AI Networking

Introduction

Artificial Intelligence (AI) has emerged as a revolutionary technology that is transforming various industries and aspects of our daily lives. The rapid arrival of real-time gaming, virtual reality, generative AI and metaverse applications are changing the way network, compute, memory, storage and interconnect I/O interact for the next decade. As AI continues to advance at an unprecedented pace, the network needs to adapt to the humongous growth in traffic connecting hundreds of processors with trillions of transactions and gigabits of throughput. As AI moves out of labs and research projects toward wide adoption, it will demand significant computing resources. Recent developments are merely building blocks for things to come over the next decade. We see AI clusters growing substantially over the coming years.

A common characteristic of these AI workloads is that they are both data and compute-intensive. A typical AI workload involves a large sparse matrix computation distributed across hundreds or thousands of processors – CPUs, GPUs or TPUs. These processors compute intensely and then exchange data with their peers. Data from the peers is reduced or merged with the local data and then another cycle of processing begins. In this compute-exchange-reduce cycle, any slowdown can detrimentally impact the job completion time.

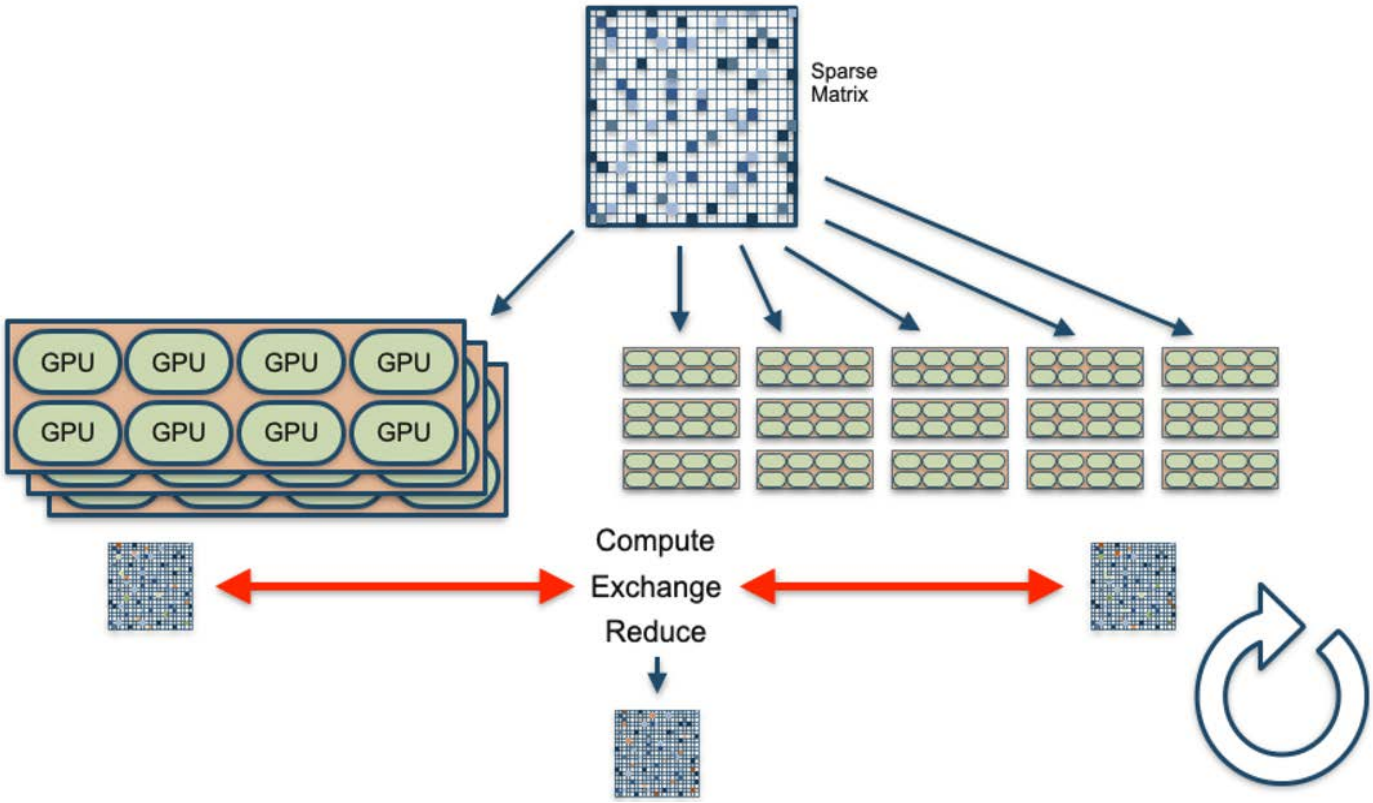

Figure1 - Compute-Exchange-Reduce Cycle

TCP/IP and RDMA

In TCP/IP sockets, data has to be copied from the user space to the kernel space before reaching the network driver and then the network. When working with large volumes of data associated with AI applications, the CPU can become the bottleneck.

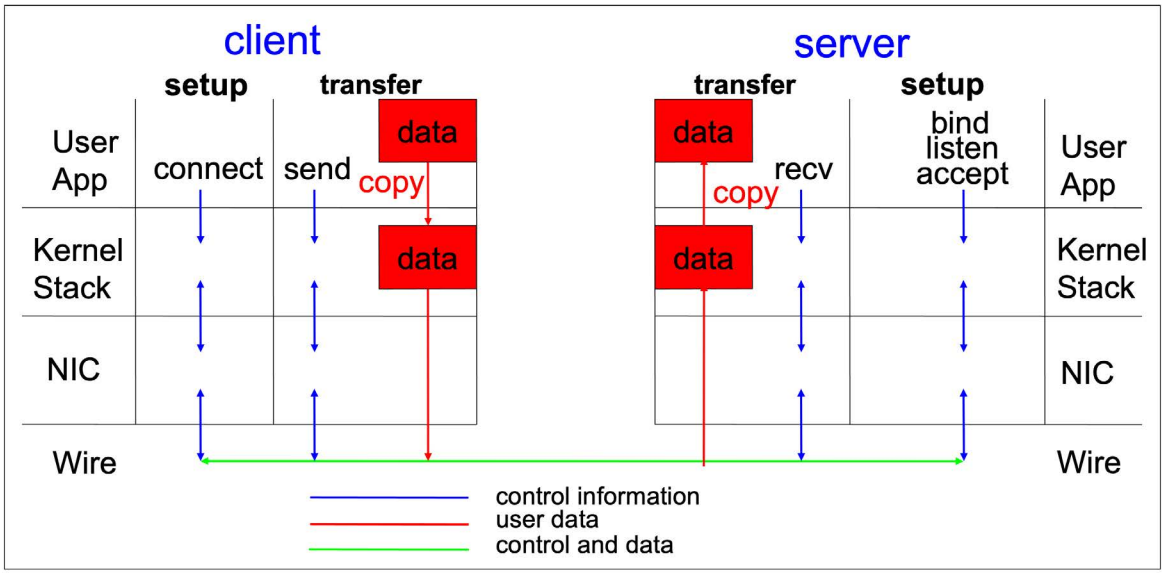

Figure 2: TCP/IP Transfer

This is where Remote Direct Memory Access (RDMA) comes in. RDMA is ubiquitous in high performance computing systems as it enables the exchange of data in main memory without relying on the kernel. RDMA helps improve throughput and performance resulting in faster data transfer rates and lower latency between RDMA enabled systems as it lowers the number of CPU cycles involved.

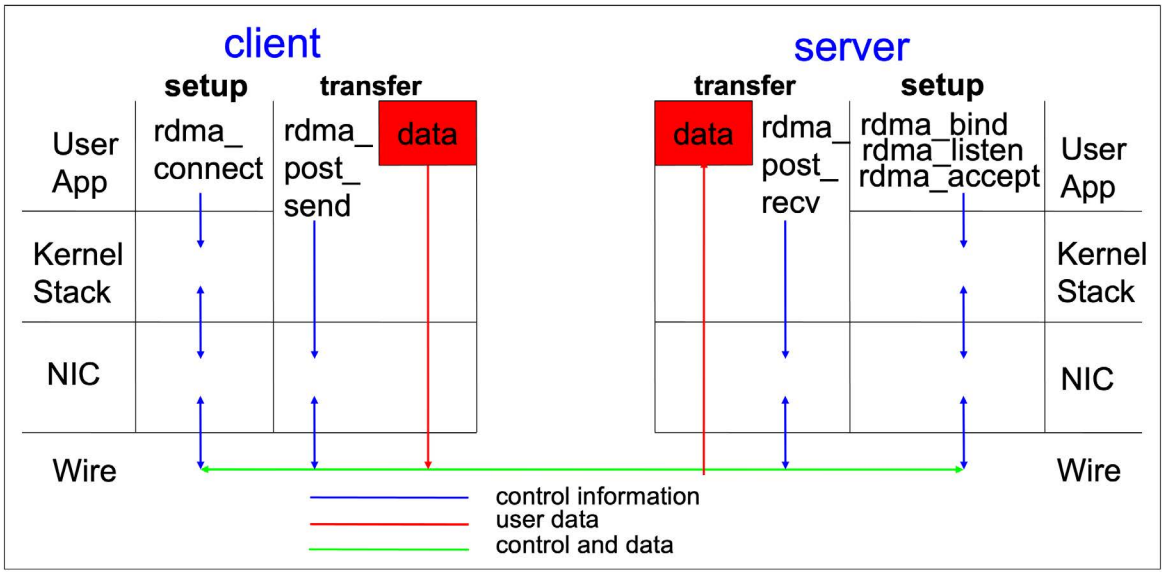

Figure 3: RDMA Transfer

The semantics of RDMA transfers are defined by the InfiniBand Verbs software interface. This includes registration of memory blocks, the exchange of descriptors, and the posting of RDMA read and write operations. This interface is independent of Infiniband as a physical transport layer.

RoCE defines how to transport an InfiniBand payload over an Ethernet network. RoCEv2 extends this scalability and functionality further by allowing traffic to be routed and enables scaling RDMA over Ethernet.

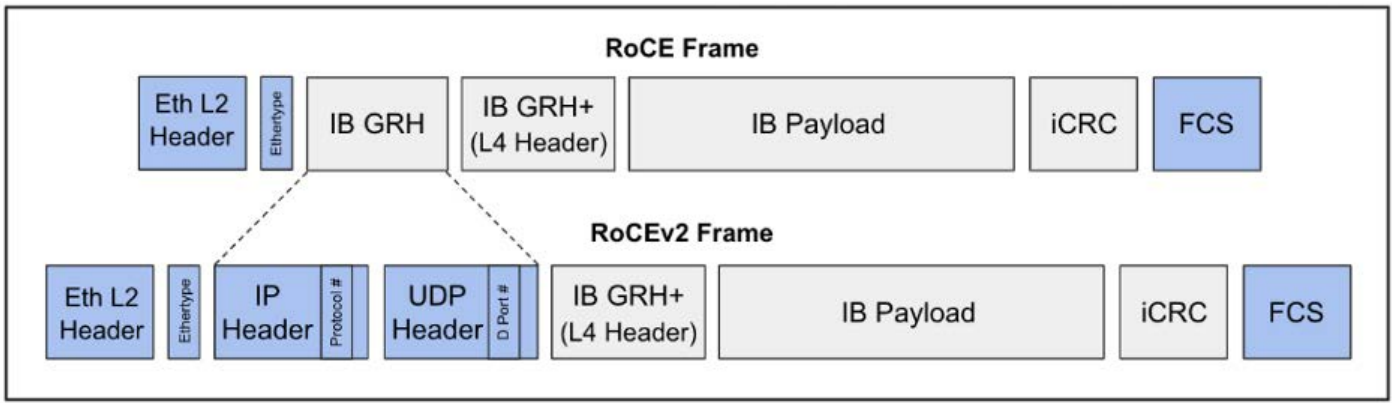

Figure 4: RoCE and RoCEv2 Frame Format

Interconnect for AI Networks

Ethernet is ubiquitously deployed in Data Centers, Backbone, Edge, Campus Networks with varying use cases from very low speeds to high speeds of 100G, 200G, 400G, 800G today and 1.6T in roadmap. Infiniband, on the other hand, is a network technology commonly used in HPC clusters. As mentioned earlier, AI/ML workloads are network intensive and are different from those of traditional HPC.

Further, with the explosion of Large language Models (LLMs) there is a constant demand for more GPUs and storage capacity. Modern AI applications require large clusters with thousands of GPUs & storage devices and these clusters have to scale to tens of thousands of devices as the demand grows. With GPU speeds doubling every other year, it's critical to avoid both compute and network bottlenecks through scalable network design. While the application teams focus on compute capacity, the network teams have to carefully evaluate the interconnect based on several factors:

Performance

One of the key metrics to measure performance of an AI Cluster is job completion time. To achieve the ideal performance, the network has to be lossless, non blocking and deliver linerate link utilization. As discussed later, with proper congestion control mechanisms and efficient load balancing techniques, RoCE can deliver the best performance required for AI workloads.

Bandwidth and Speeds

As the training jobs are becoming larger, it is important to deliver a faster network. This can be done more efficiently with higher density switches with faster port speeds. With merchant silicon ethernet solutions, the bandwidth of the network can be doubled every 2 years while reducing the cost / bit and power / bit.

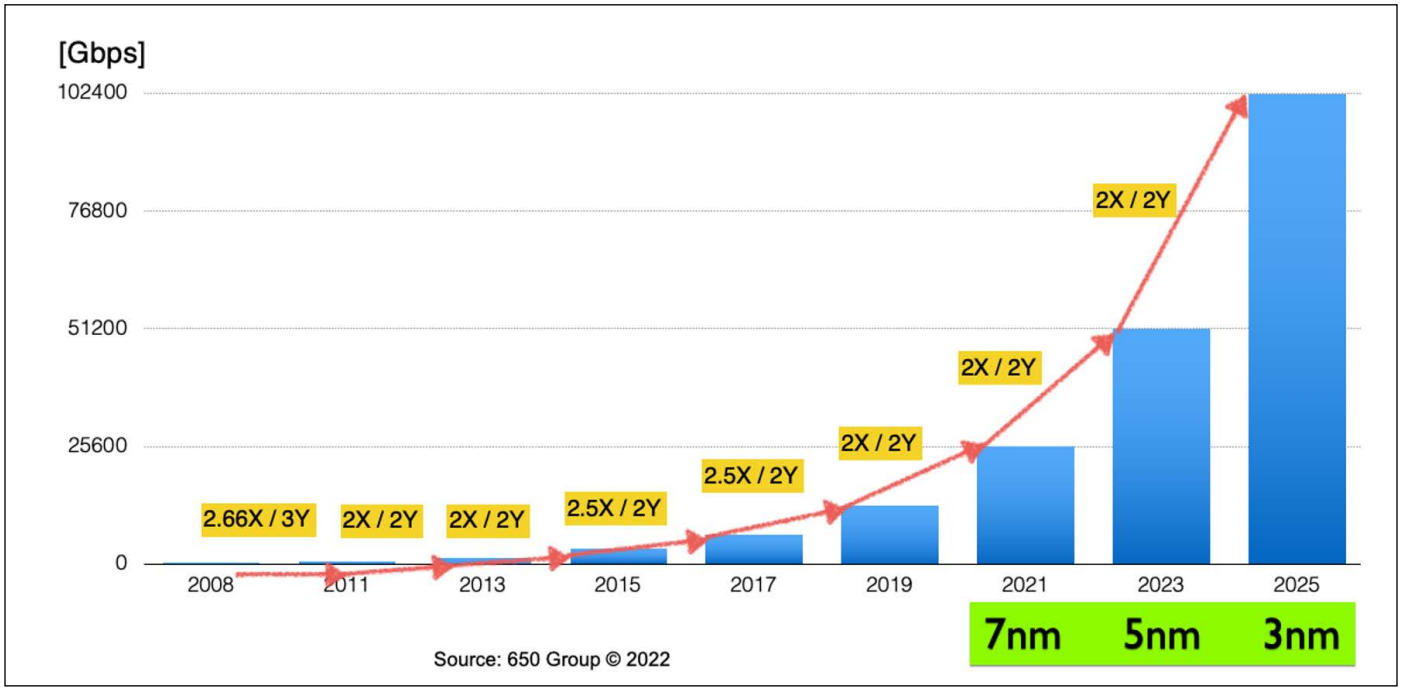

Figure 5: Single Chip Ethernet Switch Silicon Through 2025

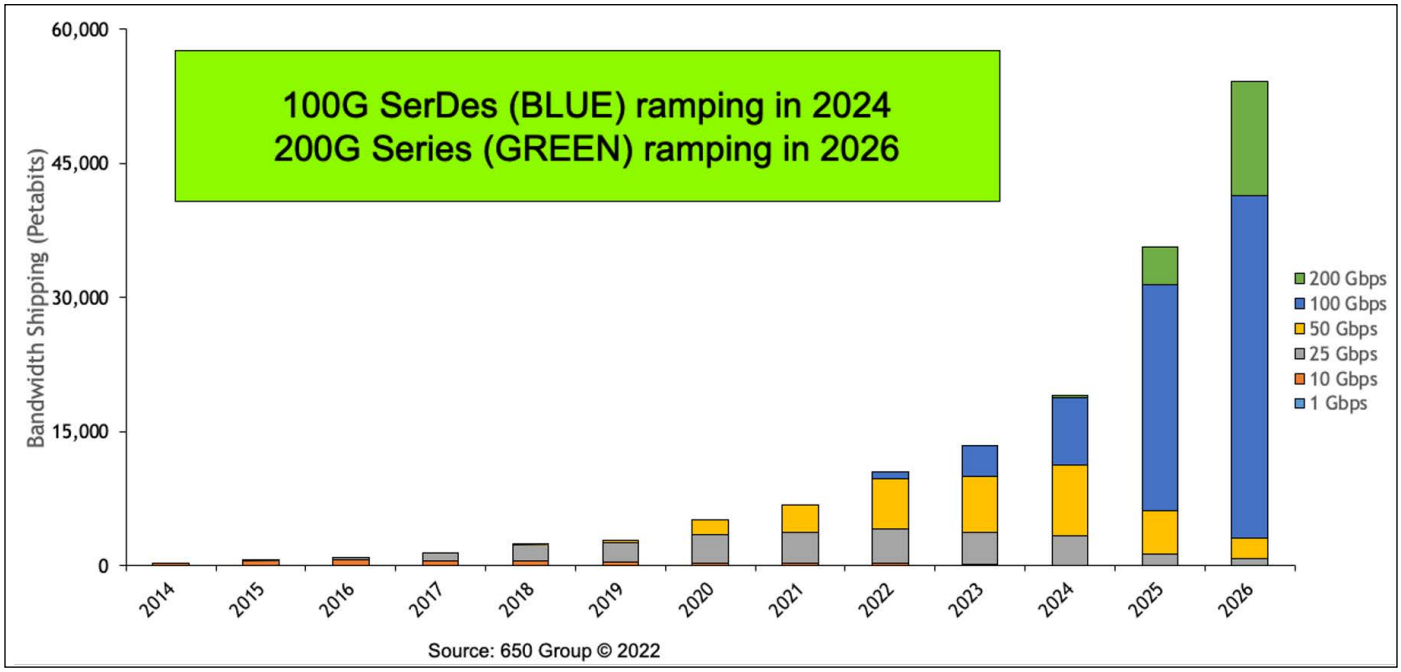

Figure 6: Data Center Ethernet Switching Bandwidth Growth, by SerDes Speed

Lossless Network

While faster speeds are useful, a lossless network is paramount to the job completion time. Infiniband applies credit-based flow control to avoid packet loss. A sender waits to transmit packets until it receives credits from the destination host indicating available buffers. Through the use of Explicit Congestion Notification (ECN) and Priority Flow Control (PFC), Ethernet can also operate as a lossless channel. These mechanisms apply back pressure to senders to avoid overrunning host or switch buffers. A reliable transport through IB flow-control or Ethernet with ECN/PFC is critical for maximizing RDMA performance.

Scalability

Increasing model size in LLMs has led to reliable and predictable improvements in capability. This in-turn is driving for larger LLMs and which inturn is driving a larger interconnect for AI clusters. In short, the scalability of the network is a very important consideration.

Ethernet has proven its ability to scale in the largest cloud networks in the world. Network teams have been able to adopt cloud designs and build their distributed network with CLOS architectures running Border Gateway Protocol (BGP).

On the other hand, Infiniband's control-plane is centralized through a single Subnet Manager that discovers the physical topology and sets up forwarding tables and QoS policies on each node. It periodically sweeps the network and reconfigures devices based on topology changes. This works well on small clusters, but can become a bottleneck at scale. There are complex after thought solutions that act as a patch. However, the distributed control-plane in Ethernet scales beyond the maximum subnet size of 48000 of Infiniband and offers more resiliency by avoiding single points of failure.

Resilience

When Infiniband's Subnet Manager fails, the entire subnet can go down. Infiniband does have some techniques that allow continuous forwarding in certain situations, but ultimately the control plane is still centralized and brittle. A full failover to the backup Subnet Manager will involve some downtime, and the downtime goes up the larger the subnet is (more state to transfer, larger sweep across the nodes). From talking to customers, it can be 30 seconds to several minutes. In some use cases, customers might be able to live with it but with large AI/ML workloads such failures will significantly impact job completion times and overall performance. With a distributed scalable architecture using ethernet and features such as Arista's SSU, link and node failures will have minimal to no impact to the overall performance of large AI networks.

Network Management

Over the last decade, network teams have adopted cloud principles to operate and manage their infrastructure as a singular unit. As AI clusters are becoming more common, network teams wouldn't want this cluster to be an island and would prefer to treat it as part of their common infrastructure. Additionally, ethernet is used ubiquitously across Data Center, Campus, Backbone, Wan, Edge and the network teams have built a strong expertise around ethernet.

Visibility

Telemetry and visibility are extremely important for automating the network and taking actions seamlessly. Network teams would want to extend the current tools, process and solutions used for their Data Center general compute and storage to their AI clusters as well.

Interoperability

OAI networks often interface with a variety of storage and general compute infrastructure. Ethernet-based AI networks enable efficient and flexible network designs that avoid pipeline bottlenecks through these various systems. While IP traffic can be transported over a physical Infiniband network, all servers must either have an Infiniband HCA or go through an Infiniband-to-Ethernet gateway that drastically limits throughput into and out of the IB network.

Open

Ethernet has a very strong ecosystem with multiple silicon vendors, system vendors, optics vendors and drives towards open and standard based solutions that can interoperate across vendors. InfiniBand falls behind significantly with a limited choice and locked-in solution.

In summary, Ethernet is considered the best solution for AI networking due to its scalability, interoperability, reliability, cost-effectiveness, flexibility, and familiarity. Its proven track record, widespread adoption, and support for high-speed networking make Ethernet a compelling choice for organizations looking to build efficient and scalable networking infrastructure to support their AI workloads.

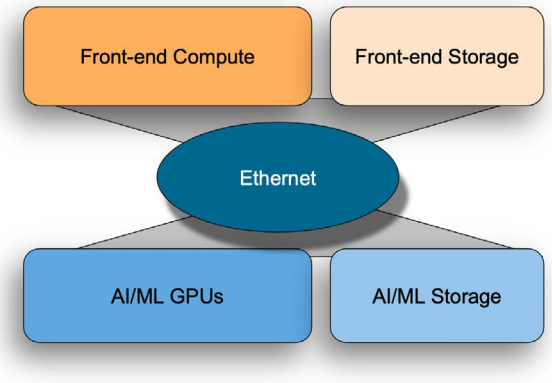

Figure 7: Ethernet for Compute, AI/ML, and Storage

Let us take a look at the key requirements for AI workloads using Ethernet. The network needs lossless transport in support of RoCEv2, Quality of Service (QoS) to prioritize control traffic, adjustable buffer allocation, effective load balancing and real-time monitoring.

Arista EOS

Modern AI applications need a high-bandwidth, lossless, low-latency, scalable, multi-tenant network that can interconnect hundreds and thousands of GPUs at speeds of 100Gbps, 400Gbps, 800Gbps and beyond. With support for Data Center Quantized Congestion Notification (DCQCN), Priority Quality of Service (QoS) and adjustable buffer allocation schemes, EOS provides all the necessary tools to achieve a premium lossless, high bandwidth, low latency network.

Through the support of Data Center Quantized Congestion Notification (DCQCN), EOS provides an end-to-end congestion control scheme using a combination of Priority Flow Control (PFC) and Explicit Congestion Notification (ECN) to support RDMA over Ethernet. Without visibility into network traffic and buffer utilization, configuring appropriate PFC and ECN thresholds can be challenging. Arista EOS®(Extensible Operating System) offers in-depth visibility into workload traffic patterns using the **AI Analyzer** and **Latency Analyzer** features.

AI Analyzer monitors interface traffic counters at intervals of microseconds, while Latency Analyzer tracks interface congestion and queuing latency with real-time reporting. AI Analyzer and Latency Analyzer help correlate the performance of the application with network utilization and congestion events, allowing PFC and ECN values to be optimally configured to best suit the requirements of the application.

With GPU clusters, data is transferred between nodes using a small number of queue pairs. This translates into a small number of high bandwidth traffic flows at each switch. Due to lack of entropy in the packet headers, it is easy for these flows to collide and cause congestion, driving up the job completion time. EOS takes real time traffic utilization of the network links into account and balances flows uniformly across them, avoiding network hotspots. EOS also offers source-interface based hashing to prevent traffic deceleration in non-oversubscribed networks. Traffic flows arriving on host interfaces can be directly hashed to designated uplinks, avoiding traffic fan-in and collisions. Additionally, load-balancing in EOS can also be configured to use user defined fields in the packet header to add further entropy. These result in less congestion in the network, fewer ECN marked packets, fewer pause frames, and higher aggregate throughput across nodes resulting in shorter completion times for the workloads.

Not all RDMA applications behave alike. Some are extremely latency sensitive while not being fixated on throughput while others require the highest possible throughput while willing to trade off on the latency front. Most applications fall somewhere in between the above mentioned types. With tools like QoS classification, scheduling and adjustable buffer allocation schemes, EOS allows customers to gain complete control of the network so they can tailor it to meet the requirements of the application. With support for VxLAN and EVPN, EOS addresses the need for scalable multi-segmentation by allowing several such applications to run in a single network.

In addition to strong software features, there is a need for reliable, best of breed hardware.

Platforms

The bandwidth and scale requirements for AI networks will vary from customer to customer and application to application. One size does not fit all. By leveraging the best-in-class merchant silicon packet processors, Arista Networks offers the best of breed ethernet AI leaf and AI spine systems for every AI network in the world.

AI Leaf - 7060X5

The 7060X5 Series are high density and power efficient fixed configuration data center switches for 800G, 400GbE, 200GbE and 100GbE; optimized for Hyperscale Cloud, Artificial Intelligence and Machine Learning environments employing high network radix. They provide a compelling solution for the largest hyperscale cloud and IO intensive environments, with consistent low latency combined with proven visibility, traffic instrumentation and automation features.

Represented as AI Leaf the 7060X5 provides up to 32 ports of 800G in a single rack unit enabling efficient high radix clusters that maximize the 25.6T bandwidth performance of compute and storage by eliminating bottlenecks between leaf and spine tiers. The 7060X5 series also supports up to 64 x 400G or 256 x 100G using breakouts, doubling current densities with new, cost effective, 800G optics.

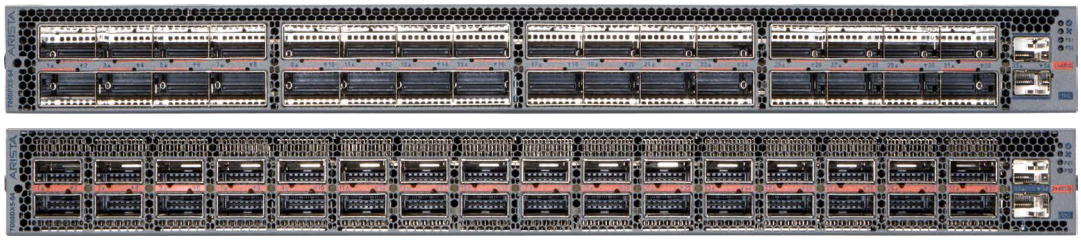

Figure 8: Arista 7060X5-64: 32 x 800G QSFP-DD or OSFP800 ports, 2 SFP+ ports

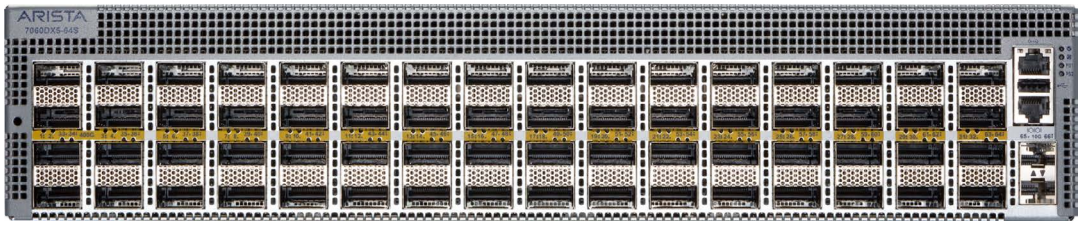

Figure 9: Arista 7060DX5-64S: 64 x 400GbE QSFP-DD ports, 2 SFP+ ports

The Arista 7388X5 Series deliver high density 200G and 400G optimized for Hyperscale Cloud, Artificial Intelligence and Machine Learning environments employing high network radix. The 7388X5 is a modular system built on a single 25.6Tbps high capacity packet processor for data intensive workloads requiring consistent low latency, with a flexible choice of industry standard interfaces, and significant improvements in power consumption and system density.

Figure 10: Arista 7388X5: 128 ports of 200G or 64 ports of 400G

AI Spine - 7800R3

The Arista 7800R3 Series of purpose-built modular switches deliver the industry's highest performance scaling to 460 Tbps of system throughput to meet the needs of the largest scale data centers and high performance compute networks. The Arista 7800R3 Series delivers non-blocking switching capacity that enables dramatically faster and simpler network designs for data centers while lowering both capital and operational expenses.

The 7800R3 has some key characteristics that makes it an ideal platform for AI Networking:

Virtual Output Queuing (VoQ): a distributed scheduling mechanism is used within the switch to ensure fairness for traffic flows contending for access to a congested output port. A credit request/grant loop is utilized and packets are queued in physical buffers on ingress packet processors within VoQs until the egress packet scheduler issues a credit grant for a given input packet.

Cell Based Fabric: A cell-based fabric takes every packet and breaks it apart into evenly sized cells before evenly "spraying" across all fabric modules. This spraying action has a number of positive attributes making for a very efficient internal switching fabric with an even balance of flows to each forwarding engine. Cell-based fabrics are considered to be 100% efficient irrespective of the traffic pattern.

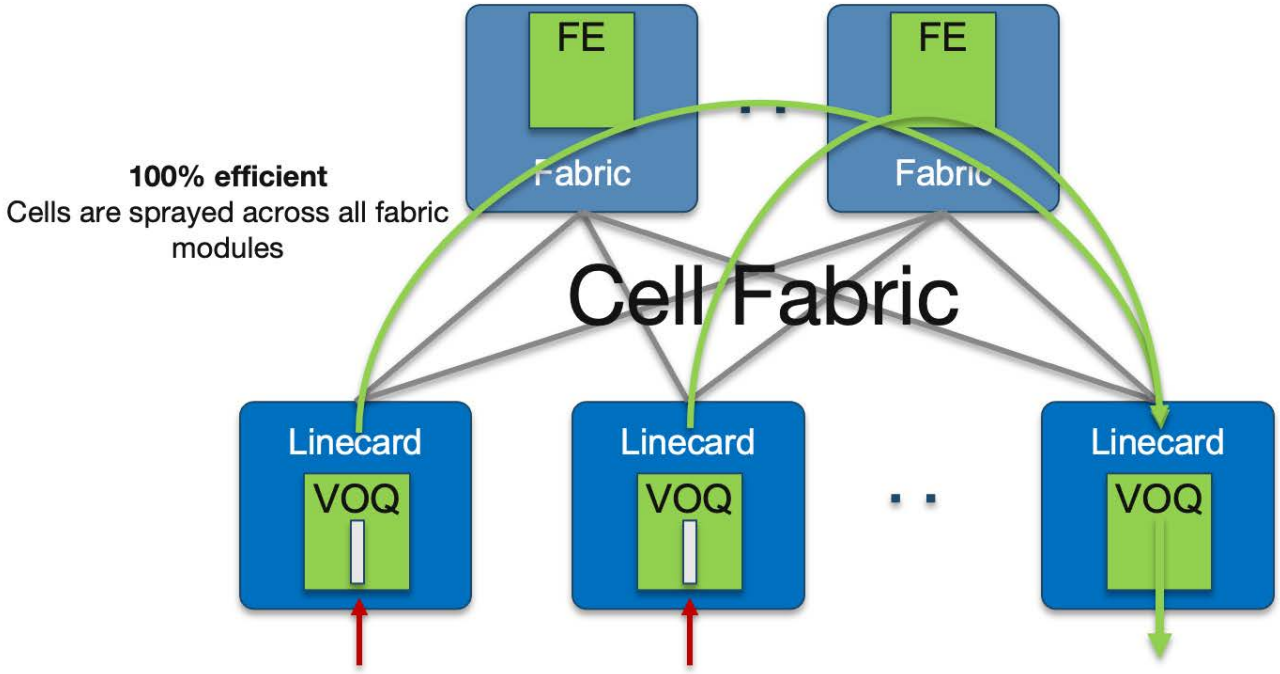

Figure 11: Cell-based Fabric Architecture

This spraying behavior makes a cell fabric inherently good at dealing with mixed speeds. A cell-based fabric is not concerned with the front panel connection speeds, making mixing and matching 100G, 200G and 400G of little concern.

Moreover, the cell fabric makes it immune to the "flow collision" problems of an Ethernet fabric. Because a flow uses all paths to reach its destination, there are no internal hot spots in the network, making the cell fabric especially well suited to the "elephant flow" heavy traffic that is common to AI/ML applications.

Deep packet buffering: The 7800R3 series line cards utilize on-chip buffers (32MB) in conjunction with flexible packet buffer memory (8GB of HBM2 per packet processor). The on-chip buffers are used for non-congested forwarding and seamlessly utilize the HBM2 packet buffers for instantaneous or sustained periods of congestion. Buffers are allocated per VoQ and require no tuning. It's further worth noting that during congestion, packets are transmitted directly from the HBM2 packet buffer to the destination packet processor.

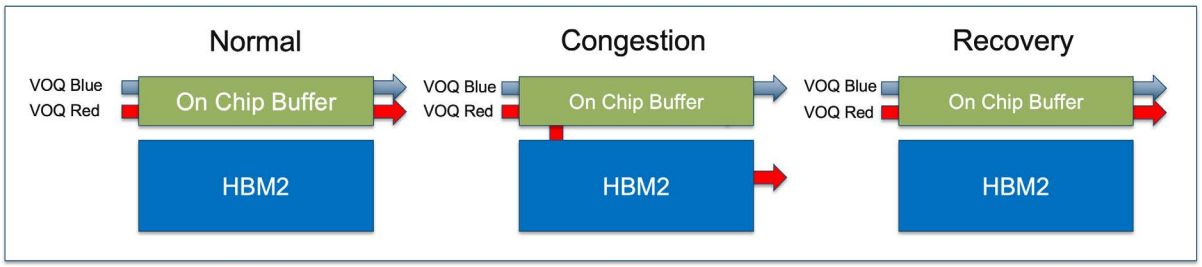

Figure 12: Packet buffer memory access

HBM2 memory is integrated directly into the Jericho2 packet processor this provides a reliable interface to the Jericho2 packet processor and eliminates the need for additional high-speed memory interconnects as does HMC or GDDR. This results in upwards of a 43% reduction in power utilization than the equivalent GDDR memory.

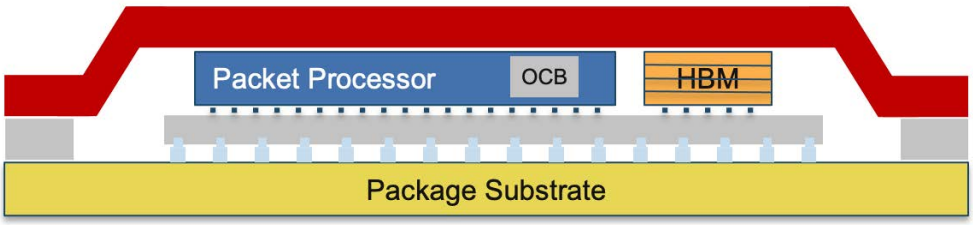

Figure 13: HBM memory packaging integration

Predictable Performance: A combination of Advanced Queuing Credit schedulers with Virtual Output Queues (VOQs) and deep buffers (for congestion avoidance) on a cell-based platform makes 7800R3 a lossless system. Cell-based systems give you more predictable performance under any load and the addition of Virtual Output Queue (VOQ) helps protect against packet loss during congestion. These two capabilities coupled with a deep buffer platform guarantee the lossless transport for RoCEv2 in GPU interconnects with AI/ML workloads.

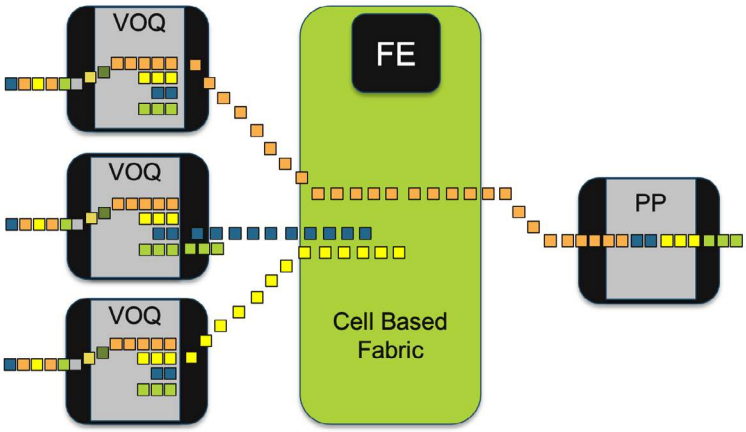

Figure 14: Credit Based VoQ Architecture

Density: The 7800R3 Series are available in a choice of 4, 8, 12 and 16-slot systems that support a rich range of line cards providing high density 100G and 400G with choice of forwarding table scale. At a system level, the 16-slot Arista 7816R3 scales to 460 Tbps and enables 576 x 400G in a 32 RU front to rear power efficient form factor, providing industry-leading performance and density without compromising on features and functionality.

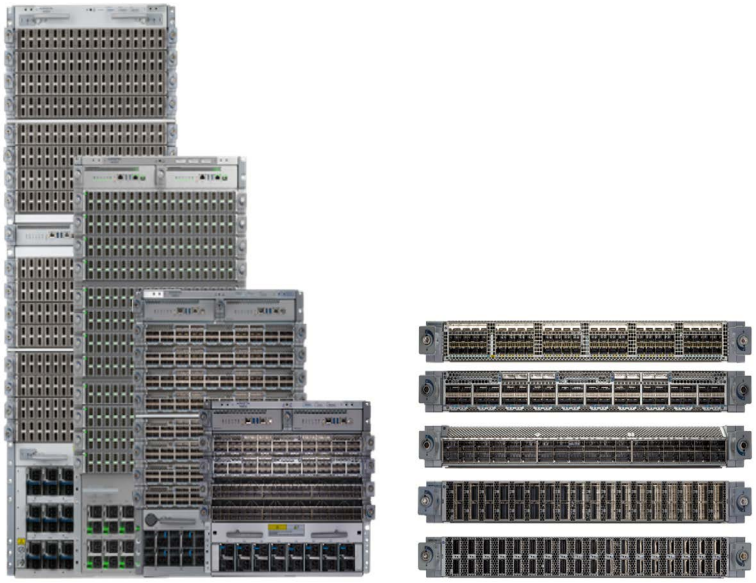

Figure 15: Arista 7800R3 Series

Flexibility & Efficiency: All components in the 7800R3 series are hot swappable, with redundant supervisor, power, fabric and cooling modules with front-to-rear airflow. The system is purpose-built for data centers and is energy efficient with typical power consumption of under 25 watts per 100G port and 50W per 400G port for a fully configured chassis.

All of these attributes of Arista 7800R3 combined with the strong feature set of EOS make the 7800R3 an ideal platform for building reliable and highly scalable data center networks and High Performance Network.

Design Based on AI App Size

Over the years, new technologies and applications such as Server Virtualization, Application Containerization, Multi-Cloud Computing, Web 2.0, Big Data and High Performance Computing (HPC) have significantly changed the east-west and north-south traffic patterns within the data center. To optimize and increase the performance of these new technologies, a distributed scale-out, deep-buffered IP fabric has proven to provide consistent performance that scales to support extreme 'East-West' traffic patterns. Customers have successfully built small to large data center cloud networks using IP/Ethernet to support modern application and network requirements.

Historically, AI/ML applications could co-exist in the IP fabric in conjunction with other applications. However, due to the significant growth in AI/ML applications and their associated complexity from the adoption of special purpose GPUs, DPUs and TPUs, we recommend designing a dedicated network for these applications. It will allow operators to tune the network to better handle unique traffic patterns that come with modern AI/ML workloads.

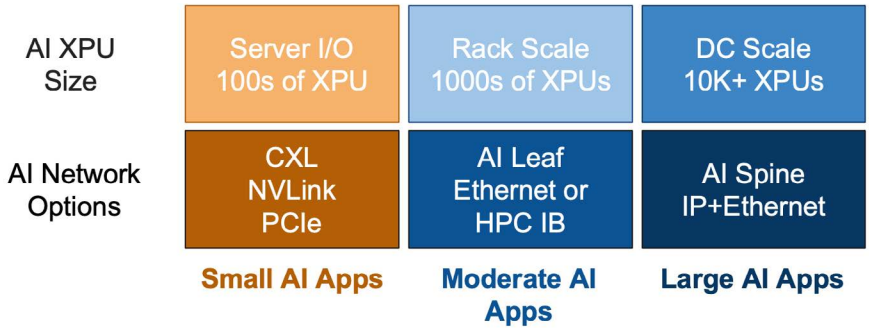

Figure 16: AI Network Design Guidelines

For smaller AI applications requiring interconnecting GPUs confined to a single rack, load/store interconnects like PCIe, CXL, and other proprietary options like NVLink can be considered to move data efficiently at low latency and power consumption. However, these solutions quickly become cost and power hungry rendering them unfeasible for connections across racks. For applications requiring interconnecting hosts across racks, Ethernet or InfiniBand become the choice of protocols.

Small AI Applications

A single Arista DCS-7060DX5-64S or DCS-7388X5 switch with 64 x 400G ports or 128 x 200G ports can effectively interconnect GPUs across a few racks. In this design, each GPU can communicate with all other GPUs in a non-blocking configuration at a predictably low latency. This option requires minimal tuning simplifying operations and management.

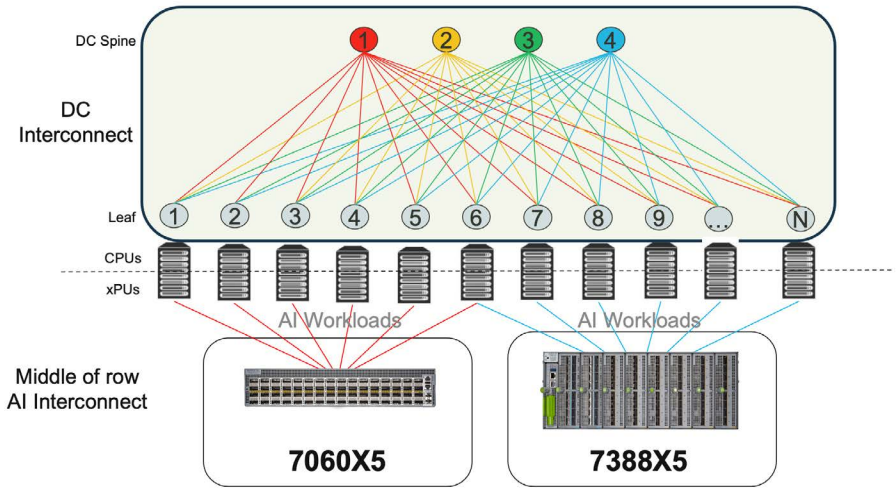

Figure 17: Middle of row AI Interconnect

Moderate AI Applications

A single Arista 7800R3 switch with support for 576 x 400G ports can act as a simple, out of the box AI spine interconnect to support moderate sized AI applications. Since this design provides a consistent, single hop between the end hosts, it drives down the latency and power requirements. The 7800R3 with its cell-based, non-blocking VOQ architecture, enables a single, large, lossless network without any configuration or tuning. A single hop solution ensures ECN and PFC configurations are required only on the host facing ports, allowing GPUs to send and receive line rate data at all times.

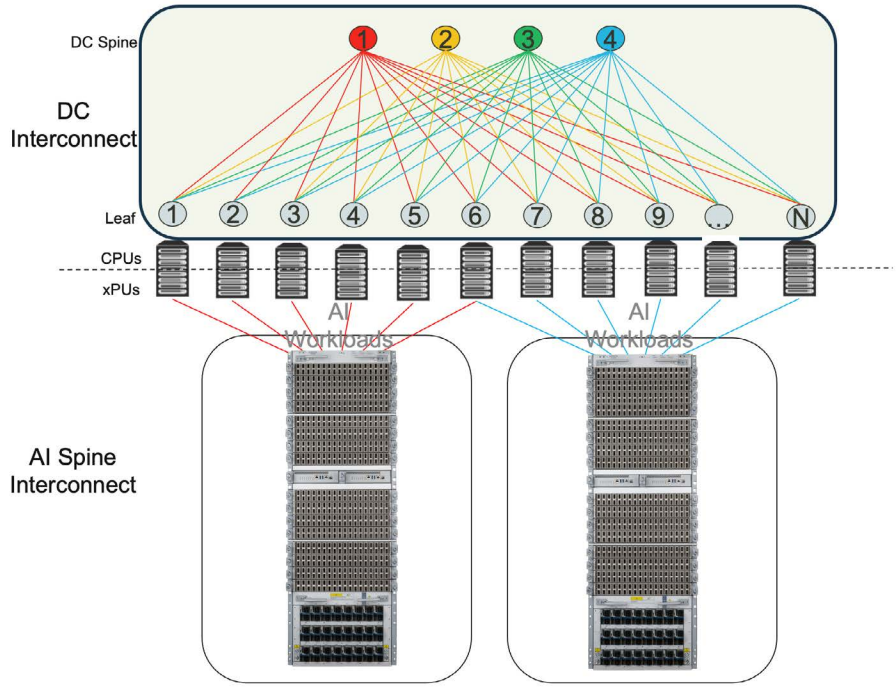

Figure 18: AI Spine Interconnect

Large AI Applications

For large scale AI applications, requiring tens of thousands of GPUs to be connected in data centers, Ethernet becomes the most viable option. Arista's Universal Leaf and Spine design offers the most simple, flexible and scalable architecture to support AI workloads at data center scale. This design allows more than 18,000 x 400G end hosts to be interconnected while keeping the latency predictive and low. In such a design, Arista EOS' intelligent load-balancing capabilities that take real time traffic utilization of the network into consideration to uniformly distribute traffic flows can be leveraged to avoid flow collisions. Arista EOS' advanced telemetry options like AI Analyzer and Latency Analyzer make it simple for network operators to determine optimal PFC and ECN configuration thresholds to allow GPUs to exchange line rate throughput across the network while preventing packet drops.

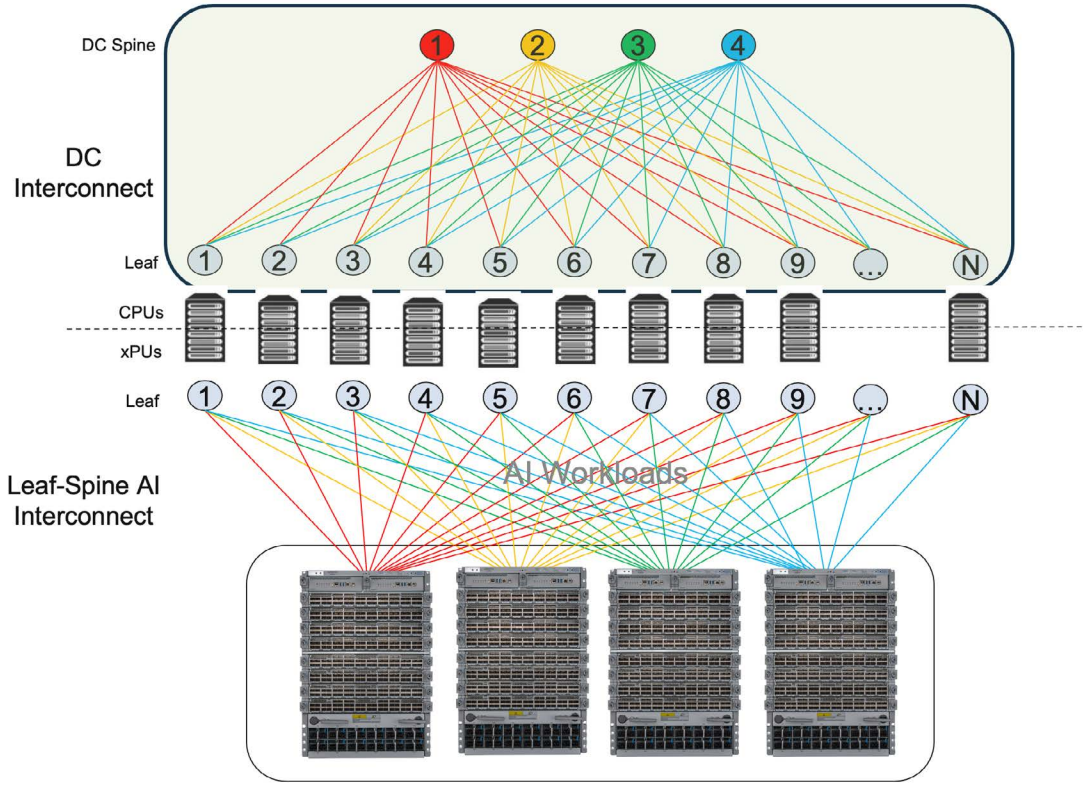

Figure 19: Leaf-Spine AI Interconnect

The Universal Leaf and Spine design provides an ideal solution for AI models currently requiring a few hundred GPUs and offers the flexibility to scale out to tens of thousands of GPUs in the future with consistent performance.

Storage for AI networks

The amount of data used by AI has exponentially grown as businesses attempt to improve the accuracy of their models. In the training phase, large datasets are needed to improve the accuracy of the AI model. As such, organizations are forced to manage massive collections of data, starting with Petabytes. This puts a lot of strain on the network handling transfer of data between GPUs and the storage nodes. A dedicated storage network is recommended to avoid expensive and in demand GPUs from idling for data due to network bottlenecks. To allow efficient movement of data, most GPUs enable a direct data path between their memory and remote storage using NVMe-oF.

Arista's Universal Leaf and Spine architecture using 7280R3 series as Leaf switches and 7800R3 as Spine switches are proven best in class solutions for storage networks using protocols like NVMe/ROCE and NVMe/TCP. Arista's 7280R3 and 7800R3 series switches with integrated VOQ mechanism and deep packet buffering capability can be used to design cost-effective, high performance storage solutions at scale.

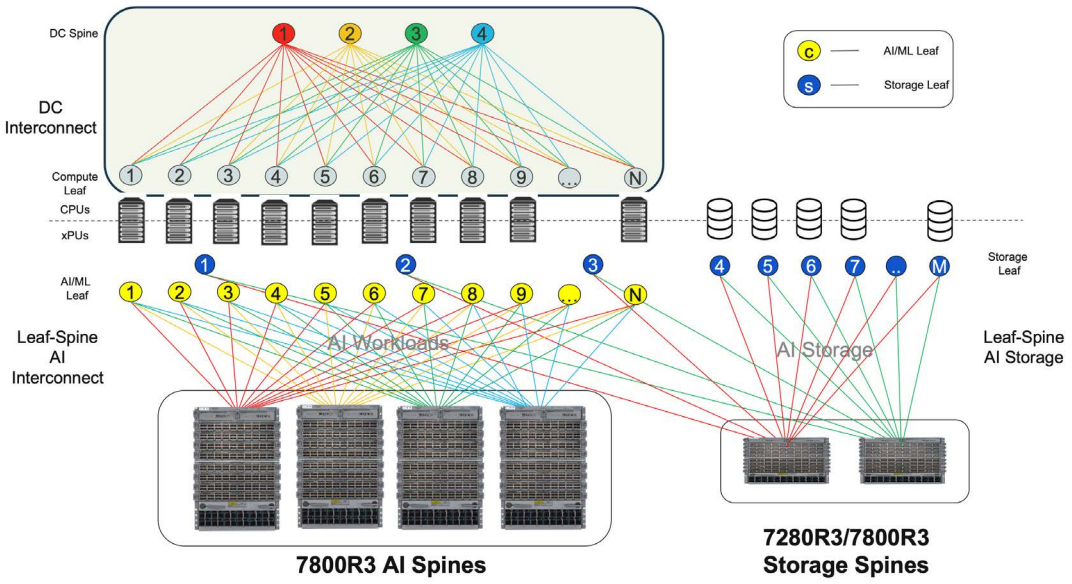

Figure 20: Storage for AI

Conclusion

Arista provides the best solution using IP/Ethernet switches for GPU and Storage interconnects driving AI/ML workloads. Exponential growth in AI applications requires standardized transport such as Ethernet to build a power efficient interconnect and overcome administrative, scale-out complexities of traditional approaches. Building an IP/Ethernet architecture with high-performance Arista switches maximizes application performance while optimizing network operations. The 7800R3 AI spine & 7060 AI leaf combined with EOS innovations is an ideal choice for modern AI applications.

Reference

- RDMA – http://www.rdmaconsortium.org/
- RoCE - https://cw.infinibandta.org/document/dl/7781 - "InfiniBand Architecture Specification Release 1.2.1 Annex A17: RoCEv2". InfiniBand Trade Association.
- Arista L3LS Design Deployment Guide
- Arista 7800R3 Switch Architecture WP
- Arista UCN Deployment Guide

Santa Clara—Corporate Headquarters
5453 Great America Parkway,
Santa Clara, CA 95054

Phone: +1-408-547-5500
Fax: +1-408-538-8920
Email: info@arista.com

Ireland—International Headquarters
3130 Atlantic Avenue
Westpark Business Campus
Shannon, Co. Clare
Ireland

Vancouver—R&D Office
9200 Glenlyon Pkwy, Unit 300
Burnaby, British Columbia
Canada V5J 5J8

San Francisco—R&D and Sales Office
1390 Market Street, Suite 800
San Francisco, CA 94102

India—R&D Office
Global Tech Park, Tower A, 11th Floor

Marathahalli Outer Ring Road
Devarabeesanahalli Village, Varthur Hobli
Bangalore, India 560103

Singapore—APAC Administrative Office
9 Temasek Boulevard

#29-01, Suntec Tower Two
Singapore 038989

Nashua—R&D Office
10 Tara Boulevard
Nashua, NH 03062