OVERVIEW

Traditional storage networks and costs are growing at an exponential rate. New requirements for compliance and new applications such as analytics mean that ever-increasing volumes of unstructured data are collected and archived.

Legacy storage networks cannot meet the need to scale-out capacity and reduce capital and operational expenditures. In response to this challenge, new storage architectures based on Ethernet have evolved.

How will this change the storage network? Market research firm IDC predicts that Fibre Channel will remain at the core of many data centers to support traditional mission-critical mainframe and UNIX-based applications. Most future IT asset deployments will leverage 10 Gigabit Ethernet (10GbE) (and soon 40 Gigabit Ethernet [40GbE]) for the underlying storage interconnect for newer applications.

Arista Networks has responded to the need for a new IP Ethernet storage fabric. Arista provides industry-leading products and solutions that deliver operational and infrastructure efficiencies that have been previously unavailable.

IP ETHERNET STORAGE CHALLENGES

IP Ethernet storage has a very different set of challenges from legacy storage transport. Figure 1 shows these three common challenges:

- Storage networks require massive east to west performance and scalability.
- TCP incast occurs when a distributed storage and compute framework such as Hadoop is implemented.
- Storage and compute devices interconnect at different speeds.
NEED FOR MASSIVE EAST TO WEST SCALABILITY

At the very core, storage is simply data. The more data that can be moved from host to target, the more efficiently the applications will run on the hosts. This data often traverses multiple links, and legacy networks are not designed to provide a non-blocking deterministic latency solution. Traditional three-tier designs can be massively oversubscribed in the data center because they were designed primarily for north to south traffic.

Storage data I/O must be resilient and ensure reliable transmission to prevent loss or corruption. Fibre Channel was designed as a lossless medium that uses a buffer-to-buffer credit system for flow control. Ethernet was designed as a best effort transport that relies on upper-layer protocols to recognize data loss and handle retransmission. File-level protocols such as iSCSI use TCP, which can detect packet loss and retransmit. However, even in this scenario, any packet loss is undesirable and dramatically effects performance.

TCP INCAS

TCP incast is a many-to-one communication problem that can occur in data networks. As distributed storage and compute frameworks such as Hadoop have been deployed, a single host request for data can be served simultaneously by tens of storage nodes. This response oversaturates the connection to the host. With the drive toward virtualization and maximizing efficiency, storage devices often service many hundreds or even thousands of hosts. If many of these hosts must access the storage device simultaneously, a massive incast condition happens at the storage device. In the incast communication pattern, a receiver issues data requests to multiple senders. When the senders receive the request, they concurrently transmit a large amount of data to the receiver. The data from all senders traverses a bottleneck link in a many-to-one fashion. This condition often results in a microburst that overwhelms a single host or switch port. Traffic is dropped, which causes TCP timeouts and retransmits and degrades application performance.

STORAGE AND COMPUTE DEVICES INTERCONNECTING AT DIFFERENT SPEEDS

Bottlenecks can occur in networks where all devices are connected at different speeds, and today many networks are a mixture of 1Gbps, 10Gbps, 40Gbps, and soon 100Gbps. This speed mismatch can cause a condition where a server is connected at 1Gbps but the storage system is connected at 10Gbps. In this scenario, a single request is sent at 1Gbps, but the responses come back at 10Gbps and must be serialized to 1Gbps. This massive speed mismatch can also lead to buffer exhaustion in a switch and result in dropped packets, which degrades the performance of applications.

Figure 1: Challenges with legacy network IP Ethernet storage
ARISTA IP ETHERNET STORAGE NETWORK SOLUTIONS

Arista Networks is the leader in building scalable high-performance data center networks. To meet the challenges of transporting storage on an IP Ethernet network, the Arista solution includes these features and technology to assist with the IP storage challenges, as shown in Figure 2:

- Massively scalable Leaf/Spine non-blocking topology
- Deep buffering technology and lossless delivery
- Operational visibility into the network to provide an early warning of congestion

MASSIVELY SCALABLE LEAF/SPINE TOPOLOGY

Arista provides an industry-leading solution that offers a superior price to performance ratio when designing networks to support IP Ethernet storage. The Arista solution ensures effective use of all available bandwidth in non-blocking mode and provides failover and resiliency when any individual chassis or port has an outage condition. Multichassis Link Aggregation (MLAG) and Equal Cost Multipath routing (ECMP) provide standards based, non-proprietary, multipath technologies at Layer 2 and Layer 3. These technologies currently scale linearly to more than 50,000 compute and storage nodes, both physical and virtual.

The use of next-generation multicore server CPUs with dense virtual machines (VMs) and storage, make the use of this type of uncompromised Leaf/Spine topology critical. In addition uplink, downlink, and peer ports that are not oversubscribed and are all switched at wire speed are very important.

The Arista Leaf/Spine topology effectively solves the IP Ethernet storage challenge of massive east to west traffic by delivering a solution with deterministic latency characteristics and any-to-any non-blocking host communication.

BUFFERING TECHNOLOGY AND LOSSLESS DELIVERY

Storage devices and hosts are often connected at different speeds. In these situations, switch port buffering is a very important consideration. The Arista 7048 and 7500 switches are designed to appropriately allocate bandwidth to all traffic flows in the data center. The Arista 7048 switches have 768MB of packet buffer memory, which allows the architecture to scale in congestion scenarios. All ports can simultaneously buffer up to 50ms of traffic without any drops. The deep buffer architecture of the Arista 7048 switch also provides a platform for lossless asymmetric connections when 10GbE nodes, such as storage devices, communicate with 1GbE nodes, such as servers, in the data center.

Figure 2: Modern IP Ethernet storage design
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 and thus is well suited for storage networks.

DCB has two important features: Data Center Bridging Exchange (DCBx) and Priority Flow Control (PFC). These features enable Arista switches to carry storage data more reliably. DCBx is an exchange protocol that capable switches can use to exchange data center bridging capabilities to ensure consistent configuration and support for lossless transport.

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. This Quality of Service (QoS) capability eases congestion by ensuring that critical storage I/O is not disrupted or corrupted and that other non-storage traffic that is tolerant of loss can be dropped.

The Arista Latency Analyzer (LANZ) is a pro-active, event-driven solution that is designed to provide real-time visibility of congestion hot spots 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.

**OTHER ARISTA OPERATIONAL ADVANTAGES**

In addition to the features already described, Arista solutions also provide many other operational advantages, including:

- Industry-leading power efficiency per 10GbE port at less than 3.5W
- Single binary image across all platforms
- Open APIs for third-party integration
- Advanced Event Management for customizable actions based upon specific events
- Zero-touch provisioning (ZTP) for rapid deployment and expansion of storage clusters
- VM Tracer for dynamic and automated provisioning of VLANs in virtualized environments
- Hardware virtual extensible LAN (VXLAN) support to allow for seamless integration with virtualized compute infrastructure

Visit Arista Networks for more information on each of these features and functionality.

**CONCLUSION**

Arista provides the best solution for IP Ethernet storage. Exponential growth in storage needs requires an efficient transport like Ethernet to overcome the expense, interoperability, administrative complexity, and scale-out challenges of traditional approaches. Building an IP Ethernet storage architecture with high-performance Arista switches maximizes application performance. Arista switches deliver a highly resilient and available network, deep buffers to address TCP incast, and the operational flexibility and extensibility in a lower-cost solution to meet the demands of data growth and next-generation storage networks.
APPENDIX A – STORAGE TECHNOLOGIES OVERVIEW

STORAGE AREA NETWORKS
A storage area network (SAN) is an architecture whereby servers access remote disk blocks across a dedicated interconnect. Most SANs use the SCSI protocol to communicate between the servers and the disks. Various interconnect technologies can be used, and each of them requires a specific SCSI mapping protocol.

A SAN is a specialized network that enables fast, reliable access among servers and external or independent storage resources. A SAN is the answer to the increasing amount of data that must be stored in an enterprise network environment. By implementing a SAN, users can offload storage traffic from daily network operations while establishing a direct connection between storage elements and servers. SAN interconnects tie storage interfaces together into many network configurations and across large distances. Interconnects also link SAN interfaces to SAN fabrics.

Building a SAN requires network technologies with high scalability, performance, and reliability to combine the robustness and speed of a traditional storage environment with the connectivity of a network.

FIBRE CHANNEL PROTOCOL
Today, the majority of SANs use the Fibre Channel Protocol (FCP) to map SCSI over a dedicated Fibre Channel. Enterprises that deploy Fibre Channel deploy multiple networks including the LAN network and the dedicated Fibre Channel network. Typically the LAN uses Ethernet technology, which is a basic component of 85% of all networks worldwide, and is one of the most ubiquitous network protocols in existence.

iSCSI
Internet Small Computer Systems Interface (iSCSI) has gained traction and attention as data centers strive to reduce costs for robust storage. iSCSI rides on IP Ethernet transport, which alleviates the complexity of a separate traditional Fiber Channel SAN. iSCSI is an ideal solution for many small and medium enterprise organizations. iSCSI relies on TCP/IP protocols, which makes it a natural communication for private and public cloud communications.

The performance advantages of iSCSI are compelling. Storage arrays must keep up with new multicore processors and stack software that can generate a million iSCSI I/O operations per second (IOPS) when connected to a nonblocking storage access switch from Arista.

NETWORK ATTACHED STORAGE
An increasingly popular method for consolidating storage resources is network attached storage (NAS). A NAS appliance is a server that supplies file-based data storage services to other devices on the network. NAS is a remote file system I/O where the file request is redirected over a network.

NAS is recognized for three principal benefits, which in combination lower the overall total cost of ownership (TCO):

- Storage consolidation
- Deployment simplicity
- Ease of management

NAS systems have evolved to support, via a standard Ethernet network, the storage tiering, high performance, and high availability that had previously only been available in SANs. This support, combined with its TCO advantages, has made NAS an increasingly adopted solution in the enterprise.

Distributed storage architectures that have evolved from the original NAS scale-up concept have become known as scale-out NAS storage systems.

FIBRE CHANNEL OVER ETHERNET
The Fibre Channel over Ethernet (FCoE) protocol essentially is an encapsulation of FCP over Ethernet. FCoE enables enterprise customers who are accustomed to Fibre Channel to run the Fibre Channel Protocol directly over their LAN Ethernet network. They can consolidate their LAN and storage network over the same network infrastructure. FCoE is aimed at organizations that want to keep a high-end Fibre Channel SAN, yet are interested in LAN and SAN convergence.

For FCoE to be efficient, interoperability must exist with converged network adapters (CNAs) and gateways from different vendors. No two FCoE implementations are alike and interoperability is not guaranteed. Nevertheless, Arista switches do support FCoE standards and can carry FCoE traffic.
Figure 3 compares block-level storage protocols. Fibre Channel, iSCSI, and FCoE all use SCSI I/O commands, but they differ in the layers that encapsulate and communicate between initiators and targets. Each has several pros and cons, but the current consensus is that they are all viable and are close in performance, efficiency, and latency.