ARISTA

White Paper

AI Networking

Emergence of Artificial Intelligence (AI)

Artificial Intelligence (AI) has emerged as a revolutionary technology that is transforming many industries and aspects of our daily lives from medicine to financial services and entertainment. The rapid evolution of real-time gaming, virtual reality, generative AI and metaverse applications are changing the ways in which network, compute, memory, storage and interconnect I/O interact. As AI continues to advance at unprecedented pace networks need to adapt to the colossal growth in traffic transiting hundreds and thousands of processors with trillions of transactions and gigabits of throughput. As AI quickly moves out of labs and research projects toward mainstream adoption it demands increases in network and 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 training workload involves billions of parameters and a large sparse matrix computation distributed across hundreds or thousands of processors – CPUs, GPUs or TPUs. These processors compute intensively 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, approximately 20-50% of the job time is spent communicating across the network so bottlenecks have a substantial impact on job completion time.

Networking for AI

Ethernet has come a long way since its invention by Bob Metcalfe, first introduced as a memo in 1973 and commercialized in 1980. Since then, the technology has evolved multiple times and has been extended all over the world. It's grown in speeds from the initial 10 Mbps to now 800 Gbps per port, with 1.6 Tbps on the distant horizon, and has evolved to support campus switching, data center switching, storage networking, high-frequency trading (HFT), high-performance computing (HPC), voice telephony and streaming video, cloud computing, and even wide area networks (WANs).

Now, Ethernet has evolved further into the best option to power the artificial intelligence (AI) workloads that promise to revolutionize the way we work, play, and learn. Ethernet is deployed today at large scale for high-performance AI networks, ready right now. And Arista is helping lead the further evolution of Ethernet to enhance AI networking even more.

Rise of the AI Center

Network operators have a choice when building new back-end networks for AI workload training; to reinforce historical silos of technology and staff skills, or to embrace a more modern approach to AI that accepts the collective nature of AI workloads and expands the concept across a holistic, inclusive AI solution. InfiniBand is the standard-bearer for a siloed approach to AI, while Ethernet represents the unifying option that expands AI from back-end to front-end networks for consistency and coordination.

Traditional solutions for building high-performance computing (HPC) clusters have relied upon InfiniBand, and some customers have considered extending that to AI training in new back-end networks. However, that approach fundamentally introduces network silos as the traditional data center (and thus the front-end network for AI inference clusters) has historically been built around Ethernet. Customers will need gateways to connect these network silos, which adds complexity. With this model, there will be disparate operational skill sets for back-end AI vs. front-end AI networks, and operational silos between AI accelerators (XPUs), general compute (CPUs), networking, and storage.

The AI Center represents a better way, with Ethernet unifying all elements of the complete system with open standards at every layer. The AI Center unifies the entire front- and back-end ecosystem to deliver scale-out networking for AI with optimized performance and operations alike. It enables coordinated visibility, management, and control of AI workloads, compute, networks, and storage along with existing data center workloads and systems.

Figure1 - Compute-Exchange-Reduce Cycle

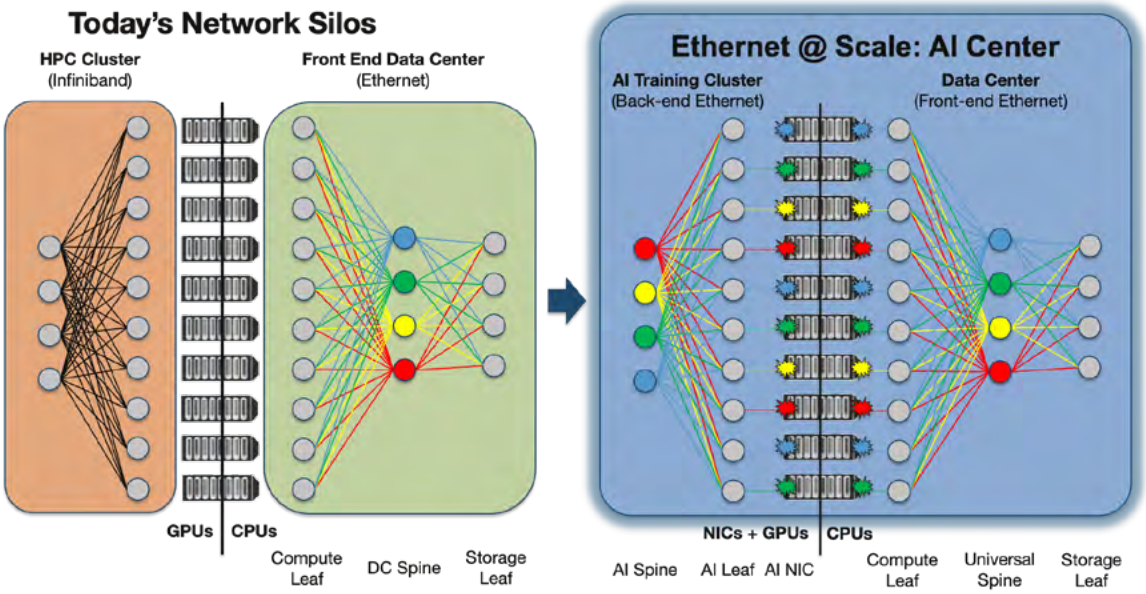

Figure 1: Rise of the Ethernet-enabled AI Center, breaking down network silos

Arista's platform is designed to help customers easily bring the unified vision and advantages of the AI Center to fruition.

Ultra Ethernet Consortium

While the current Ethernet based solution scales well, the underlying Ethernet network needs to be simplified and redesigned to accommodate higher speed and scalability to further improve the job completion rate.

To this end, Arista is proud to be a founding member of Ultra Ethernet Consortium (UEC), publicly announced in July 2023, whose other members include suppliers and operators of many of the largest AI and HPC networks today. The goal of UEC is to leverage its members' many years of experience building and operating these networks to deliver an Ethernet-based open, interoperable, high performance, full-communications stack architecture to meet the growing network demands of AI/ML and HPC workloads deployed on-premises and in the public clouds.

UEC aims to replace the legacy RoCE protocol with Ultra Ethernet Transport, a modern transport protocol designed to deliver the performance that AI applications require while preserving the advantages of the Ethernet/IP ecosystem. UEC will follow a systematic approach with modular, compatible, interoperable layers and tight integration of these layers to provide a holistic improvement for demanding workloads, while minimizing communication stack changes and maintaining and promoting Ethernet interoperability. This is to address some of the shortcomings of RDMA related to dropped packets, DCQCN, lack of multipathing, and scaling limitations pertaining to endpoints and processes. Since AI models are increasingly sensitive and valuable business assets, UEC will also incorporate network security by design making it robust to support AI/ML and HPC networks of the future.

UEC protocols are also designed to support modern HPC workloads, leveraging the same transport mechanism outlined above while preserving broadly used APIs such as MPI and PGAS. For more information on UEC, please visit http://www.ultraethernet.org/.

AI Center Powered by Arista Etherlink™

Arista Etherlink brings the AI Center into fruition, with the most comprehensive, high-performance offering for holistic AI networking. Etherlink delivers optimized performance with low power consumption for AI training clusters, enabling unprecedented coordination and visibility into AI workloads across networks, NICs, compute, and XPU resources. Etherlink is based on open standards Ethernet, is available today, and incorporates features compatible with the Ultra Ethernet Consortium (UEC). Arista Etherlink AI platforms are forward compatible with future UEC specifications, and will be upgradable to remain compliant.

Arista Etherlink for the AI Center spans a number of coordinated elements that together comprise a complete solution for optimized AI networking, greater in total than individual parts on their own:

1. AI Platforms: a comprehensive choice of AI-optimized, UEC-compatible Arista Ethernet networking systems, enabling AI clusters of any scale ranging from just 32 XPUs all the way up to 100,000+ parallel XPUs. These systems include fixed options based on Tomahawk 4 (TH4) and Tomahawk 5 (TH5) silicon in Arista 7060X series 400G & 800G-optimized switches, as well as modular variants with virtual output queuing (VOQ) based on Jericho 2C+ (J2C+) and Jericho 3-AI (J3-AI) silicon in Arista 7800R series 400G & 800G-optimized routers. Depending on the size of the desired AI cluster, these systems could be deployed standalone or in an AI leaf / spine cluster, with an additional deployment option of a single-hop scale-out design based on the 800G-optimized Arista 7700R series distributed Ethernet switch (DES).

2. AI Suite of Software: the proven, high-quality Arista EOS operating system with a diverse suite of features to optimize transport of AI workloads over Ethernet. These include RDMA-aware dynamic load balancing, advanced congestion control, and reliable packet delivery to all network interface cards (NICs) supporting RDMA over Converged Ethernet (RoCE).

3. AI Agents: coordination, control, and visibility between networks and NICs to ensure optimized performance and unified management of compute + network environments, enabled by an EOS-based AI agent deployed onto SmartNICs or servers in the future.

4. AI Observability: intended to provide deep insights and visibility into AI workload performance across the entire AI cluster, based on the combination of Arista's network data lake (NetDL) for ingesting streaming telemetry, Arista's Cloudvision network automation software, and Arista's latency, collectives, and AI analyzer features embedded into EOS on each system.

Arista Etherlink: Unprecedented Performance, Scale, Resiliency

Arista's Etherlink AI solution coordinates systems, software, and AI agents to optimize end-to-end performance with unprecedented insights and control. Key benefits include:

- Optimized performance: Etherlink delivers a high-performance AI solution, with up to 65% improved performance compared to traditional non-optimized Ethernet system performance. This performance is achieved by pairing lossless, low latency platforms with innovative RDMA-aware load balancing and congestion avoidance features for the lowest AI job completion times (JCT).

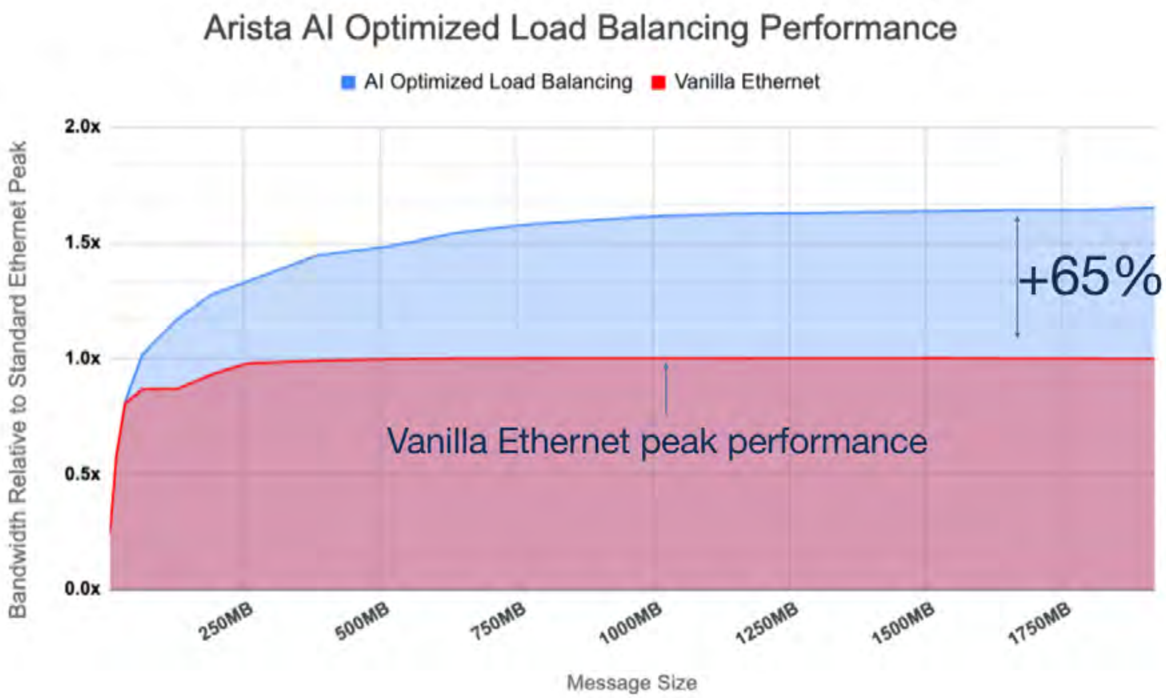

Figure 2: Arista Etherlink AI platform performance compared to traditional Ethernet systems

Performance can be characterized in a few ways; generally we are most concerned with securing peak performance in best-case conditions with all aspects of the system operating properly, however the reality is that failures will occur, especially in large-scale AI clusters with thousands of systems, NICs, and XPUs all interconnected by fragile optical cables. Performance of any AI cluster is directly affected by any component failure in the holistic system, so peak performance must also be measured by efficiency of failure recovery to return the cluster back to optimum operation. Arista's Etherlink shines here too, with up to 30x faster convergence for failure recovery than InfiniBand.

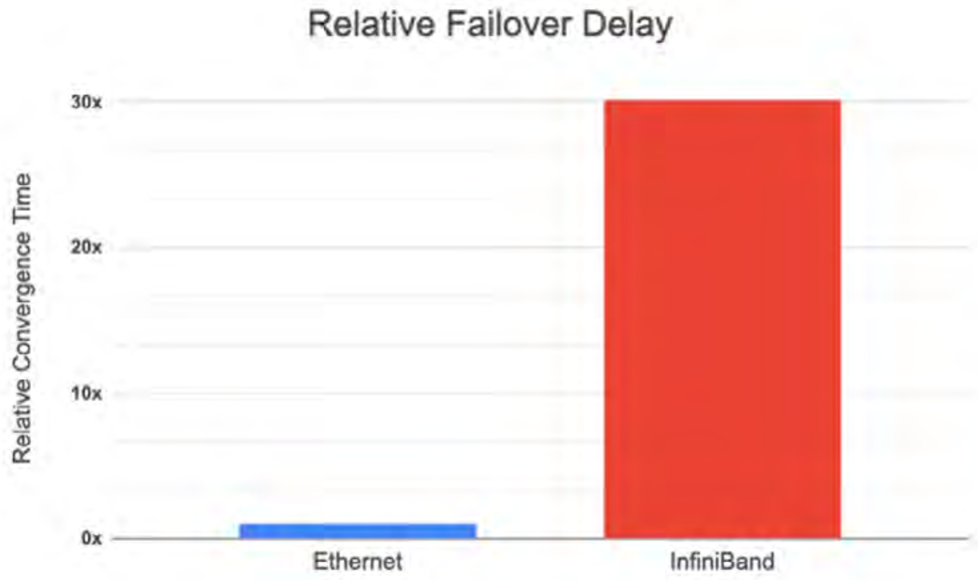

Figure 3: Arista Etherlink AI platform efficiency for failure convergence compared to InfiniBand

- Utmost flexibility and choice: Etherlink enables AI workloads of any size, and being completely based on open standards, it interoperates with any XPU accelerator, NIC, and workload. All Etherlink systems are fully forward-compatible for new Ultra Ethernet Consortium (UEC) enhancements. Further, Etherlink can be deployed in single-tier designs with just a single fixed system or modular chassis, in multi-tier leaf/spine or planar designs with in turn can be flexibly designed with fixed and/or modular systems in the leaf or spine roles, or in an innovative new single-hop distributed Etherlink switching (DES) system.
- Ethernet at scale: Etherlink offers the broadest flexibility of options with the highest scalability in the market, ranging from fixed and modular standalone systems, to single-hop distributed Etherlink switching, to massive-scale leaf/spine and planar designs with both fixed and modular system options.
  - Single systems: up to 576 ports of 800G or 1,152 ports of 400G with a single modular 7800R4 series system, or up to 64 ports of 800G or 128 ports of 400G with a single fixed 7060X6 series system.
  - Single-hop distributed Etherlink switching: up to 16k ports of 800G or 32k ports of 400G with the 7700R4 series system.
  - Two- and three-tier leaf/spine or planar topologies: over 100k XPUs with modular 7800R4 series systems in both leaf and spine roles, with lower scale options available in parallel using fixed 7060X6 series systems as the leaf and/or spine.
- Optimal power consumption: Etherlink enables significant power reduction for holistic AI clusters at both the system level and the optical interconnect level. Arista's Etherlink systems leverage best-in-class 5nm silicon to consume at least 25% lower power per Gb than prior generations of 7nm or older platforms. Further, Etherlink systems can use Linear-drive Passive Optics (LPO) to further reduce net power consumption by 50% compared to traditional pluggable optics and Active Optical Cables (AOCs). Taken individually, each of these solutions can provide compelling power reduction, but when combined, Etherlink delivers the lowest power consumption overall for AI workloads thus freeing up valuable resources for additional compute capacity.

- Sustainable quality: Etherlink is based on the high-quality EOS operating system, which yields the most rapid introduction of new AI networks for reduced time-to-train. EOS has demonstrably fewer Common Vulnerabilities and Exposures (CVEs) and defects than alternative operating systems on the market, especially compared to open source operating systems.
- Coordinated control and visibility: Etherlink provides networking teams with critical end-to-end insight and comprehensive control to ensure optimized performance across networks to NICs inside compute nodes. Often network and server teams each have only partial views of the overall AI solution performance, with each team managing configuration of their respective domains in isolation. This can lead to inadvertent configuration mismatches between QoS settings on NICs vs. the network which can lead to performance issues across the entire AI cluster. Arista's EOS-based AI Agent, an important part of the holistic Etherlink solution, helps avoid such issues to deliver optimized end-to-end AI performance by providing coordinated configuration and unified visibility across network and compute domains.
- UEC-ready now: while Etherlink utilizes well-proven, standards-based Ethernet from the past, it is also ready for the future with compatibility and upgradability to new features from the Ultra Ethernet Consortium (UEC). Ethernet enables a number of key advantages relative to alternate technologies such as Infiniband, offering high-volume deployments from a rich ecosystem of suppliers which yield cost advantages inclusive of optics and cables, as well as a long history of technology innovation to introduce features suitable for new use cases. Etherlink yields up to a 10% performance improvement for AI workloads compared to Infiniband today, and is ready now for upgrade to new UEC specifications to further the Ethernet performance advantage.

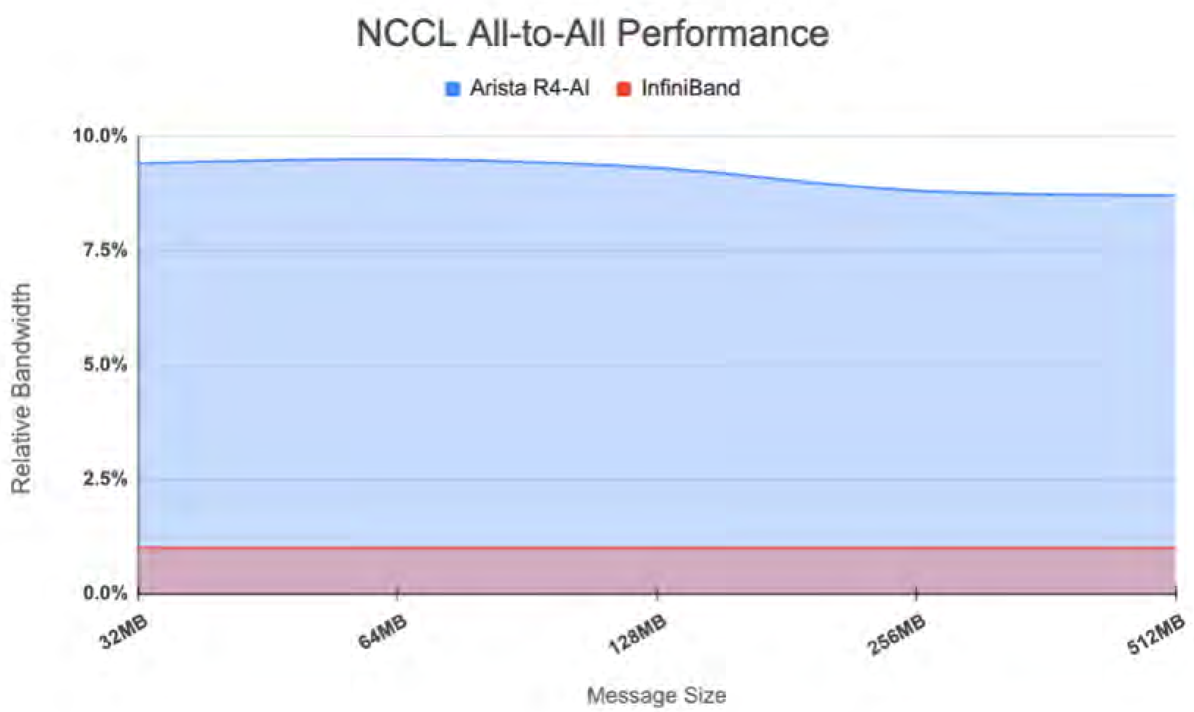

Figure 4: Arista Etherlink AI platform performance compared to InfiniBand performance for AI

Arista Etherlink AI 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. Arista Networks leverages best-in-class silicon packet processors to offer a full range of hardware systems optimized for any size of AI network. Arista offers fixed and modular systems, usable both standalone and in spine/leaf AI topologies, as well as distributed Ethernet switching for single-hop AI networks.

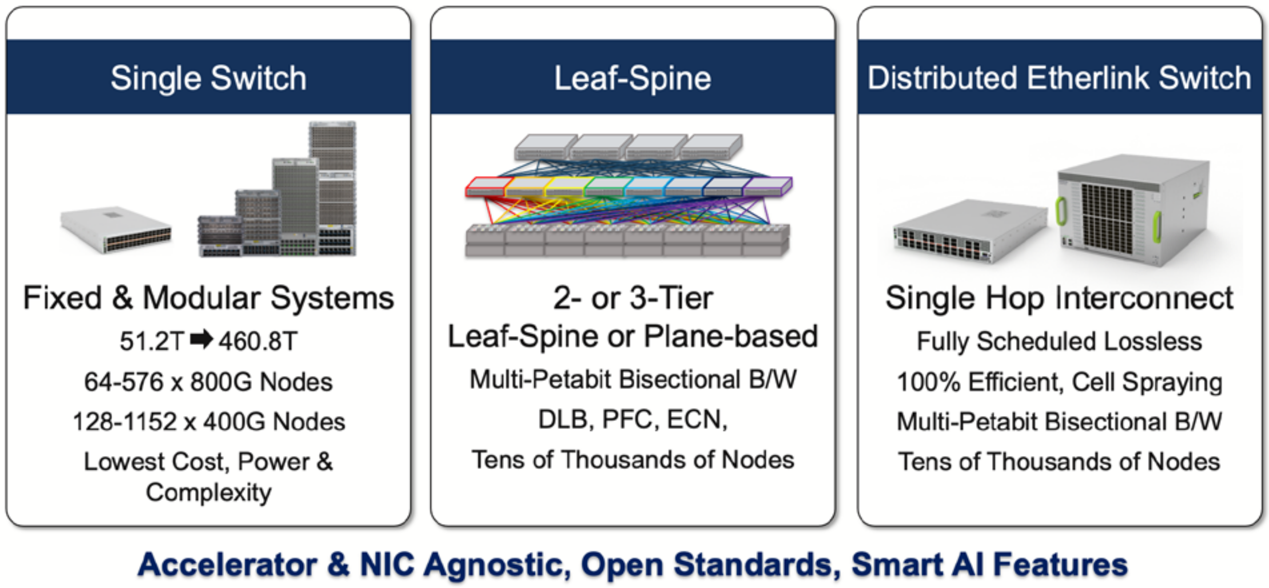

Figure 5: Arista Etherlink AI portfolio

Arista offers a full range of AI-optimized platforms as part of the Etherlink portfolio: the 7060X series fixed platforms, the 7800R modular chassis, and the 7700R distributed Etherlink switch.

7060X series: fixed AI leaf

The 7060X series deliver high-capacity, low-latency Ethernet switching in fixed form factors, ideal for use in a leaf role in high-scale AI clusters. The 7060X6 series platform is based on the latest Tomahawk 5 silicon from Broadcom, while the 7060X5 series platform is based on Tomahawk 4 silicon.

The 7060X6 series comes in two variants, offering either 51.2T of capacity with 64 ports of 800G (or 128 ports of 400G) in a 2RU form factor, or 25.6T of capacity with 32 ports of 800G (or 64 ports of 400G) in a 1RU form factor. Both systems can aggregate 200G, 100G, and 50G ports as well using breakout cables.

The 7060X6 series delivers a breakthrough in reduced power consumption for AI networking. Not only does the system consume 25% lower power per Gbps of capacity than the prior generation, but it also supports new Linear-drive Passive Optics (LPO) which can reduce net power consumption by another 50% compared to traditional optics.

As a complementary option, the 7060X5 series provides 25.6T of capacity with up to 32 ports of 800G or 64 ports of 400G in 1RU or 2RU fixed form factors.

7800R series: modular AI spine

The Arista 7800R series modular systems deliver up to 460 Tbps of capacity to meet the needs of the most demanding AI workloads and modern multiservice routing and switching use cases. 7800R systems are available in 4-slot, 8-slot, 12-slot, or 16-slot modular chassis options. In the 16-slot chassis, up to 576 ports of 800G or 1,152 ports of 400G are supported to maximize single-system density for small to medium AI workloads or as a spine in a large AI spine/leaf cluster.

The 7800R series is based on a 100% fair, cell-based switching fabric that provides maximum efficiency and built-in fabric speedup for integrated over-provisioning of ingress network ports and an overall non-blocking architecture. AI workload performance is optimized via the lossless, fully scheduled, virtual output queuing (VOQ) design inherent to the forwarding pipeline and fabric, coupled with deep packet buffers to avoid congestion and packet loss.

The 7800R4 AI-optimized linecard provides 28.8 Tbps of capacity with 36x 800G OSFP ports, based on a pair of Jericho 3-AI silicon packet processors. This linecard offers a low latency packet pipeline focused solely on AI workload needs, coupling VOQs with 32G of deep packet buffering to a streamlined feature set and scale optimized for AI. As a result, the 7800R4 AI linecard achieves very low power consumption, which can be further amplified with low power linear-drive passive optics (LPO).

The 7800R also supports 7800R3A 400G-optimized linecards, based on a pair of Jericho 2C+ silicon packet processors and offering 14.4 Tbps of capacity with 36x 400G OSFP or QSFP-DD ports.

Figure 6: Arista Etherlink AI platforms; 7060X series, 7800R series, and 7700R series

7700R4: distributed Etherlink switch

The Arista 7700R Distributed Etherlink Switch (DES) represents a new paradigm in AI-centric networking, delivering leaf and spine scalability while operating as a single logical system and presenting a single network hop to the compute cluster. Much as a 7800R4 modular chassis links linecards together with an intermediate switch fabric, so does the 7700R4 DES offer distributed, fixed leaf platforms that forward traffic to each other via dedicated switch fabric systems. Both the 7800R4 and 7700R4 share a common architecture with VOQs, deep packet buffering, and lossless cell-based fabrics.

The 7700R4 DES system might look similar to a traditional leaf/spine topology, and indeed, the cabling interconnects are similar. However, the 7700R4 enables a single-hop forwarding paradigm, distinct from leaf/spine designs which require three-hop forwarding since both the ingress and egress leaf nodes plus the spine all have their own forwarding processor. As a result, the entire 7700R4 DES system is managed as a single logical, fully-scheduled cluster and delivers 100% fair, lossless transport between all nodes in the system. There is no need for QoS features typically needed for AI networks inside the DES fabric, such as PFC/ECN or other RDMA-aware load balancing features, as DES is auto-tuned for 100% efficiency on day one.

Each fixed leaf node in the DES system offers 18 ports of 800G for AI accelerator connectivity, plus 20 ports of dedicated 800G ports for uplink to the central switch fabrics, each of which hosts 128 ports of 800G connectivity to aggregate multiple leaf nodes. The DES system thus automatically provides fabric overprovisioning for redundancy and overspeed for traffic management, with automated 100ms detection and rerouting around link or system failure in the cluster. Each DES leaf router is based on a Jericho 3-AI silicon packet processor with dedicated VOQs and deep packet buffers, with similar low latency and low power consumption characteristics as the J3-AI linecard in the 7800R4 modular chassis. Also as with the J3-AI linecard, the 7700R4 DES system supports linear-drive passive optics (LPO) for power savings of as much as 50% compared to traditional pluggable optics.

Arista EOS AI Software Suite

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. Arista EOS® (Extensible Operating System) 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 Converged Ethernet (RoCEv2). Without visibility into network traffic and buffer utilization, configuring appropriate PFC and ECN thresholds can be challenging. EOS 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.

Arista EOS-based AI Agent

AI workloads require optimized performance and availability at all times, to minimize job completion time and thus maximize utilization of expensive XPU accelerators. There is zero tolerance for misconfigurations or finger-pointing between network and server operations teams if problems arise.

Arista provides coordinated performance optimization between the networking and compute domains, along with unified visibility across the entire ecosystem to pinpoint areas to optimize or otherwise improve. The EOS-based AI Agent can reside either directly on a SmartNIC or else on a server CPU, to provide local configuration management of NICs along with streaming telemetry of NIC performance fed to directly-attached Arista EOS-based switches. This ensures the QoS parameters for AI optimization are consistently applied from the NIC to the network alike, to avoid misconfigurations which might cause performance bottlenecks without an easy-to-diagnose root cause. And with telemetry data spanning the AI NICs and the AI networking platforms, the network

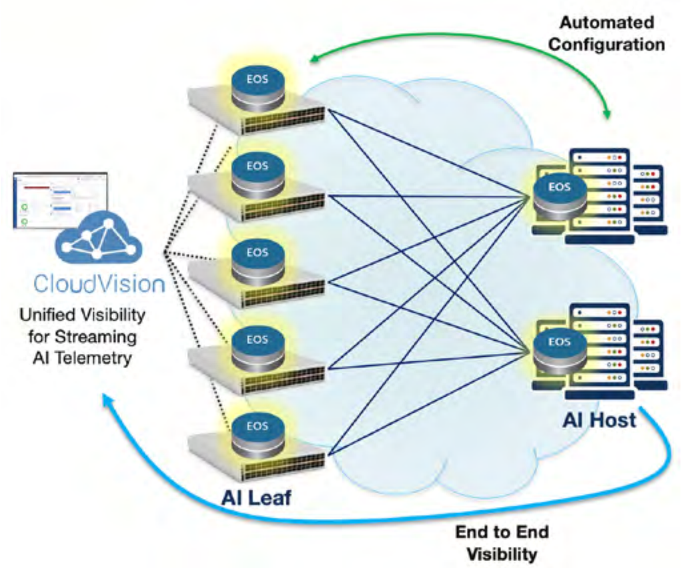

Figure 7: Arista EOS-based AI Agent with unified visibility

operations team can have comprehensive visibility into the entire traffic path with immediate insight into performance and problems.

Arista AI Observability

It's difficult to optimize AI networks without first knowing the state of that AI network, with real-time data gathered via streaming telemetry from all parts of the AI ecosystem and interrogated for insights. The more that is known, the better decisions can be made to further optimize the entire workload and infrastructure in tandem.

Arista's AI observability is designed on open standards, and is intended to present comprehensive information to network operations, sharing the right data in real time to unlock the right insights. This solution includes:

- Streaming data from Arista EOS-based AI platforms, including fixed, modular, and DES platform variants, and also from compute nodes via EOS-based AI Agents hosted on SmartNICs and/or server CPUs.
- The Arista NetDL network data lake, providing a central repository for all streaming data from all managed devices. NetDL provides a single data source of truth for analytics and forensic examination for performance monitoring, among other uses.
- Arista's AI Analyzer and Latency Analyzer features in EOS to provide real-time, detailed insight to traffic statistics, interface congestion, and queuing latency metrics at microsecond intervals to correlate AI workload behavior with network characteristics.
- The Arista Cloudvision network automation platform provides visibility into network and compute topology, health, and configurations.
- Arista's Autonomous Virtual Assistant, or AVA, works with Cloudvision and data stored in NetDL to offer a conversational assistant to network operators. This allows them to ask questions in natural language to query status of the end-to-end AI ecosystem and glean insights to potential sources of disruption.

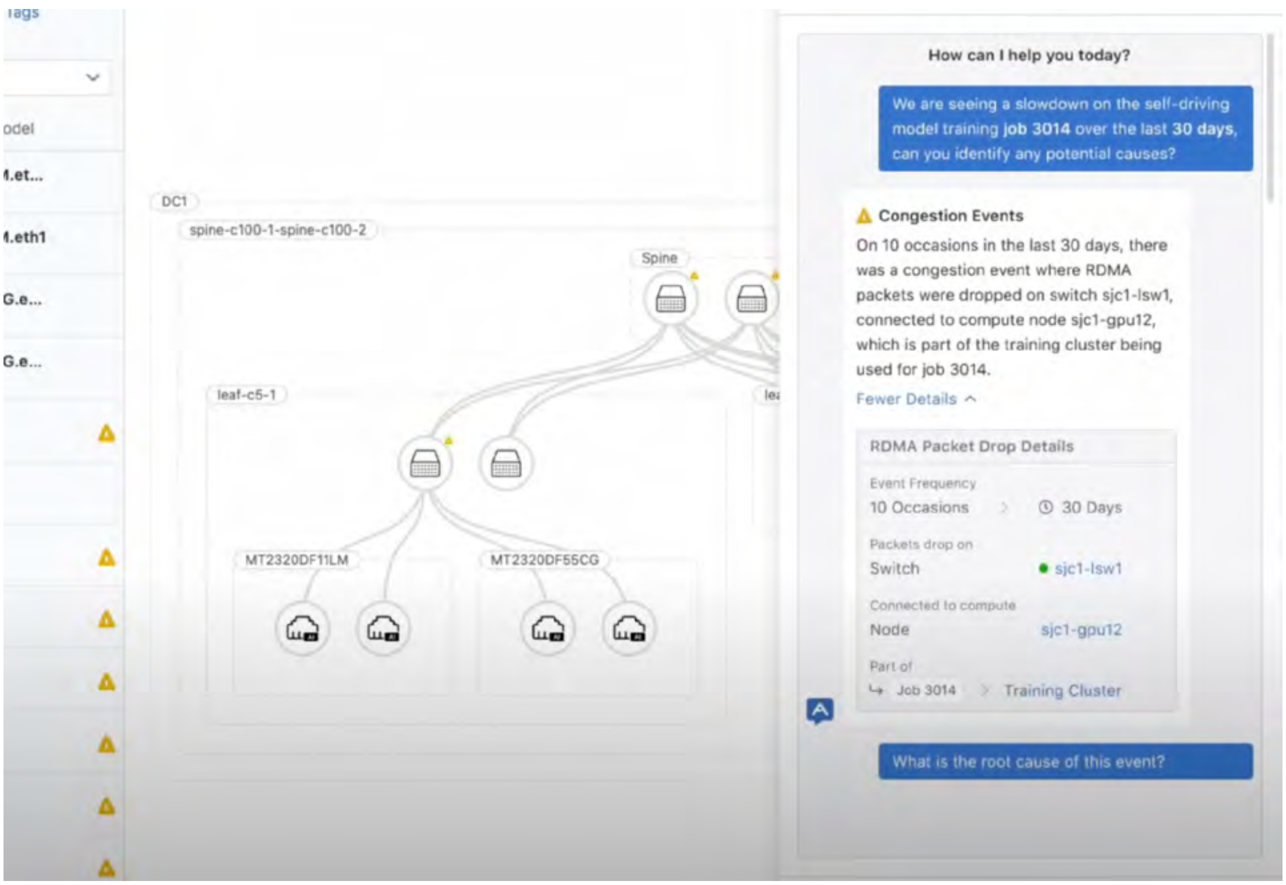

Figure 8: Arista AVA shown diagnosing AI congestion issues spanning NICs to network switches

AI Center Design Options at Scale

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.

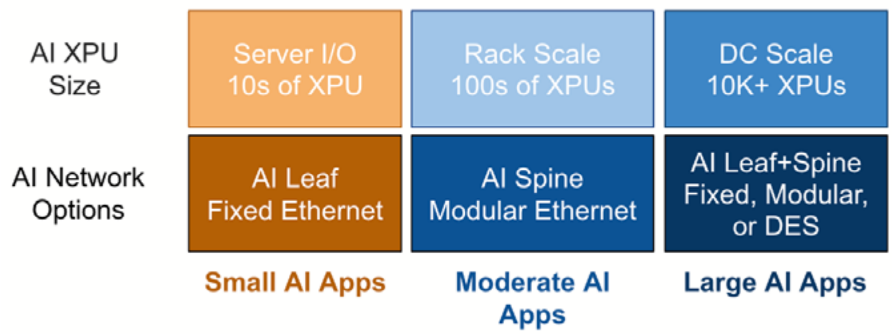

Figure 9: AI Network Design Guidelines

Low-scale AI Centers: Single Fixed AI Platform for 10s of XPUs

A single Arista 7060X6-64PE switch with 64x 800G ports or 128x 400G 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.

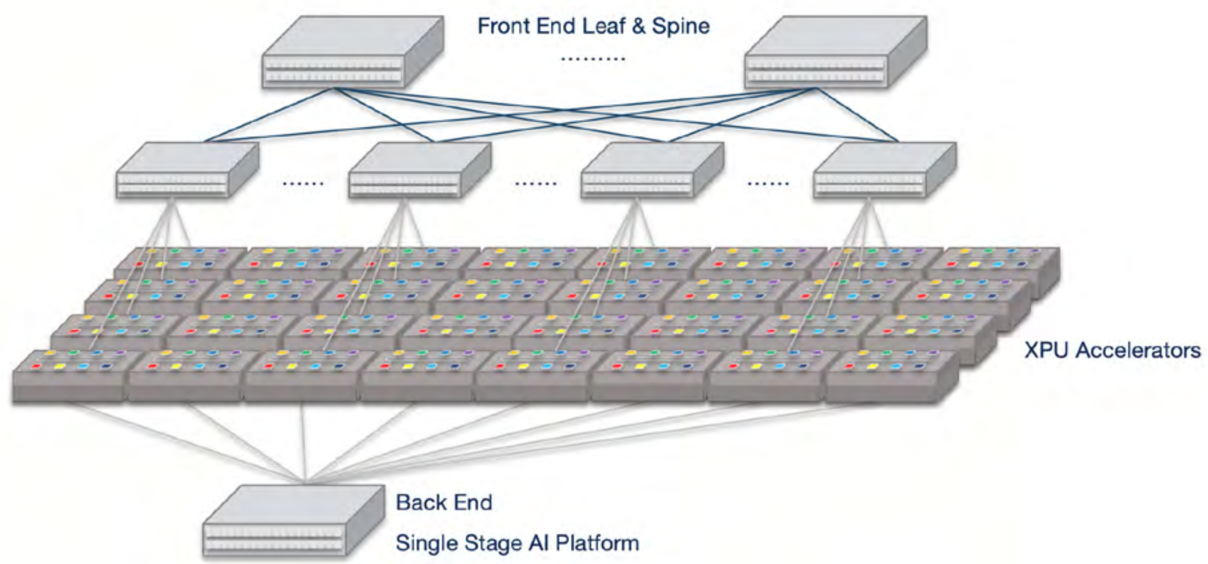

Figure 10: Low-scale AI Center, Single Stage based on one Arista 7060X6 fixed AI Platform

Moderate-scale AI Centers: Single Modular AI Platform for 100s of XPUs

A single Arista 7800R4 chassis with support for 576 x 800G or 1,152x 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 7800R4 is based on a cell-based, non-blocking VOQ architecture, and thus enables a 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.

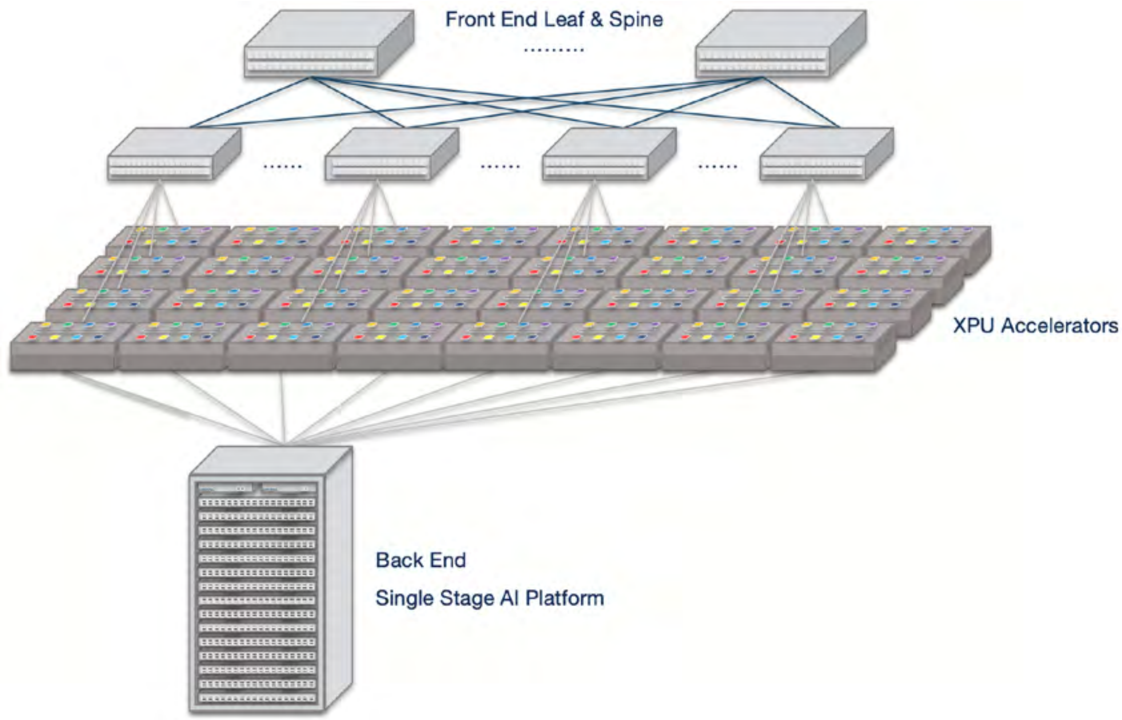

Figure 11: Moderate-scale AI Center, Single Stage based on one Arista 7800R modular chassis

Large-scale AI Centers: Single-Hop with Distributed Etherlink Switch AI Platforms for 1000s of XPUs

A new AI design option has emerged to pair large-scale aggregation of as many as 32k XPU accelerators with simplified operations without sacrificing any performance: distributed Etherlink switching (DES). This innovative design offers the scale of traditional leaf/spine networks but is managed as a single logical cluster, presenting just a single hop path to attached hosts. DES is optimized for universal 800G links, including optional support for 2x 400G host connectivity per port with different optics, including inherent fabric overprovisioning to manage transient bandwidth bursts. Because DES is managed as a single cluster, all forwarding decisions happen at the leaf layer, and the cell-based fabric offers automatic 100% fair, lossless transport and 100ms failure detection and recovery.

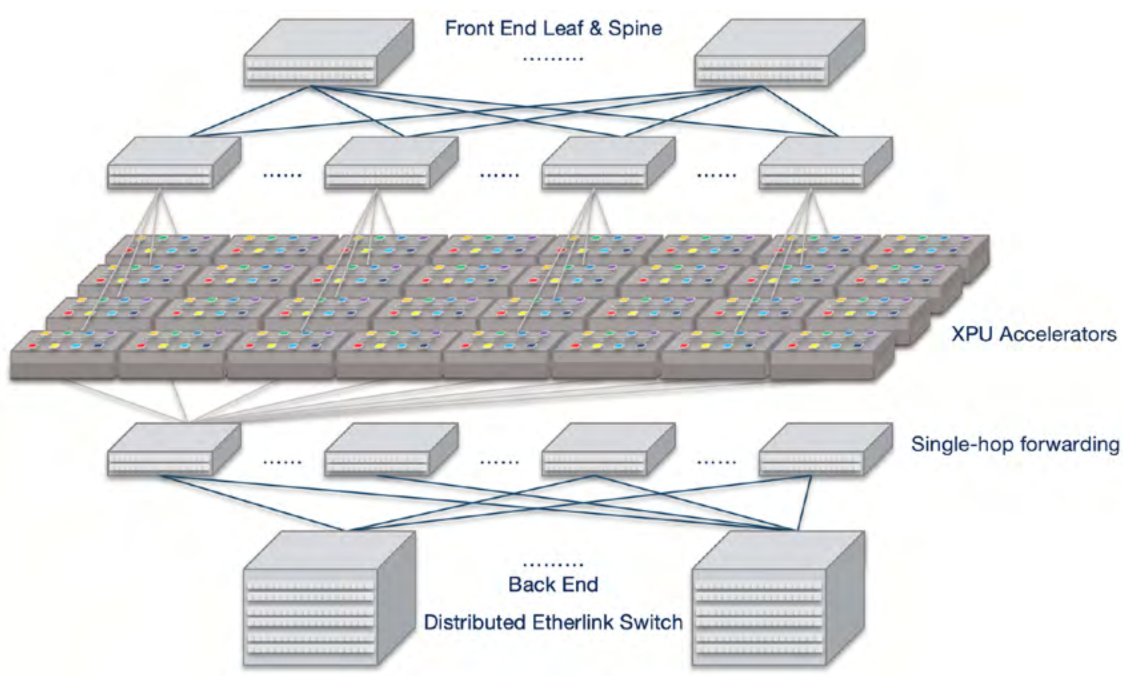

Figure 12: Large-scale AI Center, Single-hop design with 7700R4 Distributed Etherlink Switch

Ultra large-scale AI Centers: Multi-stage Leaf/Spine AI Platforms for 100,000+ XPUs

For ultra large-scale AI applications, requiring many tens of thousands of XPUs - or even as many as 100k parallel XPUs - to be connected in data centers, a scale-out leaf/spine design 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 100,000 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.

Depending on how many tens of thousands of GPUs are required for a given AI cluster, AI leaf options could span from fixed AI platforms to high-capacity modular chassis. XPU density will be maximized with modular platforms in both AI leaf and spine roles.

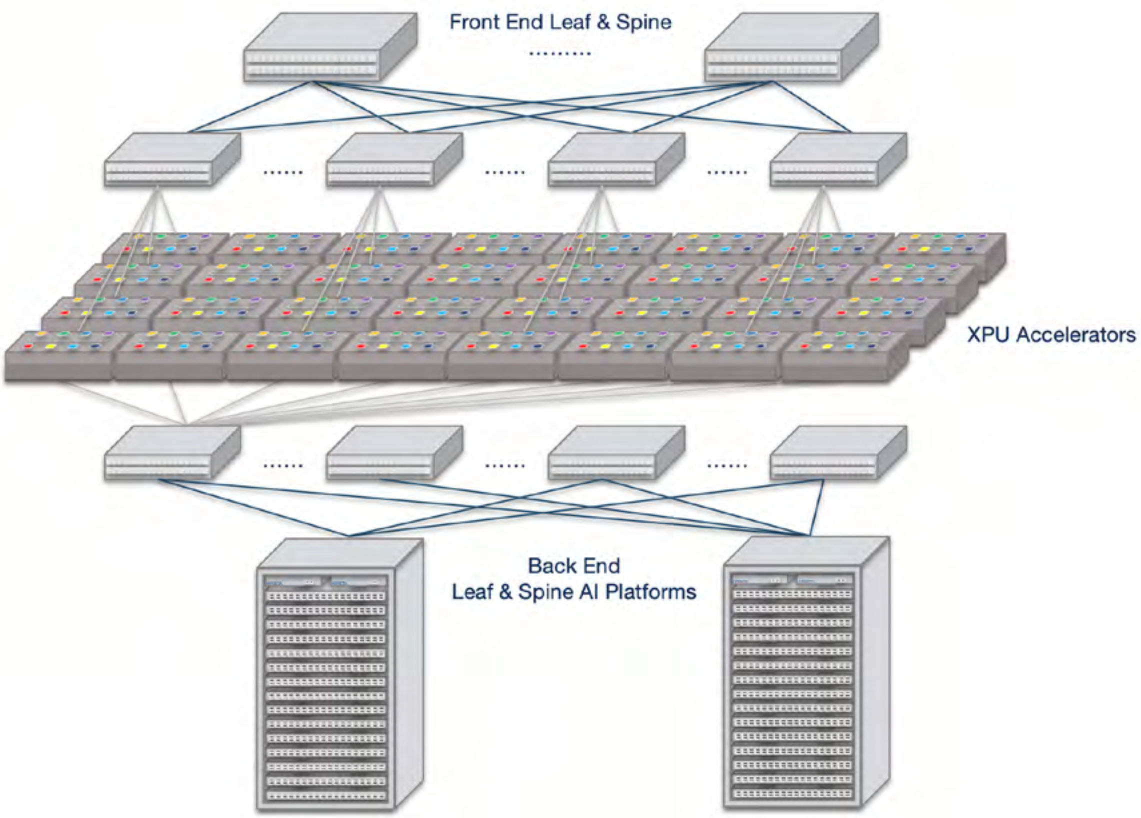

Figure 13: Ultra large-scale AI Center, Multi-stage; Arista 7800R AI spines + 7060X AI leaves

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.

Conclusion

Arista's Etherlink AI platform delivers a holistic solution for optimized AI networking, greater as a coordinated total than individual parts on their own. Etherlink comprises best of breed products spanning AI-optimized Ethernet hardware systems, EOS-based software with leadership AI congestion avoidance features, EOS-based AI agents to coordinate networking with NICs, and end-to-end observability for AI infrastructure. Etherlink is based on Ethernet, which is the optimal choice for building high-performance AI clusters today, and will continue to evolve with new enhancements from the UEC. Ethernet delivers an incremental performance improvement over Infiniband for AI workloads, and Arista's Etherlink AI platform delivers up to 65% improvement over plain vanilla Ethernet solutions with superior fault recovery to boot.

The AI Center offers an opportunity to unify the entire networking ecosystem, bringing together the new AI back-end network with existing data center systems and infrastructure. Consistency with Ethernet technology offers the opportunity to align operations with coordinated skill sets across the entire organization. Then with unified networks and unified operations, it becomes easier to integrate with other systems in the existing data center, glean new insights from the network to compute nodes with coordinated observability, and to optimize performance for AI.

Reference

- Blog: The Arrival of Open AI Networking, by Jayshree Ullal
- Blog: The New AI Era: Networking for AI and AI for Networking, by Jayshree Ullal
- Blog: The New Era of AI Centers, by Jayshree Ullal

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