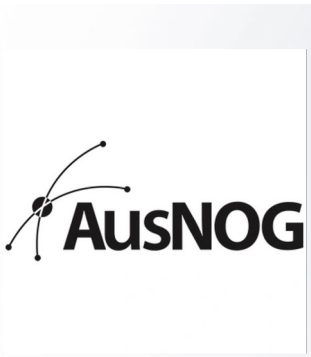


ARISTA

Practical AI Networking Innovations

Matthew Thurbon - Systems Engineering Team Lead
September 2024

turbo@arista.com

My 30 Year Journey

ALCATEL

Alcatel·Lucent

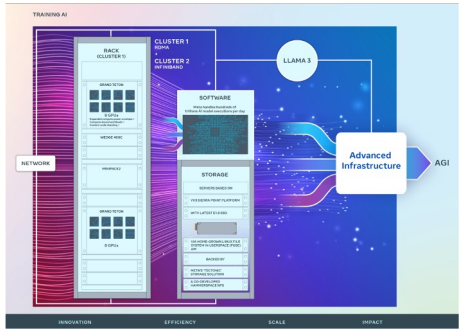
JUNIPER NETWORKS

ARISTA

Hyperscalers AI Infrastructure

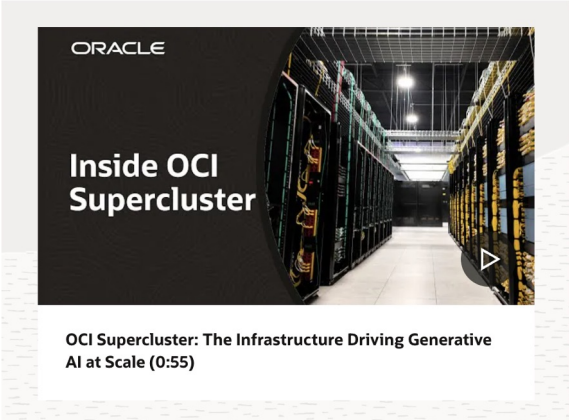
POSTED ON MARCH 12, 2024 TO AI RESEARCH, DATA CENTER ENGINEERING, ML APPLICATIONS

Building Meta's GenAI Infrastructure



By the end of 2024, we're aiming to continue to grow our infrastructure build-out that will include 350,000 NVIDIA H100 GPUs as part of a portfolio that will feature compute power equivalent to nearly 600,000 H100s.

https://engineering.fb.com/2024/03/12/data-center-engineering/building-metas-genai-infrastructure/



Run the most demanding AI workloads faster, including generative AI, computer vision, and predictive analytics, anywhere in our distributed cloud. Get the latest GPU compute, scaling up to the 32,768 GPU Oracle Cloud Infrastructure (OCI) Supercluster.

AWS Teases 65 Exaflop 'Ultra-Cluster' with Nvidia, Launches New Chips

Google Cloud claims 'most powerful' publicly available machine learning cluster

Compute

Announcing Trillium, the sixth generation of Google Cloud TPU

May 15, 2024

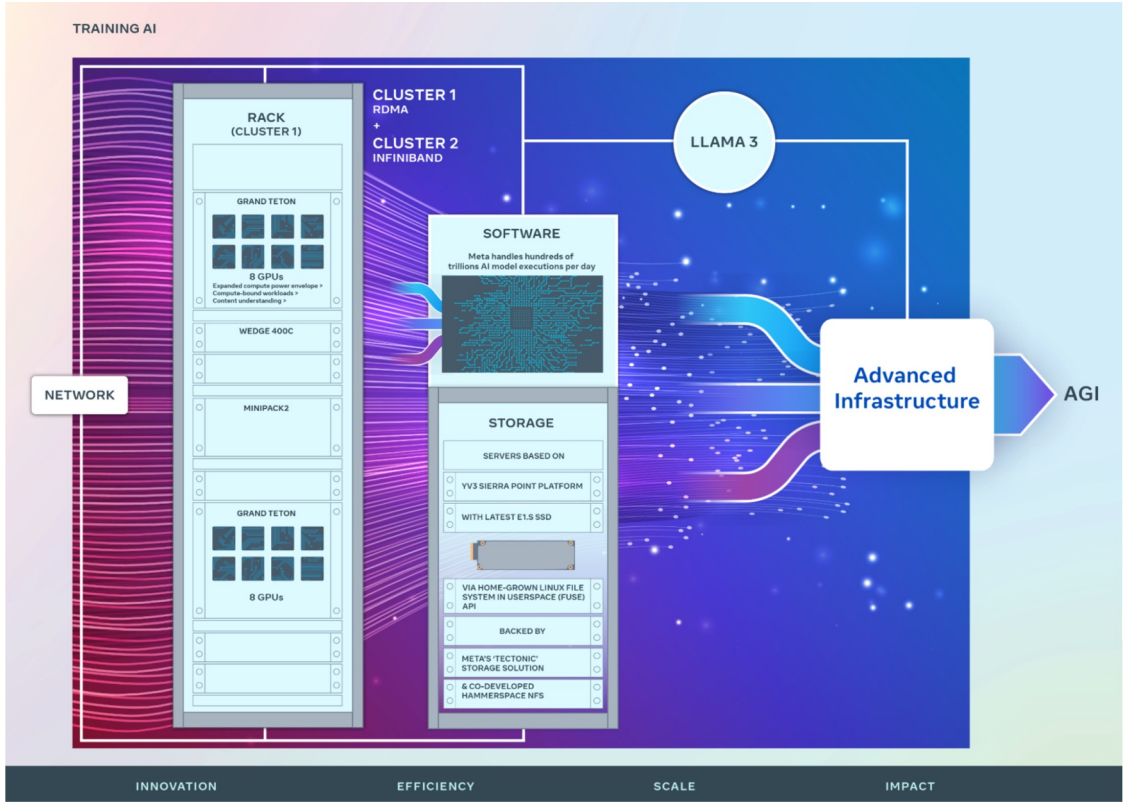
Empowering Azure Storage with RDMA : Today, around 70% of traffic in Azure is RDMA and intra-region RDMA is supported in all Azure public regions. https://www.microsoft.com/en-us/research/uploads/prod/2023/03/RDMA_Experience_Paper_TR-1.pdf

Nvidia launches AI foundry on Microsoft Azure
https://www.sdxcentral.com/articles/news/nvidia-launches-ai-foundry-on-microsoft-azure/2023/11/

Tesla Unveils Top AV Training Supercomputer Powered by
The cluster uses 720 nodes of 8x NVIDIA A100 Tensor Core GPUs (5,760 GPUs total) to achieve an industry-leading 1.8 exaflops of performance.

Meta GenAI Llama3 Infrastructure

Meta Built a 24k GPU Ethernet Cluster



"Through careful co-design of the network, software, and model architectures, we have successfully used both RoCE and InfiniBand clusters for large, GenAI workloads (including our ongoing training of Llama 3 on our RoCE cluster) without any network bottlenecks."

"With this in mind, we built one cluster with a remote direct memory access (RDMA) over converged Ethernet (RoCE) network fabric solution based on the Arista 7800…"

https://engineering.fb.com/2024/03/12/data-center-engineering/building-metas-genai-infrastructure/

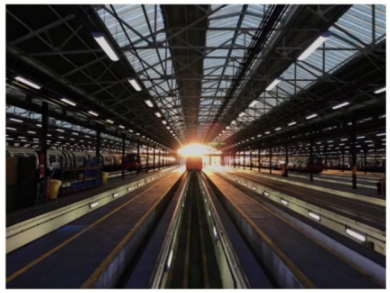
AI Applications - Driving intense network demand

Morgan Stanley – Chatbot for Financial Advisors

- Built on GPT-4
- Trained with 100,000 curated research documents
- Provides expert advise to financial advisors
- Transparently provides answers, reasoning and sources
- Currently piloting with 300 advisors, planning to rollout to 16,000 in summer

Network Rail – Training Videos

- Previously used external production and voice actors
- Generative video cut the time to produce a video by 95%
- Created 500 training videos in 6 months
- Able to quickly adapt video content rather than re-start
- Diverse avatars and voices enable inclusion

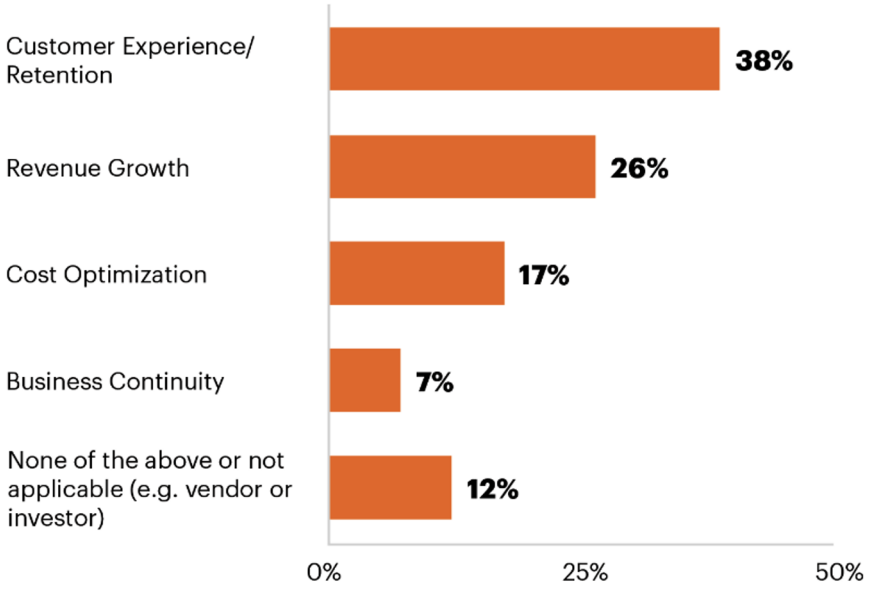


Gartner

IBM – Project Photoresist

- Identifying a new molecule typically takes 10 years and costs 10-100 million
- Project photoresist was able to reduce that to months
- IBM developed a new photoacid generator, which is used in lithography, a key process in developing computer chips
- 100 times faster than traditional methods.

Primary Focus of Generative AI Initiatives



gartner.com

Source: Gartner
© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. 2323718

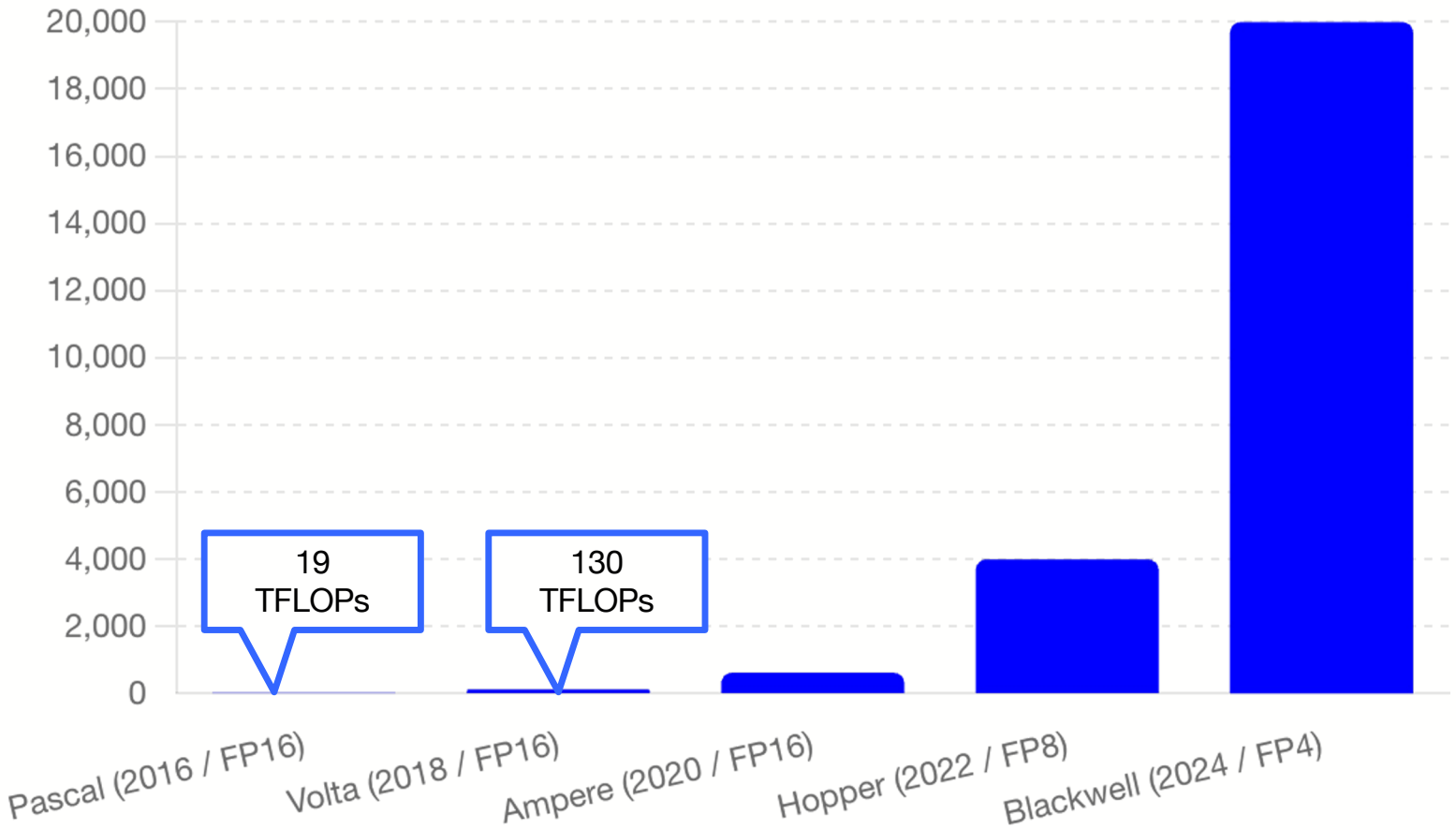
Gartner

ARISTA

Improved Performance per GPU - Nvidia

From Pascal to Blackwell, GPU performance has increased by 1,053x

Lower Floating Point (FP) results in lower precision and lower range but also lower bandwidth and memory requirements



ARISTA

Blackwell Creates First-Ever FP4 Generative AI Image

FP4-quantized model produces the “4-bit Bunny”

Model using FP16

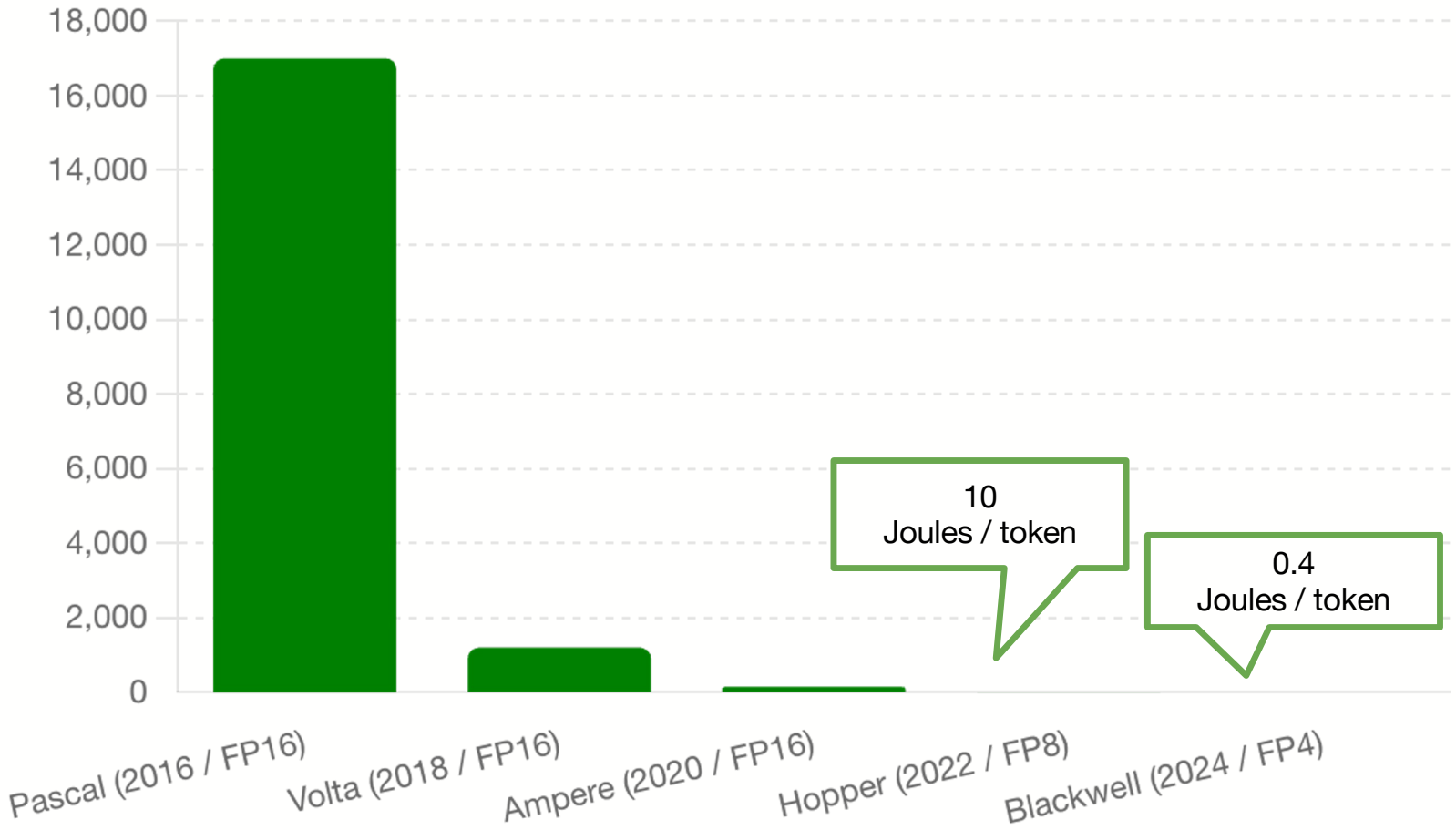
Model using FP4

Prompt: Close up photo of a rabbit, forest in spring, haze, halation, bloom, dramatic atmosphere, centered, rule of thirds, 200mm 1.4f macro shot

Improved Energy Consumption per Token - Nvidia

Current LLMs average
3 tokens / word
4 characters / token

From Pascal to Blackwell, the energy consumption for token generation has dropped by 45,000x

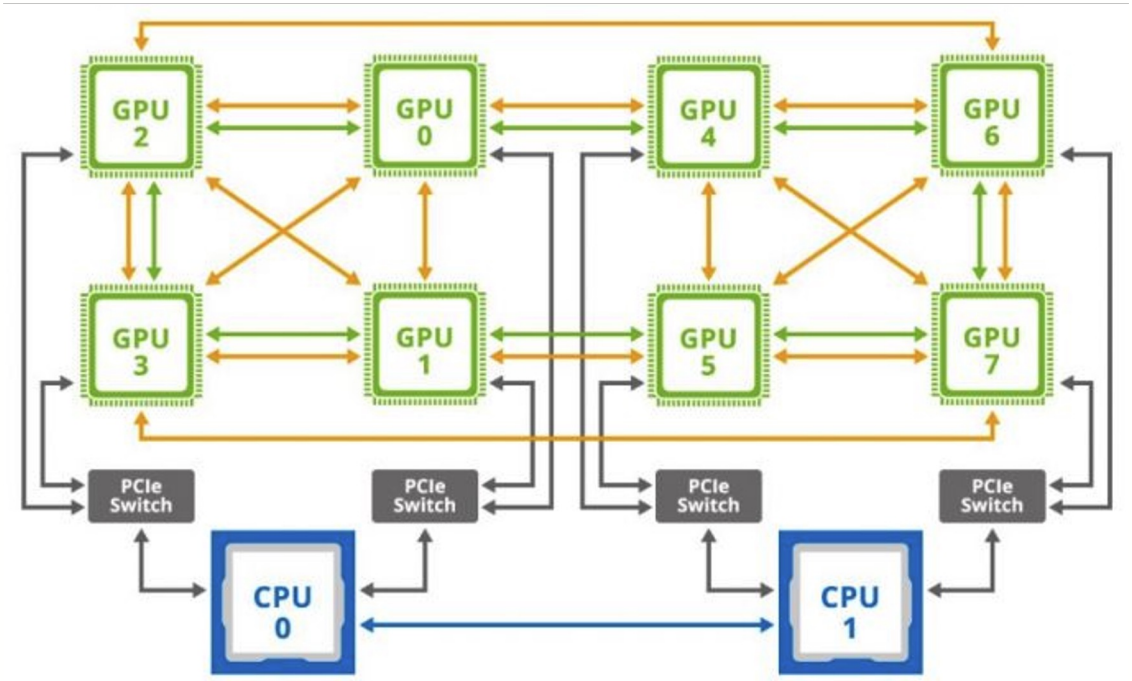


ARISTA

What do these servers look like - NVIDIA example

DGX = an NVIDIA built system
HGX = an OEM built system
(SuperMicro, Dell, HPE)

GPU collective = NCCL

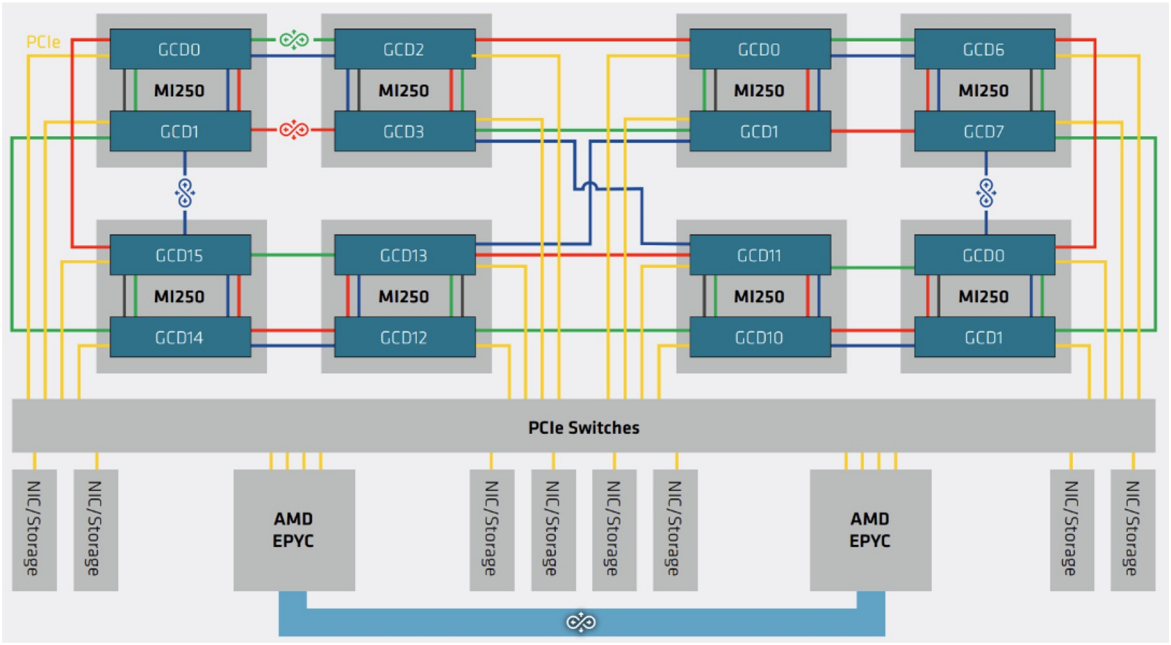


Internal CPU <-> GPU and GPU <-> GPU
is a mix of PCIe and NVLink

What do these servers look like - AMD example

SuperMicro, Dell, HPE

GPU collective = RCCL



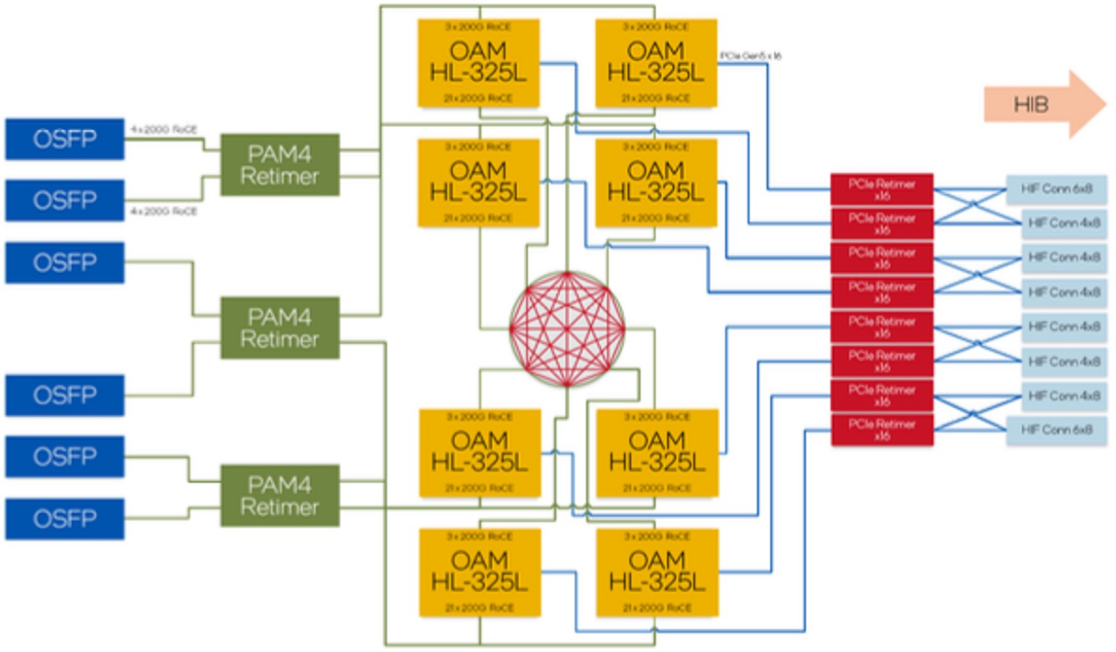
Internal CPU <-> GPU and GPU <-> GPU
is a mix of PCIe and Infinty

ARISTA

What do these servers look like - Intel example

SuperMicro, Dell, HPE

GPU collective = HCCL



Internal CPU <-> GPU and GPU <-> GPU
is a mix of PCIe and Alternative Pathways

ARISTA

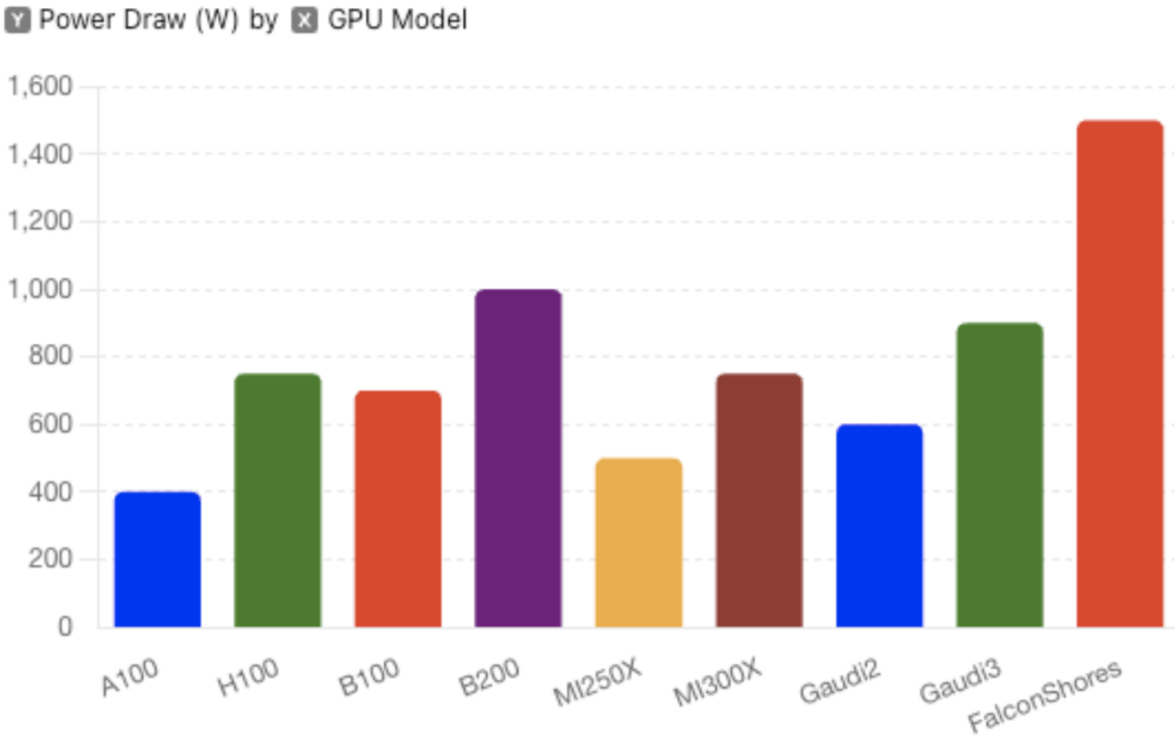
Power Requirements

Rough Calculations:

8 x 750W GPU + 500W server
= 6,500W Host

Normal DC Rack = 6-10kW
GPU Racks = 16-20kW

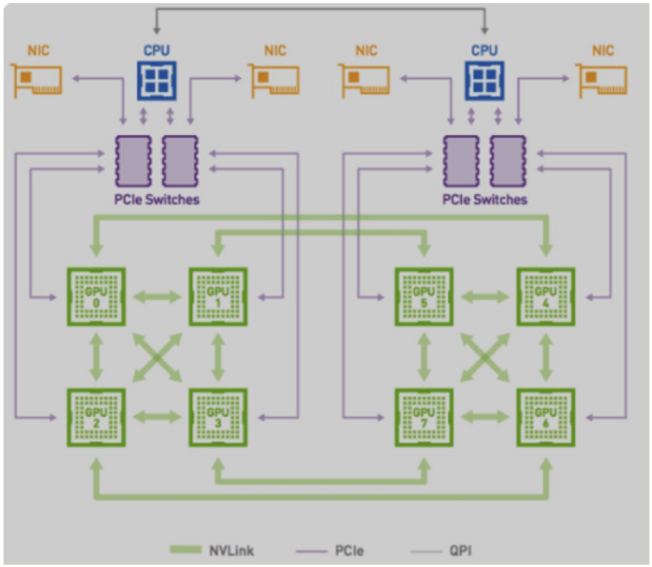
Average DC capacity ~10MW
15k GPUs =18MW
100k GPUs = 150MW



ARISTA

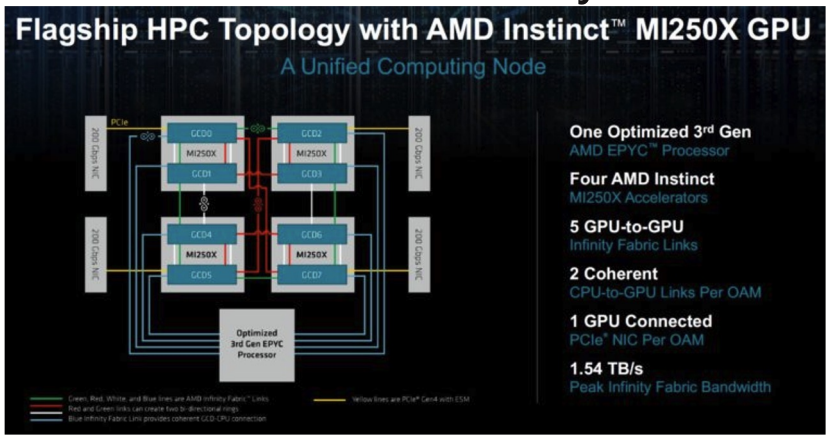
Quick note of “networks” involved GPU-to-GPU inside the host

NVIDIA = NVLink

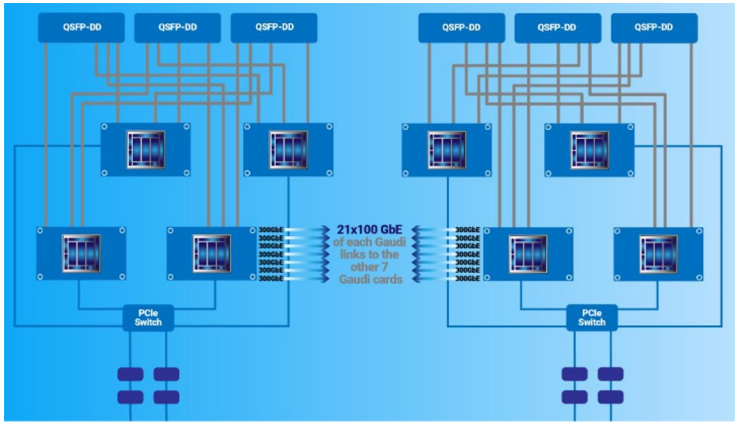


https://www.hpcwire.com/2024/05/30/everyone-except-nvidia-forms-ultra-accelerator-link-ualink-consortium/

AMD = Infinity



Intel



ARISTA

Options for xPU Connectivity

xPU is the collective name for:

- **GPU:** optimized for parallel processing, making them ideal for training large AI models with uniform data. (NVidia)
- **TPU:** specialized for TensorFlow-based AI workloads (developed by Google for neural networks), offering extreme efficiency for specific tasks.
- **DPU:** Deep learning (or Data) Processing Unit, optimised for data and often embedded in SmartNICs (NVIDIA Bluefield2 /AMD (Pensando) / Marvell)

There are three methods to connect xPU's:

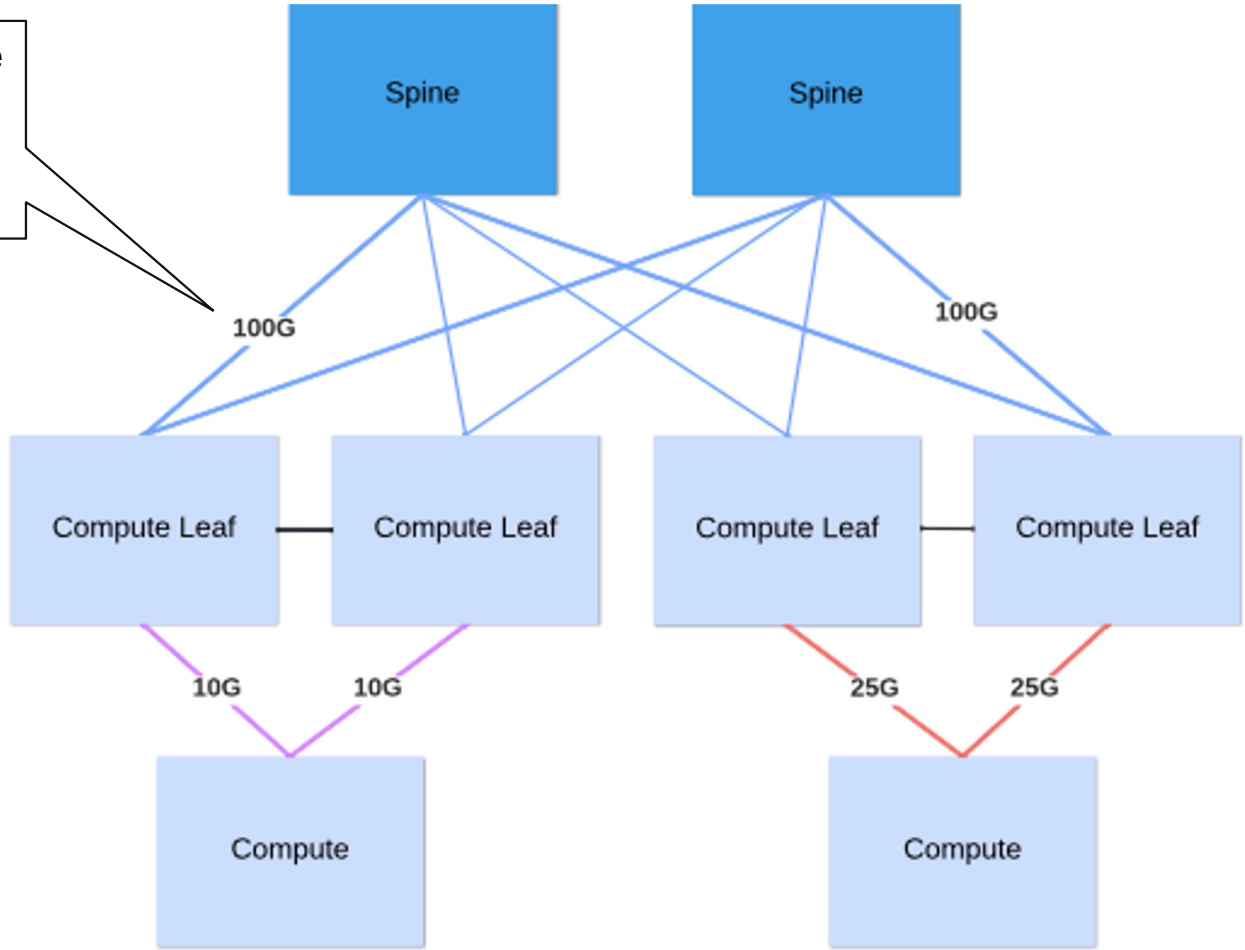
- ➔ **PCI Bus:** A standard server can usually support 4-8 GPUs across the PCI bus.
- ➔ **xPU to xPU Interconnect:** Nvidia created NVLink, a GPU-to-GPU connection that can currently transfer data at 1.8 terabytes per second between GPUs. There is also an NVLink rack-level Switch capable of supporting up to 576 fully connected GPUs in a non-blocking compute fabric (single computational domain). There is currently work in progress to develop Ultra Accelerator Link (UALink), establishing an open industry standard that will enable AI accelerators to communicate more effectively.
- ➔ **Server-to-Server Interconnect:** Ethernet or InfiniBand (Nvidia) can connect servers that contain xPUs. This connection level is often called scale-out, where faster multi-GPU domains are connected by Ethernet/IP or IB fabric networks to form large computational networks. Recently, the Ultra Ethernet Consortium (UEC) was formed to deliver everyone else's "InfiniBand."

Frameworks for Managing GPU's

- NCCL - NVIDIA Collective Communication Library
- RCCL - AMD ROCm Collective Communication Library
- PyTorch - Opensource
- Tensorflow - Opensource developed by Google
- Keras - Opensource
- CNTK - Microsoft Cognitive Toolkit
- SageMaker - AWS

Surprise! I've got 200GbE / 400GbE xPU NICs, where should I plug them in?

The average Enterprise network focused on supporting CPUs today



ARISTA

Networking for AI - What comprises an AI Infrastructure?

Front End network

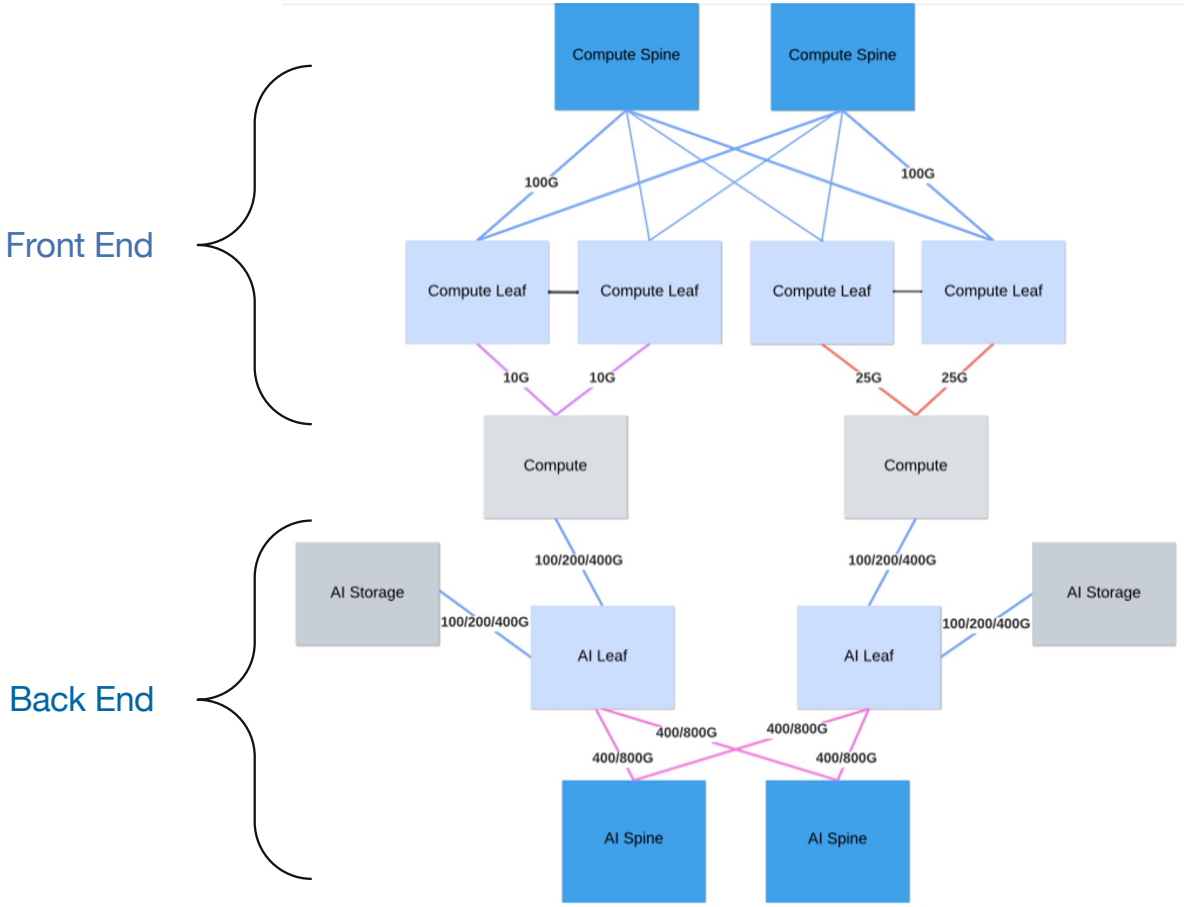
Back End network

Compute network

Storage network

In-band Management network

Out-of-Band Management network



ARISTA

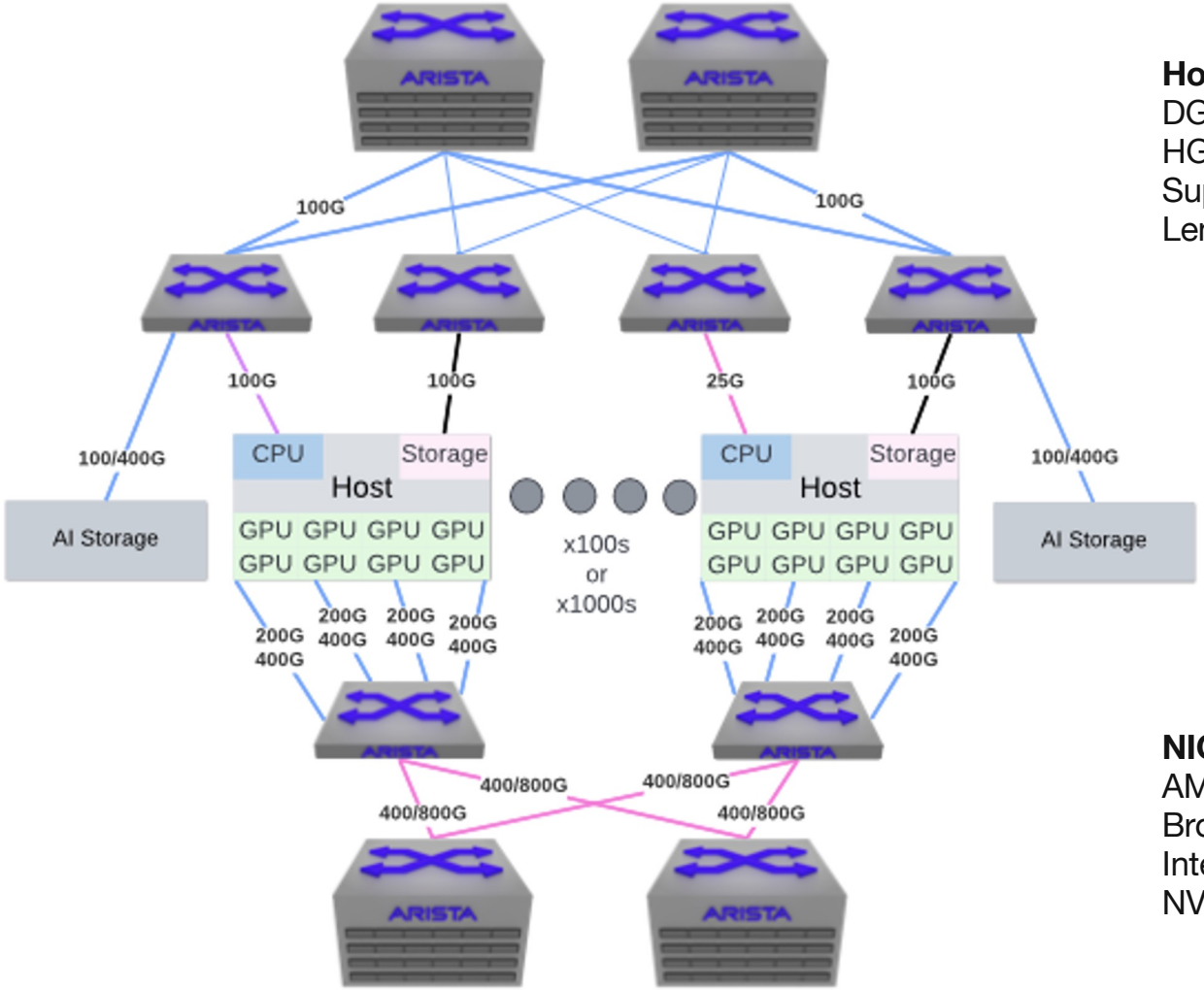
Networking for AI - What do these clusters look like?

Front End

CPU

Back End

GPU



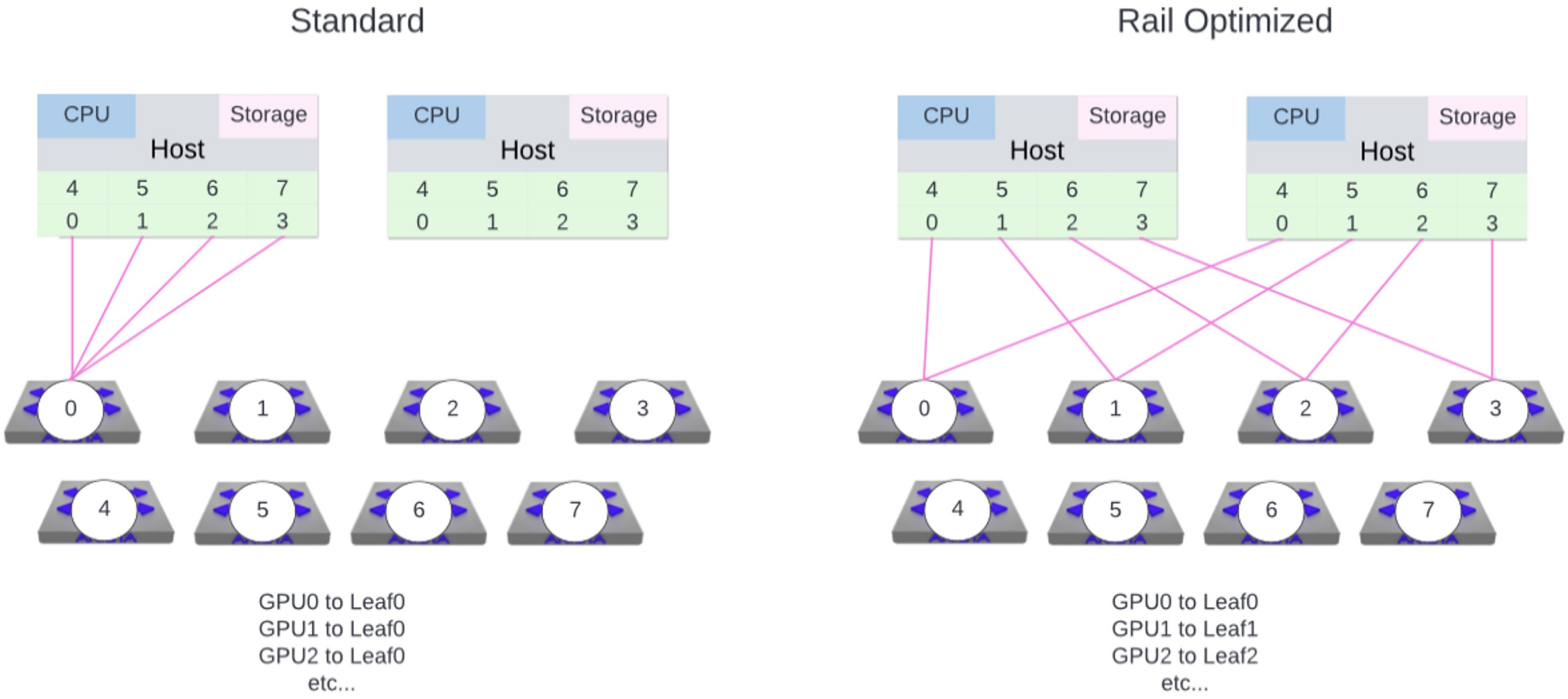
Hosts
DGX = NVIDIA
HGX = 3rd party + NVIDIA GPU
SuperMicro
Lenovo

Storage
WEKA
Vast
Intel
DDN

NICs
AMD
Broadcom
Intel
NVIDIA

Networking for AI - GPU Connectivity Options

A rail is comprised of GPUs that have the same rank within different hosts/clusters and are connected to the same network switch.



ARISTA

Networking for AI - GPU Connectivity Options

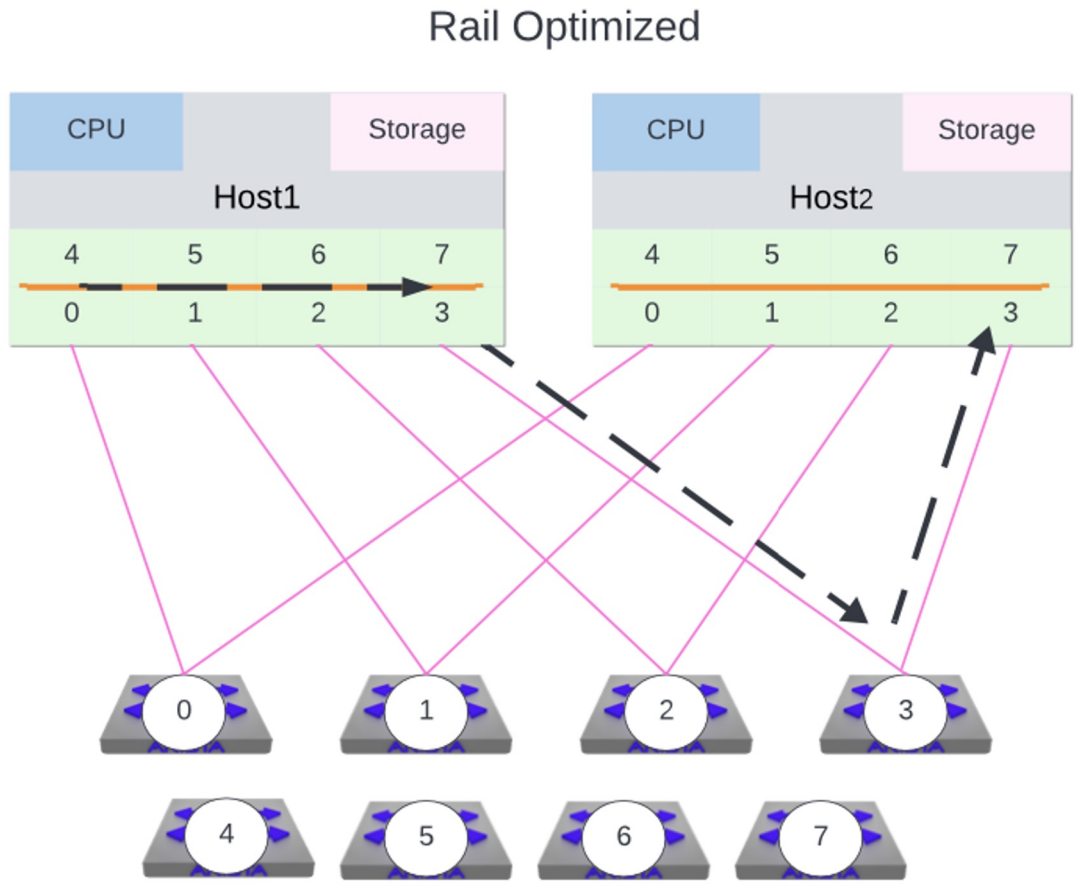
A rail is comprised of GPUs that have the same rank within different hosts/clusters and are connected to the same network switch.

GPU0 in Host1 needs to communicate with GPU3 in Host2:

GPU0 will leverage the NVLink pathway internal to Host1 to reach Host1's GPU3.

Host1's GPU3 will then send that traffic to Leaf3.

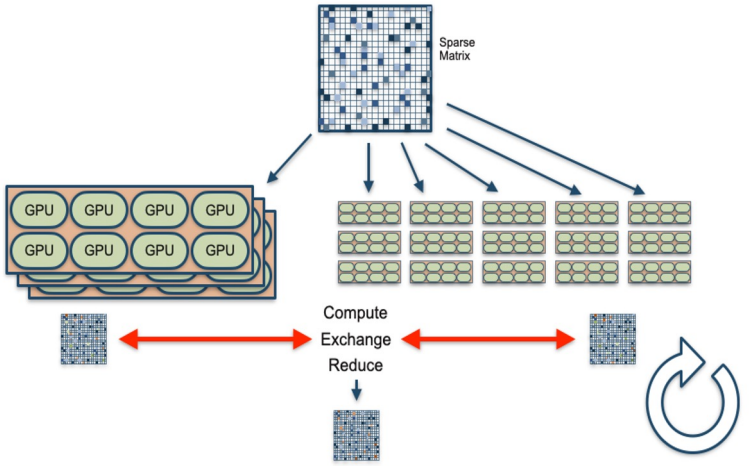
Leaf3 sends the traffic down to Host2's GPU3.



ARISTA

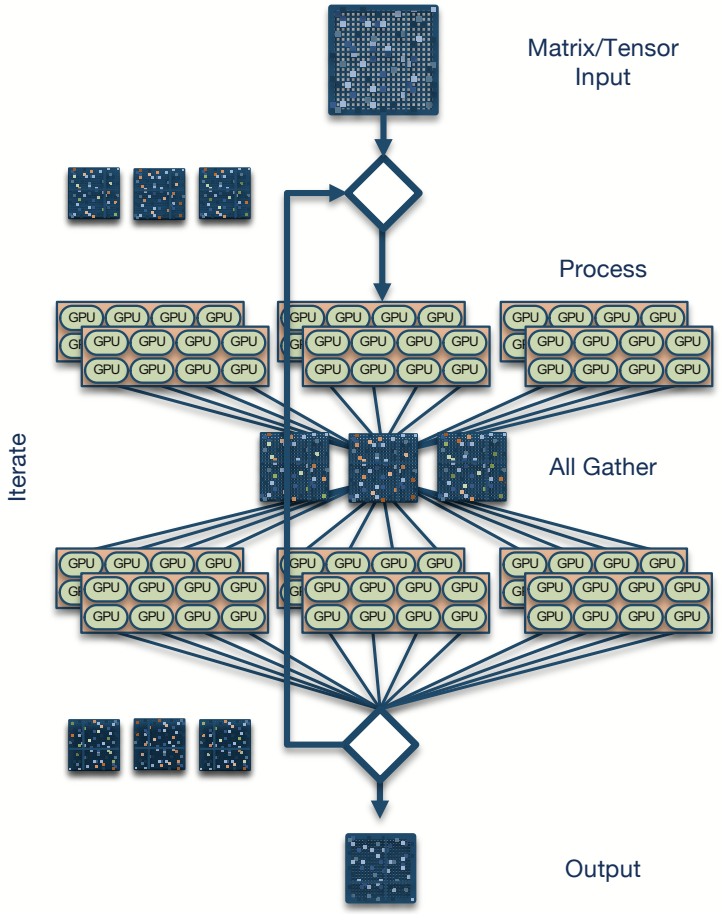
AI Workloads

- AI workloads are extremely computational intensive and exceed the memory available on a single GPU therefore the workload is split across a cluster of GPU's
- These GPU's perform execute algorithms and calculations and then share results amongst each other. There are several ways of doing this commonly used term is "Collective Operation" they include Broadcast, Reduce, Ring and All-Reduce
- All GPU's have to wait until all GPU's have shared their messages and passed on results, the workload is stalled until this happens, commonly used term is "Barrier"
- GPU's within a host have their own internal fabric



What's different about AI workloads?

- Most AI workloads use collective communication for parallel computation
- Use RDMA (kernel bypass) for better performance
- Traffic characteristics
  - Tight time synchronization between traffic flows
  - Small number of large sized flows ~2 flows per NIC
  - Very little entropy
  - Short periodic bursts of network activity followed by compute
- Highly susceptible to collisions
- Job Completion Time (JCT) - key metric measures time between:
  - Launching a job
  - Processing the data across the cluster of GPU/Accelerator resources
  - Posting the results
- Time Spent in Networking (TSN)
- Slowest member determines performance

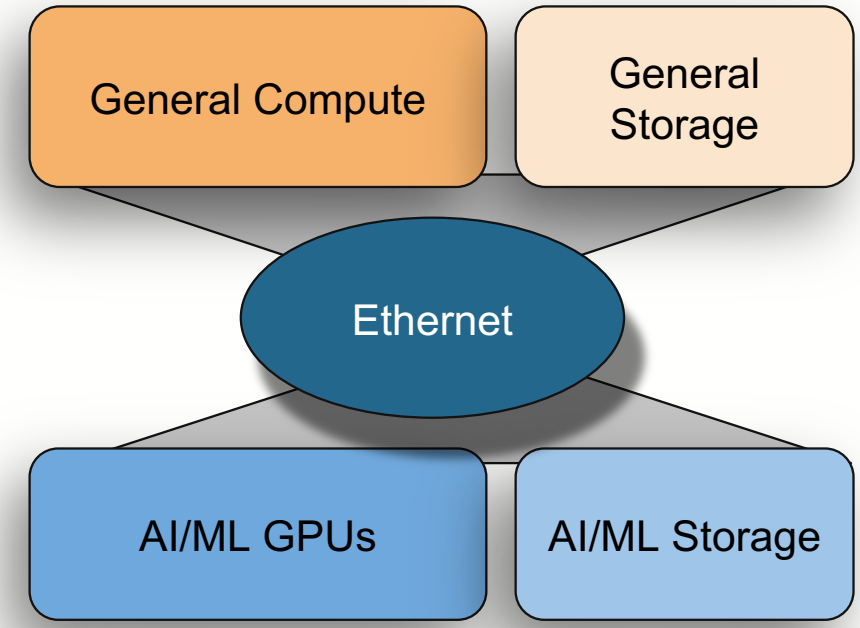


AI Training Workloads Are Highly Co-ordinated - Sensitive to Delayed Jobs

ARISTA

Requirement for AI Networks

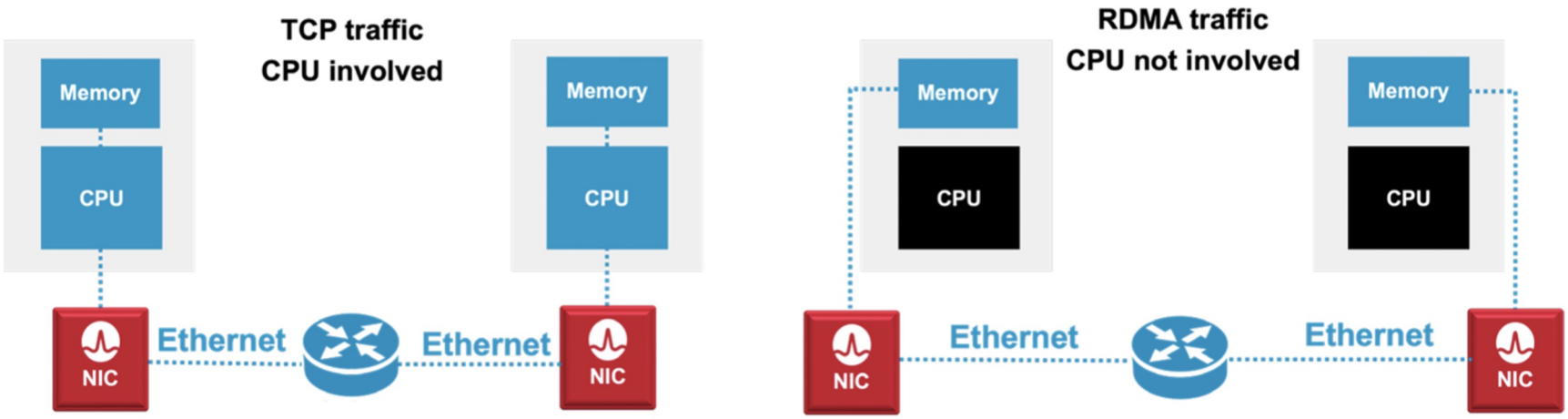
- Scale Out Capacity
- A fast, lossless network with zero oversubscription and often over provisioned.
  - for many forms of communication
    - ALL-REDUCE, BROADCAST, ALL-TO-ALL, RING
  - Graceful handling of large synchronized flows
- Fast transfer from host to network (RDMA over Converged Ethernet)
- Visibility and telemetry
  - to identify bottlenecks in the network or application
- Reliable and Resilient
- Open and Interoperable



AI Workloads require a dedicated high performing lossless network

RDMA

Remote Direct Memory Access (RDMA) is the remote memory management capability that allows server-to-server data movement directly between application memory without any CPU involvement.



ARISTA

RDMA Over Converged Ethernet (RoCEv2)

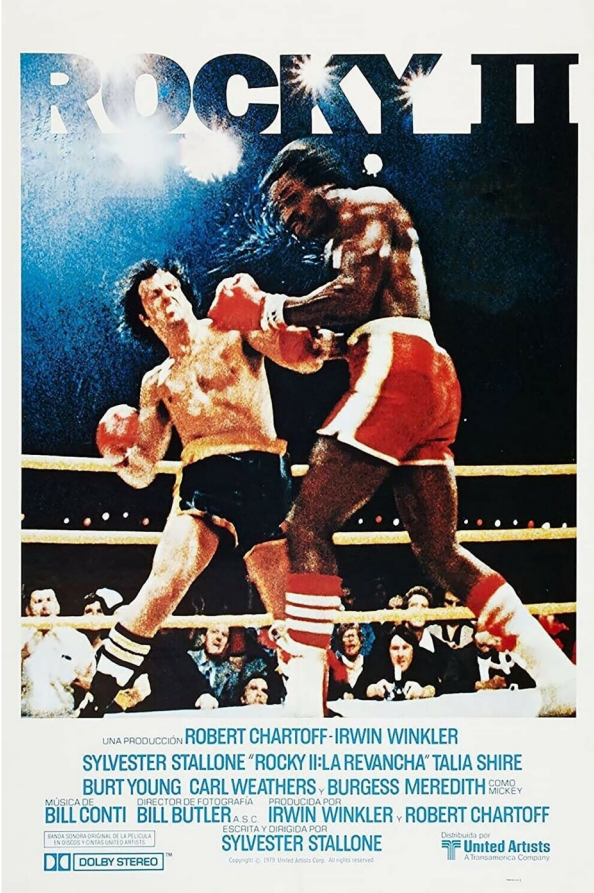
RoCE (RDMA over Converged Ethernet) is a network protocol that allows RDMA over an Ethernet network.

RoCEv2 extends the functionality and scalability of ROCE, adding a UDP/IP header

Broadcom/Intel/Nvidia Ethernet Adapters support RoCEv2 in hardware and allows for higher throughput, lower latency, and lower CPU utilization, which are critical for AI/ML, Storage, and High-Performance Compute (HPC) applications.

RoCEv2 provides the following advantages:

- Operation on routed ethernet networks, ubiquitous in large scale-out leaf-spine data centers
- IP QoS – The DiffServ code point (DSCP)
- IP congestion control – Explicit congestion notification (ECN)
- UDP header provides more entropy for better ECMP



ARISTA

Infiniband versus Ethernet

Proprietary NVIDIA

Current products top out at 400GbE

Centralized Control Plane

“Subnet Manager” controls entire network

Single hop

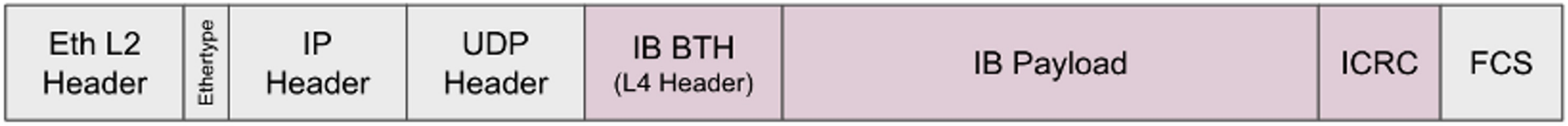
Industry Standard

800GbE and beyond

Distributed Control Plane

Routing protocols make local decisions

ROCEv2 brought Layer 3/IP support (multihop / scale-out)



RoCEv2 frame: InfiniBand over Ethernet

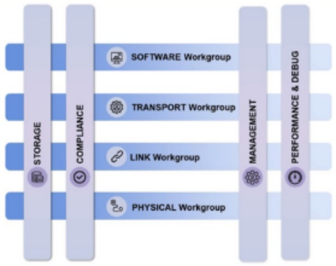
Ultra Ethernet Consortium (UEC)

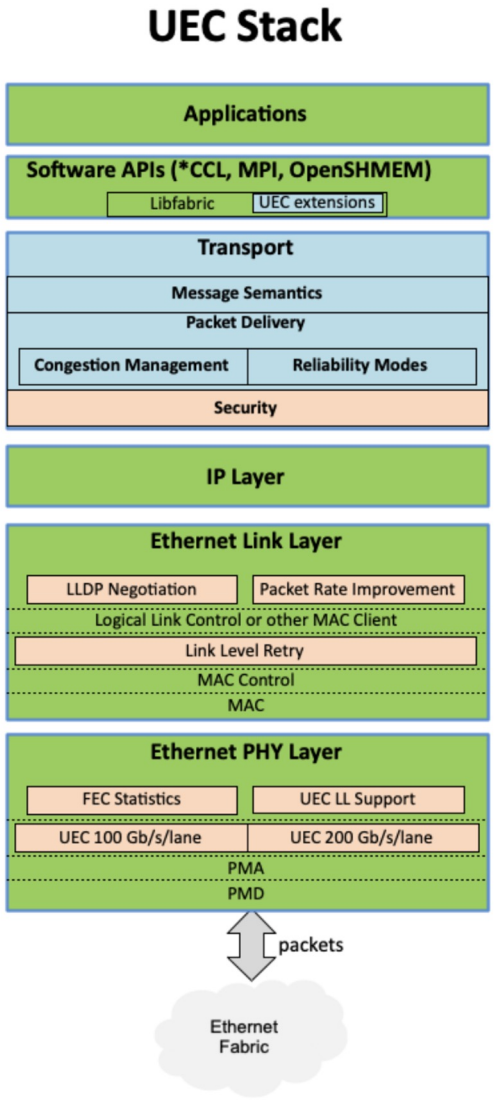
- Formed in the 2023, the UEC aims to develop a new standard for interconnection for AI and HPC data center requirements
  - 55+ members
- Next generation AI/HPC transport
  - Multi-Pathing and Packet Spraying
  - Flexible Ordering
  - HPC/AI optimized congestion control
  - End-2-End telemetry
  - Security

Steering Members

AMD ARISTA BROADCOM CISCO EVIDEN an atos business

Hewlett Packard Enterprise intel Meta Microsoft ORACLE





Ethernet Powers the Future for Generative AI Clusters

Ethernet Advantages over Infiniband for GenAI Networking Fabrics

Most Consistent Architecture

- Ethernet is native to all other systems in a data center ecosystem
  - Data warehouses
  - OLTP systems
  - Systems mgmt.
  - All others
- Infiniband stands alone on a unique island

Highest Performance for Lossless AI

- Highest performance for lossless, low latency AI job completion time
- Highest capacity for massive-scale GenAI training clusters, using 800G capable switches with clear path to 1.6T

Open Innovation for the Future

- Ethernet offers a rich, open ecosystem with multiple vendors
- Infiniband forces lock-in
- Ultra Ethernet Consortium (UEC) is advancing state of the art for GenAI flow control, scale & security

Ultra Ethernet Consortium

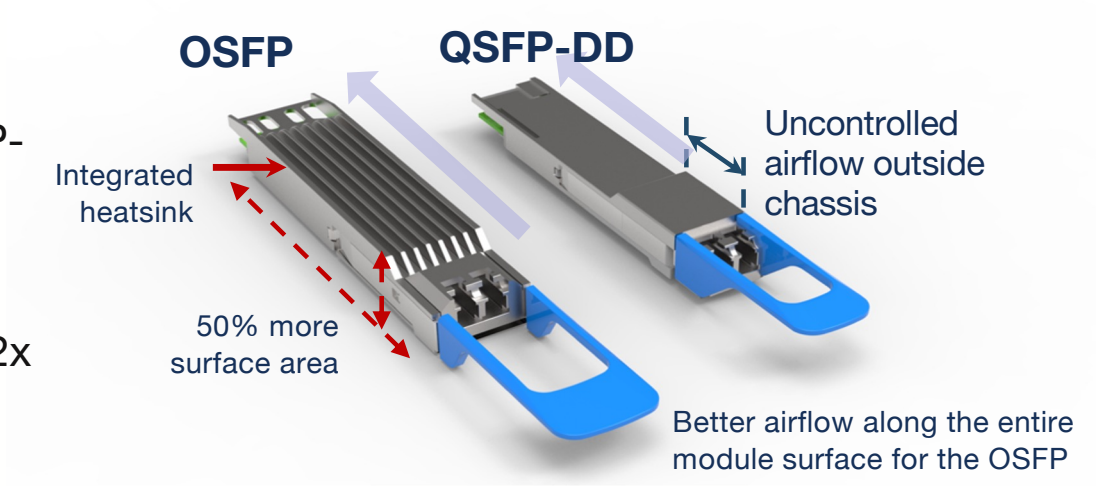
ARISTA

What's different about AI Connectivity?

- Higher speeds to the NIC
  - 200G and 400G connectivity to the NIC is common
  - Serdes speeds at 50G PAM-4 or 100G PAM-4
- More frequent use of AECs, AOCs and Optics for NIC Connectivity
  - Copper DACs still the cheapest option, but limited reach at higher speeds
  - Connecting NICs across multiple racks may require AOCs or optics
  - Conversion between 50G PAM-4 / lane and 100G PAM-4 / lane may be required
- Connectivity between switch vendors
  - Understand the NIC port Ethernet speed, electrical lane speed, and physical form-factor
  - Multiple options for the same Ethernet speed!!
  - "400G" port could mean:
    1. QSFP-DD or OSFP port using 8x 50G electrical lanes
    2. OSFP Riding Heat Sink (RHS) port using 4x 100G electrical lanes
    3. QSFP-DD port using 4x 100G electrical lanes
    4. QSFP112 port using 4x 100G electrical lanes

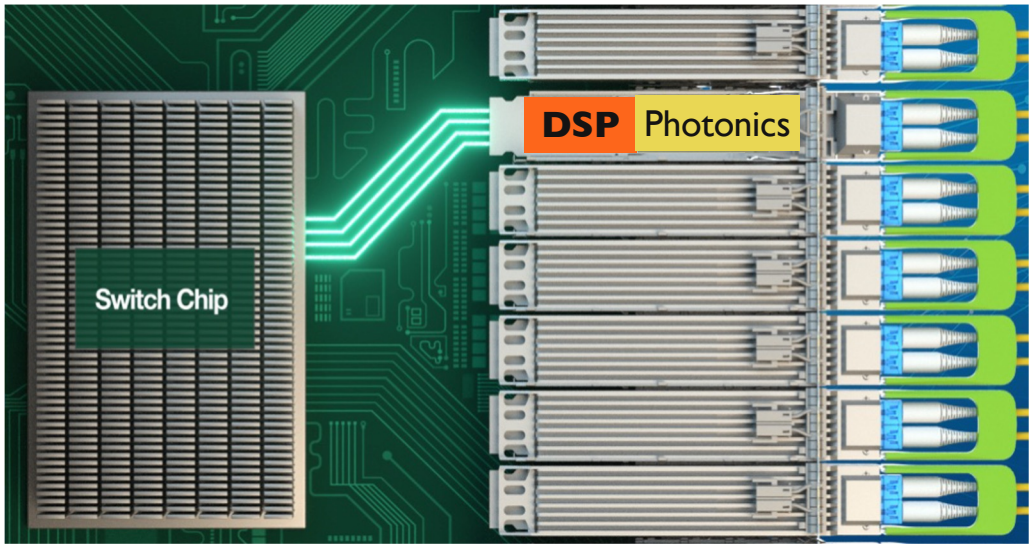
OSFP Versus QSFP-DD 400G - Thermal Performance

- Integrated heatsink directly attached to temp sensitive components
- ~50% Greater surface area and volume
- Better airflow across entire surface of the module
- OSFPs operate ~10 to 15C cooler than QSFP-DDs for equivalent platforms
- Lower operating temperatures - dramatic increase in optics reliability
- +10C temp increases optics failure rate by ~2x
- Easier to cool - Lower system fan speeds
- ~10% - 25% less overall system power
- OSFP has clearer roadmap to 800G, 1.6T (OSFP-XD) and beyond
- Further efficiencies with upcoming Linear-Drive Pluggable Optics (LPO)

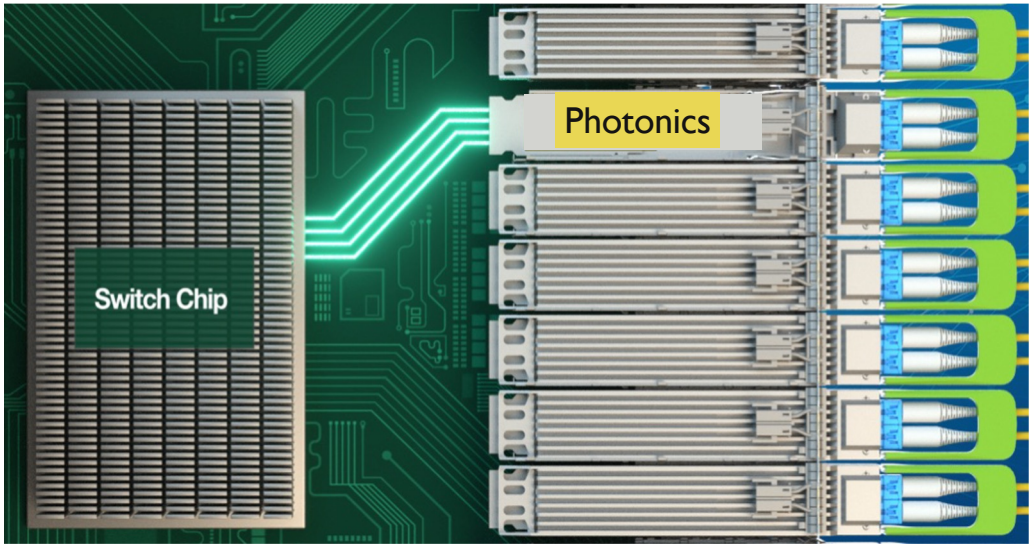


Linear-Drive Pluggable Optics (LPO)

Traditional pluggable optical modules



LPO Modules



- LPO means no DSP in the transceiver module
- How is this possible?
  - Broadcom Tomahawk5 and Jericho3 Switch Silicon serdes use advanced DSP technology
  - Requires expert system design and careful serdes tuning by networking vendors

LPO-MSA

Acceliink · AMD · ARISTA · BROADCOM · CISCO · eoptolink · Hisense Broadband · INNO LIGHT · intel · MACOM · NVIDIA · SEMTECH

ARISTA

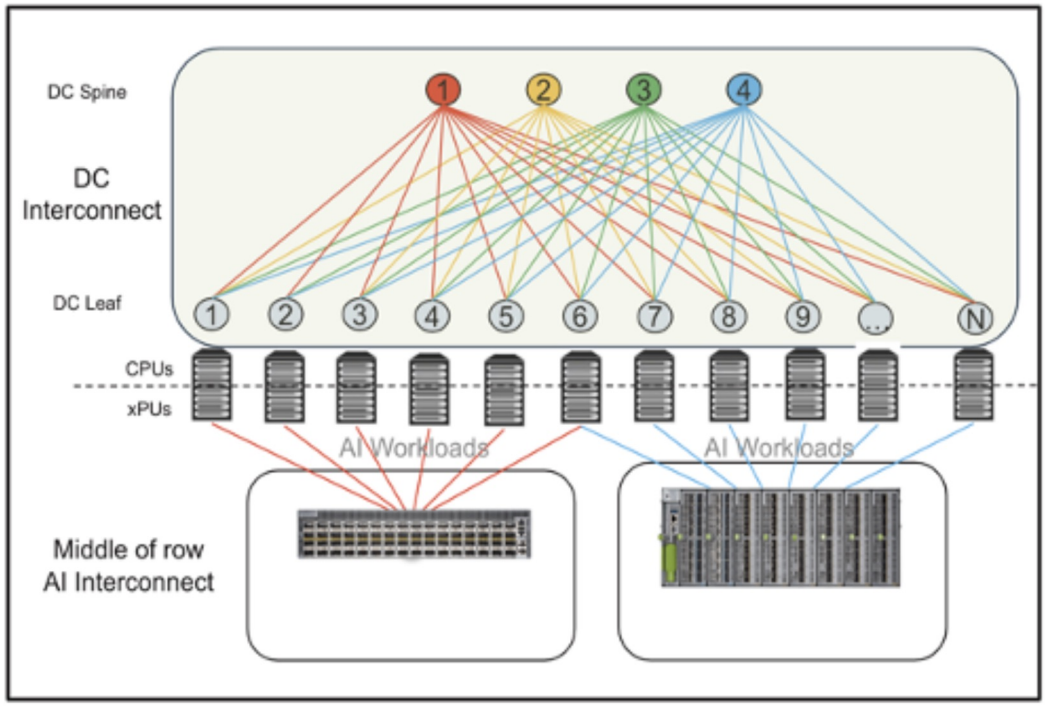
“Traditional” DSP Optics & LPO

Feature	“Traditional” DSP Optics	Linear Pluggable Optics (LPO)
Power	~14W - 16W for 800G modules	~7W for 800G modules
Latency	~100ns delay through DSP	~1ns delay through module
Cost	$X	~$0.6X
Technology Maturity	Mature, in high volume	New. Volume in late 2024 / Early 2025
Ecosystem	Multi-vendor, plug and play, standards based	LPO MSA just announced LPO Modules will require tighter integration with the system Restricted to TH5/J3 platforms

LPO offers significant power & cost advantages and improved MTBF
Most 800G optics sold in 2024 will be DSP based optics.

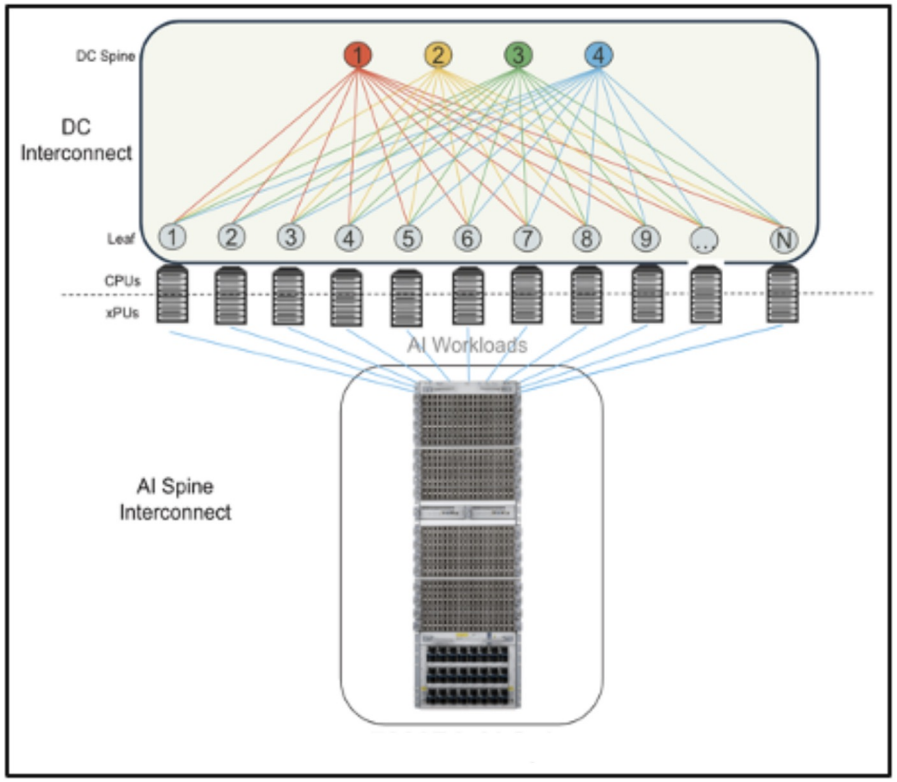
Network Design for “Small” AI Applications

- Scale - 64 x 400G or 128 x 200G xPU NICs
- Broadcom Tomahawk4
  - 64 x 400G QSFP-DD or OSFP
  - 25.6Tbps
- No Flow Collisions - Single ASIC
- Handling incasts at the receiver:
  - ECN (Explicit Congestion Notification)
  - PFC (Priority Flow Control)
  - Standard buffer requires moderate tuning



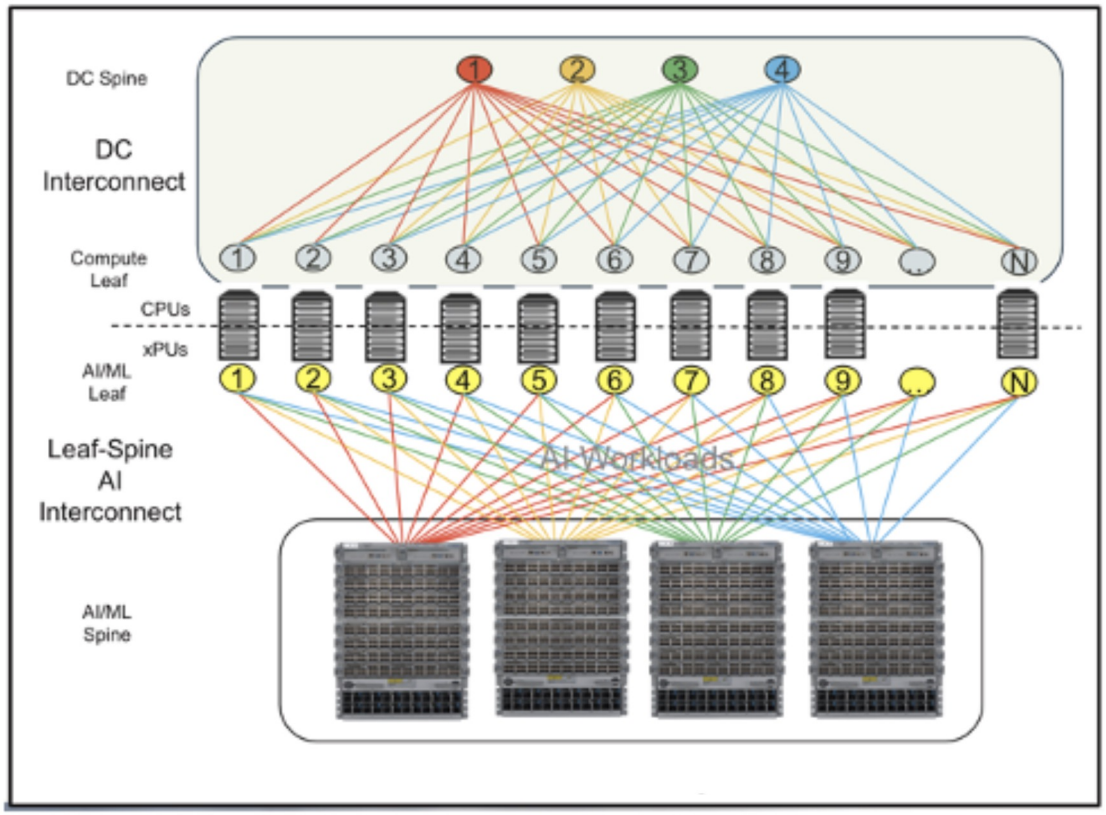
Network Design for “Moderate” AI Applications

- Scale - 576 x 400G xPU NICs
- Broadcom Jericho2/2C+
  - 576 x 400G QSFP-DD or OSFP
  - 230 Tbps
- AI Leaf and Spine in single chassis
- No Flow Collisions between line cards and fabric - overprovisioned.
- High availability:
  - Redundant Fabric, Supervisor, Fan, PSU
- Handling incasts at the receiver
  - ECN and/or PFC
  - Deep buffer requires minimal tuning



Network Design for “Large” AI Applications

- Scale - 18,432 x 400G xPU NICs
- AI Spine - Broadcom Jericho2/2C+
  - 576 x 400G QSFP-DD or OSFP
  - 230 Tbps
- AI Leaf - Broadcom Tomahawk4
  - 64 x 400G QSFP-DD or OSFP
  - 25.6Tbps
- Flow Collisions between AI Leaf and Spine
  - Requires Dynamic Load Balancing (DLB), Source LB
- High availability AI Fabric:
  - Requires ~1:1.2 over-provisioning on AI Leaf
- Handling incasts at the receiver
  - ECN and/or PFC
  - AI Leaf standard buffer requires extensive tuning



Thank You

ARISTA

AI ready networks

https://www.arista.com/en/solutions/ai-networking