AI Networking

Introduction

The rapid arrival of real-time gaming, virtual reality and metaverse applications are changing the way network, compute memory and interconnect I/O interact for the next decade. As the future of metaverse applications evolve, the network needs to adapt to the humongous growth in traffic connecting hundreds of processors with trillions of transactions and gigabits of throughput. As compute power continues to evolve, AI has found its way out of research labs and is powering a lot of the technological progress today. 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 TPU. 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.

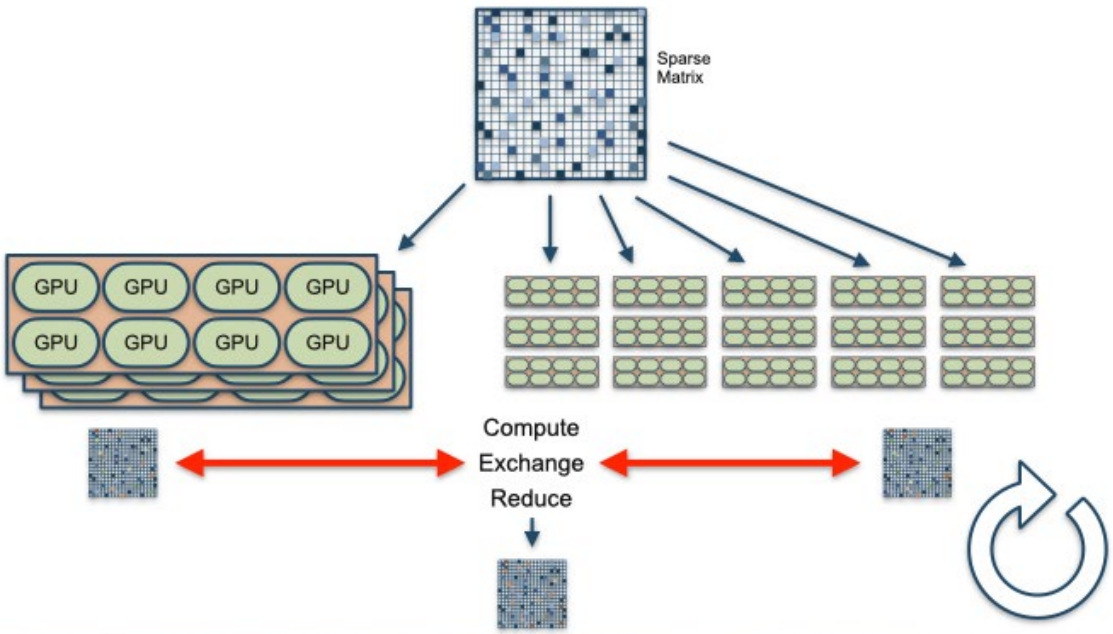

Figure 1: Compute-Exchange-Reduce Cycle

TCP/IP and RDMA

In TCP/IP, 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.

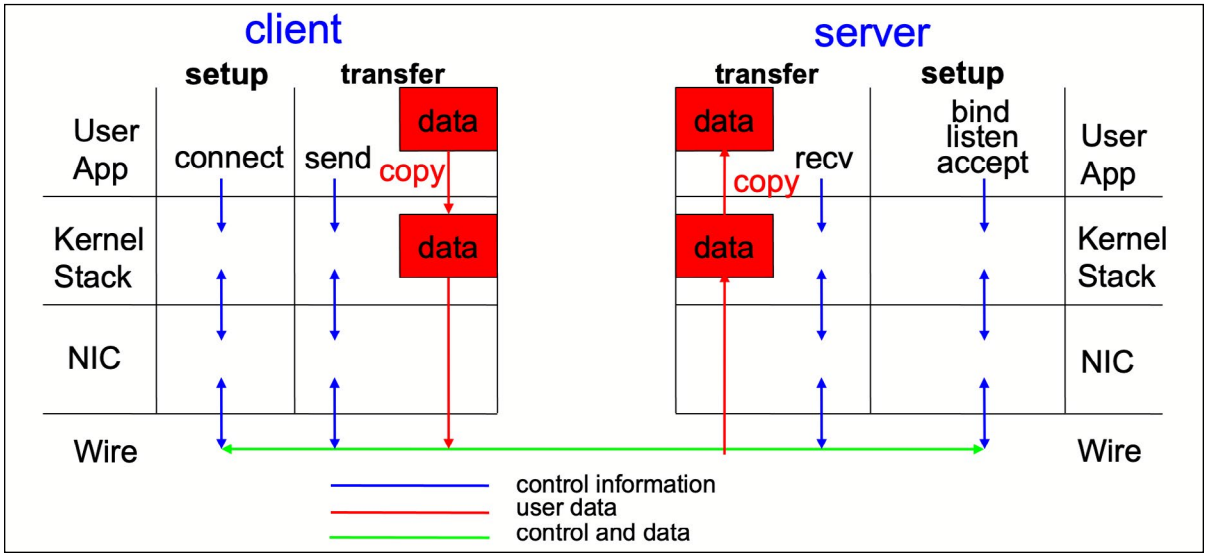

Figure 2: TCP/IP Transfer

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.

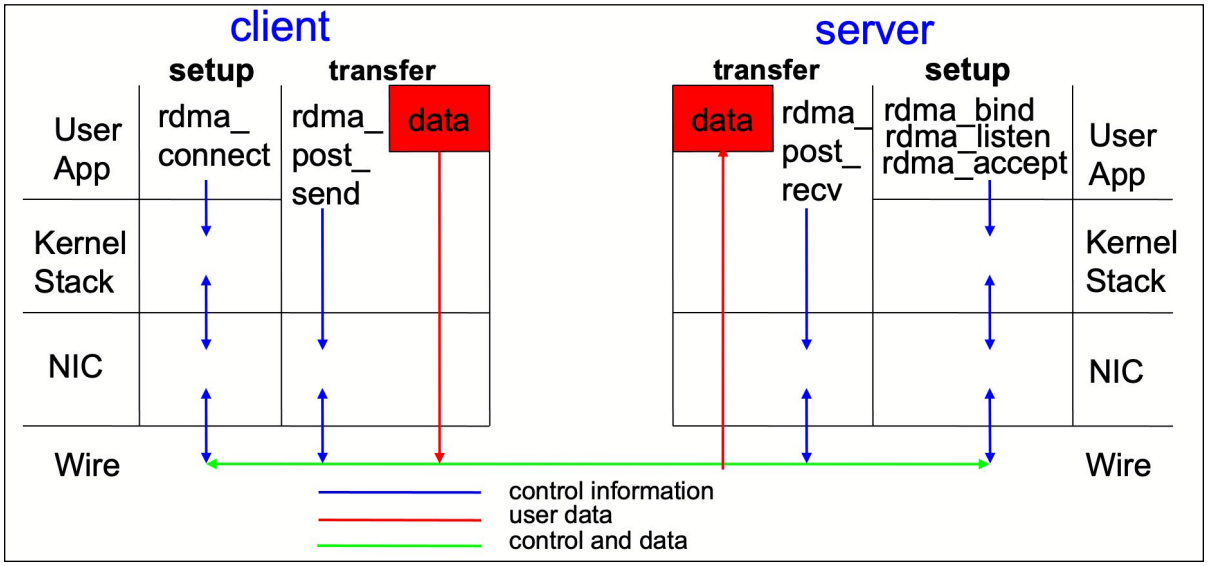

Figure 3: RDMA Transfer

InfiniBand

Traditionally, InfiniBand was the preferred choice of transport for RDMA-based workloads. It worked well in research labs and supercomputers where all the nodes fit within a single data center. With GPU and NIC speeds doubling every few years, modern AI applications require large cluster sizes to interconnect hundreds of nodes and scale them to tens of thousands of nodes as the demand grows. Although the link bandwidth for InfiniBand is increasing, there are several downsides to InfiniBand such as the limited number of silicon suppliers, open solutions for InfiniBand, cost per bit, power per bit and most importantly the failure to address the need for multi-tenancy. Running several applications on a common fabric helps optimize cost and performance while keeping design, testing and operations simple.

Ethernet

Using RDMA over Converged Ethernet (RoCE), AI workloads that were previously confined to InfiniBand networks can now leverage the economies of scale that Ethernet brings to the table. 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.

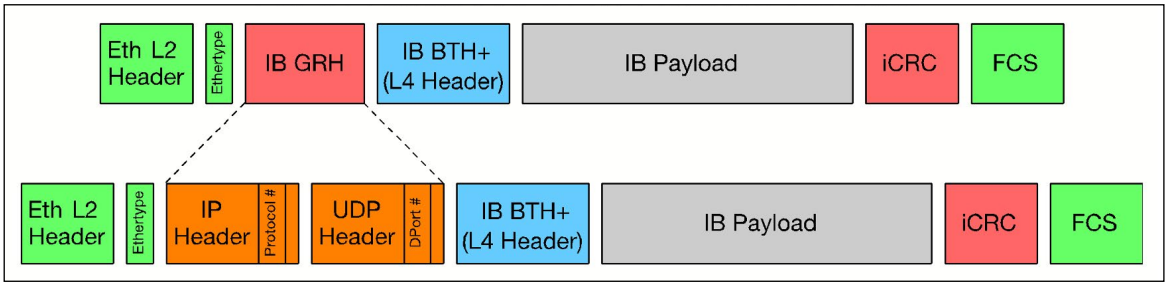

Figure 4: RoCE and RoCEv2 Frame Format

By its very design Ethernet supports scalable multi-tenancy and is the most popular networking technology in the world. It enjoys tremendous supplier and silicon diversity with many companies actively investing in switching systems and NIC technologies. With Moore's law pushing silicon processes from 7nm to 5nm to 3nm, Ethernet is emerging as the clear winner.

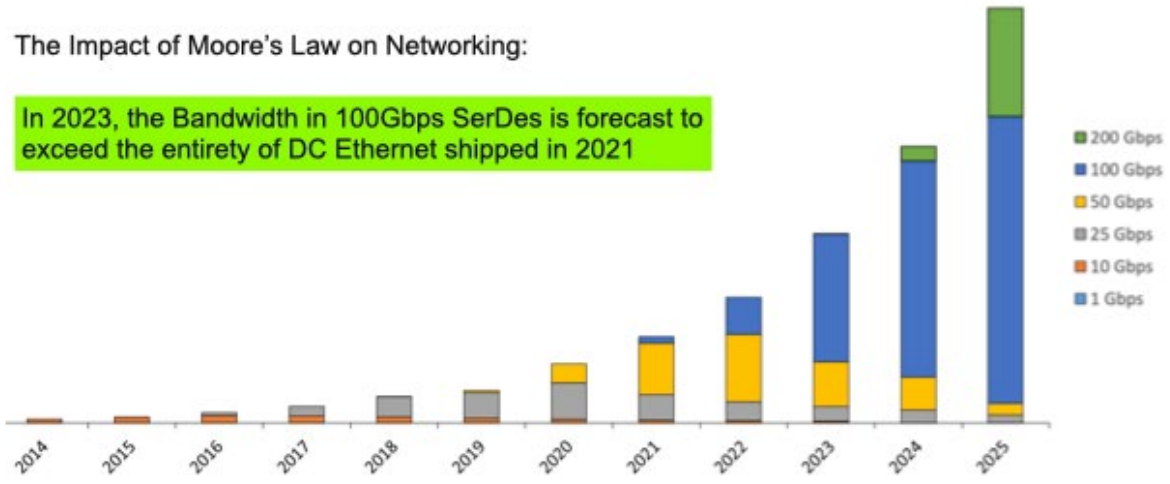

Figure 5: Data Center Switching Bandwidth Growth, by SerDes Speed

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 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 buffer utilization, configuring appropriate PFC and ECN thresholds can be challenging. Arista EOS (Extensible Operating System) offers an easy solution called Latency Analyzer (LANZ) which tracks interface congestion and queuing latency with real-time reporting. This helps correlate the performance of the application with network 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. This results 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 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.

Arista Networks has always leveraged best-in-class merchant silicon packet processors for leaf and spine systems. The network silicon within the Arista 7800 Series is no exception and utilizes the latest high capacity multi-chip system packet processor, the Jericho2 from Broadcom.

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.

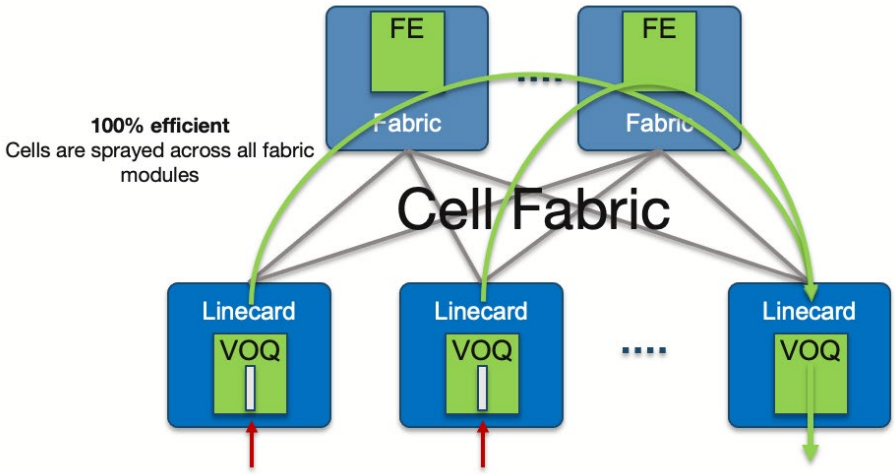

Figure 6: Cell-based Fabric Architecture

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 with Jericho2) 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.

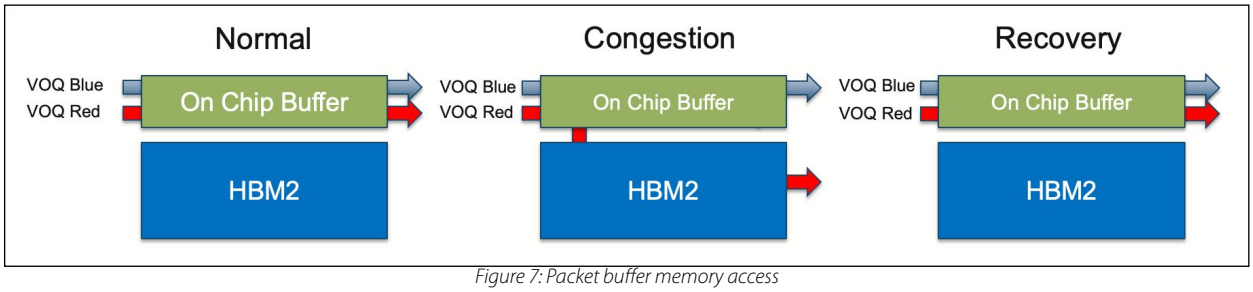

Figure 7: Packet buffer memory access

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.

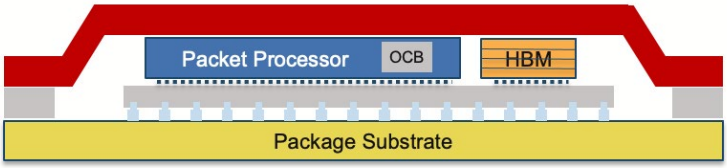

Figure 8: HBM memory packaging integration

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.

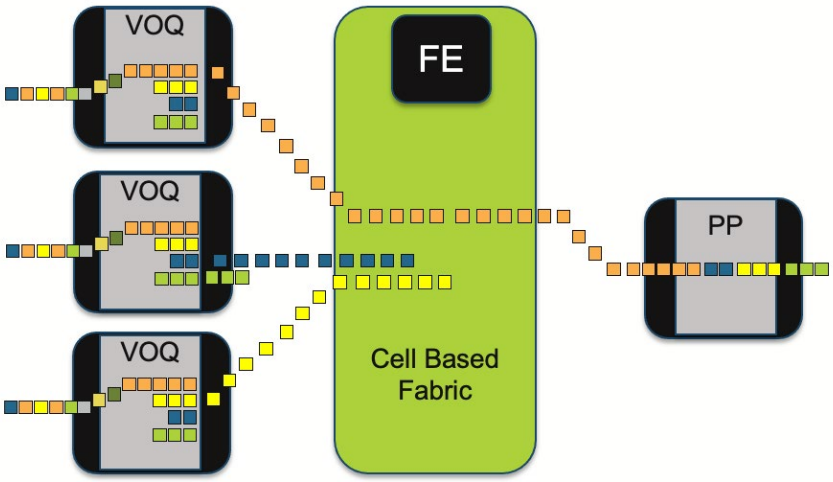

Figure 9: Credit Based VoQ Architecture

Density: The 7800R3 Series are available in a choice of 4, 8 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 with a fabric that scales to 460 Tbps 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.

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 Networks.

Reference Architecture

Over the years, traffic patterns within the data center and out of the data center have changed. This was primarily driven by new technologies and applications such as Server Virtualization, Application Containerization, Multi-Cloud Computing, Web 2.0, Big Data and High Performance Computing (HPC). To optimize and increase the performance of these new technologies, a distributed scale-out, deep-buffered IP fabric has been proven to provide consistent performance that scales to support extreme 'East-West' traffic patterns. Customers have been successful in building small to large data center cloud networks using IP/ethernet to support their application and network requirements.

Historically, AI/ML applications could be supported in the IP fabric in conjunction with other applications. However, there has been a significant growth in AI/ML applications and its associated complexity. AI/ML applications have been driving the adoption of special purpose GPUs, DPU and TPUs to support their complex computational requirements. Modern AI/ML workloads powered by GPU clusters come with unique traffic patterns and it would be ideal to design a dedicated network for these applications.

Design 1

A dedicated Leaf and Spine architecture using the principles of Arista Universal Cloud Network design would allow the network to scale to hundreds of racks while keeping the latency predictive and low. In such a design, special care must be taken to ensure traffic flows are uniformly distributed across the Leaf and Spine links. Arista EOS' intelligent load-balancing capabilities can be leveraged to avoid flow collisions. QoS classification, ECN and PFC thresholds need to be appropriately configured on all the switches to avoid packet drops. With visibility into buffer utilization, Arista EOS' LANZ capability can be utilized to determine the ECN and PFC configuration thresholds to avoid packet drops while keeping the throughput high.

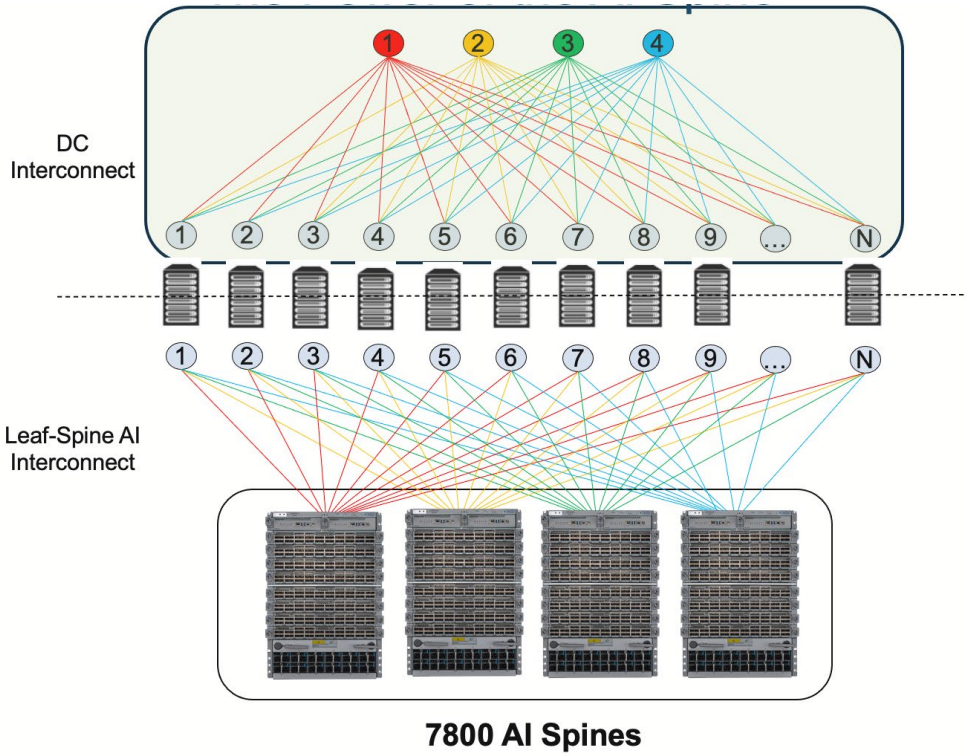

Figure 10: Leaf-Spine AI Interconnect

An Arista Leaf Spine design is an excellent choice to serve AI applications that need horizontal scalability.

Design 2

A simplified out of the box solution that can scale to hundreds of end points is to connect all the GPU modes directly into the 7800R3 AI Spine. This architecture provides a consistent single hop between all end points, further driving down latency and power requirements. 7800R3 with its Non Blocking VOQ Architecture and Cell based architecture enables a single large lossless network without any configuration or tuning. The existence of a single hop between the end points simplifies ECN and PFC configuration.

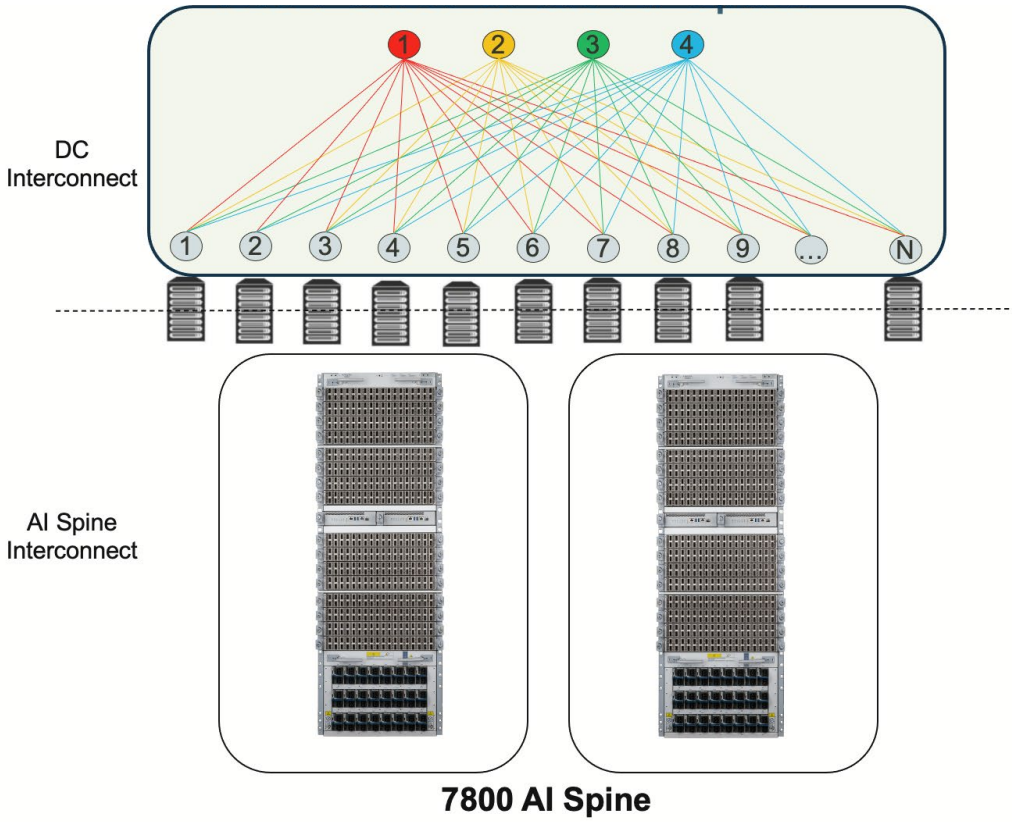

Figure 11: Arista DCS-7500R3-36CQ-LC module architecture

Conclusion

Arista provides the best solution using IP/Ethernet switches for GPU 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 7800 AI spine 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