World Class, High Performance Cloud Scale Storage Solutions

Arista and EMC ScaleIO

The universal adoption of the Internet and the explosive use of mobile computing with smartphones, tablets and personal computers have caused a tidal wave in the amount of information we generate, consume and want to access instantly 24 x 7. IDC predicts that the Digital Universe will be at 16ZB (16 Million PB) by 2017 and grow to 20ZB by 2020.

Planning for and managing this explosive growth in Storage has become a challenge for all IT departments. IT departments are beginning to develop scalable storage solutions that take advantage of the rapid improvements in storage (such as Flash memory) as well as build data analysis, backup, restoral and archiving solutions that meet the commercial, financial and regulatory needs of their organizations.

Faced with continuing volume growth a Large Financial Institution (referred to as LFI in the rest of this paper) undertook a study to identify the best architecture and solution for the future of their storage infrastructure. The key objective for LFI was to achieve the agility and economic benefits enabled by cloud technologies, as building the cloud is the focus of their strategic vision. LFI is actively pursuing implementation of cloud technologies, on the compute, storage and network fronts. From a storage perspective, the goal is to provide a common pool of capacity that could be scaled on demand and software defined, managed and provisioned in concert with their Software Defined Data Center (SDDC) vision. In addition to a sound ROI LFI's key requirements were Scalability, Elasticity and Performance.

LFI already had hands-on experience with running block storage over IP. Utilizing iSCSI for intra-cabinet access to flash arrays allowed IP as a storage transport to prove itself in a demanding environment. However, iSCSI was not seen as a strategic protocol due to the lack of resiliency and scale inherent in point to point architectures. Distributed architectures, such as EMC® ScaleIO® represent technologies far more inline with LFI's vision as they offer the scale, performance, and reliability necessary to bring LFI's storage architecture to the next level.

While moving towards commodity hardware offered economic benefits, and software-defined solutions provided agility, the network had to be built to scale. The proliferation of 10/40G Ethernet provided the means to build a storage network architecture, which could provide better resiliency than Fibre Channel, at greater scale, and lower cost. To ensure that the network would meet their needs, extensive product selection research and testing was performed by their architecture and engineering organization.

LFI "Storage at Scale" Solution and Architecture

LFI selected EMC ScaleIO and Arista Networks to work with them to develop a Software-Defined Storage (SDS) Solution and ensure that they could quickly deploy a robust, scalable high performance solution with superior reliability and high availability.

EMC ScaleIO - Software Defined, Scale-Out SAN

EMC ScaleIO is a software-only solution that turns existing server storage into shared block storage. This empowers IT organizations with a software-defined, scale-out SAN that delivers flexible and scalable performance and capacity on demand. It is hardware-agnostic and designed with enterprise-grade resilience as a must-have, making it an ideal solution for Public Hosting, Service Providers and Enterprise customers.

Figure 1: ScaleIO Benefits

ScaleIO provides the following benefits:

- **Massive Scalability** - Scales-out in both capacity and compute starting from 3 servers to over 1,000 servers. Storage grows automatically as application servers are added.
- **Extreme Performance** - Parallel I/O architecture ensures workloads are evenly shared among all servers eliminating storage entry point bottlenecks, and performance scales linearly. Performance is optimized automatically whenever rebuilds and rebalances are needed, with minimal or no impact to applications and users.
- **Unparalleled Flexibility** - Deployment options include two-layer or hyper-converged and customers can leverage mixed server brands, configurations, OS platforms, and media types. Customers can dynamically add, move or remove storage and compute resources on the fly with no downtime.

- **Enterprise Grade** - Provides multi-tenant capabilities and enterprise features like QoS, snapshots, and thin provisioning. There is no single point of failure as ScaleIO provides data protection and resiliency via two copy mesh mirroring.
- **Compelling Economics** - Converges compute and storage resources of commodity hardware into a single layer, enabling customers to maximize local storage and simplify management. Makes additional resources (power, cooling) and a dedicated SAN fabric unnecessary, reducing TCO by over 60% compared to traditional SAN.

ScaleIO consists of three primary components:

1. ScaleIO Data Client (SDC) is the "client" component, which reads data from remote disks and writes data to them.
2. ScaleIO Data Server (SDS) is the "server" component, and it receives connections from the SDCs to write to and read from local disk.
3. The Metadata Manager (MDM) is the brain behind the operation and communicates with both SDC and SDS components to create a map of where the data is stored.

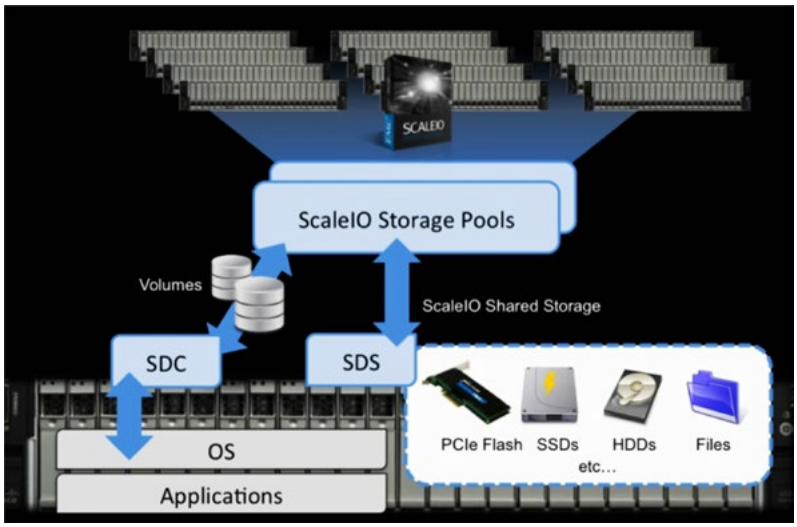

Figure 2: ScaleIO software deployment

The ScaleIO software can be deployed using two different deployment models.

1. Fully Hyper-converged deployments run both the SDC and SDS components on all nodes creating a hyper-converged compute and storage infrastructure.
2. Dedicated mode separates the two components; dedicated hosts are used to run the SDS components while the SDC component runs on the compute nodes. Dedicated mode is similar to a traditional SAN environment but uses an IP infrastructure for all communications and uses cost efficient commodity hardware for the storage nodes.

EMC ScaleIO Solution

LFI's selection of ScaleIO was driven by many considerations:

- Operates at L3 over a scalable leaf/spine networking model
- Provide a "pay as you grow" model
- No need for dedicated storage components
- Take advantage of standard of the shelf hardware and leverage rapid cost declines
- Promise of linear investments tied to growth with predictable costs.

Arista Networks Solution

Recognizing that Storage Defined Solutions are a vital element of the emerging Software-Defined Data Center, Arista developed and introduced industry leading network switches purpose built for demanding storage environments.

The Arista 7280SE, and 7500E switches are purpose built with deep buffers and the ability to fairly allocate bandwidth to all traffic flows in the data center. The Arista 7280SE switch provides unparalleled performance for IP storage networks and provides 9GB of buffer capacity in a compact 1U form factor. Its Virtual Output Queue (VoQ) architecture ensures fairness and lossless port-to-port forwarding in extreme conditions and offers unsurpassed performance for massive intra-cabinet IP storage flows as well as incast towards and from the spine.

Figure 3: Arista 7280SE-72 Switch

Why Purpose Built Deep Buffer Switches with VOQ Architectures

The distributed nature of massive scale-out IP storage architectures requires a fundamental new approach on how to deploy storage networks in today's new virtualized Software defined Data Centers. The three most important challenges are:

1. Managing for massive East - West traffic volumes, performance and scale.
2. Managing TCP incast, a many to one communications challenge.
3. Eliminating bottlenecks that occur when devices connect at different speeds.

Figure 4 illustrates these three new challenges:

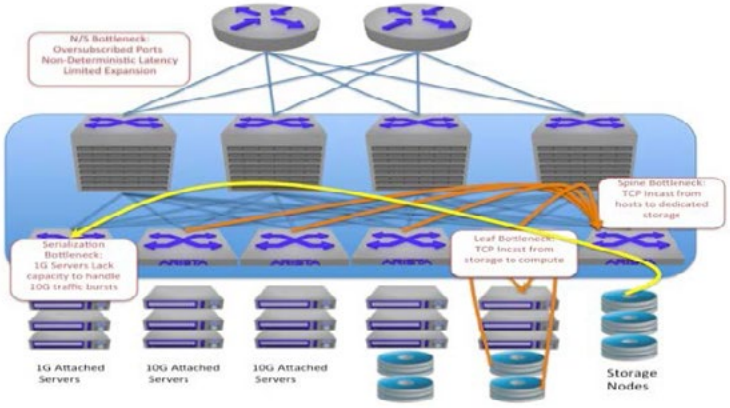

Figure 4: Buffering challenges in a leaf-spine network

East/West Traffic Scaling

Handling large volumes of east/west traffic with minimal packet loss is mandatory for high performance storage networks and switches with deep buffers are essential in ensuring high performance.

TCP Incast

TCP incast is a many-to-one communication problem. As distributed storage frameworks are deployed, a single host request can generate simultaneous responses from many storage nodes and saturate the host connection often creating a microburst that overwhelms a single port. Widespread adoption of virtualization exacerbates this situation as storage devices may have to service hundreds or even thousands of hosts. Deep port buffers neutralize the effect of a microburst by queuing the traffic rather than dropping it. Arista switches are purpose built to provide deep buffers and manage TCP incast at both the spine and leaf layers of the network.

Storage and Computer Devices Interconnecting at Different Speeds

As servers and storage networks scale to 40G it is common to see a mix of 1Gbps, 10Gbps, 40Gbps, and 100Gbps in the Data Center. Bottlenecks can occur when devices are connected at different speeds. For example, a single request at 1Gbps gets responses at 10Gbps and these must be serialized to 1Gbps. Speed mismatch can lead to buffer exhaustion, especially at the leaf layer where speed differentials frequently occur. Deep buffered switches like the Arista 7500E and 7280SE are required where speed differentials occur in the network.

Managing and Taming Latency

The Arista Latency Analyzer (LANZ) pro-active, event-driven software provides nanosecond resolution and real-time visibility of congestion hot spots and their impact on application performance and latency. LANZ visibility into the network can ensure a lossless transport.

Arista solutions also provide important operational advantages, including Arista's Extensible Operating System (EOS®) 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 and Hardware virtual extensible LAN (VXLAN) support for seamless integration with virtualized compute infrastructure.

After significant research, LFI concluded that Arista's 7280SE and 7500E switches were the best in the industry that could reliably provide the scale, buffering, and instrumentation necessary for high performance IP storage across a leaf/spine topology.

Proposed Network Architecture

LFI adopted a very extensive evaluation and design process to build their IP storage network. After discussions with other enterprises and cloud providers who had deployed software defined storage networks the LFI team elected to adopt ScaleIO's two-layer mode of operation and build a dedicated leaf/spine network for SDS->SDS communications.

Between SDS nodes, there are some operations, which need very high bandwidth and can be unpredictable. These operations include:

1. Writing backup copies of blocks to another SDS node
2. Re-building blocks from a failed node
3. Re-balancing the load as systems come on and off line

By segmenting the storage network from the existing IP Fabric using a 4-way Arista 7500E spine with 40Gb/s uplinks from dual top of rack switches per rack, the bandwidth and resiliency provided exceeds that of traditional FC SANs and takes their storage architecture to the next generation.

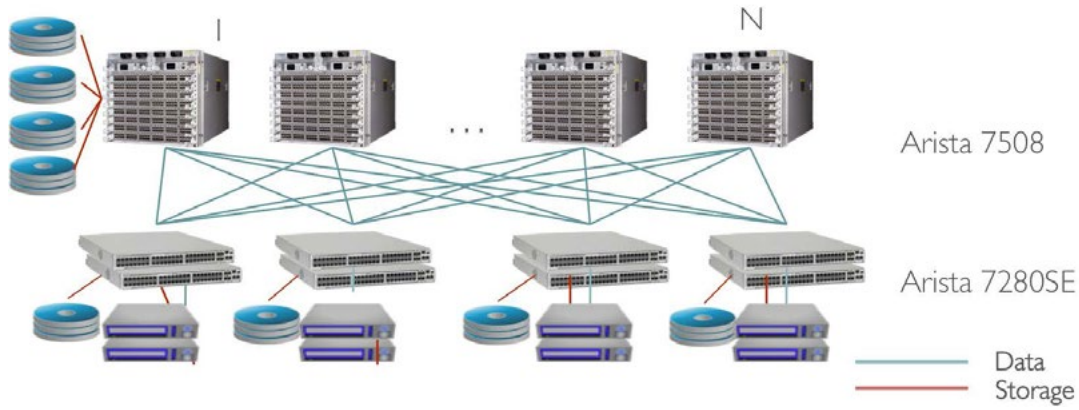

Figure 5: Proposed network architecture

On the front end of the network, both dedicated and shared NICs and switches were proposed for the SDC - SDS traffic. Since the 7500E is a non-blocking VOQ Architecture based switch, and was already widely deployed in the LFI network, it was used as a spine for the front-end traffic. No contention for bandwidth between storage and non-storage traffic would arise except in the case of shared leaf switches for compute/storage as the uplinks to the spine from the SDS nodes, and dedicated storage client switches are dedicated to storage traffic alone. To accommodate the shared link scenario, a simple QoS policy was proposed to give priority to ScaleIO traffic. L3 ECMP is used on both the front end (connecting to existing L3 ECMP spine) and back end networks for scale and simplicity.

Solution Testing

After product selection was complete, and the architecture was proposed, LFI, along with EMC and Arista, planned, developed and installed a large scale Proof of Concept system to simulate the real world network and focus on qualification and quantification of ScaleIO performance across a deep-buffered leaf/spine network.

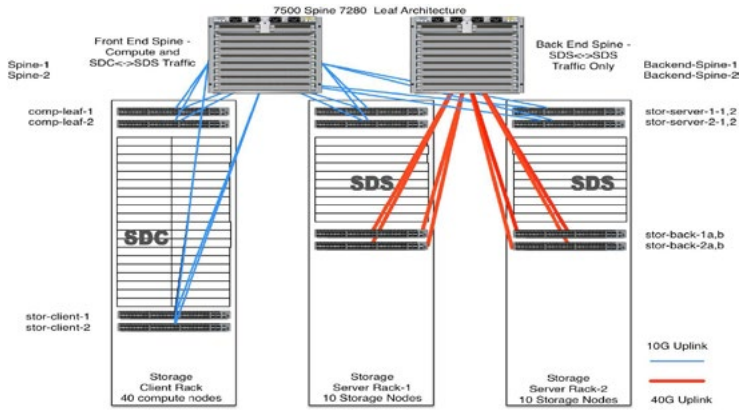

Figure 6: Arista/ScaleIO Lab topology

A full range of tests was identified to develop baseline system performance, scale and evaluate impact of network traffic congestion and failures. Unlike many network based proof of concept tests, which treat the network as an independent unit operating in a vacuum, the LFI ScaleIO Arista POC focused on application performance across a real-world network. To achieve this, instead of using traffic generators, servers running ScaleIO software were used in the POC. The full list of devices used to build the test architecture is listed below in Table 1.

Table 1: Test Equipment Overview

Device Class	Function	Type	Quantity	Disk	Network Bandwidth	Comments
SDC	Compute w/remote Storage	Server	40	10TB 7200RPM SATA	40 Gb/s via 2*2 port 10G NICs	
SDS	Dedicated Storage	Server	24	10TB 7200RPM SATA	40 Gb/s via 2*2 port 10G NICs	
Spine	Compute traffic and SDC->SDS traffic	Arista 7280SE Switch	2	N/A	640Gb/s	All connectivity at 10G
Compute- Leaf	SDC rack ToR for SDC<->SDS and background traffic – shared NIC model	Arista 7280SE Switch	2	N/A	640Gb/s	Used primarily for NIC contention testing
Stor-Client	SDC rack ToR for SDC<->SDS traffic – dedicated NIC model	Arista 7280SE Switch	2	N/A	640Gb/s	
Stor-Server	SDS rack ToR for SDC<->SDS traffic	Arista 7280SE Switch	4	N/A	640Gb/s	
Backend Spine	Dedicated spine for SDS<->SDS traffic	Arista 7280SE Switch	2	N/A	640Gb/s	40G Downlinks to Stor-back switches
Stor-Back	SDS rack ToR for SDS<->SDS traffic	Arista 7280SE Switch	4	N/A	640Gb/s	40G uplinks to Backend Spine

The SDC nodes are x86 servers running the SDC software. They are running RedHat Enterprise Linux (RHEL) 6.x and their intent is to serve as compute nodes. The SDS nodes are identical from a hardware perspective, but only run the SDS software, and in three instances, the MDM as well. The Spine, Compute-Leaf, Stor-Client, and Stor-Server switch types are all connected to the front-end network, and enable the communications between the SDC and SDS. This infrastructure, with the exception of the SDS pods already exists in LFI's network. The links between the Spine, Stor-Client, and Stor-Server switches are all dedicated. The SDS nodes themselves also connect to the backend network via the Stor- Back and Backend-Spine switches. This network is utilized exclusively for SDS->SDS traffic such as data backup, rebalance, and rewrite operations. The connectivity from the SDS nodes to the Stor-Back switches is dedicated routed networks to ensure that no fate sharing exists between the different paths out of the rack.

The test suites were designed to anticipate worst case scenarios, push system performance to the limit and ensure that the system operated correctly. Two areas of focus overall in the testing were network and software performance.

For the network, basic functionality was tested first to determine maximum performance capabilities of the network and of the ScaleIO components.

Thereafter, various congestion scenarios were evaluated where competing non-storage traffic was intermixed with storage traffic to create contention and stress on both the network and servers.

Destructive testing in which all network components (links and switches) were failed and restored to service was also implemented and the impact to the system was recorded. Bandwidth utilization, packet discards, and queue-depth were monitored on the switches in conjunction with IOPs, application latency, and application queue depth from the ScaleIO side. By juxtaposing these results, an accurate picture of the application's performance on the network was developed.

Both the hardware performance (i.e. the servers with the constraints of the disk IO capability), and the ScaleIO software performance were measured.

In the end, the test

- Proved performance and resiliency of ScaleIO across a leaf/spine architecture
- Proved benefit of deep buffers at both the leaf and spine layers of the network
- Built confidence in the resiliency and capabilities of IP based SAN
- Moved forward the extensive design efforts based on empirical evidence of ScaleIO behavior and performance

Congestion Testing

Write Operations

During write operations, heavy buffer utilization is seen on the Stor-Client switches, which connect the compute hosts to the spine. These switches are dedicated leaf switches for SDC->SDS communications. Oversubscription at the leaf->spine causes congestion during heavy utilization and this congestion was seen on multiple interfaces simultaneously. This effect was predicted based on the test topology where a single rack of devices running SDC communicated with multiple racks of devices running SDS. The oversubscription ratio on the leaf can be adjusted to minimize congestion. However increasing traffic and unpredictable bursts make oversubscription more problematic underlining the need for deep buffers in the architecture.

During this particular write operations test, there was no background traffic. Only ScaleIO was running and despite the congestion IOPs remain high, with no discards, and with application latency remaining consistently low.

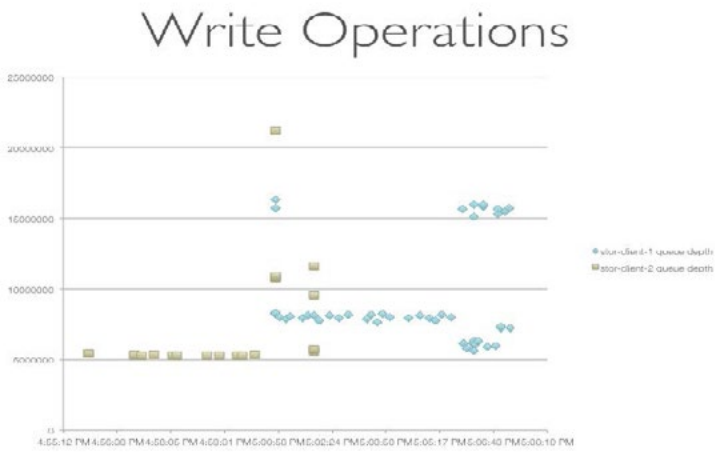

Figure 7: Queue Depth during write operations

Read Operations

During Read operations, heavy buffer utilization is seen on the spine switches. As with the behavior seen in the "write" tests, this is expected based on the network layout where there are two racks of SDS's with 4 leaf switches and 160Gb/s of uplink capacity transmitting to one rack of SDCs with 80 Gb/s of uplink capacity. Oversubscription in the spine->leaf direction causes congestion during heavy utilization and this congestion is seen on multiple interfaces simultaneously. This is expected in the "read" case based on the topology as multiple racks of SDS each with 80Gb/s of uplink bandwidth (total of 160Gb/s) are transmitting to a single rack of SDC with 80Gb/s of uplink bandwidth.

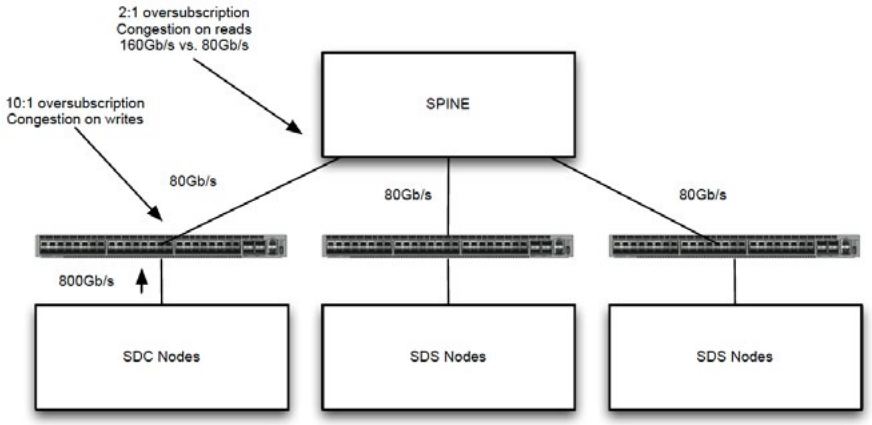

Figure 8: Read Operations

Both the "read" and "write" scenarios, illustrate the behavior predicted during the initial architectural discussions and illustrated in Figures 4 and 6. As with the "write" scenario, in this test, there is no background traffic, only ScaleIO is running. IOPs remain high, but are more variable per flow. There are no discards, and latency is consistent within 1.5ms at the application layer. The queue depth measurements in Figure 7 represent values at given moments in time across multiple interfaces. With only ScaleIO traffic, running at 45% of line rate, 10MB of buffer utilization is seen at multiple time periods across multiple interfaces simultaneously.

As is evidenced in the "read" scenario, when the distributed ScaleIO software is receiving data from a higher bandwidth source, i.e. the SDS nodes which have 160Gb/s of uplink capacity vs. the SDC nodes which only have 80Gb/s of uplink, congestion is likely to be seen at the spine layer. As the ratio between the different node types widens, buffering requirements will increase. As storage capacity and network scale increase, in terms of the number of racks, similar stress will be placed on the buffers of the spine.

Figure 9: Congestion during network reads

The "write" test highlighted a different issue, but still one, which can be understood and planned for. When oversubscribing the uplinks as is common in leaf/spine topologies, queue-depth must be carefully monitored using Arista Latency Analyzer (LANZ). As IO capabilities of the devices increase (bandwidth and IOPS, using flash for example), the likelihood of experiencing drops on the uplinks increases dramatically with oversubscription Arista's deep-buffered 7280SE switches accommodate these bursts of traffic and absorb them without drops without having to resort to more expensive alternatives such as eliminating oversubscription.

While shallow-buffered switches may seem to offer cost savings at first glance, when comparing a 3:1 oversubscribed 64 port deep buffer switch which costs $30,000 to a 1:1 non-oversubscribed 64 port shallow buffer switch which costs $20,000, the story changes. The shallow-buffer switch can connect 32 servers at a cost of $625 per server connection, whereas the deep buffer switch can connect 48 servers at the same $625 per server connection resulting in a 50% capacity increase for the same price per usable port.

As the 7280SE switches did not discard packets during these tests, additional congestion in the form of non-storage traffic was added to ensure that buffers were overrun and quantify the effects of discards. During a simultaneous read/write test, it was observed that experiencing congestion to the point of minimal discards (a max of 200 packets/second discarded) led to major variability in the system's performance from the standpoint of application latency, a critical factor in storage performance.

Figure 10 shows application response time during the test with only ScaleIO when there were no drops.

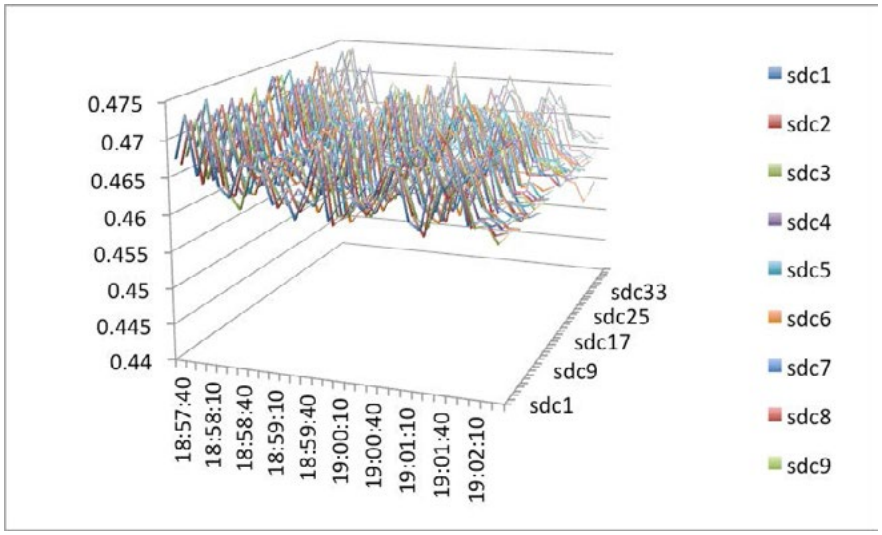

Figure 10: Application response time (ms) with no drops

As can be seen in figure 10, the application response time is consistent over time for each SDC, with variance measured at approximately .015 ms. In contrast, when drops occur, certain flows are affected rather dramatically with response times spiking up to 50ms! The affected flows are not predictable, which is consistent with the TCP bandwidth capture effect described in the Arista Networks whitepaper "Why big data needs big buffers" which can be found at http://www.arista.com/assets/data/pdf/Whitepapers/BigDataBigBuffers-WP.pdf

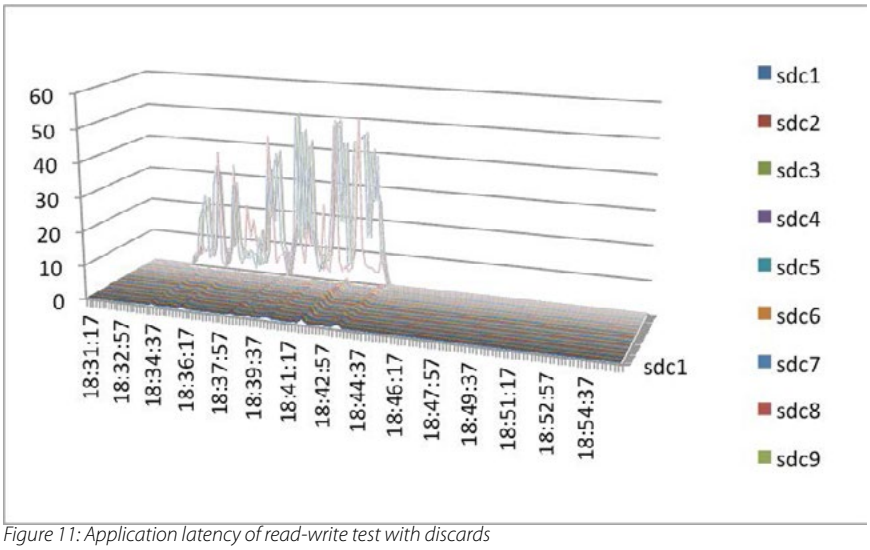

Figure 11: Application latency of read-write test with discards

As the preceding results show, big buffers are a critical component of deploying a high performance IP storage architecture with ScaleIO. The software is more than capable of generating enough traffic to saturate the highest performance switches, especially given the parallel nature of the connections the software launches.

Failure Testing

During failure testing, all components in the network were failed and then restored to service and the impact to the system as a whole was monitored. During the failure testing, no lasting impact to the service was observed. Both the network and software were able to recover within seconds in the worst case to all failures introduced. The application response time in Figure 10 illustrates this quick return to normal service once all retransmits are completed after failure.

As shown in figure 12 an Arista deep buffered network incurring multiple and repeated failures is still able to support a vastly improved ScaleIO experience when compared to a network which experiences discards due to buffer exhaustion as shown in figure 12.

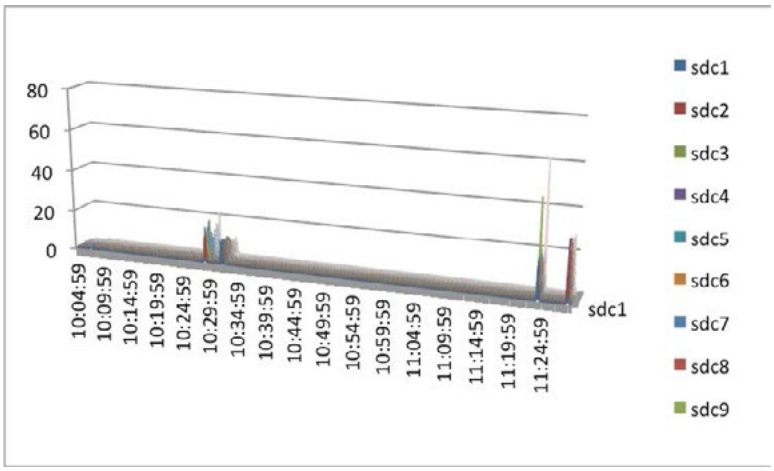

Figure 12: Application response time over multiple network failures

This is attributed to a solid, field proven network design, extremely low re-convergence numbers in the Arista switches, and ScaleIO's reliance on TCP to retransmit the few packets lost in flight. Perhaps the most pronounced aspect of incurring network failures was the bandwidth shift towards the reduced set of network components available post failure. Buffer utilization on the spine switches passed the trigger threshold during multiple failure testing scenarios despite ScaleIO traffic being run at a relatively low rate with no other network traffic. This highlights the need to design the network and ensure adequate buffer capacity for a worst-case situation.

Interestingly failures caused spikes in traffic utilization, not just where the failure occurs, but throughout the network. For example, a front-end failure caused massive spikes in bandwidth on the back-end network utilized for communications between SDS nodes. While the maximum re-convergence for the network was always in the low single digit seconds at the worst case, the buffer utilization resulting from high bandwidth spikes, primarily on the back end, persists as shown in figure 13.

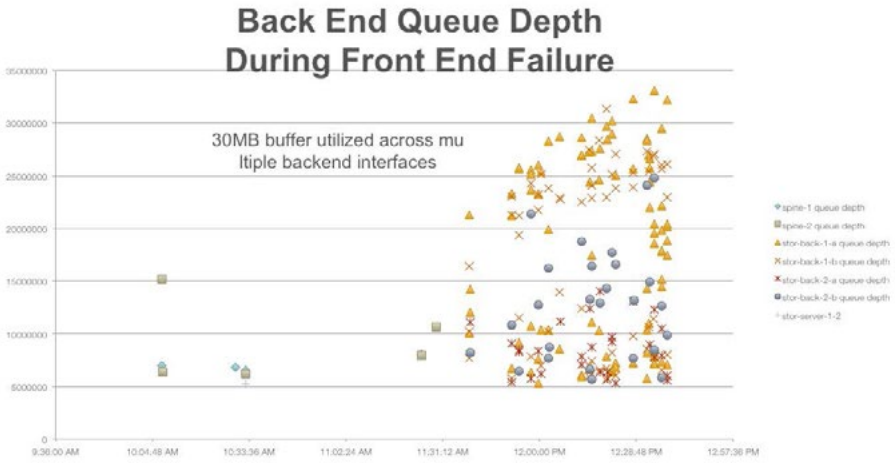

Figure 13: Buffer utilization after front-end failure

Findings

The Architecture designed by LFI, EMC ScaleIO and Arista was able to scale and meet the design objectives of developing an agile, scalable, and resilient, high performance storage environment.

The Arista Spine Leaf Network accommodated worst-case "read" and "write" loads with no packet drops due the deep buffer and VOQ architecture inherent in the 7500E and 7280SE switches. There was no congestion observed in the testing except during congestion testing and the buffers on the 7500E and 7280SE managed bursts during failover events without loss. 52 MB of buffer was utilized per interface in many of the tests.

ScaleIO Software performed at speed and at scale without bottlenecking the performance at any point. The maximum IOPS using 7200 RPM discs was a 25k per SDC. The SDC throughput could have reached 55k IOPS per node before becoming bandwidth limited. The total throughput measured 1M IOPS for the 20 nodes SDS cluster.

Final Design

Based on the test results as described, LFI decided on a design leveraging the dual spine topology as tested. As was tested, dedicated nodes will run the SDS software. These nodes are to be built using commodity server hardware. Dedicated mode deployment is enhanced, by running a discrete network to handle inter-SDS traffic. By virtue of the existence of dedicated SDS nodes, the bandwidth from the spine to leaf in the SDS pods is completely protected. The oversubscription ratios were reduced to match the oversubscription ratios deployed in their existing Arista IP fabric, i.e. 3:1. This changes the uplink bandwidth from 80 Gb/s per rack to 160Gb/s per switch for a total of 320Gb/s per rack. Dedicated networks are used from the SDS nodes to the Storage-Leaf switches. The Arista 7280SE-64 was chosen as the leaf switch for the network, and the Arista 7508E is used for the existing IP Fabric spine, as well as the new IP Storage Spine. This choice was justified based on the heavy buffer utilization seen in the testing, along with the zero drop environment they enabled. The topology is defined in Figure 14.

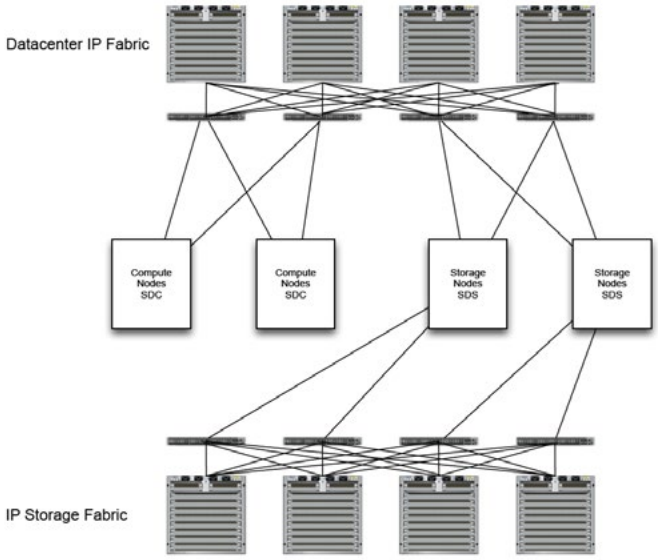

Figure 14: Final design

40 Gb/s uplinks are used for the IP Storage Fabric providing 160Gb/s of bandwidth out of each back-end leaf. To protect traffic on the compute leaf switches, a dedicated queue is allocated for ScaleIO traffic on both the spine and leaf switches, so that at the end of the day, a highly resilient dedicated path is provided from end to end, between all nodes in the storage topology. In the end, all storage traffic follows either a dedicated physical path, or a dedicated queue across a shared path.

The bandwidth available is head and shoulders above what's available in their legacy FC SAN, and the storage network has no fate sharing to increase reliability. The limitations of SAN environments have been eliminated without sacrifice to any of the features and functions available to a legacy SAN, all while increasing performance, and decreasing cost.

Conclusions

LFI, ScaleIO and Arista have demonstrated that world class scalable storage solutions can be cost effectively built using commodity hardware while achieving high performance, high throughput, reliability, availability, and agility. By using Arista 7500E spine and 7280SE leaf deep buffer "gold standard networking for IP Storage" together with EMC ScaleIO Enterprise IT departments can now achieve Cloud Scale agility, performance and economics for their Software-Defined Data Centers.

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