Arista 7060X6 Series:
Accelerated DC and AI Networking at 800G

Introduction

As Artificial Intelligence (AI) models continue to grow in scale, capability and complexity, new demands are being placed on both the compute and network infrastructure supporting the clusters. Optimizing the network for distributed computing to support large language AI models, generative AI, training data, inference and storage access along with higher performance XPU accelerators has increased network scale and interconnect bandwidth requirements; underscoring the need for robust, reliable, lossless and scalable networking for AI in both the backend and front end network fabric. The network infrastructure is critical to support the compute capacity, minimize Job Completion Times (JCT) for AI training models and ensure efficient utilization of computational resources and work cycles.

Front-end Datacenter networks are also witnessing a growth in total capacity, while being constrained by physical space, finite power and cooling combined with the need to leverage consistent architectures across all layers of a network. These challenges are addressed through the use of higher capacity systems that consume less power per Gbps, allow consolidation of multiple devices and are available in flexible data center optimized solutions.

The 800G Arista 7060X6 switch series are an integral component of the Arista AI Etherlink portfolio, and are optimized to meet the throughput, scale and performance demands of the rapidly evolving challenges from AI and Datacenter networks. Offering up to 64 ports of 800G or 128 ports of 400G in a compact 2 RU form factor, the Arista 7060X6 series provides 51.2 Tbps of throughput (102.4 Tbps full-duplex) while supporting a suite of features designed to maximize the performance and simplify the management of the most demanding AI networks. Additionally, the fully shared buffer architecture lends itself perfectly to absorb microburst traffic patterns that are typical of AI workloads. Support for advanced telemetry, monitoring and rich instrumentation enables the users to monitor and analyze their network, and give them the ability to proactively or reactively remediate any possible issues. These systems also support the Linear Pluggable Optics (LPOs), which can reduce the total system power consumption by up to 50%, and its new chassis design raises the bar for serviceability and reliability.

This whitepaper explores the Arista 7060X6 product family in detail. It covers forwarding architecture and hardware capabilities, as well as its unique features designed to ensure seamless execution of AI workloads.

800G Ethernet & AI Workloads

AI workloads are extremely compute and data intensive given the trillions of parameters that have to be computed, and the computation results that have to travel across the cluster. The data sets that these AI models have to process are so large, that computations are run in parallel to reduce overall completion time. Some of the typical approaches to achieve this involve data parallelism, pipeline parallelism or tensor parallelism over hundreds or thousands of processors. At each stage, each processor performs extensive computation on its data by using widely available collective operations, and then shares the results with other nodes in the cluster. Since the collective calculation is widely distributed across the cluster, the next step cannot begin until all processors have a unified view of the data, and thus an AI cluster is only as fast as the slowest task being completed.

Consequently, AI workloads put unique demands on the network infrastructure, requiring lossless transfer of highly correlated traffic flows with high levels of uptime and reliability. This high level of throughput is achieved by transferring data between memory on accelerators using RDMA (Remote Direct Memory Access). With RDMA, the transfer rate is limited by the lowest bandwidth among the accelerator memories, internal buses, and network interfaces. To optimize these resources, the bottleneck in an ideal AI training job should be the computation, and not any link contention within the RDMA chain.

The key elements from a network perspective in any AI cluster synthesize down to being able to cope with synchronized, low entropy, bursty traffic patterns, and avoiding link contention and congestion issues. The nature of low entropy traffic flows renders traditional hashing protocols to be suboptimal, and necessitates newer and advanced methods of load balancing techniques, that are supported on the Arista 7060X6 series. We will take a closer look at these mechanisms later in this whitepaper.

Today's accelerators support 400Gbps of throughput, with individual hosts often reaching 3.2 Tbps across a cluster of 8 accelerators. This unprecedented amount of bandwidth necessitates denser, higher bandwidth networks. 800G Ethernet is ideally suited to this challenge.

A single 800G Ethernet port can be broken out into 4x 200G ports for wider reach in inference deployments or lower-bandwidth training scenarios, or into 2x 400G for connecting to accelerators. These ports can also be broken out into 8x 100G, to provide backward compatibility as well as support the maximum wide radix in typical deployments. 800G is also the natural step for the next generation of accelerators themselves.

Ethernet's proven, distributed control-plane protocols scale to millions of nodes; and the Ethernet based Arista 7060X6 series brings the density as well as the feature set to the most demanding AI deployments.

Arista 7060X6 Overview

The 7060X6 is the latest generation of Arista's 7060 product family of high performance AI and Datacenter switches. Part of the Arista AI Etherlink portfolio, this platform series has doubled the bandwidth of its predecessor, the Arista 7060X5 series, to 51.2 Tbps. It builds upon prior iterations of systems offering up to 64 ports of 800G, with larger fully-shared buffers of 165 MB, dedicated resources for advanced monitoring and telemetry, low latency, a consistent pipeline and feature set, and a variety of new enhancements to improve the performance and efficiency of AI workloads. Some broad categories of enhancements to power AI workloads include, among other features, cognitive routing, advanced queueing and congestion control, a suite of load balancing offerings, and fast link failover - all of which contribute to making the AI clusters more scalable and reliable.

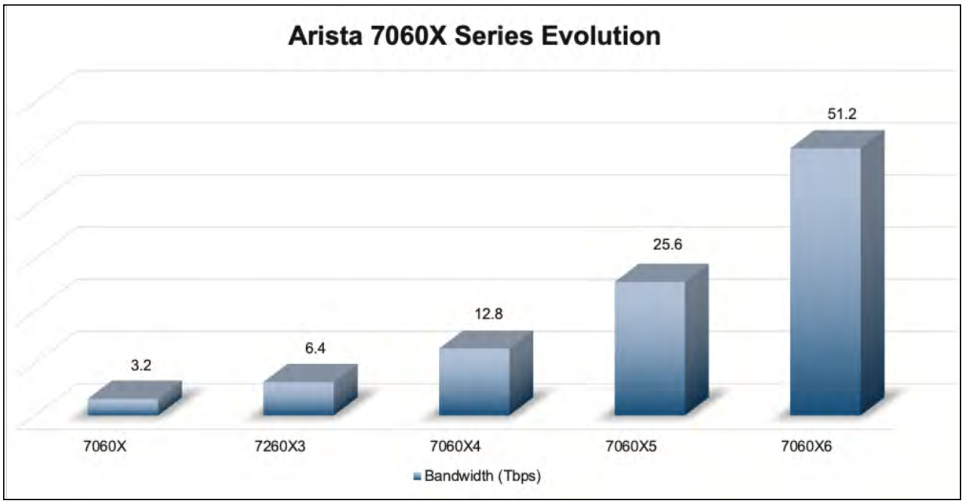

Figure 1: Bandwidth increase offered by the Arista 7060X series over the generations

The 7060X6 series platforms are also Ultra Ethernet Consortium (UEC) ready. This support will further enhance the RDMA properties on the 7060X6 series, and also allow for higher scalability in AI deployments, by tuning across multiple layers to enhance both AI and high-performance computing (HPC) workloads. AI-optimized, configuration free congestion control mechanisms like Incast management and support for out of order packet delivery will also be desirable effects born out of UEC compatibility.

Platform Choices

There are two main options available within the Arista 7060X6 series - a 64x 800G port option, which is the DCS-7060X6-64PE and a 32x 800G port option, which is the DCS-7060X6-32PE. The Arista 7060X6 series of switches are available with Octal Small Form Factor Pluggable (OSFP) form factor ports for 800G connectivity. Compared to the QSFP-DD form-factor, OSFP has improved thermal performance at higher speeds like 800G owing to its integrated heat sink design, and can support denser, dual-LC optics commonly used in 800G break-out applications.

Figure 2: DCS-7060X6-64PE and DCS-7060X6-32PE options available within the 7060X6 series

Table 1: Ordering SKUs for the Arista 7060X6 series

Description	Ordering SKU
64 x 800 GbE ports (OSFP)	DCS-7060X6-64PE
32 x 800 GbE ports (OSFP)	DCS-7060X6-32PE

The table below provides a side-by-side comparison of the options available in the Arista 7060X6 portfolio:

Table 2: Comparison between the DCS-7060X6-64PE and DCS-7060X6-32PE

Specification	DCS-7060X6-64PE	DCS-7060X6-32PE
Maximum 800G ports	64	32
Maximum 400G ports	128	64
Maximum 200G ports	256	128
Maximum 100G ports	320	256
Latency	From 700 ns	From 700 ns
Packet Buffer	165 MB	84 MB
System Memory	32 GB	32 GB
Flash Storage Memory	240 GB	240 GB
Power Consumption (Typ/Max)	672W/2219W	TBD
Power Supply	PWR-2421-HV-RED	PWR-2011-AC-RED
Fan	FAN-7021H-RED	FAN-7011H-F
Rack Units (RU)	2	1

Forwarding Architecture - A Day in the Life of a Packet

The Arista 7060X6 series leverages the Broadcom Tomahawk5 chipset. The packet forwarding pipelines can be broadly classified as the Ingress Pipelines and the Egress Pipelines, both of which interface with a consolidated Memory Management Unit (MMU).

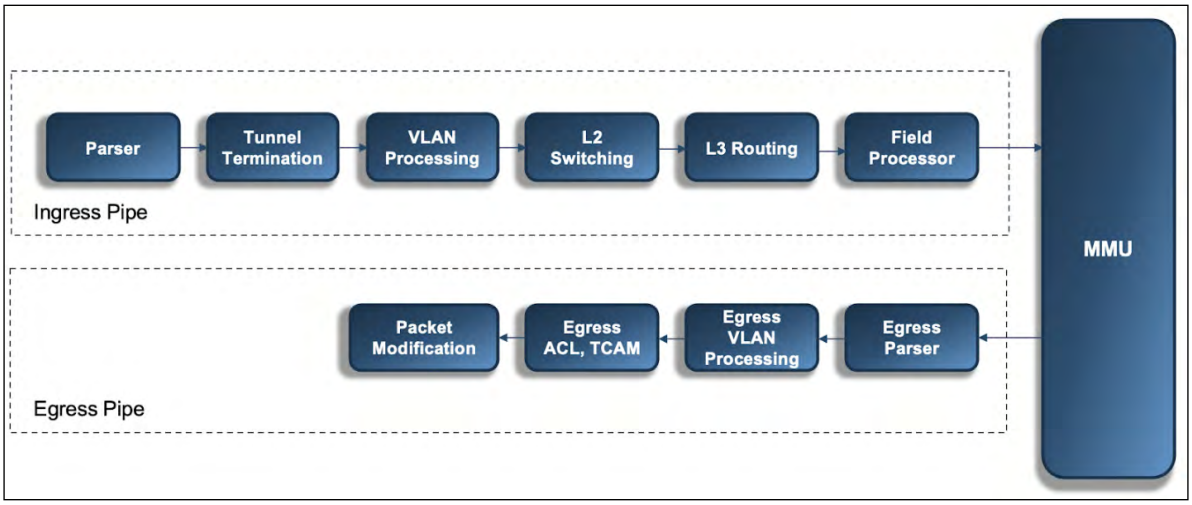

Figure 3: Packet processing pipeline inside the 7060X6 systems

- The Parser block examines the ingress packets from all the front panel ports, and extracts all the necessary information from the packet headers. This information is then stored for the reference of the various packet processing engines that might require it.
- The Tunnel Termination block determines whether any tunneled packets (MPLS, GRE, etc.) need to be terminated on that particular device, and the packet is forwarded internally based on this determination. As the name suggests, the VLAN Processing block filters any VLAN packet contents present in the packet, and is integral in identifying the L2 identity of any packet.
- Next, the L2 Switching block performs lookups like the MAC Source address lookup and the MAC destination address lookup. This enables the device to populate or refresh L2 tables.
- If the ingressed packet has an L3 header, the L3 Routing block performs source and destination based IP lookups for IPv4 and IPv6 packets, including Unicast and Multicast forwarding.
- The Field Processor block performs packet filtering based on the contents of the standard packet fields like source MAC address, destination MAC address, source IP address, destination IP address, TCP header fields, etc. before handing off the packet to the MMU.
- After the MMU stage, the packet is then transferred to the Egress pipeline, where further processing and packet modifications take place on the packet, before it exits out of the system.

Fully Shared Buffer Allocation

The 7060X6 system has 165MB of fully shared buffer. The architecture comprises of 32 Ingress Pipes (IP0 - IP31) that connect to 32 Egress Buffers (EB0 - EB31), split across two Ingress Traffic Managers (ITM0 and ITM1). The Scheduler ensures that packet priorities are appropriately honored.

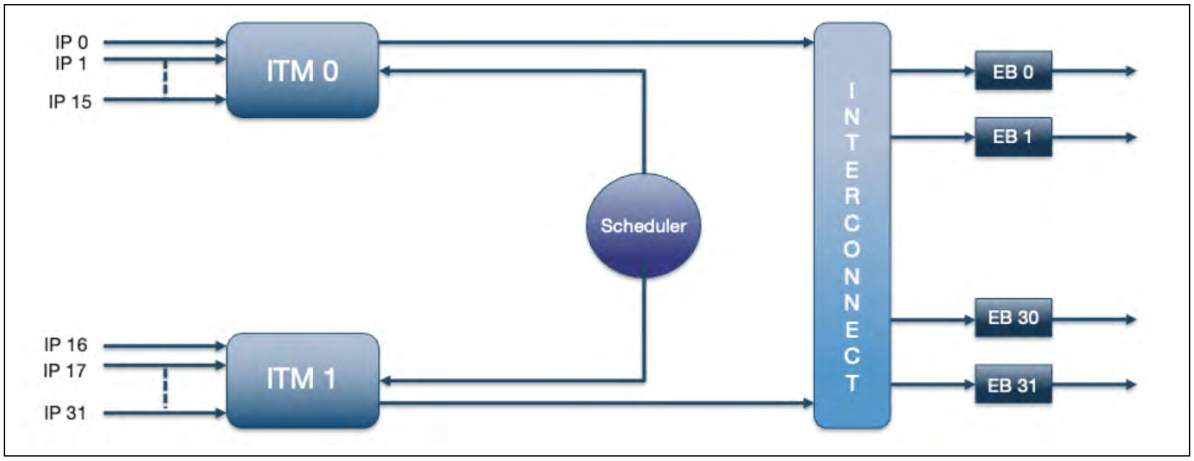

Figure 4: Connectivity between Ingress Pipes and Egress Buffers

The system memory on the 7060X6 systems is split into three main types: Reserved Memory, Shared Memory, and Headroom Memory.

Reserved Memory

This memory is reserved on a per-port per-priority group/queue basis, and can only be used for the entity it was reserved for.

Shared Memory

When the reserved memory is used up, this chunk of memory is used, with thresholds determined dynamically or based on user-tunable parameters.

Headroom Memory

This memory is used for lossless traffic classes, when Priority Flow Control (PFC) is enabled. This chunk of memory is used to absorb packets in-flight when all the shared memory for an ingress port's priority-group is exhausted, until the sender reacts to Pause / PFC frames. This becomes especially important in the context of AI workloads.

To prevent a single entity from consuming all of the memory in a service pool of the shared memory, the concept of "dynamic memory threshold" is also used, that controls the shared memory usage on a per queue and per priority group of a given port. The amount of memory an entity can consume is based on the current total usage and a user-configurable alpha value that is set per queue/priority group. The net effect being that the higher the demand on the shared memory, the lesser the memory that each individual queue is allowed to use. This results in a fair allocation of memory when multiple queues require access to the buffer, and reduces the need to carve out reserved memory for individual ports.

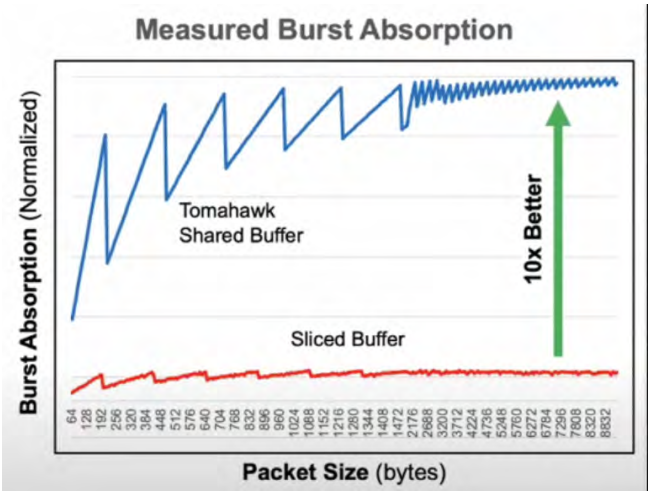

Figure 5: Illustration from Broadcom highlighting better burst utilization from a shared buffer architecture

Studies from Broadcom have shown that the fully shared buffer leveraged by the Arista 7060X6 systems is 10x better at burst absorption (an important parameter in AI workloads) as opposed to a sliced buffer architecture, wherein the buffer is carved out into predefined chunks of memory, and each port only has access to one particular such chunk.

Hardware Design

The Arista 7060X6 series is optimized for power, cooling and system reliability. In this section, we will take a closer look at some aspects that make these systems highly efficient and reliable.

Removable Supervisor

The Arista 7060X6 2RU platforms include replaceable management modules that improve serviceability.

Figure 6: Removable Supervisor module on the 7060X6-64PE

In typical fixed systems devices, components such as the CPU, DRAM, Flash and SSD are common causes of failure modes. On systems with fixed management cards that cannot be removed, the failure of one of these elements requires a customer to replace the entire device. That involves unplugging front panel cabling, removing optics, removing the system from a rack and replacing it with an equivalent device, followed by recabling etc.

This multi-step process is time consuming and introduces the possibility of errors. With the 2RU 7060X6 systems, the management card is a field replaceable element. Customers can easily replace the card without the need to replace the entire system. Additionally, as the SSD and Flash may contain customer configuration information, this allows for sensitive data to be wiped more easily.

Power Optimization

Power is a major component of any Datacenter design, and is particularly challenging from a facilities and cost standpoint in large AI clusters. As the network devices explode in radix and capabilities, it places a heavier load on the power and cooling of the systems. Any power savings directly translate into more capacity for valuable compute resources. The Arista 7060X6's silicon architecture substantially reduces system power requirements, and its higher density can reduce the total number of switches required in a given design, furthering the power savings within a deployment. Furthermore, the use of LPOs (described in more detail below) also results in significant power savings on a system-wide level.

The ability of the 7060X6 systems to support a wide radix alongwith 51.2 Tbps of bandwidth essentially allows it to consolidate within a single device the throughput of 6 equivalent devices of the prior 25.6 Tbps generation. This translates to a 6x reduction in the number of devices needed across the network for achieving the same bandwidth, and a corresponding reduction in the space and power needed. Over the span of the entire network, these power savings quickly add up.

The 6:1 device consolidation also lends itself quite well to AI workloads, as the 6 device cluster to get 51.2 Tbps of throughput would have also added an extra hop for the traffic to traverse through, rendering the network suboptimal in the absence of fine-tuned load balancing parameters.

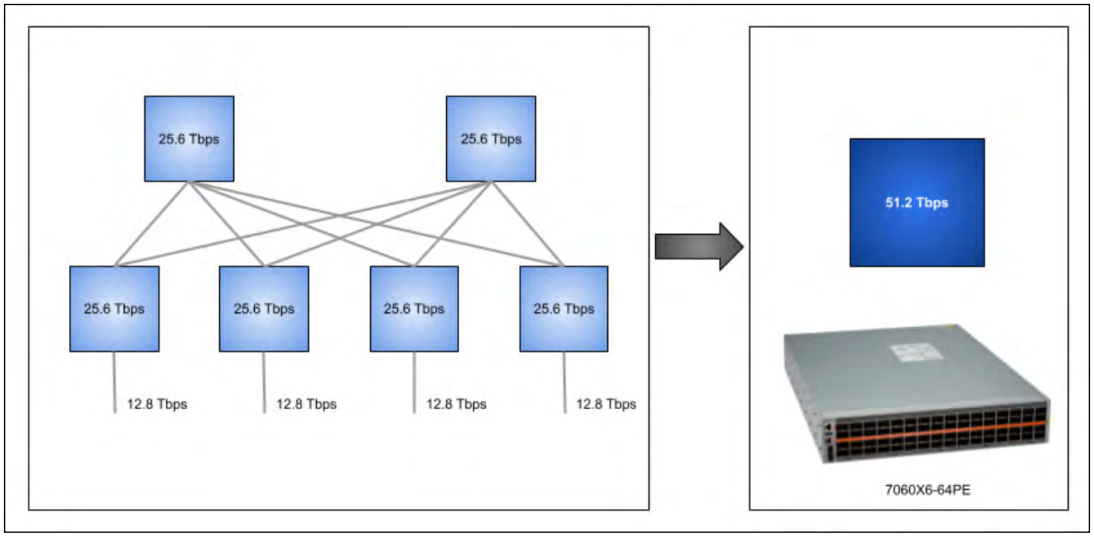

Figure 7: Consolidating a 6 device leaf-spine cluster with a single 7060X6 series device

Linear Pluggable Optics

With the introduction of 800G Ethernet and plans for 1.6T and beyond, the complexity and power requirements of optical transceivers are also escalating. High data rate transceivers (implementing PAM-4) typically contain Digital Signal Processing (DSP) circuitry to ensure the signal reaches the chipset without any meaningful degradation in signal integrity. While DSPs are required for reliable operation, implementing them in the transceiver increases power requirements, latency and module cost.

Linear Pluggable Optics represent an architectural shift in transceiver design. Leveraging the powerful 100G SerDes in the 7060X6 chipset, the DSP logic moves from the transceiver module into the switch chip itself, leaving only linear components on the pluggable optic without compromising on the signal strength and integrity as it moves from the front panel to the switch chip. This improves thermal efficiency, reduces total system power consumption by around 40%, and substantially increases the reliability of each optic. With AI clusters routinely scaling beyond tens of thousands of ports, these efficiency gains have a material impact on cluster capacity and compute utilization.

Latency and Throughput

The Arista 7060X6 systems exhibit cutting edge latency, ideal for deployment in AI environments. Latency is especially critical when you want to ensure shorter Job Completion Times across the entire cluster. The higher radix of up to 64x 800G ports per system provided by the 7060X6 devices, also translates to lower hops the traffic has to take across the cluster.

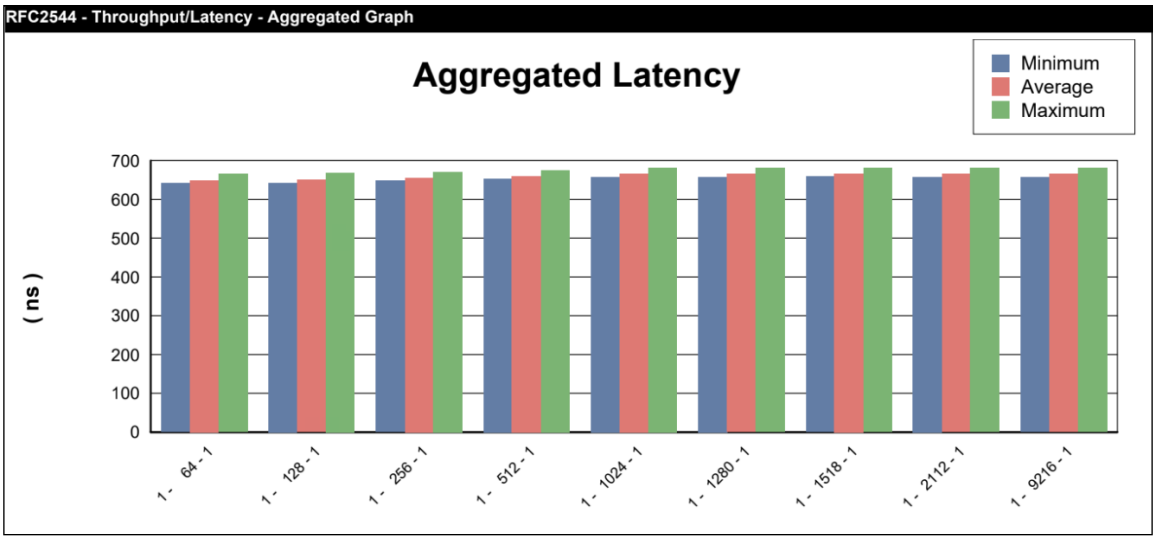

Figure 8: Aggregated Latency on the Arista 7060X6 systems over a range of packet sizes

The 7060X6 systems support both Cut-through and Store-and-forward modes of packet forwarding. While running the industry-standard RFC2544 Latency and Throughput test suites, around 650 ns was typically observed as the system wide latency. Having industry-leading latency metrics results in increased efficiency, scalability and performance in AI deployments.

AI Networking Innovations

The Arista 7060X6 systems bring to the network a highly optimized feature set, rendering these systems to be used as smart switches in the most demanding AI deployments. In this section, we will take a look at some of these innovations and feature sets.

EOS

Arista's Extensible Operating System (EOS®) provides the software foundations for the 7060X6 smart switch platforms. Designed for resiliency and programmability, EOS has a unique multi-process state sharing architecture that separates state information and packet forwarding from protocol processing and application logic. This unique architecture of EOS enables large-scale automation and facilitates the extreme scalability required in hyperscale AI clusters. The same single Arista EOS image runs across the entire portfolio of networking offerings, resulting in peerless operational management and efficiency. With extensibility at every level, no other operating system provides so many tools and methods to extend the power of the network and bring them to the customer's fingertips.

Load Balancing

Highly correlated, line-rate traffic flows are a defining characteristic of AI training jobs. Since incast with multiple line-rate flows will result in buffering and eventually packet loss, effective load balancing strategies are crucial to maximizing job performance, and is one of the most important problems to solve in an AI deployment. The 7060X6 offers a suite of load balancing techniques and enhancements to avoid traffic conflicts and maximize the overall network throughput.

Dynamic Load Balancing (DLB)

Traditional load balancing divides traffic across links based on fields in the packet headers. This works well for many smaller flows distributed across a range of sources and destinations, but hash collisions can result in link oversubscription when other paths have excess capacity. Dynamic Load Balancing (DLB) improves network utilization by choosing paths based on real-time traffic statistics. A new flow will be assigned the path with the lowest utilization across the port group, rather than based on a static hash.

DLB as a feature has been around for a couple of generations of merchant silicon. However, the latest Arista 7060X6 series powered by EOS, provide many new enhancements around DLB as follows:

- Reactive load-balancing (auto-rebalancing): This updates the egress links for active flows when congestion is detected
- Flow Monitoring: Randomly samples packets undergoing DLB and copies them to the CPU. This feature also samples packets to the mirror port or to the CPU when macroflow assignment or reassignment occurs
- Fast Link Failover: Automatically steers traffic around failed links in under 500ns

Source Interface based RDMA Load Balancing

In some planar-based cluster designs, it can be more efficient to connect specific accelerators to shared spine switches rather than using a fully-connected topology. Source Interface RDMA Load Balancing facilitates this by pinning RDMA traffic from an accelerator interface to a specific spine uplink.

Congestion Control

Congestion is also one of the key problems to solve in any AI deployment. As discussed earlier in this whitepaper, a cluster is only as fast as the slowest job getting completed on it. As a result, implementing effective congestion control mechanisms that will ensure no bottlenecks are occurring within the network are important to avoid tail latencies.

The two primary techniques of achieving this are via Priority Flow Control (PFC) and Explicit Congestion Notification (ECN), that provide granular parameters to control to achieve effective congestion control. Together, they are an integral component of the Data Center Quantized Congestion Notification (DCQCN) suite of features. They can be viewed as complementary feature sets, providing mechanisms for both the network and hosts to signal back pressure to avoid packet drops.

Priority Flow Control (PFC)

Priority Flow Control is a link-layer flow control mechanism which may be used by an overwhelmed network node to request the transmitters to stop transmission for a specified period of time. It does so by using special frames known as PFC frames, thus, relieving congestion at the receiver node. The ingress port sends PFC frames to its peer when it encounters congestion and the ingress shared limit is reached, and in response, the peer pauses transmission to alleviate the congestion.

Explicit Congestion Notification (ECN)

Explicit Congestion Notification is an extension to TCP/IP that provides end-to-end notification of impending network congestion, prior to loss. It does so by manipulating bits 0 and 1 (known as the ECN bits) in the ToS byte of the IP header. The ECN bits can have four possible values when considered together:

- 00 - (default) indicates the packet is non-ECN capable
- 01 - indicates the packet is ECN capable
- 10 - indicates the packet is ECN capable
- 11 - indicates Congestion Occurred somewhere in the network

On receiving packets marked with the ECN bits set (11), the receiving host generates a Congestion Notification Packet (CNP) which is sent back to the transmitting host. The CNP carries information identifying the specific flow (known as Queue Pairs) for the sender to know which exact flow experienced congestion. The sender then rate-limits that particular flow until congestion is relieved, thus achieving effective end-to-end congestion control.

The Arista 7060X6 series also supports multiple ECN enhancements like:

- Latency-based marking (based on switch latency and buffer status)
- Throughput-based marking
- Dynamic marking (based on instantaneous total available buffer)

Automation, Visibility and Performance Monitoring

Visibility and telemetry become especially important in network deployments, where any sort of congestion hot-spots can lead to inefficient utilization of highly valuable resources. With that in mind, the Arista 7060X6 series supports advanced telemetry and visibility features that make these platforms perfect for deployment in AI and Datacenter networks.

CloudVision

Arista CloudVision is the foundation for network provisioning, monitoring and automation needs of modern AI networks. The Arista 7060X6 series can stream all the data elements that exist on the platform in real time to CloudVision, via a centralized state-based mechanism called NetDL. This data can then be analyzed in real-time to react to network events, and can be integrated with Arista's Autonomous Virtual Assistant (AVA) to proactively avoid potential network performance issues. CloudVision also alerts the user of network congestion hotspots that occur, that can then be investigated and remediated. Further, if any config changes cause unexpected issues in the network, CloudVision provides the capability to roll-back to a previously known golden configuration.

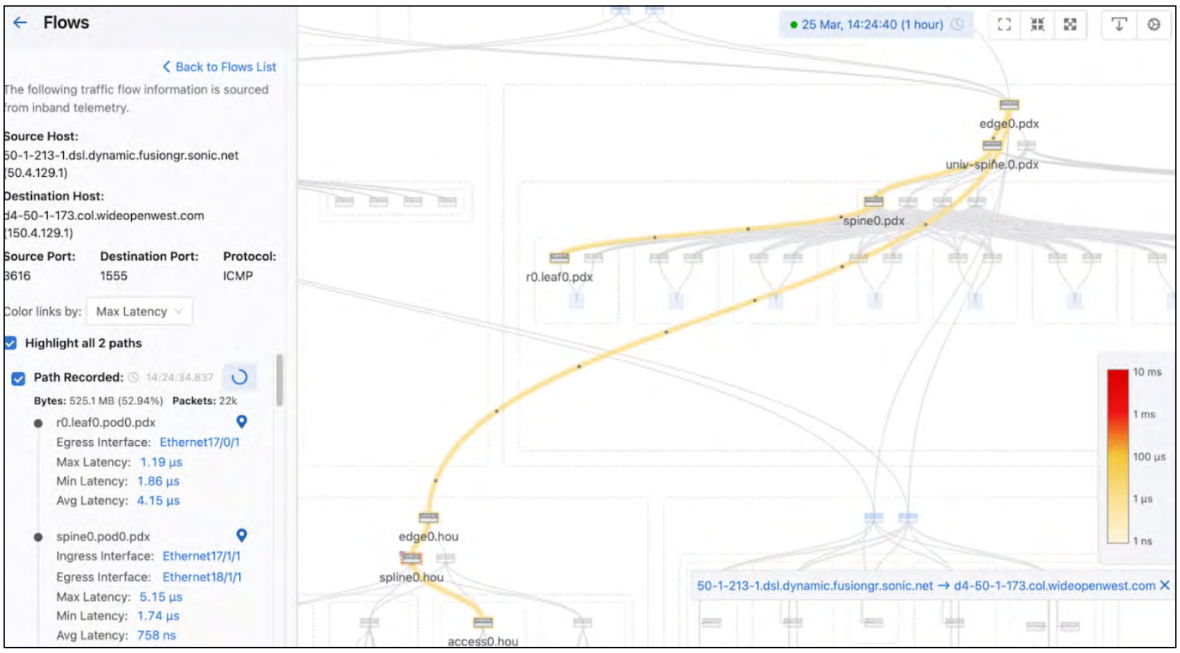

Figure 9: Arista CloudVision demonstrating network congestion hotspots

Automation

CloudVision's abstraction of the physical network to a broader, network-wide perspective allows for a more efficient approach for several operational use-cases and automates them with utmost ease. The Arista Validated Designs (AVD) takes this approach one step further, to provide an extensible data model that defines Arista's Unified Cloud Network architecture as code that can be automated and validated.

AI Analyzer

AI/ML traffic patterns exhibit unique ramp up behavior in very short intervals of time. Traditional software-based traffic counters therefore do not lend themselves to examining these unique flows efficiently. The AI Analyzer is an integrated hardware capability that enables the collection of ECMP member utilization data, aggregated over extremely short periods of time, as granular as 100 microseconds. This allows the Arista 7060X6 series to effectively analyze these traffic patterns. The results can then be applied to fine tune dynamic load balancing workloads uniformly across the member links, to optimize AI/ML applications.

Figure 10: AI Analyzer being used to apply optimal hashing on the member links carrying AI traffic

AI Agent

The Arista AI Agent takes the aforementioned network provisioning and monitoring aspects from a device-wide level to a network-wide level, to provide an effective end-to-end solution. Supported alongside the Arista 7060X6 series, the AI Agent is essentially a software daemon that runs on supported servers or Network Interface Cards (NIC's) to provide seamless Workload Integration. The Arista 7060X6 switch connected to the CloudVision server communicates with this AI Agent, giving it control over portions of the server's configuration and real-time monitoring of status.

Latency Analyzer (LANZ)

Arista Networks' Latency Analyzer (LANZ) is a family of EOS features that provide enhanced visibility into network dynamics. At its essence, LANZ tracks interface congestion and queuing latency with real-time reporting. With LANZ application layer event export, external applications can predict impending congestion and latency events can be used to proactively make traffic routing decisions with visibility into the network layer.

Ultra Ethernet Consortium (UEC)

The Ultra Ethernet Consortium (UEC) is a group of organizations with a mission to optimize and improve the performance of Ethernet for AI and HPC applications. As a founding member of the UEC, Arista is at the forefront of bringing these enhancements to our customers; and the Arista 7060X6 series of platforms is designed to be fully compatible with future UEC networks.

7060X6 Cluster Design

Over the course of this whitepaper, we took a look at the scaling requirements of typical AI clusters, and how the Arista 7060X6 is the platform of choice for the most demanding AI clusters. The following diagram offers a reference design architecture to support a scale of 8192 accelerators.

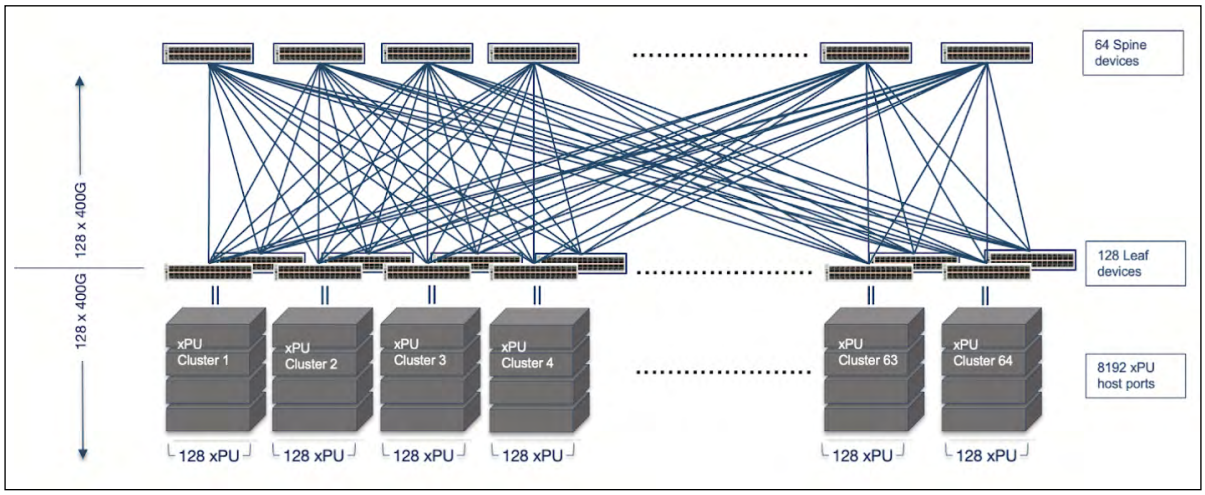

Figure 11: Reference design architecture to support 8192 accelerators using Arista 7060X6 systems

Deployed in a two-tier architecture, the design has 128 7060X6 systems deployed in an AI Leaf role, and 64 7060X6 systems in an AI Spine role, to create a fully balanced network.

We can further expand the design to incorporate a third tier of 7060X6 systems (AI SuperSpines) to support scale up to 32K accelerators.

Conclusion

As accelerator bandwidth increases and AI clusters continue to expand in size, the underlying network infrastructure must also evolve. Together with higher throughput and increased port density, the network platforms need to handle the unique requirements of line-rate RDMA training traffic, while simultaneously being power efficient and easy to manage. The Arista 7060X6 fulfills these requirements for AI deployments, by offering industry-leading radix combined with innovative traffic management features to improve job completion times. Its shared-buffer architecture offers flexibility for training and inference workloads, and support for Linear Pluggable Optics can reduce the network power requirements by up to 50%. Additionally, all of these benefits also translate equally well to the traditional Datacenter deployments, where the Arista 7060X6 can assume the roles of a DC leaf or a DC spine with ease. Overall, the Arista 7060X6 series combines high-performance, robustness and efficiency, making it an ideal platform for the most demanding network architectures.

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