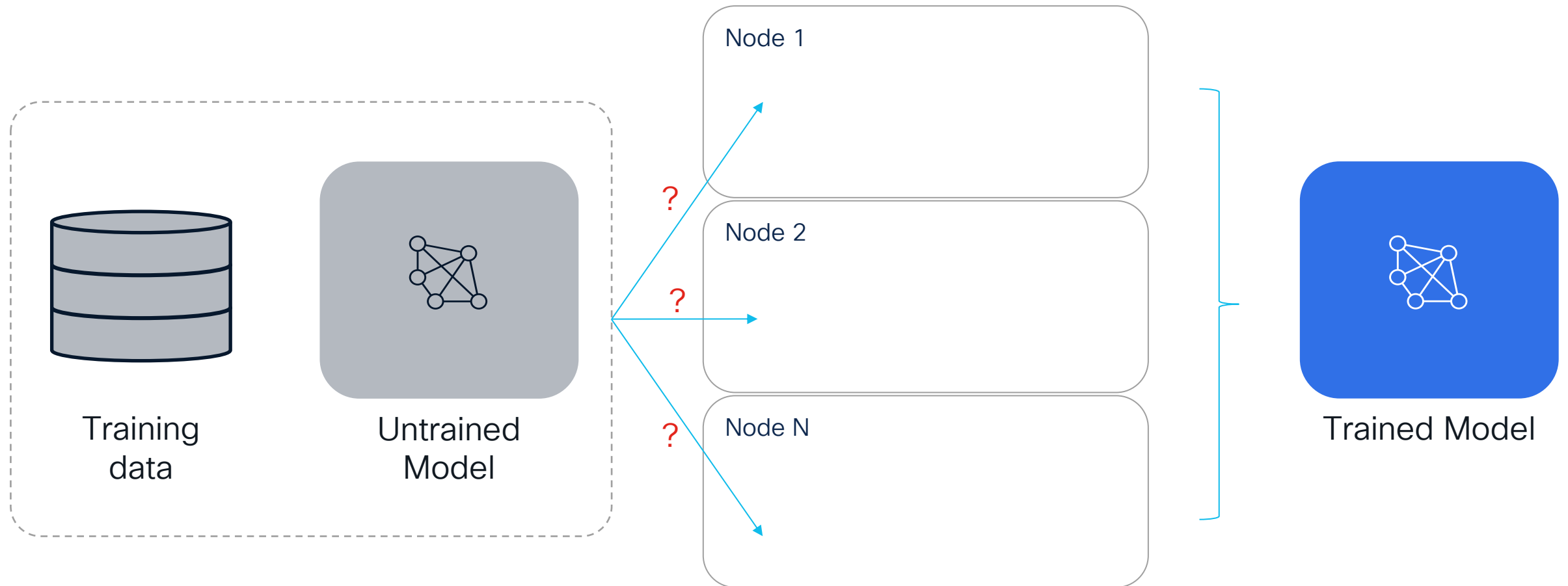


Model Training



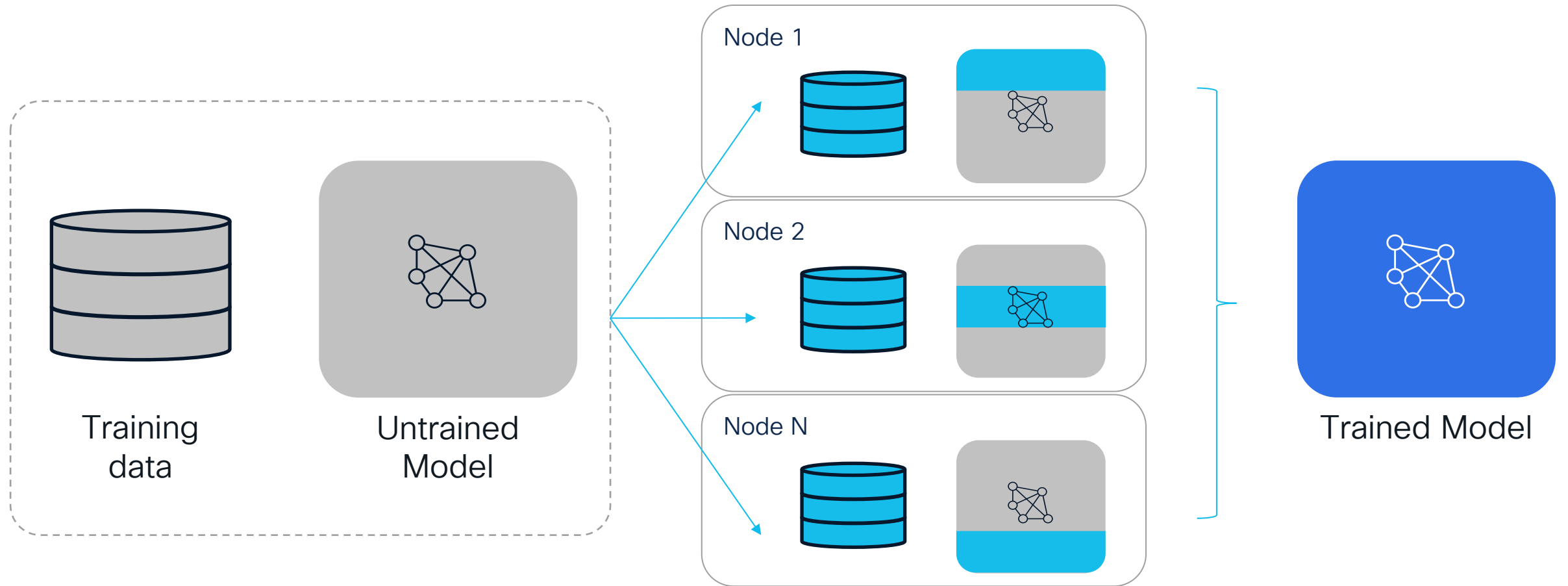
Optimiser phase

Putting the “Distributed” in Distributed ML Training



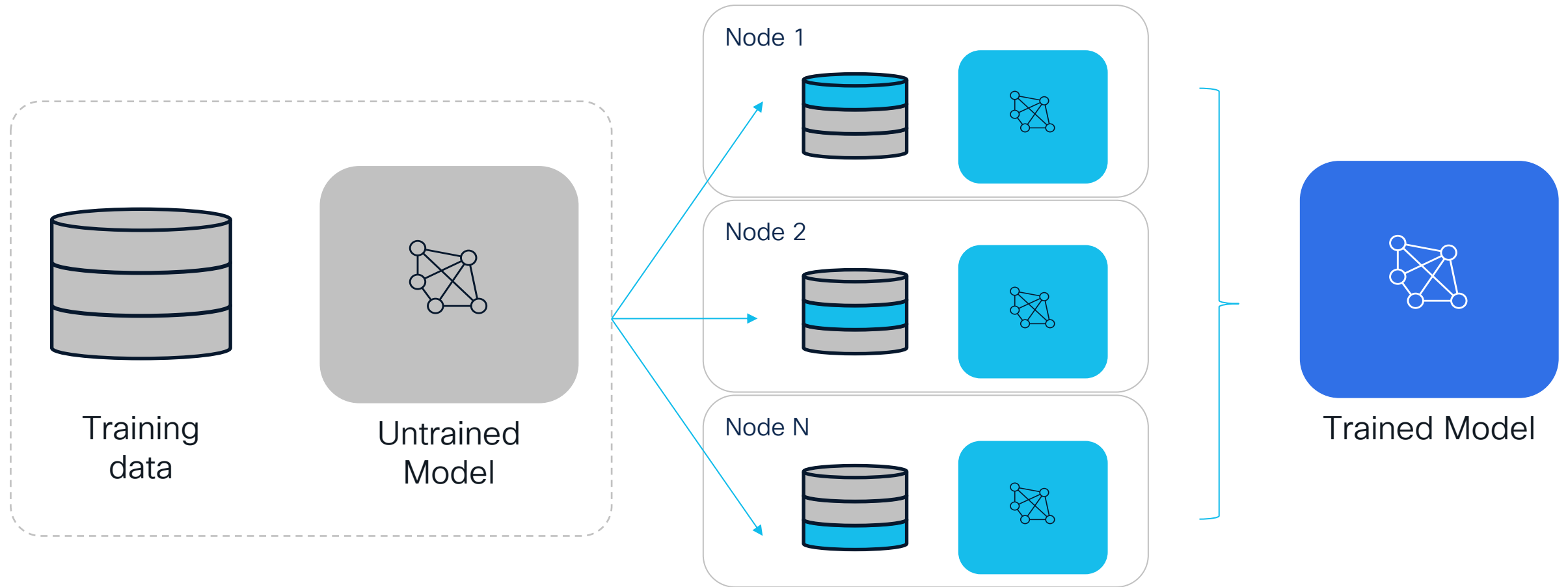
Putting the “Distributed” in Distributed ML Training

Model Parallelism



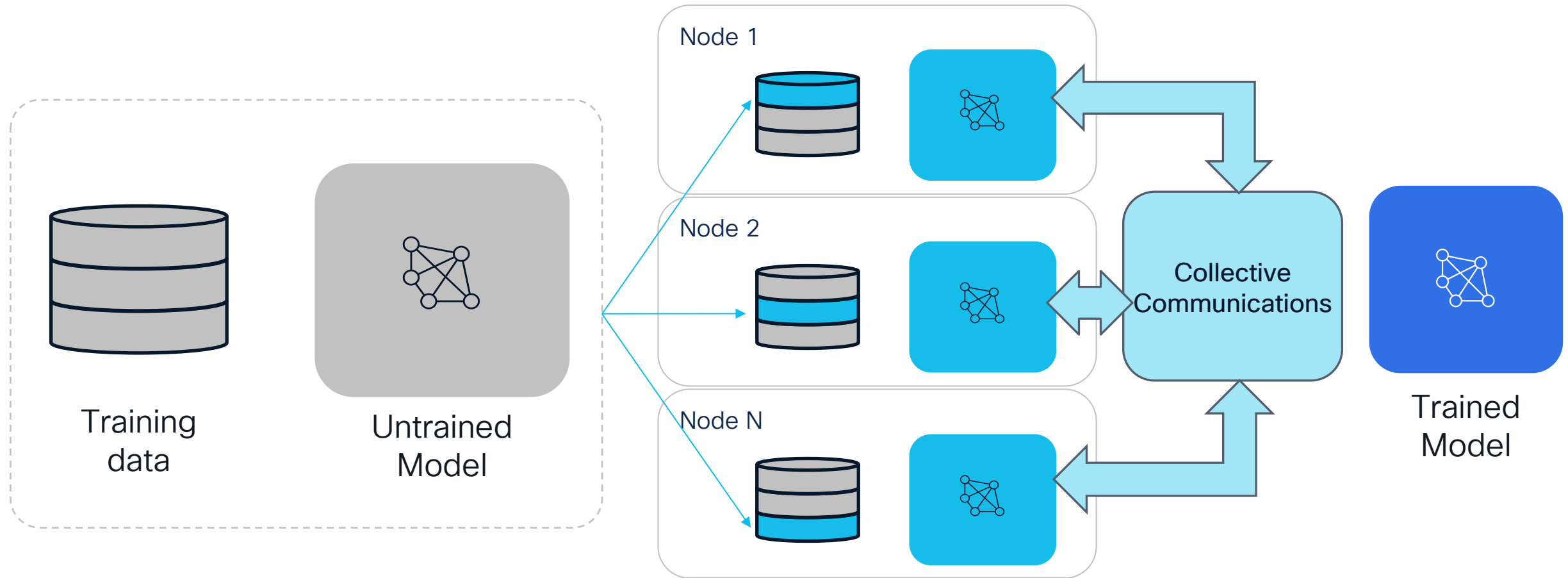
Putting the “Distributed” in Distributed ML Training

Data Parallelism

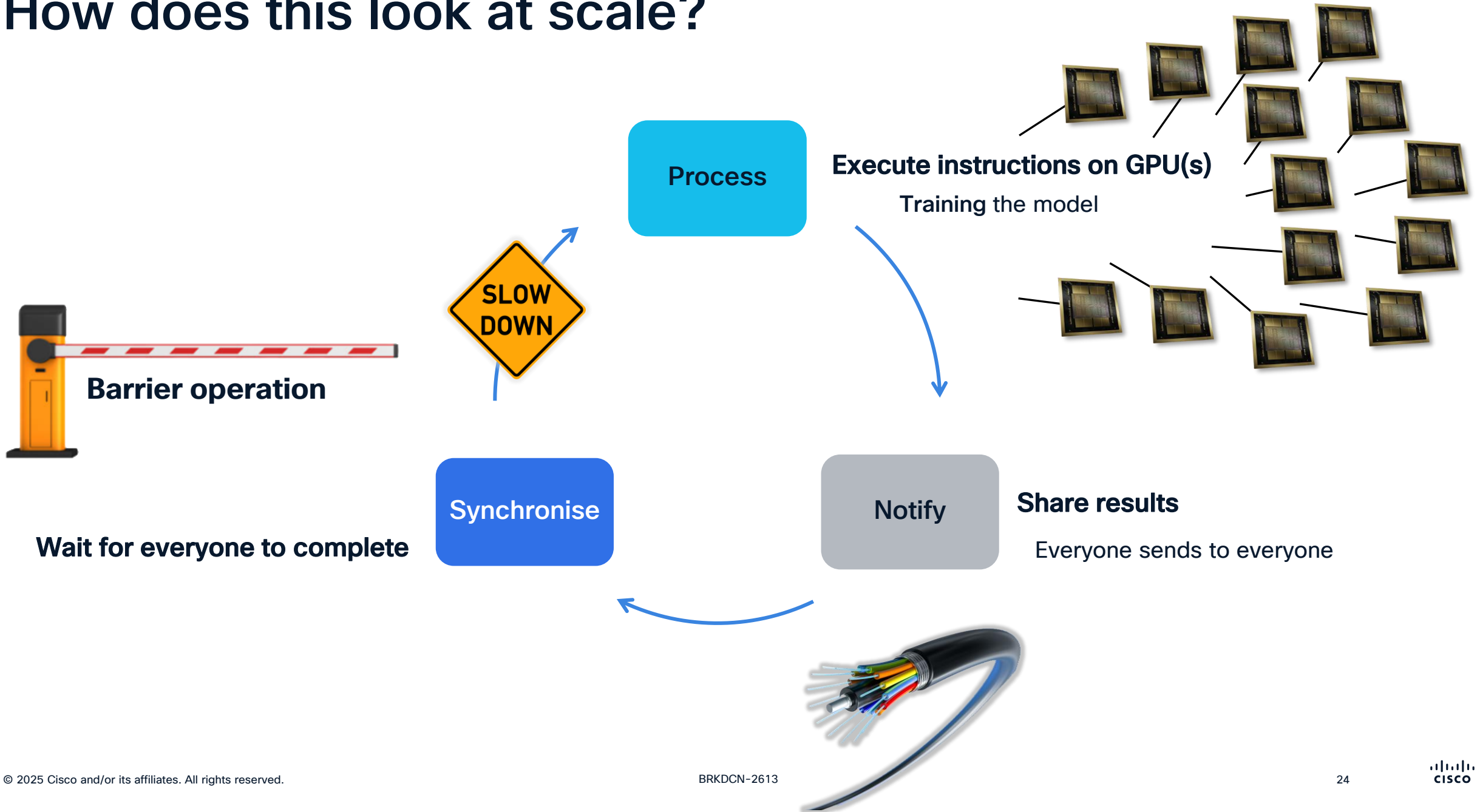


Putting the “Distributed” in Distributed ML Training

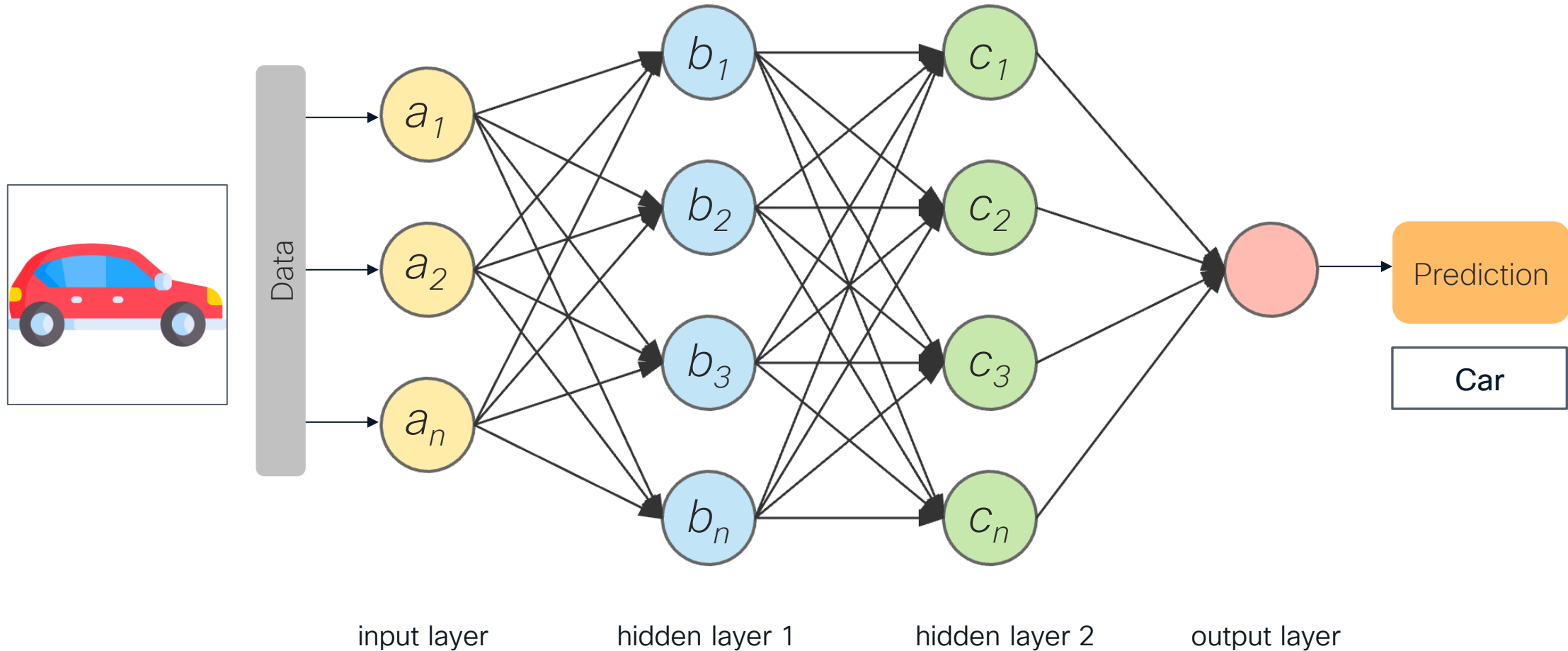
Node communication



How does this look at scale?



Model Inferencing



Inferencing (Feed forward phase only)

Inferencing at scale

Distributed inference across nodes

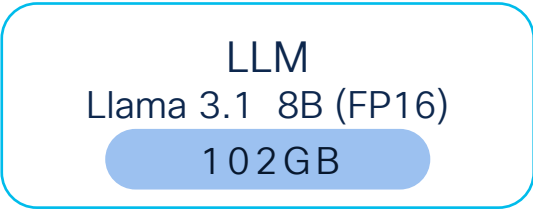
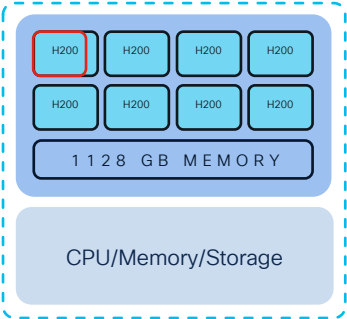
For details on GPU sizing check out BRKCOM-2115

Use case:

Corporate Chatbot that needs to serve 50 concurrent users
Max Input tokens per query – 10000
Max Output tokens per query – 4000

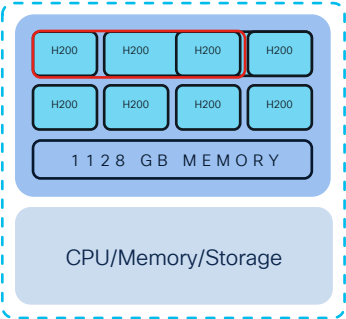
Jargon:

LLM – Large Language Model
Token – numeric representation of letters.
Token = 3-4 letters/0.75 words
1 H200-NVL GPU has 141GB of memory



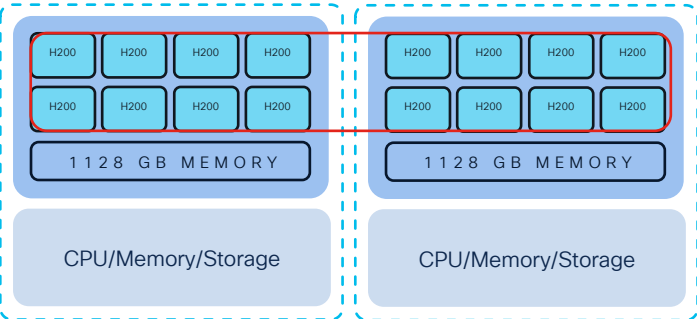
Host

Intra-host and inter-host GPU interconnect not required



Host

Inter-host GPU interconnect not required but intra-host GPU is required



Host

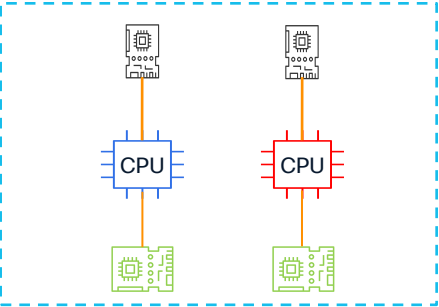
Host

Intra-host and inter-host GPU interconnect required

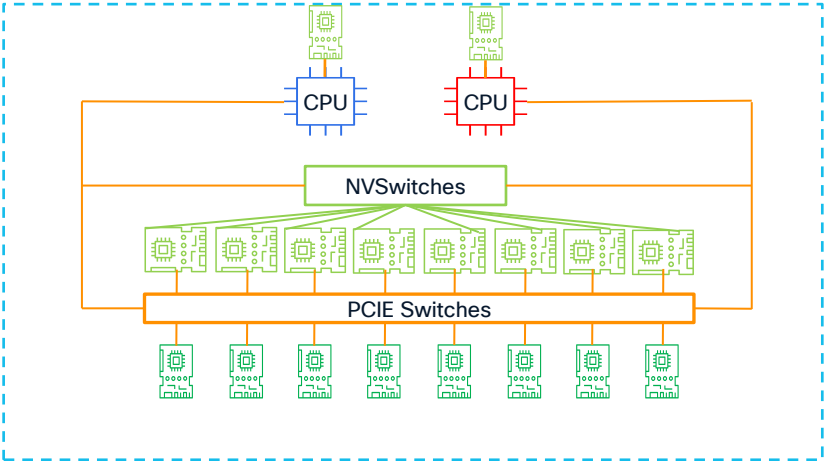
Compute Infrastructure

Types of GPU servers

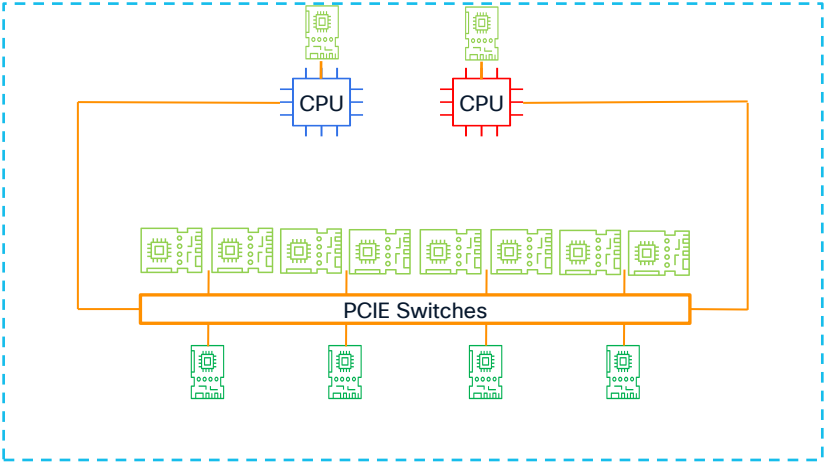
Today's focus



Traditional Server



HGX 8 SXM GPUs with NVLink Switch



MGX 8 PCIe GPUs with NVLink

PCIE

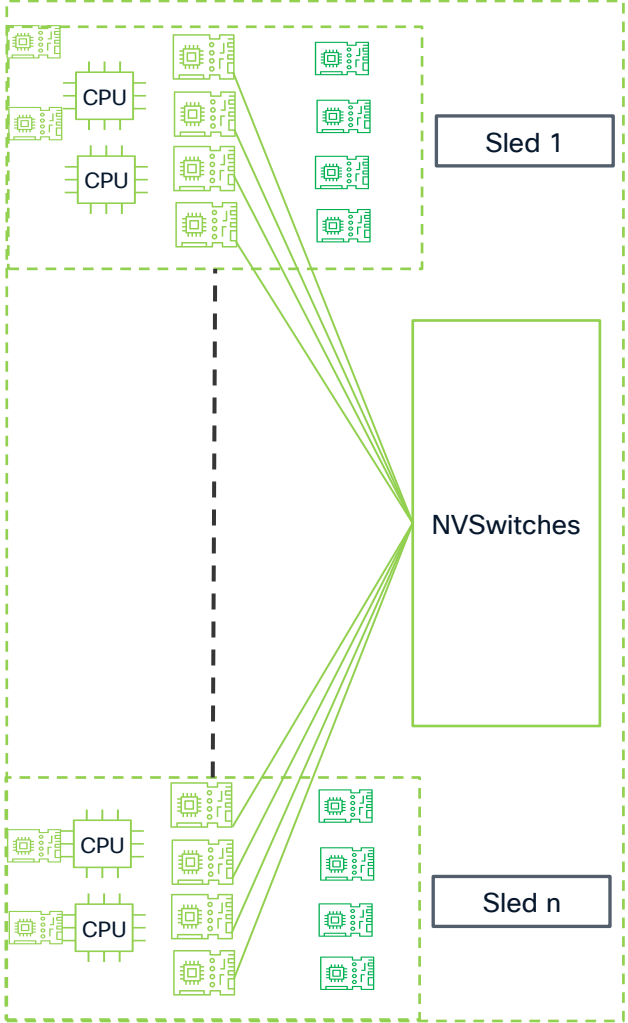
GPU

Cisco

NvLink

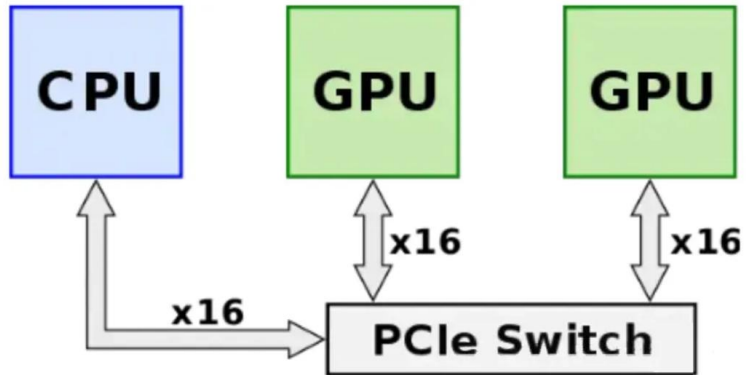
NIC

Nvidia

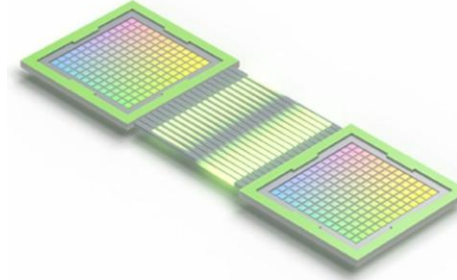


Rack scale - NVL-36/72

Why NvLink?

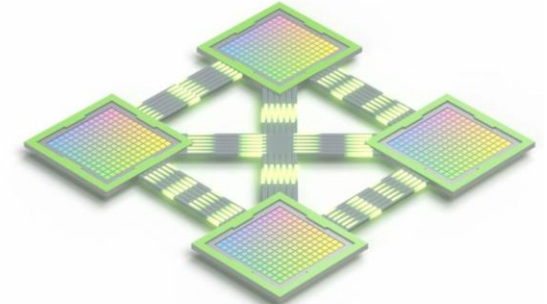


<https://naddod.medium.com/unveiling-the-evolution-of-nvlink-670dd538b237>



NVIDIA H100 PCIe with NVLink GPU-to-GPU connection

<https://blogs.nvidia.com/blog/what-is-nvidia-nvlink/>



NVIDIA H100 with NVLink GPU-to-GPU connections

Connectivity Speeds

2014

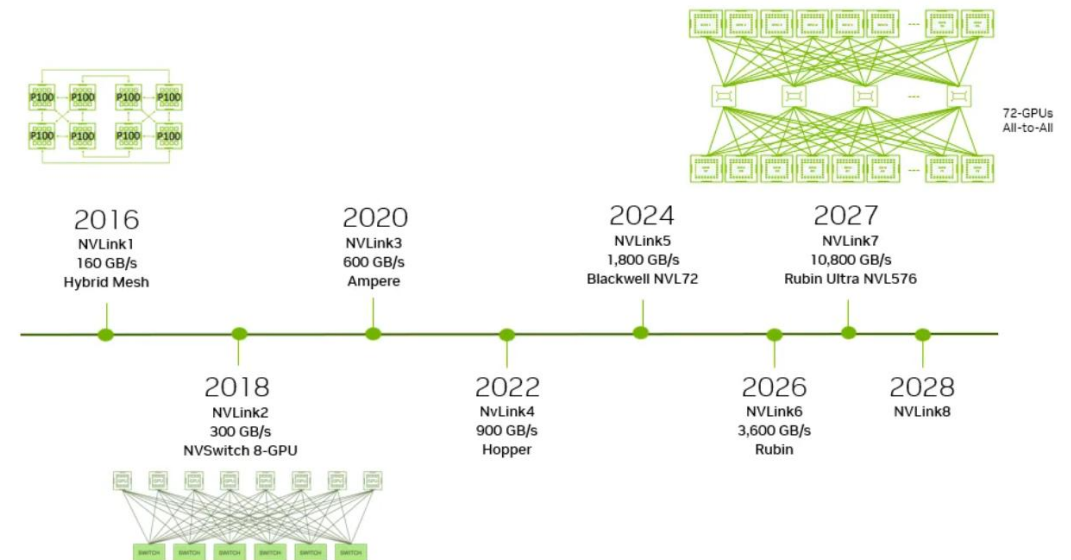
PCIe Gen 3 – 32GB/S

NVLink Gen1 – 160GB/S

PCIe Gen 5 – 128GB/s

NVLink Gen 4 (H200) – 900GB/s

NVLink Gen 5 (B200/B300) – 1800GB/s



What GPUs are available to us?

PCIE



NVIDIA L4

- 24GB GDDR6 memory
- No Nvlink support
- Great for computer vision and edge processing



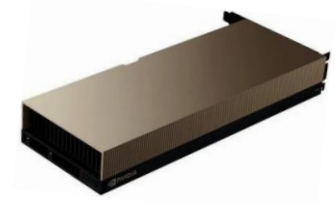
NVIDIA RTX6000 BW

- 96GB GDDR7 memory
- No Nvlink support
- Great for Graphics and rendering
- VDI
- Small/medium LLMs



NVIDIA H200 NVL

- 94GB HBM3
- Nvlink support – nearest neighbour (2-way)
- Large language model generative AI Inference



NVIDIA H200 NVL

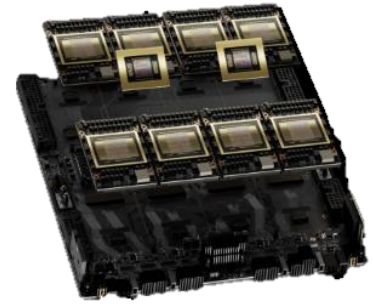
- 141 GB HBM3e
- Nvlink support – 2 way and 4 way
- Large language model generative AI Inference
- AI Training and fine-tuning

8-Way Systems



NVIDIA H200 SXM

- 141 GB HBM3e per GPU
- 8 way interconnected GPU system
- Large language model generative AI Inference
- Large scale AI training clusters



NVIDIA B300 SXM

- 288 GB HBM3e per GPU
- 8 way interconnected GPU system
- Large language model generative AI Inference
- Large scale AI training clusters



AMD MI300X/325X

- 192 GB HBM3 per GPU/250GB HBM3E per GPU
- 8 way interconnected GPU system
- Large language model generative AI Inference
- Large scale AI training clusters

GPU infrastructure to deliver AI everywhere

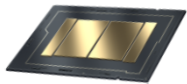
Learn more about Intersight!
BRKCOM-2540
BRKCOM-2116



Remote/Far Edge

- Predictive Maintenance
- Safety and security
- Route optimisation
- Site Documentation

- YoLO, OpenCV
- Small Language Models



CPU



L4



L40S



Edge locations

- Robotics
- OT
- Appliance protection
- Loss Prevention

- Predictive models
- Natural Language Processing



RTX PRO 6000

- Physics
- VLMs



Private/Public Cloud

- AIOps
- Customer service
- Productivity
- Digital Twins
- Sales and Finance
- Legal
- Merchandising
- Marketing
- Predictive models
- Mixture of Experts LLMs
- Vision Language Models
- Domain specific LLMs



RTX PRO 6000



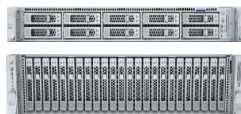
H200-NVL



H200

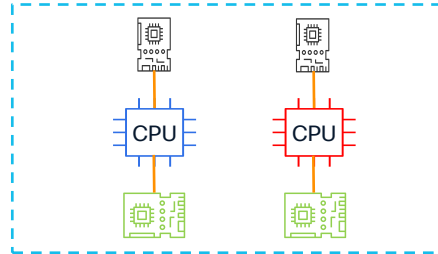


B300



Cisco Intersight

Cisco GPU infrastructure

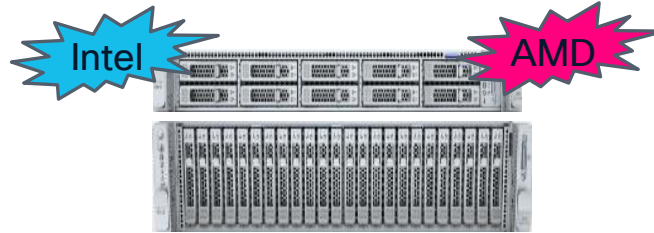


Traditional Servers



Unified Edge

Remote Edge AI/Compute/Network consolidation
3 RU for 5 mini-compute blades
Nvidia L4



Cisco C22X/24X Rack Servers

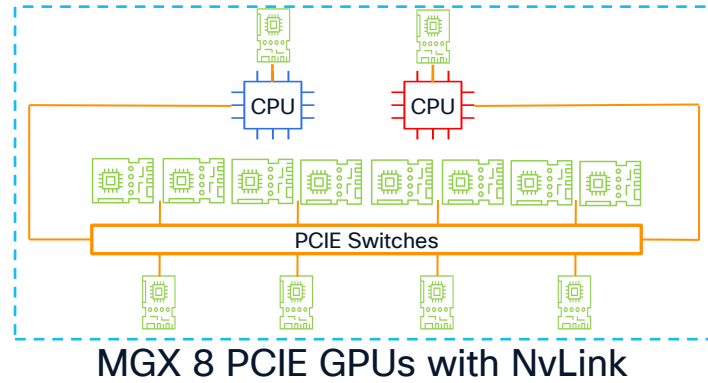
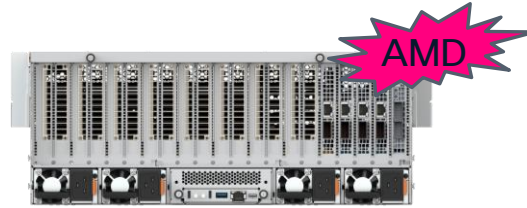
Edge/DC deployments
1RU (C22X) or 2RU (C240X)
Nvidia L4, L40S, H200-NVL



Cisco UCS-X Modular Server

Edge/DC deployments
7 RU
Nvidia L4, L40S,
2 X H200 or 2 X L40S per x440p blade
4 X H200 or 4X RTX6000 per x580p blade
2 way NV-bridge
Up to 8 H200s or RTX6000 per chassis

Cisco GPU infrastructure



C845A PCIe Dense GPU node

Edge*/ Datacentre

4RU

8 x Nvidia H200-NVL

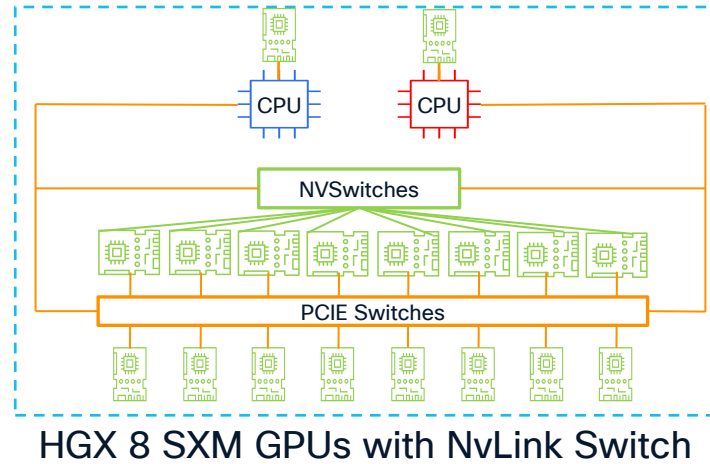
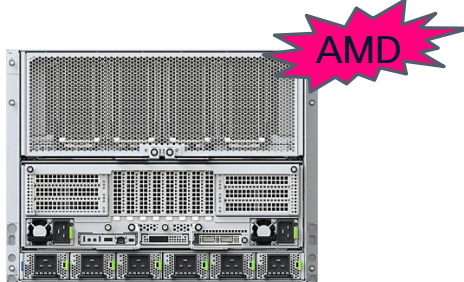
8 x Nvidia RTX Pro 6000

4 way NV-bridge

Flexible configuration (start with 2 GPUs)

* Provided sufficient Power is available.

Cisco GPU infrastructure



C885A SXM Dense GPU Node

Datacentre
8RU
8 x Nvidia H200
8 x Nvidia MI300X



C880A SXM Dense GPU Node

Datacentre
10RU
8 x Nvidia B300

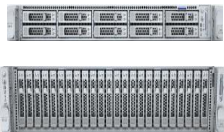
Cisco GPU infrastructure

Traditional Servers



Unified Edge

Remote Edge
AI/Compute/Network
consolidation
3 RU for 5 mini-compute blades
Nvidia L4



Cisco C22X/24X Rack Servers

Edge/DC deployments
1RU (C22X) or 2RU (C240X)
Nvidia L4, L40S, H200-NVL

MGX 8 PCIE GPUs with NvLink



Cisco UCS-X Modular Server

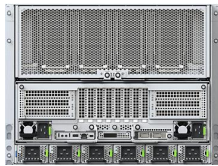
Edge/DC deployments
7 RU
Nvidia L4, L40S,
2 X H200 or 2 X L40S blades
4 X H200 or 4X RTX6000 blades



C845A PCIE Dense GPU node

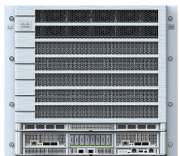
Edge*/ Datacentre
4RU
8 x Nvidia H200-NVL
8 x Nvidia RTX Pro 6000

HGX 8 SXM GPUs with NvLink Switch



C885A SXM Dense GPU Node

Datacentre
8RU
8 x Nvidia H200
8 x Nvidia MI300X



C880A SXM Dense GPU Node

Datacentre
10RU
8 x Nvidia B300

1KW<-2KW

5KW - 10KW

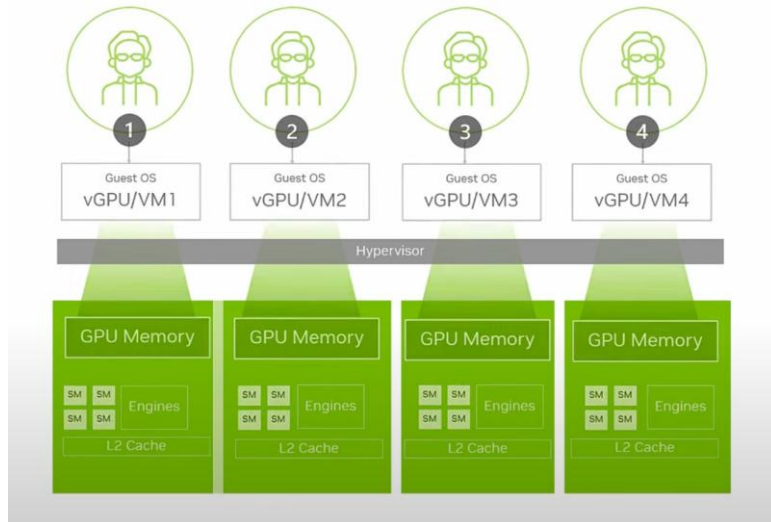
10KW - 20 KW

Slicing GPUs

What if we want to share GPU resources among different applications?

MIG-backed vGPU

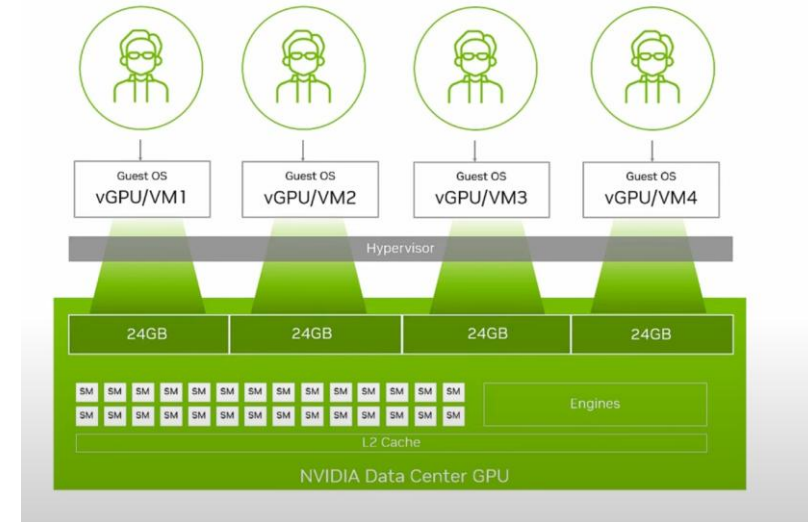
Dedicated GPU Memory, Dedicated GPU Resources



Source: [GPU Sharing for Compute and Graphics: Universal MIG with NVIDIA vGPU on RTX PRO 6000 Blackwell](#)

Time-Sliced vGPU

Dedicated GPU Memory, Time-Shared GPU Resources



- Multi-Instance-GPU (MIG) provides a hard partition of resources, whilst time slicing works like a scheduler.
- MIG provides consistent performance but can support a limited number of clients while time slicing is generally unpredictable but can support a lot more clients.
- When using enterprise hypervisors from VMware, RedHat and Nutanix, you also need Nvidia's vGPU licenses to enable sharing of GPUs.

Do we need to modernise storage for AI?

Feature	Traditional Storage Infrastructure	Modern Storage for AI Workloads
Primary Workload	Business Applications, Databases, Virtual Machines (VMs)	AI/ML Training & Inference, High-Performance Computing (HPC), Analytics
Performance Focus	Low Latency & IOPS (Input/Output Operations Per Second)	High Throughput (GB/s) & Massive Parallelism
Key Technology	HDDs, SAS/SATA SSDs, Fibre Channel SAN, NAS	All-Flash NVMe, Parallel File Systems, Scale-Out Object Storage
Data Types	Optimized for Structured Data (e.g., databases)	Optimized for Unstructured Data (images, video, text, logs)
Scalability Model	Scale-Up: Add resources to a single controller (has limits)	Scale-Out: Add more nodes to a cluster for linear performance growth
Architecture	Siloed: Separate systems for block, file, and object data	Unified/Converged: A single platform for the entire AI data pipeline
Data Access Pattern	Sequential or random access from a few clients	Massively parallel access from hundreds or thousands of clients (GPUs)
Primary Goal	Reliability and consistency for known applications	Feeding data to GPUs without bottlenecks to accelerate time-to-insight

Compute Infrastructure for AI - Callouts

Understand the hardware requirements for the AI workload, this dictates the types of GPUs that are needed.

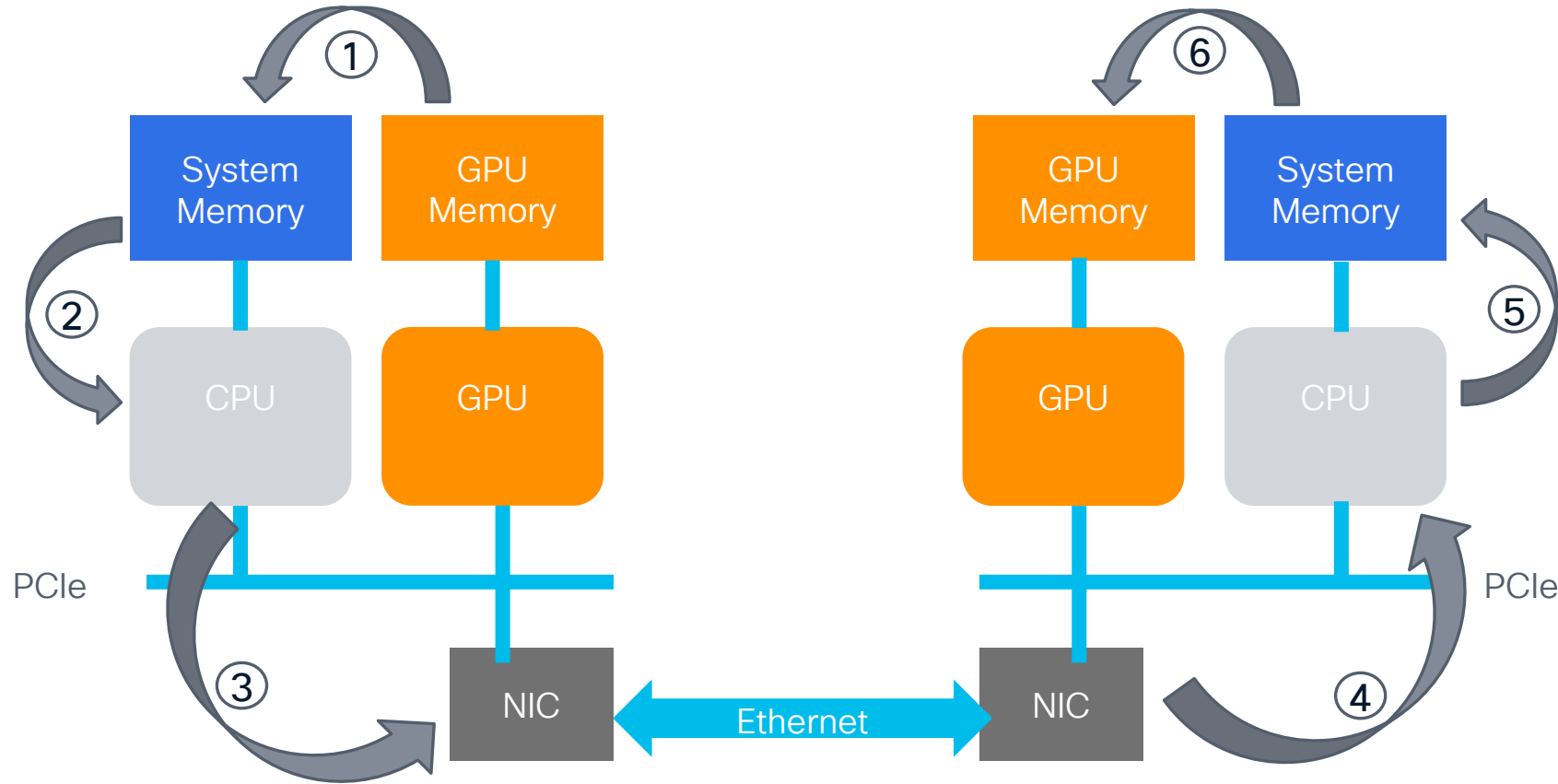
Discover what rack power and cooling are available at the Datacentre or CoLo facility. This, along with the workload requirement drives the right server platform selection.

For shared infrastructure, where fractional GPUs are required, work with your Virtualisation vendor and application owners to decide on Nvidia Licenses.

AI Backend Networks

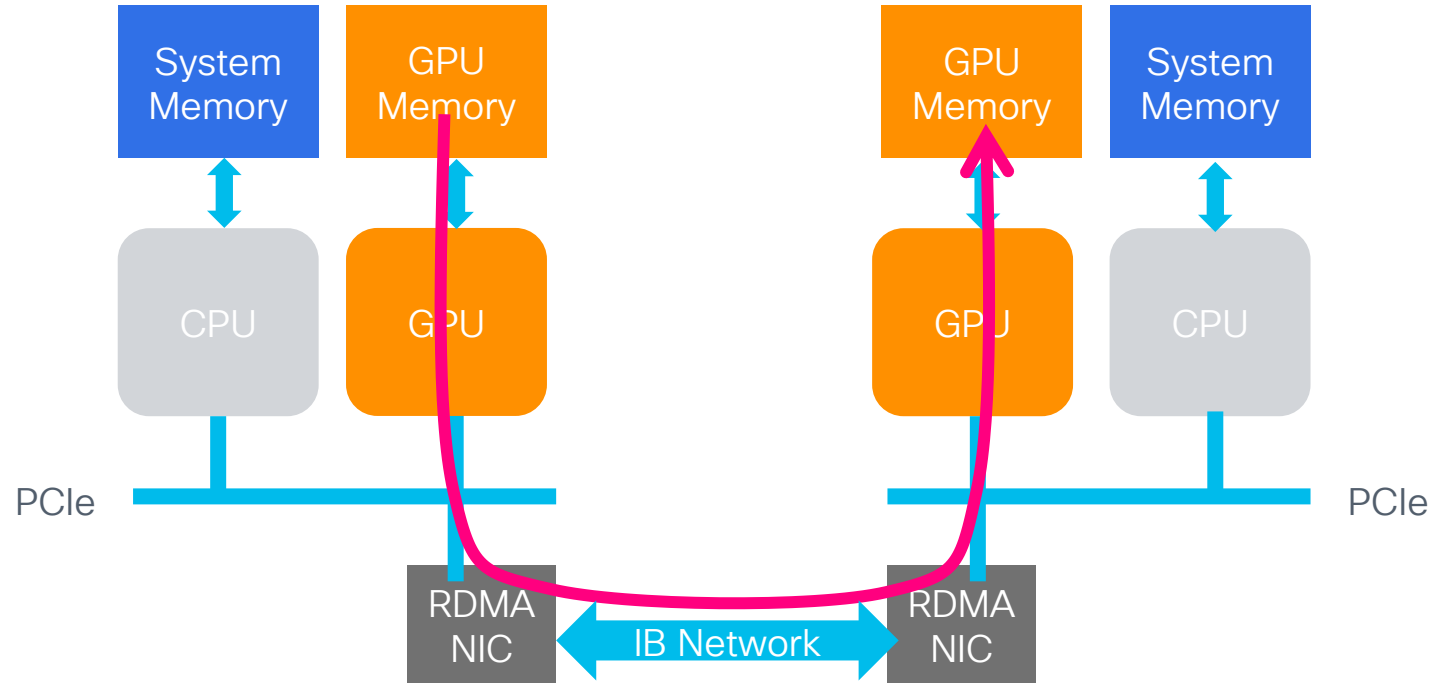
- The building blocks

Traditional Data movement is a bottleneck

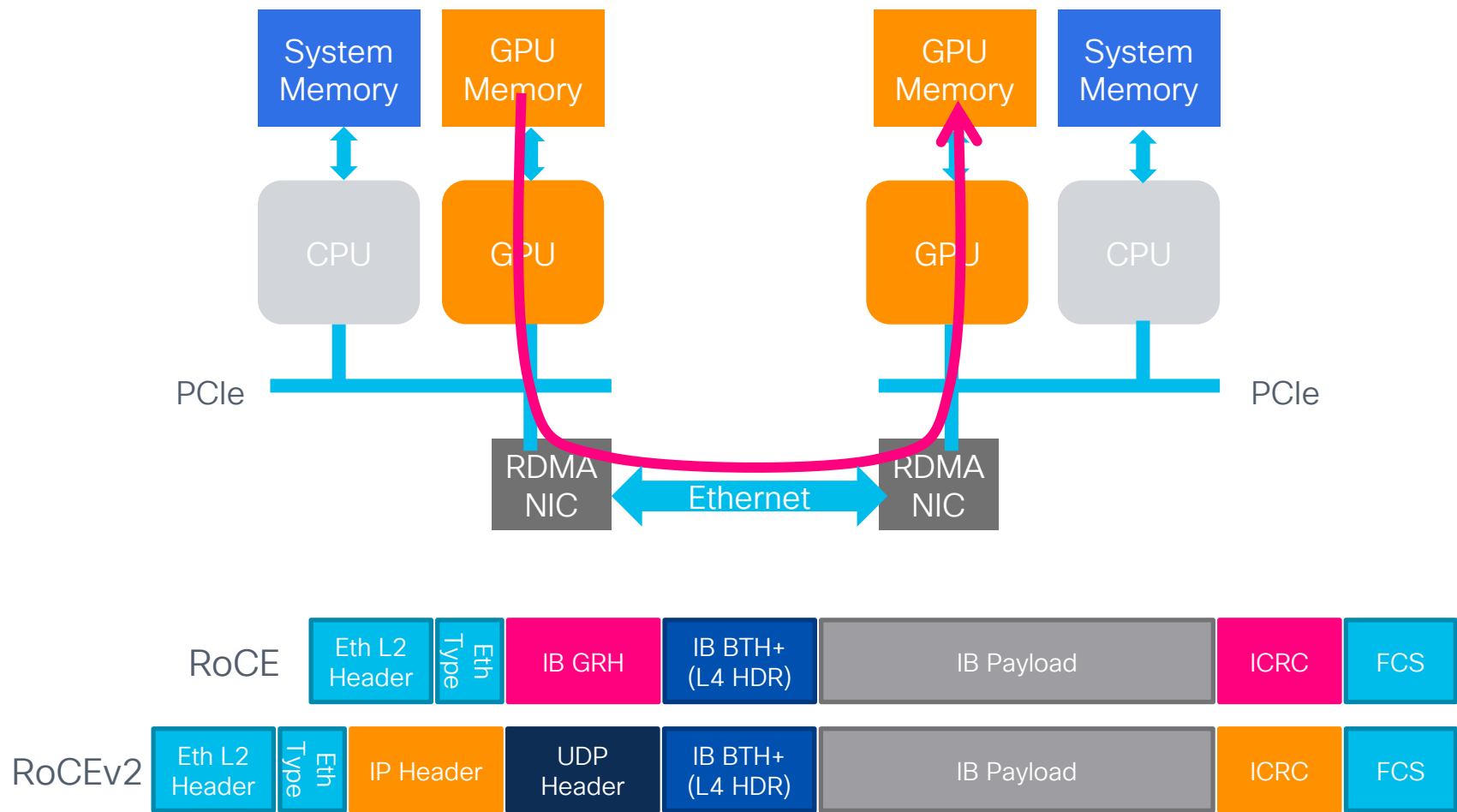


Works well for traditional architectures but think of this process when we have 10s, 100s or 1000s of GPUs communicating with one another!

RDMA (Remote Direct Memory Access)



RDMA Over Converged Ethernet – RoCE/RoCEv2

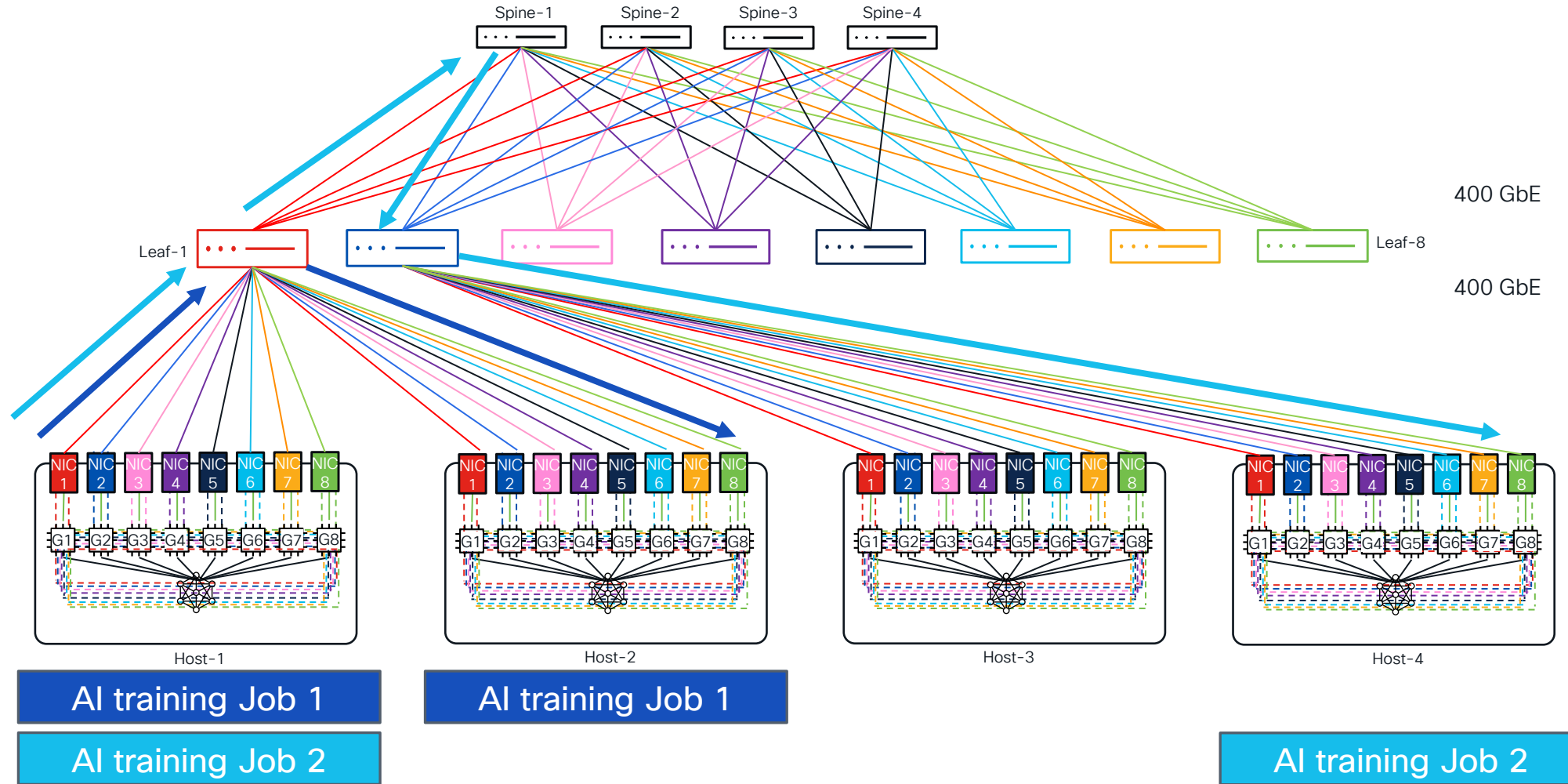


Challenge with traditional spine leaf architecture

Spine Switches: Cisco Nexus 9364D-GX2A switch (64 port x 400 GbE)

Leaf Switches: Cisco Nexus 9332D-GX2B (32 port x 400 GbE)

3:1 oversubscription (8 x 400GbE Leaf - Spine + 24 x 400GbE Leaf - Host)



Challenge with traditional spine leaf architecture

Spine Switches: Cisco Nexus 9364D-GX2A switch (64 port x 400 GbE)

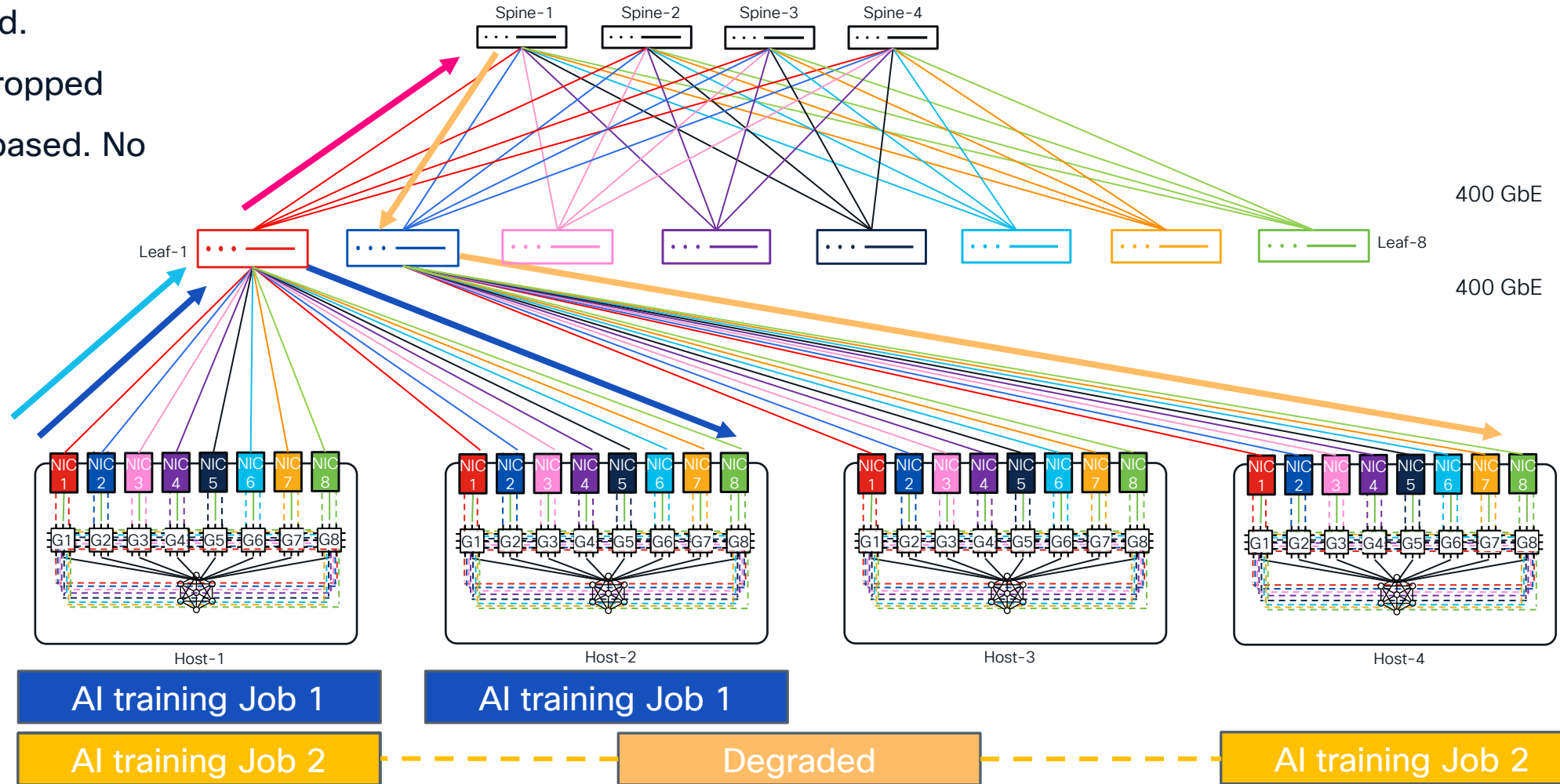
Leaf Switches: Cisco Nexus 9332D-GX2B (32 port x 400 GbE)

3:1 oversubscription (8 x 400GbE Leaf -Spine + 24 x 400GbE Leaf - Host)

Leaf -1 uplinks get
oversubscribed.

Packets get dropped

RoCE is UDP based. No
retransmits



Challenge with traditional spine leaf architecture

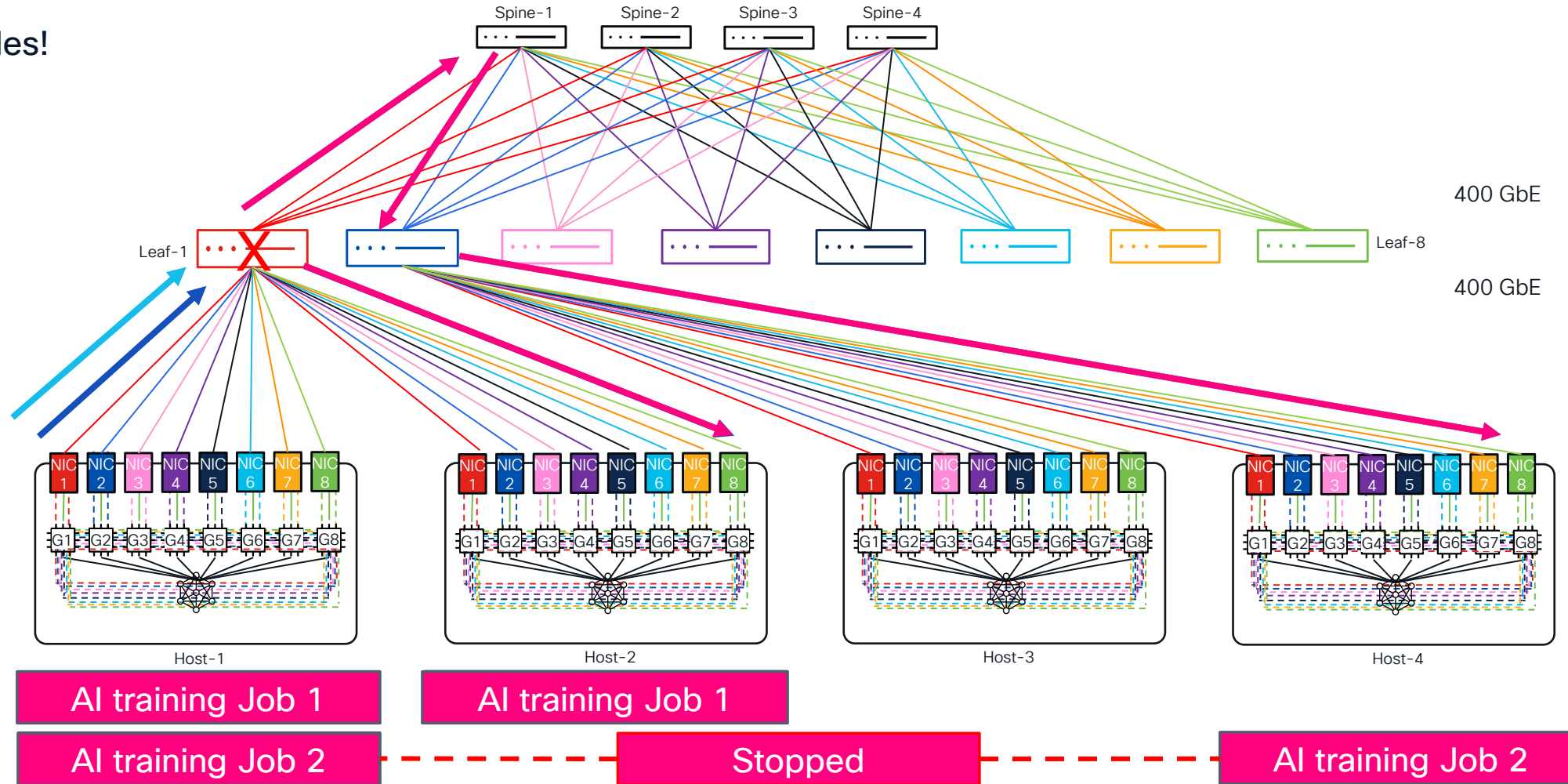
Spine Switches: Cisco Nexus 9364D-GX2A switch (64 port x 400 GbE)

Leaf Switches: Cisco Nexus 9332D-GX2B (32 port x 400 GbE)

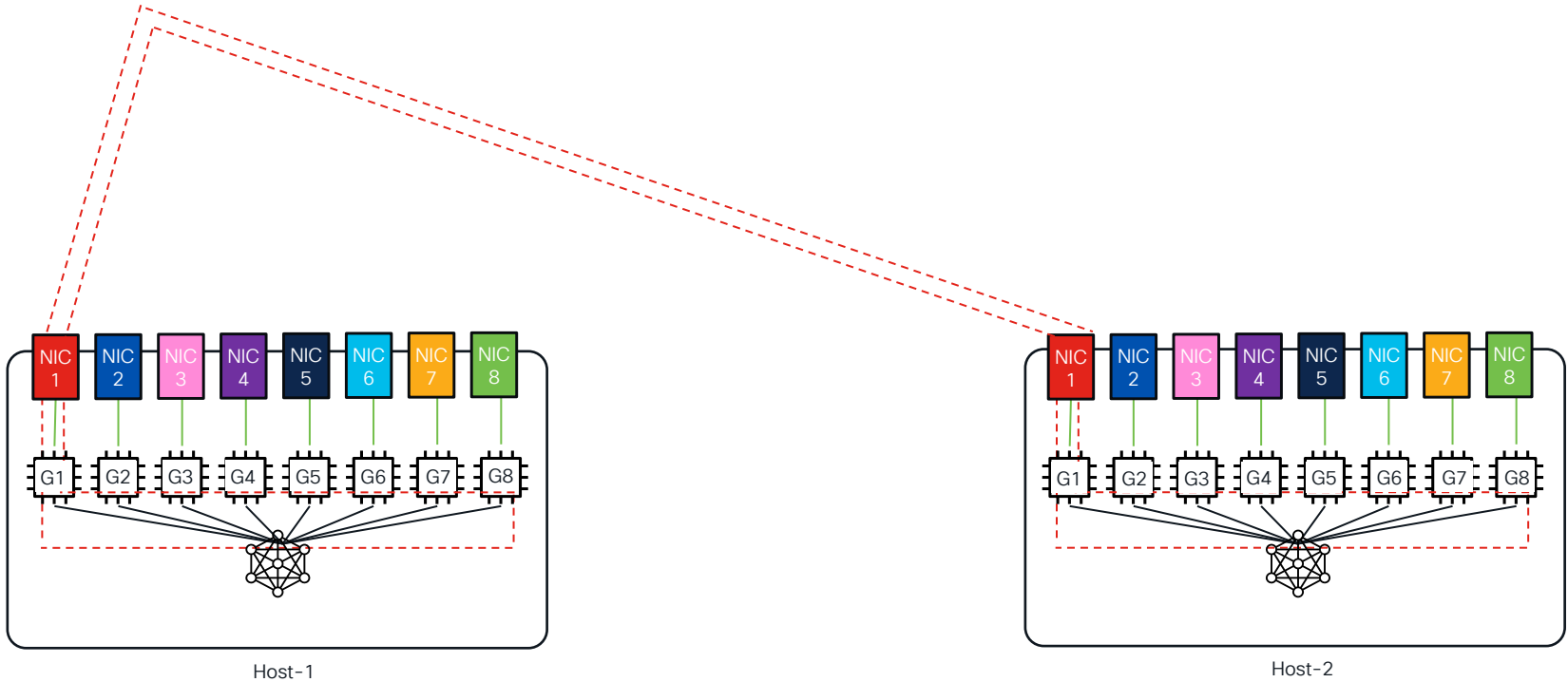
3:1 oversubscription (8 x 400GbE Leaf -Spine + 24 x 400GbE Leaf - Host)

Leaf -1 goes down

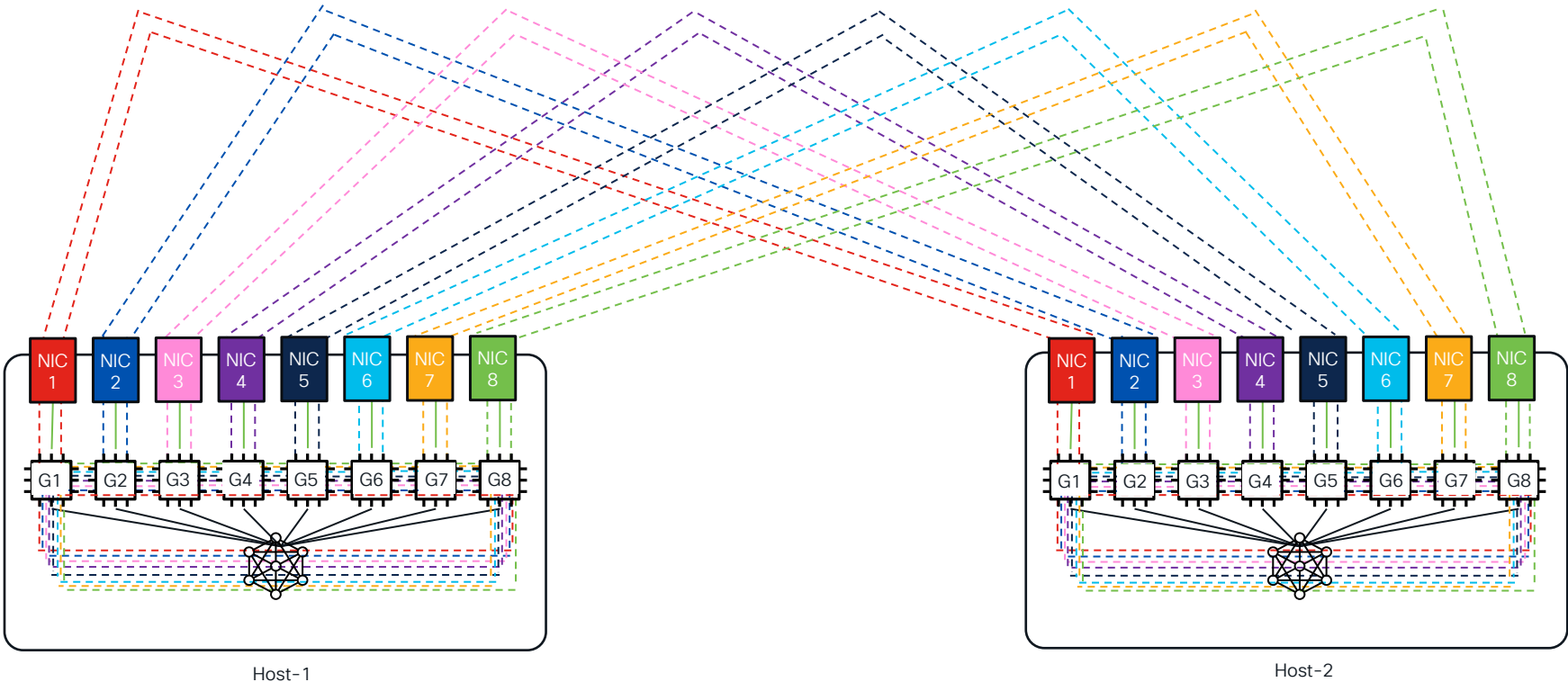
We lose 2 nodes!



Rails Architecture



Rails Architecture

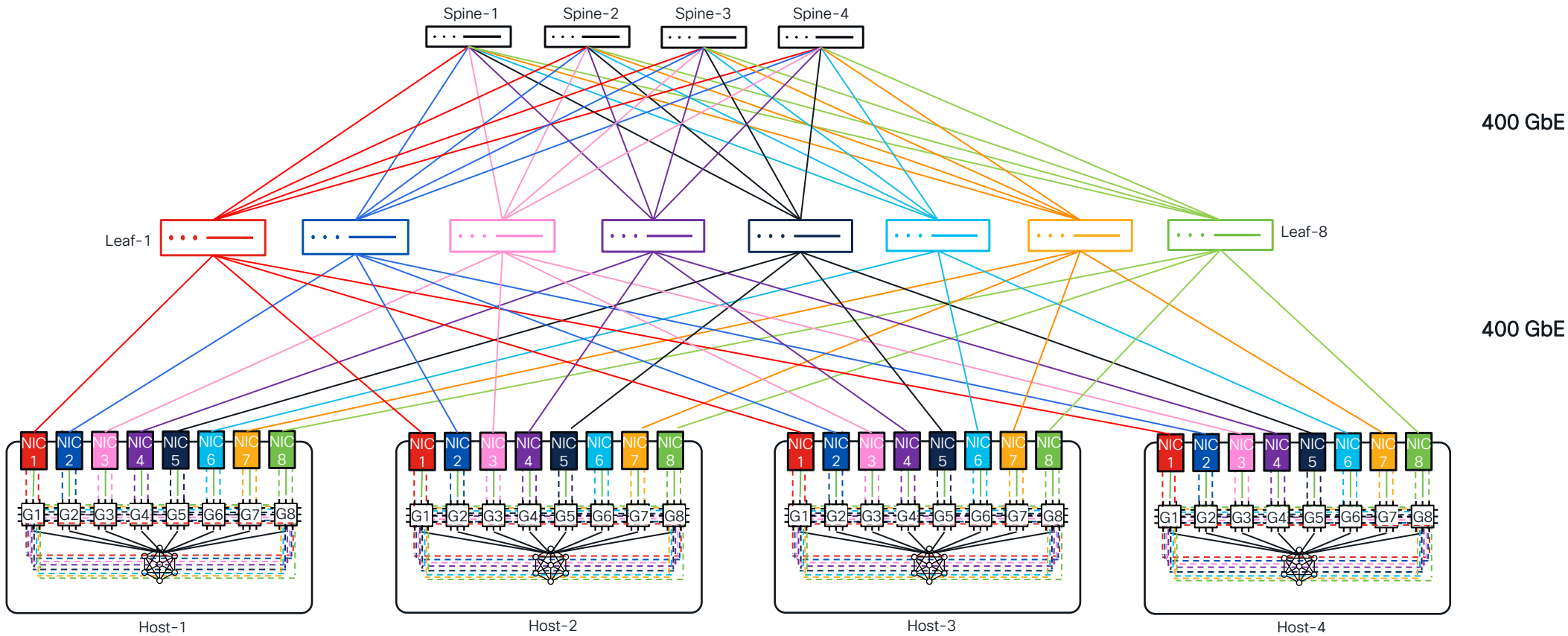


The AI Backend architecture

Spine Switches: Cisco Nexus 9364D-GX2A switch (64 port x 400 GbE)
Leaf Switches: Cisco Nexus 9332D-GX2B (32 port x 400 GbE)
Non-Blocking architecture – 16 x 400 GbE Leaf- Spine + 16 x 400 GbE Leaf - Host

Rails Optimised

Rails only



The AI Backend architecture

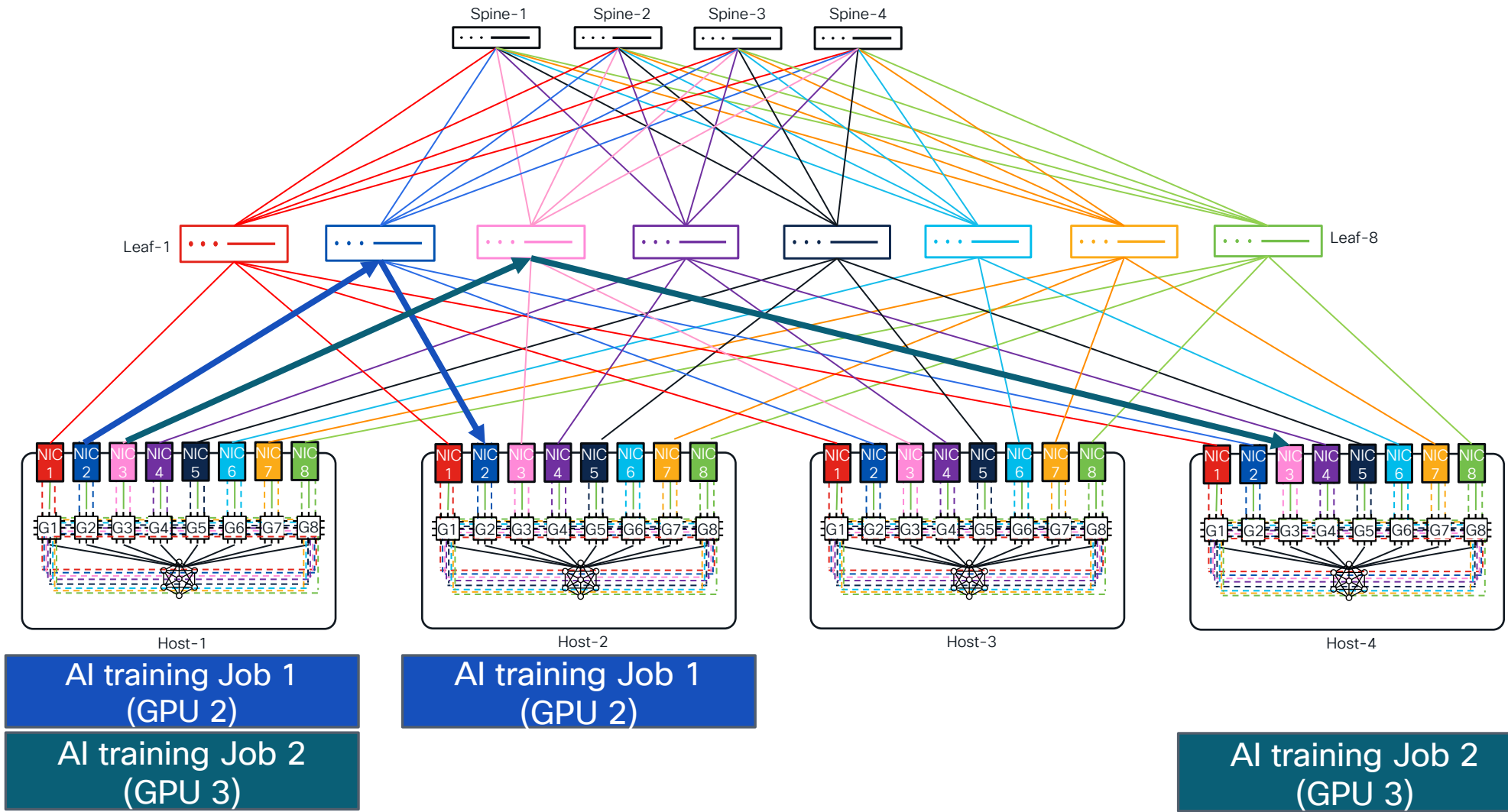
Spine Switches: Cisco Nexus 9364D-GX2A switch (64 port x 400 GbE)
Leaf Switches: Cisco Nexus 9332D-GX2B (32 port x 400 GbE)
Non-Blocking architecture – 16 x 400 GbE Leaf- Spine + 16 x 400 GbE Leaf - Host

Rails Optimised

Rails only

400 GbE

400 GbE



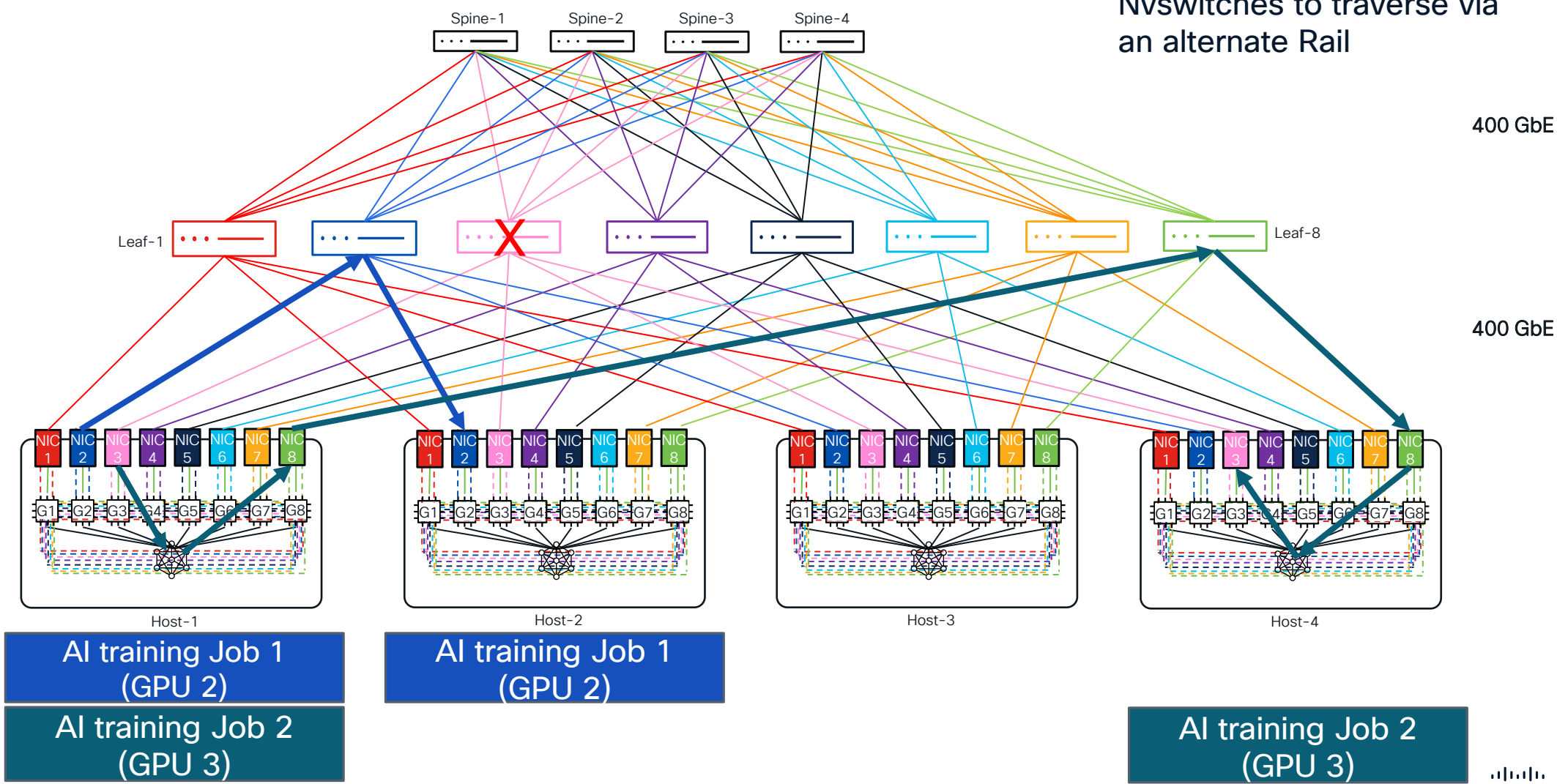
The AI Backend architecture

Spine Switches: Cisco Nexus 9364D-GX2A switch (64 port x 400 GbE)
Leaf Switches: Cisco Nexus 9332D-GX2B (32 port x 400 GbE)
Non-Blocking architecture – 16 x 400 GbE Leaf- Spine + 16 x 400 GbE Leaf - Host

What if we lose switch 3?
Traffic will use the Nvswitches to traverse via an alternate Rail

Rails Optimised

Rails only



AI Backend Networks

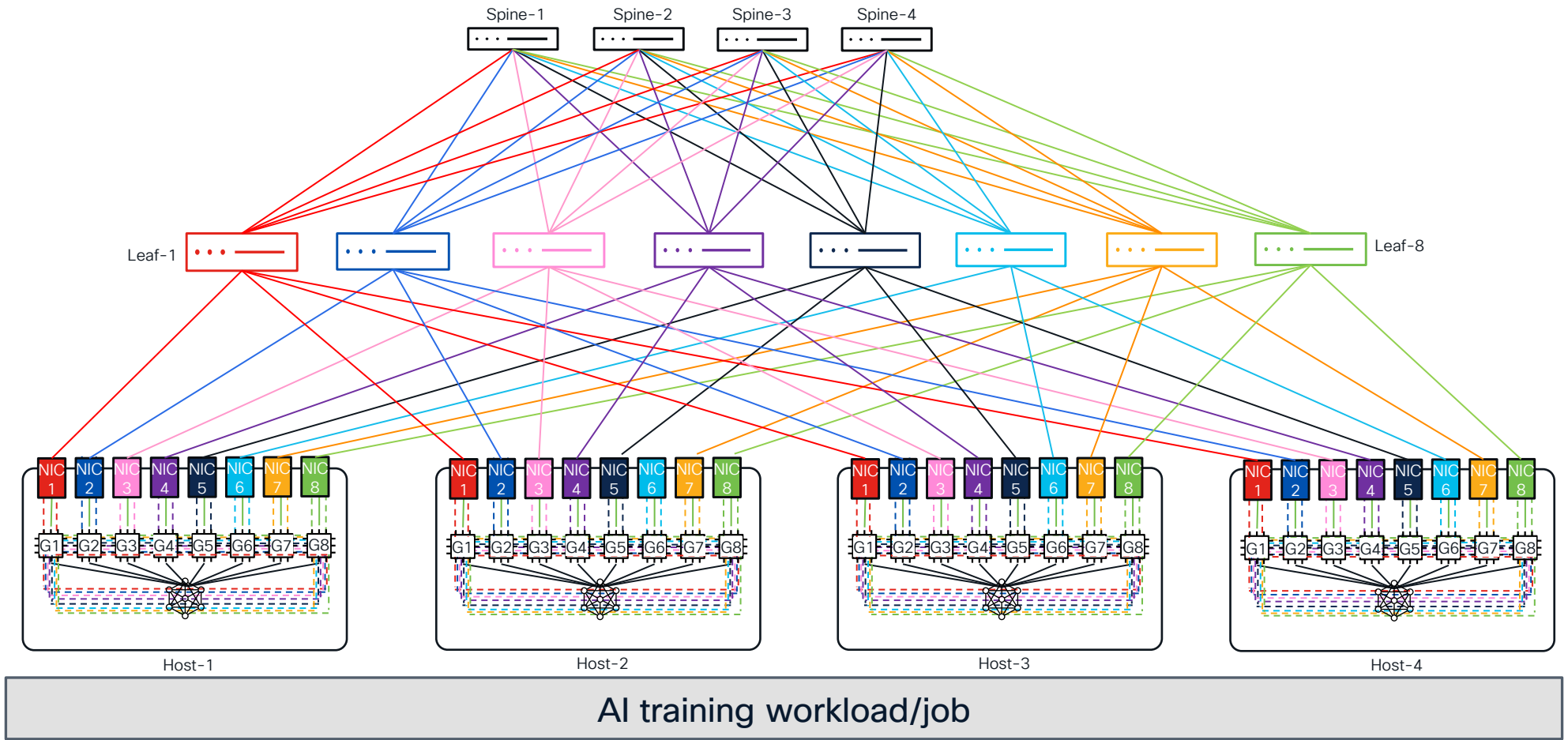
- Finetuning for optimal performance**

So this is not optimised?

Spine Switches: Cisco Nexus 9364D-GX2A switch (64 port x 400 GbE)
Leaf Switches: Cisco Nexus 9332D-GX2B (32 port x 400 GbE)
Non-Blocking architecture – 16 x 400 GbE Leaf- Spine + 16 x 400 GbE Leaf - Host

Rails Optimised

Rails only



AI/ML Backend Network Requirements

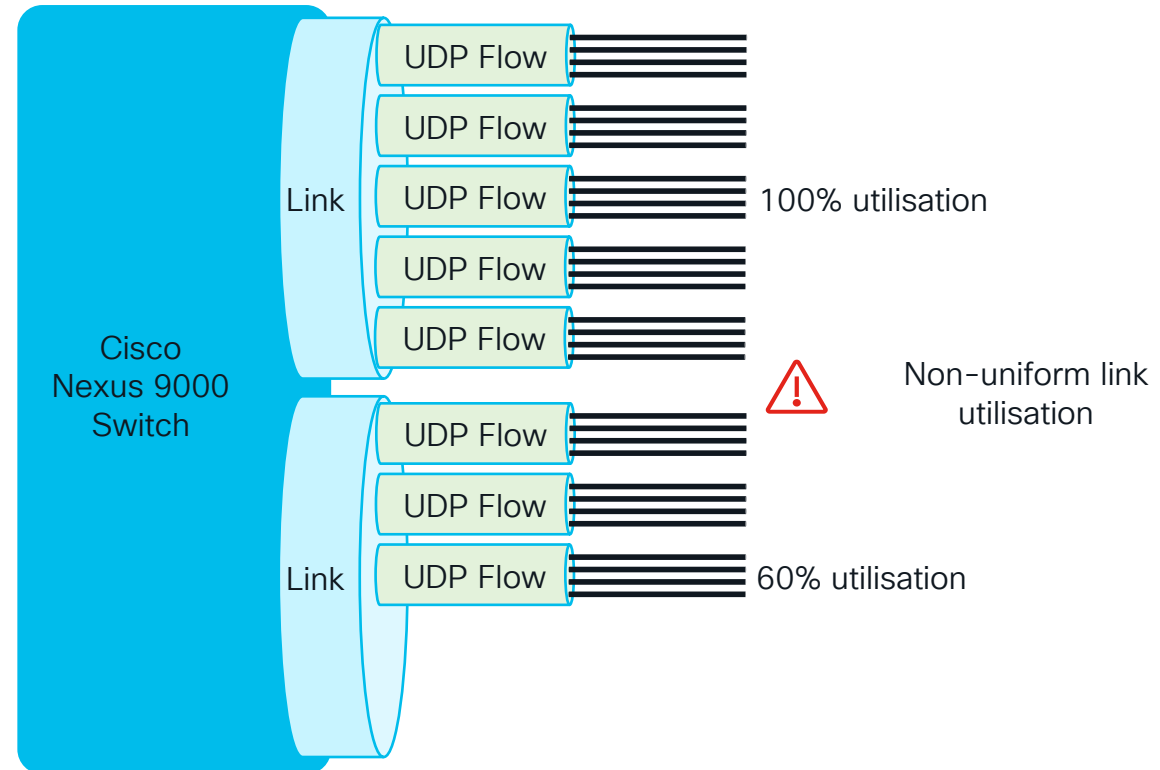
	1	2	3	4
Requirement	 Non-Blocking, High Bandwidth	 Lossless	 Low, Consistent Latency	 Operations
Driver	Data transferred during training increases at almost 6x model size	2% packet loss can cause training to be 8x slower	Increased network delay and variance degrade distributed training performance	AI Infrastructure is expensive, downtime and performance degradation are expensive
Solution	400/800G Dynamic Load Balancing Static Pinning ECMP with UDF	RoCEv2 DCQCN (PFC/ECN) Dedicated buffers	Optimised ASICs Rich QoS Intelligent Buffer Management	AI Fabric Automation AI Fabric Visibility Telemetry

Fabric Load Balancing – The challenge



*AI/ML backend
network traffic
(RoCEv2) has
very little
entropy*

Default ECMP load-balancing based on
Layer 4 UDP flow



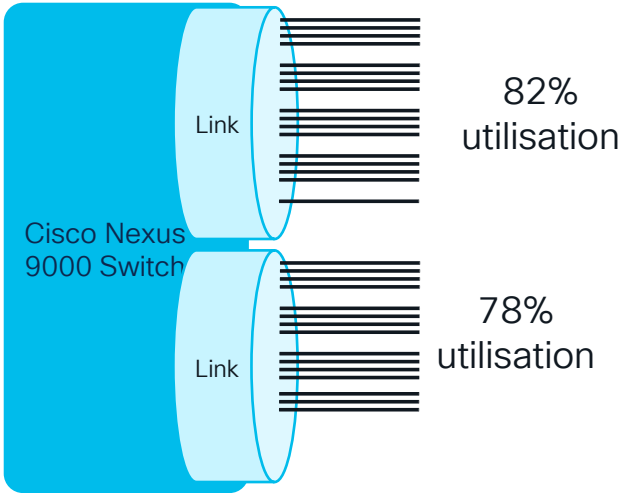
Fabric Load Balancing – The solutions

For a deeper dive,
check out
BRKDCN-1226



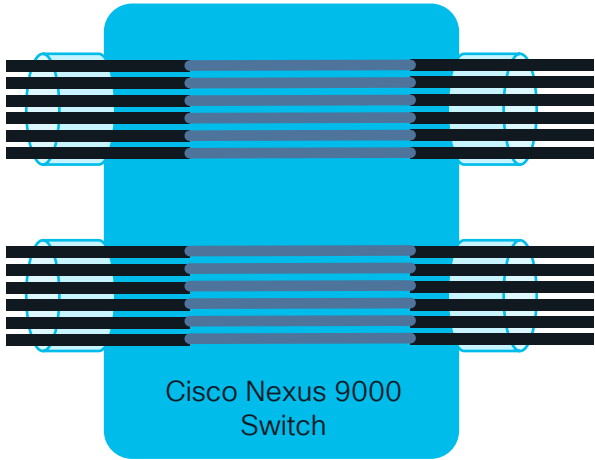
ECMP with UDF

Load-balance based on IB destination queue pair



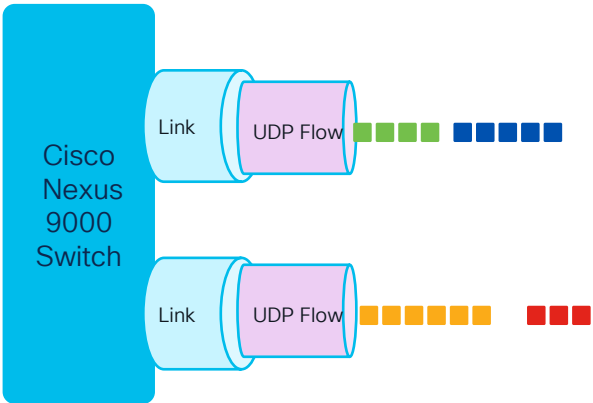
Static Pinning

Statically pin all traffic on an ingress port to an egress port and vice-versa



Dynamic Load Balancing (DLB)

Dynamically load balance **flowlets** based on **real-time** link utilisation



✓ Nearly uniform link utilisation

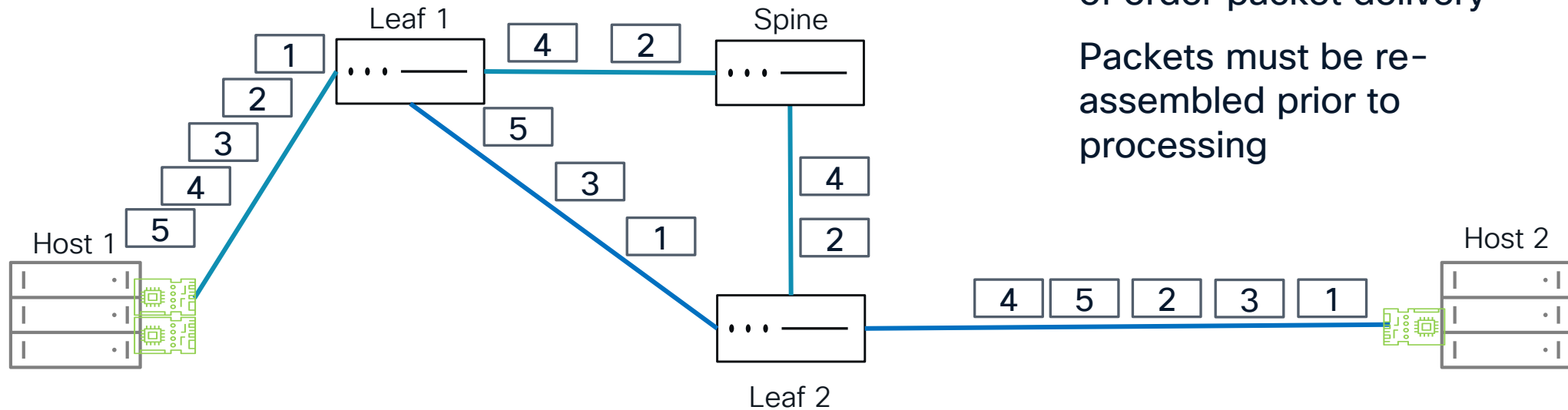
Per packet load balancing with Nvidia

For a deeper dive,
check out
BRKDCN-1226



RDMA is not tolerant to out
of order packet delivery

Packets must be re-
assembled prior to
processing



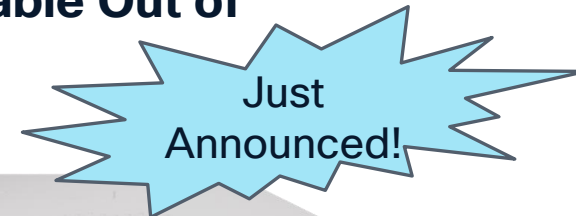
Solution: Integrate Cisco Nexus with Nvidia Bluefield3 SuperNIC to enable Out of order packet re-assembly



Cisco Nexus with Silicon One becomes the only 3rd party switching platform integrated into Nvidia's Enterprise Reference Architecture.



Cisco 9164E 800G switch with
Nvidia Spectrum 4 Silicon

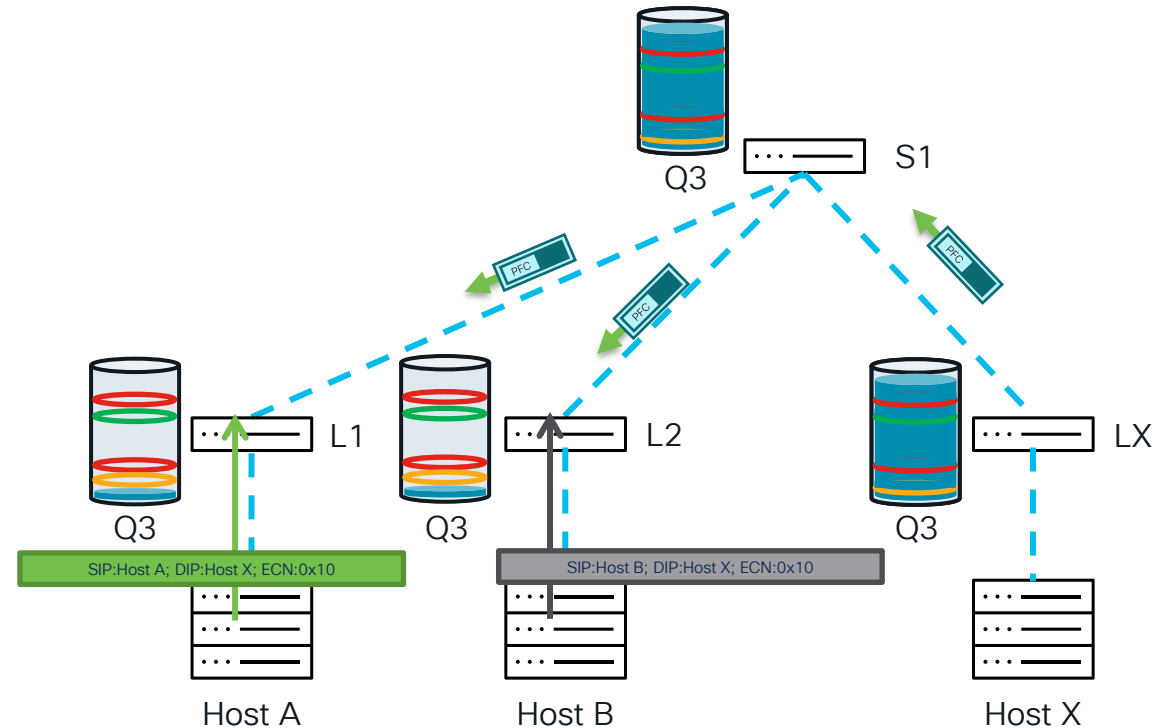


Congestion Control

For a deeper dive,
check out
BRKDCN-1226



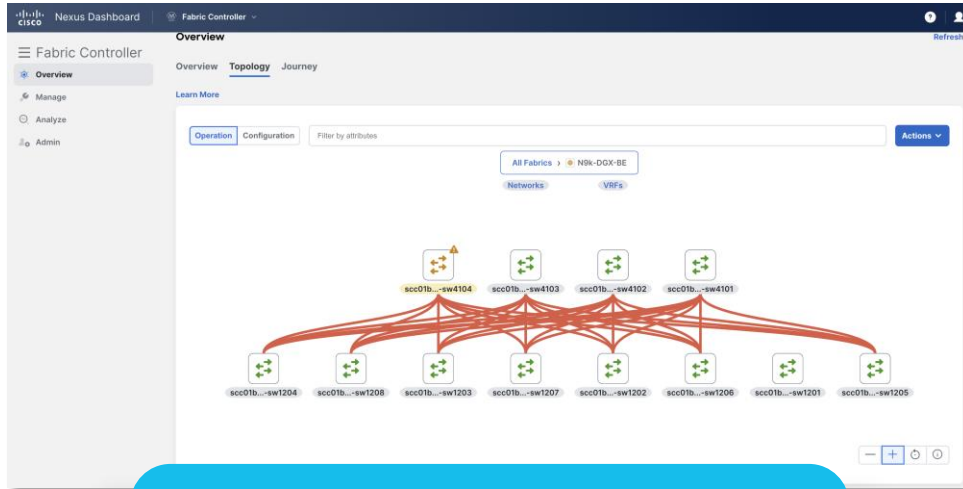
- Priority Flow Control (PFC)
- Explicit Congestion Notification (ECN)
- Data Centre Quantised Congestion Notification (DCQCN)
 - PFC + ECN



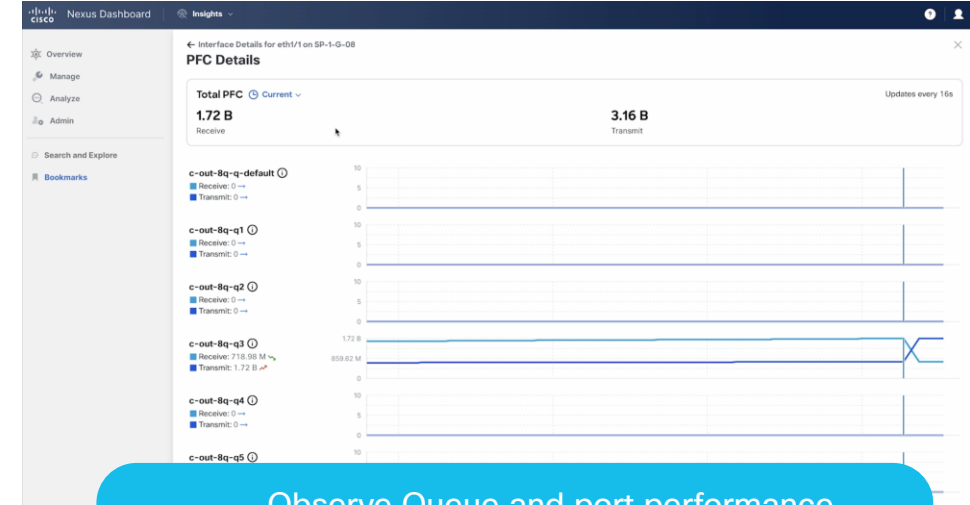
Nexus Dashboard for AI/ML Networks

Design and Operate AI/ML networks with ease

For a deeper dive,
check out
BRKDCN-1226



Design Network topologies



Observe Queue and port performance

Nexus Dashboard
Kubernetes Controller

Fabric Controller

- [Overview](#)
- [Manage](#)
- [Analysis](#)
- [Admin](#)

Manage > Templates

Templates [Learn More](#)

Name contains qps

Name	Supported Platforms	Type	Sub Type	Modified	Tags	Description	Reference Count
☐ Al.Fabric_QOS_XXXX	NBK	POLICY	DEVICE	5 months ago	QoS,AI/M	System QoS Marking and Queuing policy for NBK Cloudscale Series HW with PFC and ECN for systems with predominantly 100G uplinks	0
☐ Al.Fabric_QOS_25G	NBK	POLICY	DEVICE	5 months ago	QoS,AI/M	System QoS Marking and Queuing policy for NBK Cloudscale Series HW with PFC and ECN for systems with predominantly 25G uplinks	0
☐ Al.Fabric_QOS_A200	NBK	POLICY	DEVICE	5 months ago	QoS,AI/M	System QoS Marking and Queuing policy for NBK Cloudscale Series HW with PFC and ECN for systems with predominantly 400G uplinks	12
☐ Al.Fabric_QOS_A400_ECENonly	NBK	POLICY	DEVICE	3 months ago	QoS,AI/M	System QoS Marking and Queuing policy for NBK Cloudscale Series HW with PFC and ECN for systems with predominantly 400G uplinks	0
☐ Al.Fabric_QOS_A200_PFC_only	NBK	POLICY	DEVICE	3 months ago	QoS,AI/M	System QoS Marking and Queuing policy for NBK Cloudscale Series HW with PFC for systems with predominantly 400G uplinks	0
☐ Al.Fabric_QOS_Classification_Custom	NBK	POLICY	DEVICE	5 months ago		Custom QoS Classification using cshelp and Marking in ip-qgroups	0
☐ Al.Fabric_QOS_Queueing_Custom	NBK	POLICY	DEVICE	5 months ago	QoS,AI/M,_Custom	Custom QoS Queueing policy for interfaces. Must be used in combination with system QOS policy to Mark traffic	0
☐ interface_qos_service_policy	NBK	POLICY	INTERFACE	5 months ago			520
☐ qos_class_and_policy_map	NBK	POLICY	DEVICE	5 months ago			0
☐ ipsc_drop_threshold	NBK	POLICY	DEVICE	5 months ago		QOS No Drop Threshold Template for AI/M	0

10 items found
Rows per page: 10 1

Create and Manage AI/ML profiles

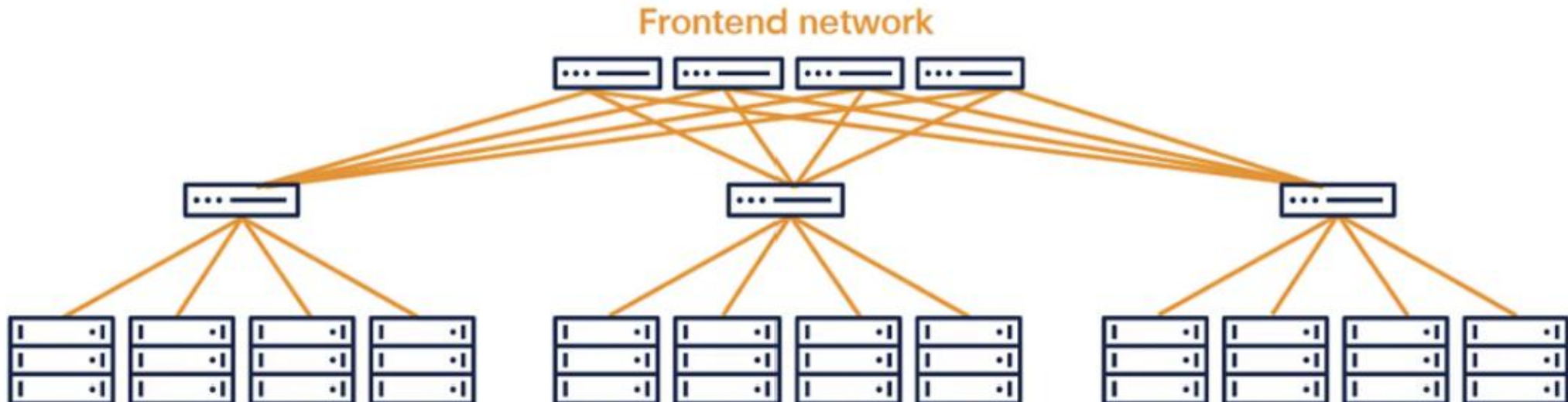


Topology view for broader troubleshooting

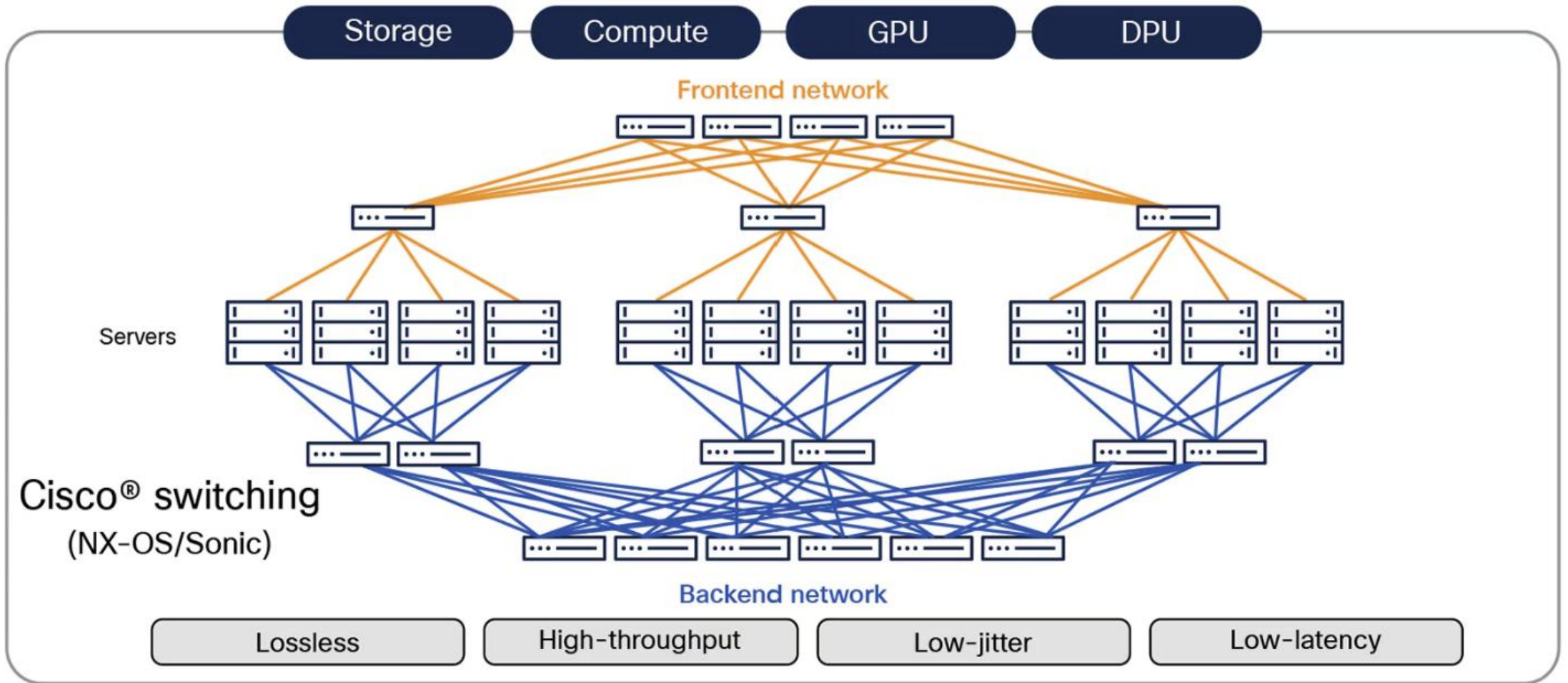
Coming
Soon!

Front-end considerations

- Uplift to high-speed Ethernet to enable fast data movement between storage, Real time stream engines and applications
- Non-blocking, particularly with high-speed AI storage is ideal but not mandatory
- Fibre Channel not acceptable for AI servers as Storage requires high speeds and is Ethernet driven. Fibre-channel support needs to be there for storage solutions for traditional applications.



Multi cluster architecture



Networking for AI – Callouts

AI workloads demand high bandwidth connectivity between multiple GPUs, resulting in the need for an AI dedicated “back-end” network.

The AI Backend network needs to be non-blocking and support features and capabilities that enable congestion detection & avoidance and low + Consistent latency

Cisco Nexus Dashboard and switches provide a rich set of capabilities that enable you to build and operate high performance AI backend networks.

Let's Review some scenarios

Example Infrastructure Requirements by Workload

Workload type	Small scale model inferencing	Large scale model inferencing	Fine Tuning	Small* model training	Large* model training
AI Accelerator (GPU)	Maybe 0-2 GPU per server	Yes 2-8 GPUs per server**	Yes 4-8 GPU per server	Yes 4-8 GPU per server	Yes 8 GPU per server/Rack Scale
Cluster size	N/A Generally, a part of the general compute infrastructure	0 - 100's** Can be dedicated or general Infra	8-80 GPUs ~10's of servers Dedicated AI Infra	20-80 GPUs ~10's of servers Dedicated AI Infra	>1000 GPUs 100s of servers Dedicated AI Infra
Intra-host GPU interconnect (e.g. NVLink)	No	Yes	Yes	Yes	Yes
Inter-host GPU interconnect ("Backend" network)	No	Needed if models span multiple nodes	Highly Likely 400-800G RoCE	Highly Likely 400-800G RoCE	Yes 400-800G RoCE/Infiniband
"Frontend" network	"Any"	100-400G	100-400G	100-400G	100-400G

* "Small" and "Large" here relate to the size of the model and the scale of inference (number of concurrent users, Token size etc)

** GPU count and Server count relate to enterprise use cases. In the case of hyperscalers, large scale inferencing can scale 1000s of GPUs.

Why security matters?

Security challenges with AI workloads

Learn more about AI Defense
BRKSEC-1554
Check out the Demo in WoS!

Application Security

- Lateral Movement
- Privilege Escalation
- Ransomware
- Container communications

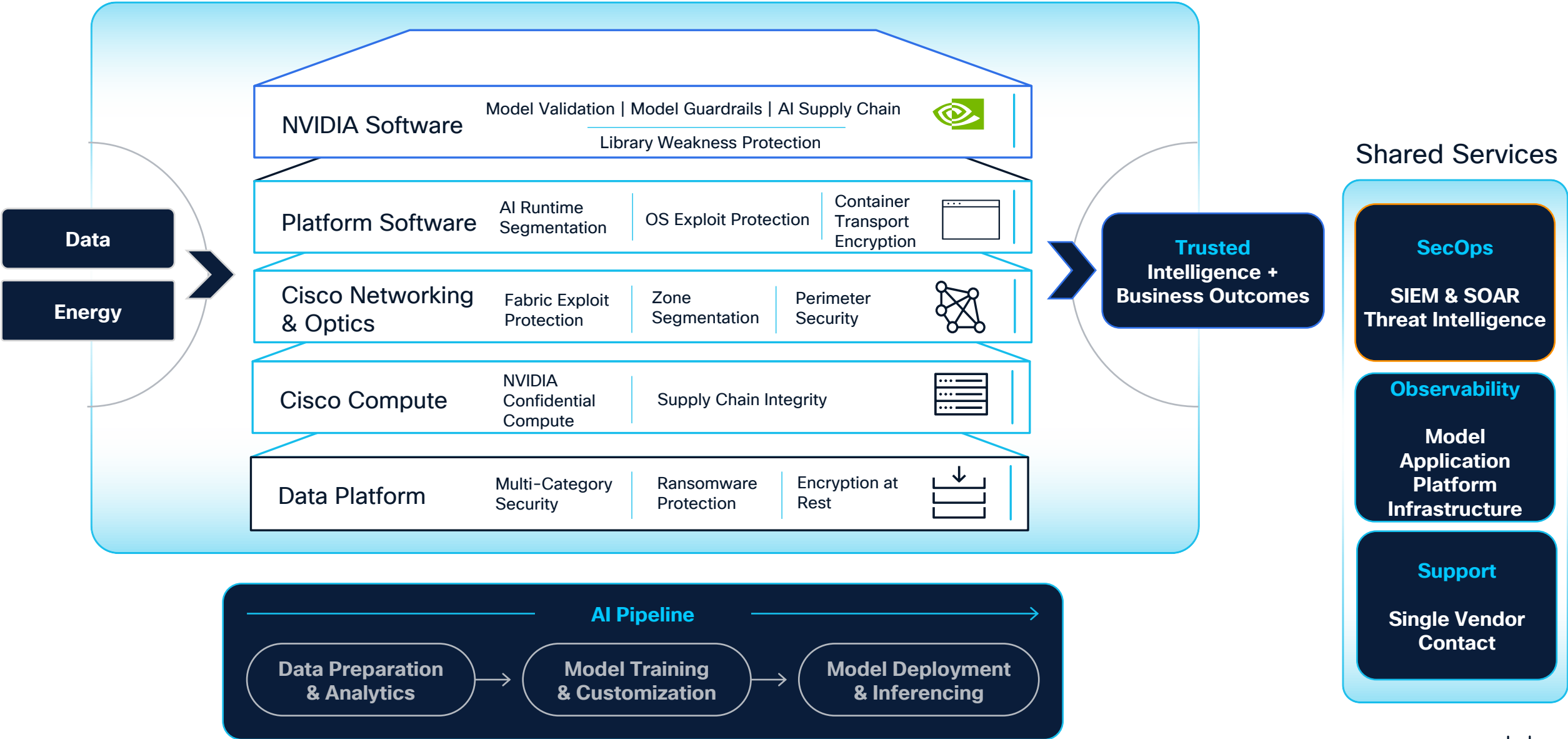
AI Security

- Data Poisoning
- Model Drift
- Prompt injection
- Misinformation
- Supply chain

Firewalls
Hypershield/Isovalent

Cisco AI Defense

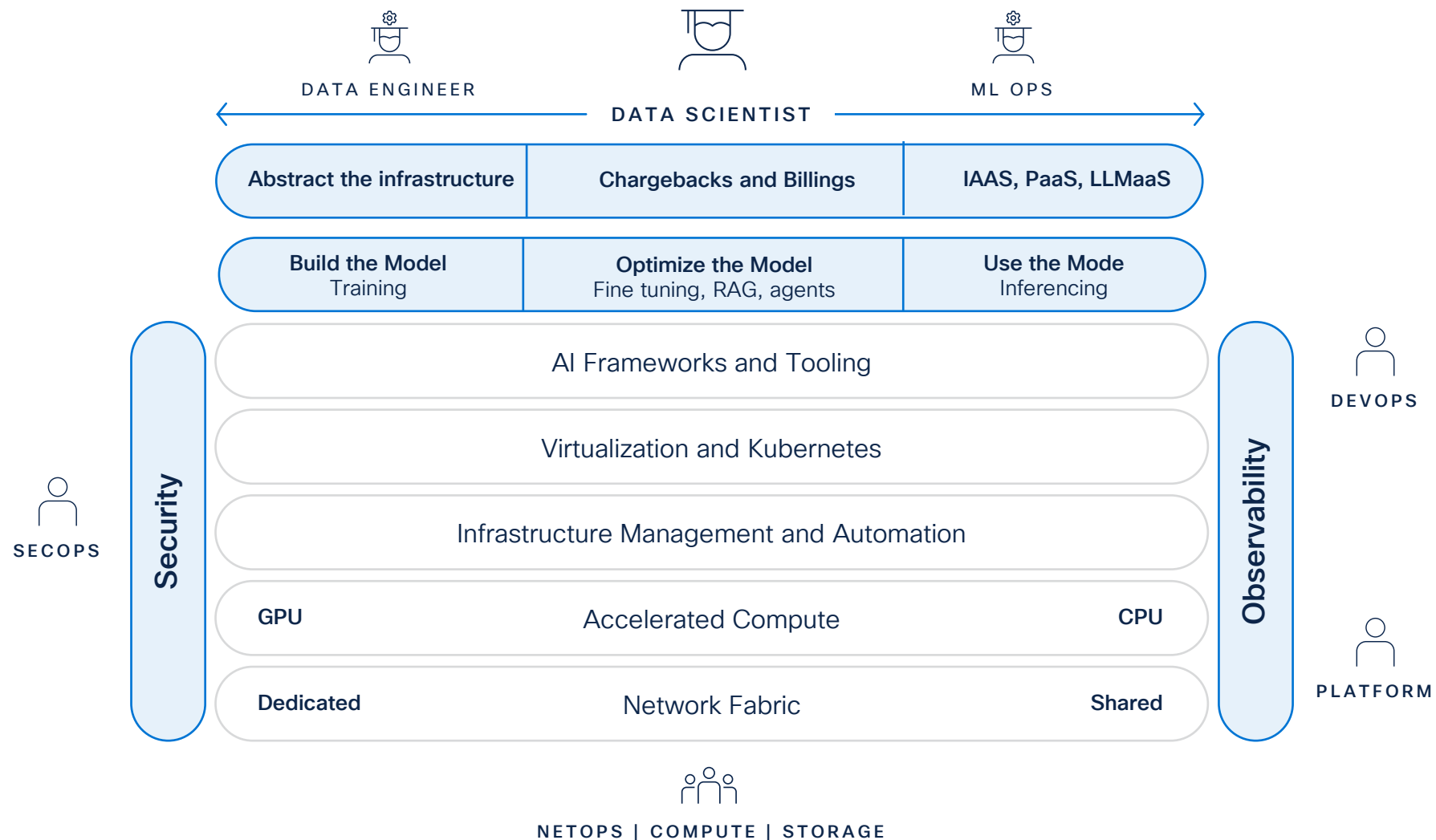
Cisco Secure AI Factory with Nvidia



The background of the slide is an abstract design featuring flowing, wavy lines in various shades of blue and white. These lines create a sense of movement and connectivity. A large, solid white rectangle is positioned on the left side of the slide, serving as a backdrop for the main text.

Bringing everything together

Generative AI – Full Stack System



Cisco AI PODs

A scalable architecture, built to support any AI workload simply & efficiently

Deploy AI with confidence
Cisco CVD, NVIDIA ERA

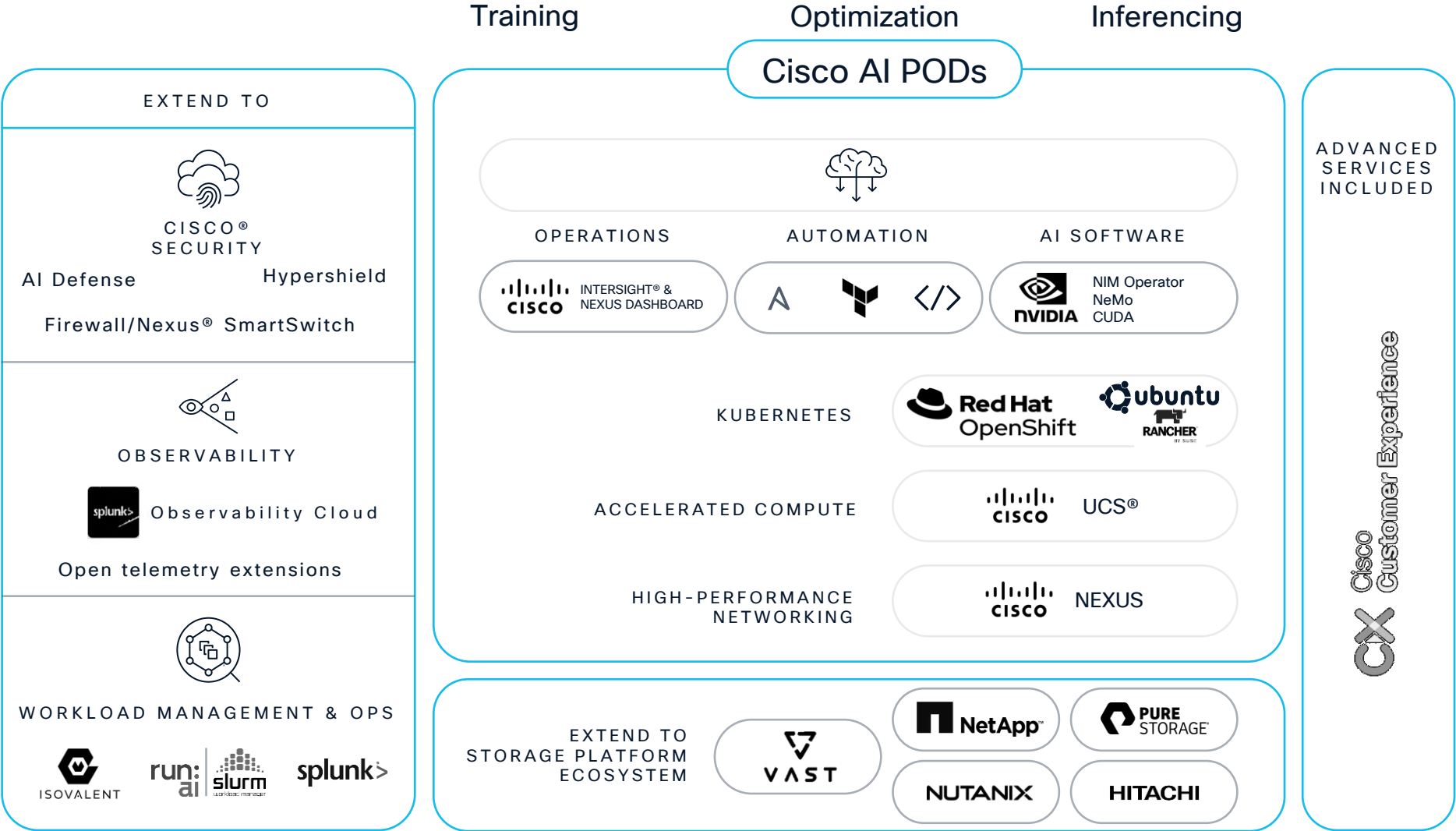
Fully supported stack
including Cisco and 3rd
party components

**Cisco CX
Success Track**

Orderable, use case
driven AI-ready
infrastructure stacks

**Inferencing.
Optimization.
Training.**

Incremental, atomic-level
–or– fabric-based
cluster scale



Conclusion

A Quick Recap



Understand your **AI requirements**

- Identify AI projects and key stakeholders
- Determine AI workload types and likely deployment locations
- Determine power and cooling capabilities of the DC or CoLo facilities.
- Understand infrastructure requirements for AI projects



Define your **AI Infrastructure Strategy**

- Don't get distracted by the latest shiny object
- Ensure Infrastructure strategy is correctly mapped to real AI requirements
- “Evolution not Revolution” – align AI initiatives with existing investments
- Build with the future in mind.



Accelerate **AI Initiatives**

- “Standardize and De-risk” – streamline AI deployments with validated reference architectures
- Leverage Cisco's experience
- Visit us at the World of Solutions

Other AI Infrastructure

Other relevant AI Infrastructure sessions:

LTSDCN-1769 – Agentic AI and the Cisco Secure AI Factory with Nvidia

BRKCOM-2115 – Building End-to-End a full-stack AI-optimised Datacentres: Architectures and Best Practices

KDDDCN-1000 – Building the AI Internet

LTSDCN-1768 – Designing and managing end-to-end AI Data Center Fabrics with Cisco Nexus and Nvidia Spectrum-X silicon

BRKDCN-1226 – Why Cisco Nexus 9000 series switches for AI/ML Infrastructure Networking

Visit the Cisco Learning team in WoS to get the latest on Cisco's AI certifications!

Complete your session evaluations



Complete a minimum of 4 session surveys and the Overall Event Survey to claim a Cisco Live T-Shirt.



Earn up to 800 points by completing all surveys and climb the Cisco Live Challenge leaderboard.



Level up and earn exclusive prizes!



Complete your surveys in the Cisco Live Events app.

Continue your education



Visit the Cisco Stand for related demos



Book your one-on-one Meet the Expert meeting



Attend the interactive education with Capture the Flag, and Walk-in Labs



Visit the On-Demand Library for more sessions at www.CiscoLive.com/on-demand

Connect with me



Thank you

CISCO Live !

