IP Storage and Next Generation Platforms

In the past few decades the most pervasive storage media, was a dedicated Fibre Channel network that connected compute to storage. There really was no other choice; it was Fibre Channel or nothing. Recently however technological innovations in both storage and networking are leading most organizations to converge their infrastructure onto an Ethernet based IP fabric. The choice to converge is typically based on cost, performance and simplicity, however this document will not cover the sometimes political nature of this decision. Instead Arista will assume the choice to converge has been made and focus on the underlying Ethernet network. This paper will examine what network technologies and topologies provide the characteristics required by these demanding applications.

Finally network connectivity speed has caught up too Moore's law, a simple observation that transistors in dense integrated circuits would double every two years. In network switches what drives this is the underlying silicon that makes up the packet processor of the switch. Switch silicon and specifically merchant silicon is finally following this same curve.

IP Based Storage

With the adoption of 25GbE and 50GbE, IP-based block, object and file storage present a better-cost per-byte performance for the data center. These are emerging technologies, which can take advantage of the multi-pathing capabilities of a leaf-spine architecture to provide scalable performance. Application performance can also be guaranteed using standard QoS frameworks to ensure the storage traffic is correctly prioritized and serviced as it traverses the leaf and spine nodes. The IP approach to storage is "Server centric" due to the fact that it can be transported over a standard IP infrastructure, including traversal of layer 3 boundaries. In other words, storage becomes another application on the network, albeit a mission critical one.

The bursty nature of the IP storage flows cannot be overlooked. These traffic bursts can result in traffic loss within the network infrastructure due to buffer exhaustion. In a leaf-spine architecture this is often observed on the spine nodes, as traffic bursts converge from multiple leaf nodes during a read from a distributed storage environment where content can span multiple racks of servers. It can also be experienced at the leaf node when a large number of storage devices are connected to a pair of leaf switches. In these scenarios, oversubscription between the leaf and spine, necessary for economic reasons, can lead to packet drops, retransmits, and an overall drop in performance. For latency sensitive storage, this can cascade into a dramatic application impact. As such careful consideration must be given for both the leaf and spine nodes. The Arista 7500R and 7280R platforms are well suited for this operation due to their large buffers. These features allow the spine and/or the leaf to absorb the microbursts and avoid packet loss. This reduces retransmissions thus improving efficiency and overall performance of the storage.

Arista's Next Generation Platforms

The Universal Spine architecture from Arista is the foundation for building high performance scalable network topologies. This architecture has been widely adopted and deployed across Arista's customer base. In choosing a platform to handle the performance characteristics of IP based storage no other product performs as well as the 7500 series of switches. The same packet processors at the heart of the 7500 Series of products are also used in the fixed configuration family of 7280 Series switches.

The Arista 7500R and 7280R Series leverage the same consistent architecture for unparalleled performance and scale in a deep buffered switch. The 7500R Series provides up to 432 100G Ethernet ports in a choice of three system capacities. Each 100G port can be broken out to provide up to 1,728 25GbE ports. A broad set of line cards provides a choice of interface types all with high bandwidth, dynamic deep buffers, and large resource scale. Additionally all are backward compatible with 7500E Series line cards.

Figure 1: Arista 7500R Universal Spine platform

The Arista 7280R Series are purpose built fixed configuration 10/25/40/25/100GbE systems built for the highest performance environments, and to meet the needs of the largest scale IP based storage. They combine scalable L2 and L3 resources and high density with advanced features to deliver scalable and deterministic network performance while simplifying designs and reducing OpEx.

The 7280R capabilities address the requirements for a lossless forwarding solution in a compact and energy efficient form factor. The broad range of interfaces and density choice provides deployment flexibility. The 7280R Series are available in a choice of models with a choice of 10GBASE-T, 10GbE SFP+ with 40/100GbE QSFP uplinks and a range of 1RU and 2RU 40/100GbE systems that offers up to 48 ports of wire speed 100GbE in a 2RU system.

The 7280R Series provide industry leading power efficiency with airflow choices for back to front, or front to back.

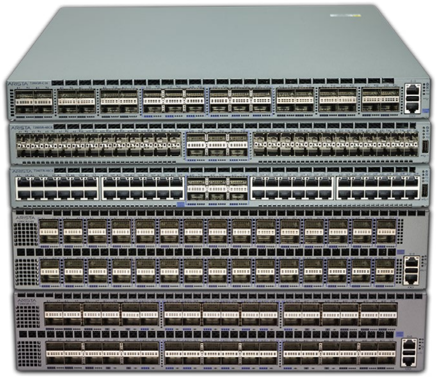

Figure 2: Arista 7280R Universal Leaf platform

IP Based Storage Architectures

When choosing an architecture for an IP/Ethernet storage network, consideration of the goals of the project must be taken into account when selecting an appropriate design.

The project goals may include: cost and space savings, performance and maintenance, data protection and data segmentation. An IP/ Ethernet storage architecture is ideal to meet all of these requirements through a collapsed, dedicated, or semi-collapsed/ hybrid approach.

Collapsed Storage Architecture

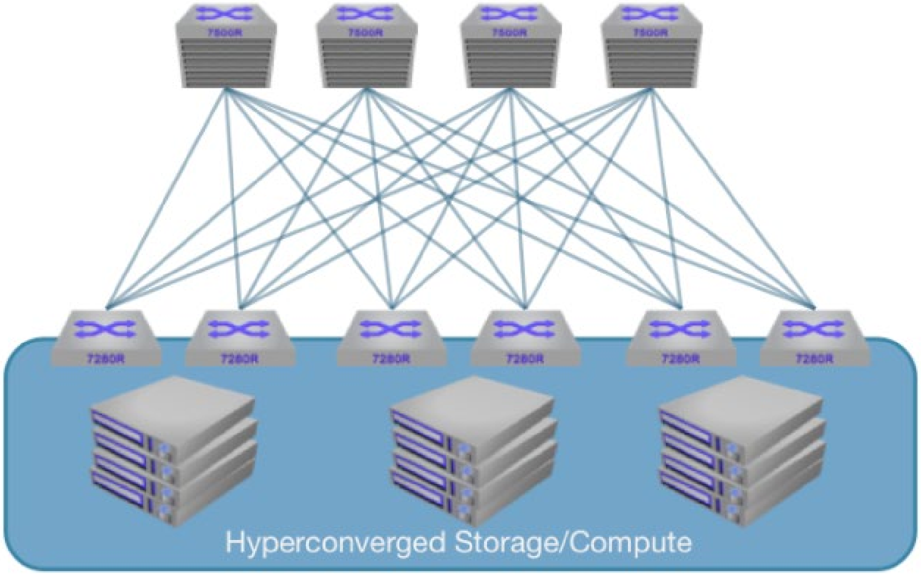

Figure 3: Collapsed Storage Architecture

Collapsed storage architecture represents the simplest topology because all traffic traverses the normal leaf/spine architecture deployed in the datacenter. Leveraging existing resources, collapsed IP Storage networks provide the most cost effective alternative for deploying IP storage, but in this design, care must be taken to ensure the proper buffering is available to handle the load without causing loss. In a collapsed architecture, storage nodes can be attached to the spine or leaf layer alongside compute infrastructure.

Dedicated Storage Architecture

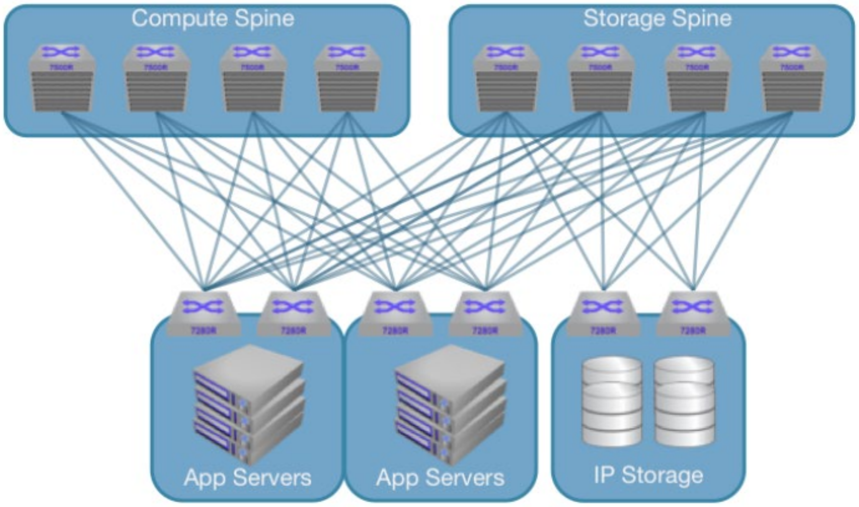

Figure 4: Dedicated Storage Architecture

Some organizations require dedicated storage networks. These can be achieved using an IP/Ethernet infrastructure, while retaining significant cost advantages over a legacy fiber channel network. HBAs and storage media are typically on the order of 90% cheaper in a converged Ethernet network versus a comparable fiber channel deployment, as such, the cost of implementing a dedicated Ethernet "SAN" can still provide up to 50% savings over a legacy alternative while providing higher performance.

This architecture consists of a dedicated leaf/ spine topology complete separate from the compute infrastructure. Compute resources have dedicated NIC's on both the compute and storage networks which ensures that compute and storage data do not contend for the same bandwidth.

Semi-Collapsed/Hybrid Storage Architecture

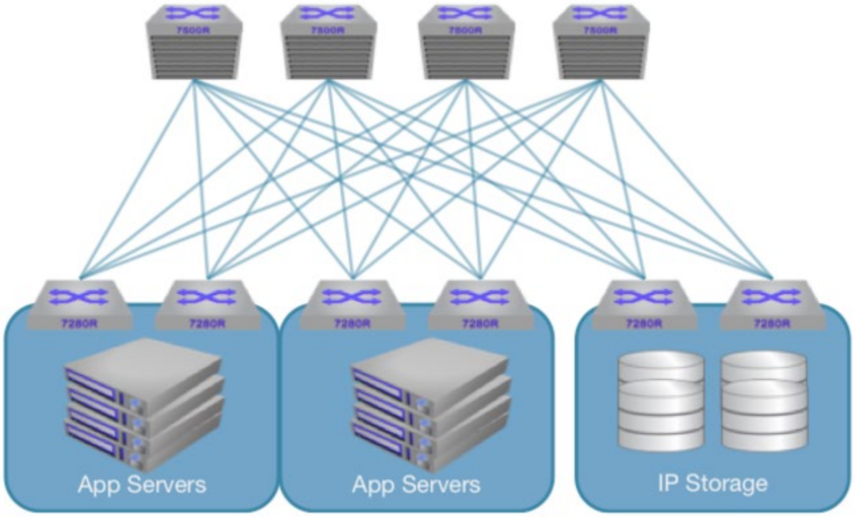

Figure 5: Semi-Collapsed Storage Architecture

In this design, a single universal spine is used for all traffic and dedicated leaf switches are used when the storage criticality dictates dedicated infrastructure. In a hybrid environment, traffic from the spine to the storage leaf switches does not contend for bandwidth with application traffic, these links become dedicated, as do the storage leaf switches themselves. To protect traffic from application leaf to storage leaf, separate classes of service can provide a simple and fair way to ensure all application performance needs are met.

Driving Storage Performance Over Ethernet Networks

Switch port buffering is a critical consideration in IP based Storage networks. The deep buffer architecture of the Arista 7280R and 7500R series switches provides a platform for lossless asymmetric connections with 10G, 25G, and 50G nodes, such as storage devices and compute nodes. Combining VOQ techniques with expanded dynamic buffering brings additional flexibility to application and overall network behavior. Large buffers mitigate the effects of congestion when traffic is bursty or highly loaded by devices simultaneously converging on common servers. An example of the latter occurs when an application server receives data from a striped bank of storage servers, and all the responses happen simultaneously. This condition, known as TCP incast, has been well researched and is well understood. Use of a deep buffered switch can absorb these transient bursts of traffic, minimizing drops, and the need for TCP retransmits, and leading to a more uniform and high performance experience. More detail on how big buffers help application performance can be found here.

For IP storage where performance is directly reflected to the underlying applications a deep buffered network architecture provides peace of mind to be able to handle all scenarios with the best possible outcome.

Virtual Output Queue (VOQ)

The 7280R and 7500R series switches employ a Virtual Output Queue (VoQ) architecture to ensure fairness and lossless port-to-port forwarding in the most extreme conditions. VOQ avoids switch fabric congestion and the "head of line blocking" problems that often plague legacy switches.

The 7280R series switch provides industry best solutions dealing with massive intra-cabinet IP storage flows, incast towards the spine, and ability to support emerging SDS architectures. Simply put, no other switch can compare for enabling scale-out storage architectures.

Buffering at the spine is also a key consideration. Table 1 contains real world buffer utilization information from spine switches on a per-port basis across multiple market segments.

Without sufficient buffering in the spine, drops will be experienced.

Table 1: Real World Spine Switch Buffer Utilization

User Segment	Environment	Max Buffer Utilization Observed
Software Vendor	Engineering Build Servers	14.8 MB
Oil & Gas	Storage Cluster - Medium	33.0 MB
Online Shopping	Big Data – Hadoop 2K Servers	52.3 MB
HPC - Research	HPC Supercomputer Center	52.4 MB
Animation	Storage Filer (NFS)	74.0 MB

DCB and PFC

In addition to well-engineered buffering technology, Arista supports other mechanisms that help to ensure the lossless delivery for IP/Ethernet storage by utilizing the IEEE standardized set of features called Data Center Bridging (DCB). DCB works to make Ethernet a lossless medium by implementing priority flow control (PFC).

PFC enables switches to implement flow-control measures for multiple classes of traffic. Switches and edge devices slow down traffic that causes congestion and allow other traffic on the same port to pass without restriction. Arista switches can drop less important traffic and tell other switches to pause specific traffic classes so that critical data is not dropped.

The architecture of a switch must be taken into consideration when deciding whether or not to deploy PFC. While a VoQ architecture such as those present in Arista's 7500R or 7280R switches can operate well in a PFC enabled infrastructure, switches which employ an internal Clos architecture for a fabric are not well suited to a PFC enabled deployment. The issue is a direct function of leveraging "switch on a chip" architectures to create a multi-stage fabric within a chassis. Consider the implications of congestion: When such a switch experiences congestion on a PFC enabled port, the port ASIC must generate a pause frame towards its upstream neighbor. In the case of an internal Clos fabric, this neighbor can quite likely be a fabric chip. This causes all traffic in the lossless class on the fabric chip to be paused. As the fabric is paused, being a natural convergence point for flows, the fabric buffers will fill up very rapidly causing drops. At this point, the fabric, experiencing congestion will send a pause to all connected interfaces, which typically translates to all ingress ports. The net effect is a PFC initiated head of line blocking across all ports on the switch. For this reason PFC is not recommended on a Clos fabric based switch.

ECN

Another method that works in conjunction with PFC is Explicit Congestion Notification (ECN) which is an extension to the Transmission Control Protocol which allows end-to-end notification of network congestion without dropping packets. ECN is an optional feature that is only used when both endpoints support it and are willing to use it. ECN operates over an active queue management (AQM) algorithm - Weighted Random Early Detection (WRED). ECN allows for hosts to learn about congested paths and react to congestion events faster than traditional methods.

Arista LANZ Provides Operational Visibility Into Congestion Events

IP/Ethernet storage network environments require deterministic characteristics with regard to throughput, performance, and latency. These characteristics are theoretical only if the network infrastructure cannot provide real time monitoring of changing traffic patterns and their effect on the design characteristics.

The Arista Latency Analyzer (LANZ) is an event-driven solution that is designed to provide real-time visibility of congestion hotspots and their effect on application performance and latency at a nanosecond resolution. The level of granularity that LANZ delivers is made possible by its unique event-driven architecture. The traditional polling model provides visibility only at discrete intervals. LANZ reports on congestion events as they occur. LANZ provides visibility into the network where IP/Ethernet storage is attached and it can be used to ensure that necessary interconnects are available to ensure a lossless transport.

Conclusion

The main goals in building next generation data centers is to remove cost and complexity and to converge on a winning long term solution. We can at least partially attribute the acceleration of IP based storage to the rise of the cloud titans, incredibly large cloud networks built with commodity servers at unprecedented scale with the most demanding performance requirements. These environments can only exist using IP storage, and the numbers show that the world is following their lead in storage, similarly to what has occurred with compute virtualization. Looking closely at industry trends in the network and storage markets a key metric, the number of fiber channel ports being shipped is shrinking dramatically year over year.

"Global Fiber Channel storage area network (SAN) revenue—including Fiber Channel switches and host bus adapters (HBAs)—is down 11 percent sequentially in 1Q15, and up 1 percent from 1Q14, to $595 million", Source: Infonetics

"In contrast with Ethernet and Infiniband, Fiber Channel switch revenue and shipments declined, both sequentially and year-over-year", Source: Crehan Research

The cloud titans are pioneering next generation datacenter environments. From the early adoption of virtualized compute, leaf/spine topologies, and now 25G/50G Ethernet and IP based storage, the path is clear. A converged, simplified architecture is the only way to maintain cost competitiveness in the datacenter. Arista Networks, with the 7500R and 7280R series switches provides the best solution for IP/Ethernet storage networks. Building an IP/Ethernet storage architecture with high-performance Arista switches maximizes application performance. Arista switches deliver the deep buffers to address TCP incast, the operational flexibility and extensibility to advance to next generation management paradigms, and the most resilient, highly available and cost effective network solutions to meet the demands of data growth and next-generation storage networks.

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 & B, 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