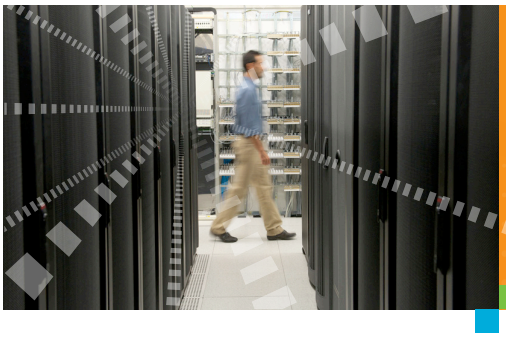
LSI
Storage. Networking. Accelerated.

White Paper



Switched SAS
Sharable, Scalable SAS Infrastructure

Abstract

As application intelligence and new techniques for safeguarding data emerge, the subject of bringing storage loser to the server arises. Direct connection to external storage has unique attributes for providing low latency access, fail-over capability, and favorable management and deployment characteristics. This paper brings to light some of the application advancements and how they align with expanded Serial Attached SCSI (SAS) protocol capabilities.

To address emerging applications that provide local/remote data replication, streamlined messaging, and increased transaction capability, a new focus has been placed on Switched SAS. This evolving technology is driving new ways to connect storage by delivering increased cable throughput, simple scalability, and do so within a familiar and tightly coupled storage model. By applying this sharable, scalable infrastructure there is no longer a concern over the familiar refrain, "purchase more servers to add more storage."

Terry Gibbons, Director Software Planning and Strategy, LSI Corporation

Terminology: neither "Switched SAS" nor "SAS Switch" are defined by the International Committee for Information Technology Standards (INCITS) T10 Technical Committee. For purposes of this discussion a SAS Switch is a managed set of SAS Expander Devices. Switched SAS is a way to conceptualize Expander functionality where T10(1) describes Expander functionality as a connection router and connection manager.

Throughout this paper, various deployment options will be discussed. To understand the deployment models, a review of the underlying technology is required. Figure 1 represents SAS Expander functionality. This is an any-to-any connection matrix noting that one could connect initiator-to-initiator but most likely has no reason to do so.

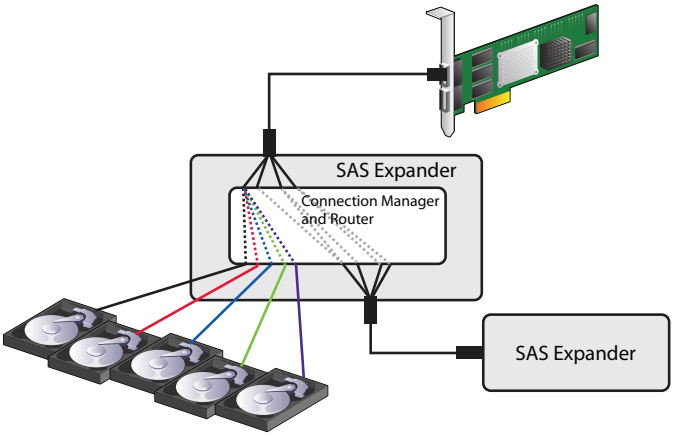
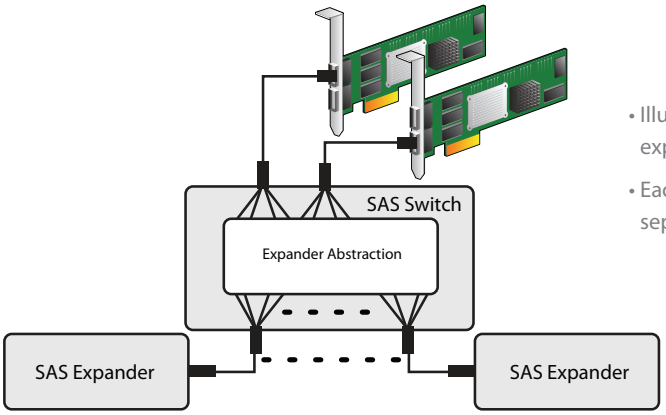


Figure 1. Expander Functionality

Figure 2 shows an outline of a SAS Switch and how it could be used in a broader topology. Note how SAS is currently deployed with four physical links (Phys) per cable and two cables (ports) per SAS Host Bus Adapter (HBA or "initiator").



- Illustrates two initiators to man expander-based storage units
- Each initiator could be in separate host

Figure 2. SAS Switch Model

Figure 3 conveys the concept of Zoning. This is a method available in SAS Expanders that allows individual devices, or groups of devices, to be hidden from selected host-based initiators. In the illustration, we see many drives isolated in Zone A to Host 1. Through the same expander (part of the JBOD), we see two drives in Zone B that are isolated to Host 2. If additional drives were added via JBOD B then they could be seen by both hosts, zoned individually, or even hidden from both hosts as a reserve pool. Drives directly attached to a host are seen only by that host.

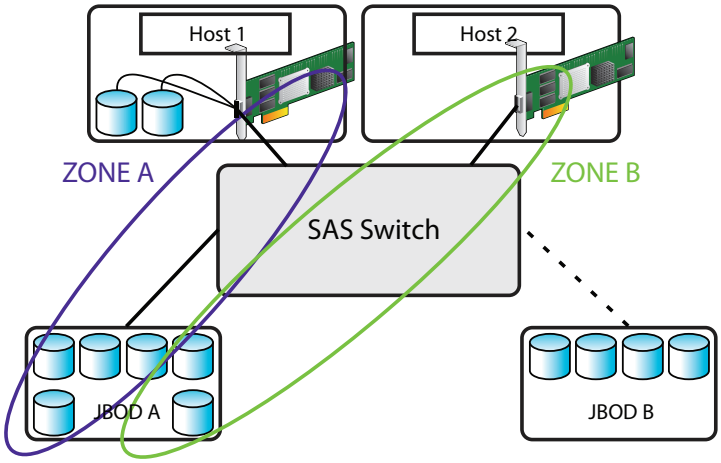


Figure 3. SAS Zoning

SAS Technology

SAS is a point-to-point technology where transactions require a complete connection from a host to a target (storage) device. SAS uses "port" to refer to a collection of links that usually represent traffic over a cable. Depending on the complexity of the topology, a transaction may require several links to connect to one another before completing a connection to a target device.

A unique feature of SAS is the bandwidth build into the cables. With each cable representing four SAS links, this represents 4800Mb/s per port, full-duplex at a 6Gb/s link rate. A 16-port SAS Switch (1U / half-width) can accommodate 768Gb/s bandwidth. Later on will be a review of total SAS Switch bandwidth and theoretical performance in server environments.

This level of bandwidth aggregation lends itself to low power consumption. While the total package of processor, protocol controller, and case + materials may be very close across SAS, FC, and GbE technologies, the fact that SAS is intended to support many links in a consolidated package provides attractive watt-per-port numbers usually below 5W.

This power profile can support 10m passive copper connections. 2010 product introductions support 25m active copper capability and future roadmaps from the SCSI Trade Association(3) indicate support for 100m optical cables. The actual/estimated power requirements for these technologies have yet to be considered as part of the public domain.

SAS zoning is not address based. Again, this follows the point-to-point concept where end-devices (initiator or target) may or may not be allowed to complete a transaction based on the sum of the physical connections that must be created.

Emerging Applications

Hardware and software applications are growing in their capability to support new features and performance criteria within many traditional computing environments. Tiered Storage (e.g. variable access requirements within databases and search engines), Distributed Applications (e.g. email applications), and Virtual Servers (e.g. server racks/pods for a Mega Datacenter) present a need for low latency and a high transaction rate while not sacrificing high availability, data integrity, or data recovery.

For purposes of this discussion, Tiered Storage is the concept of how quickly an application needs to access data. Is the data hot, warm, or cool? In this case, the emerging technology is both hardware and its access parameters set by the application. For "hot" data, like that associated with database transaction logs, high performance computing, or data mining, an SSD may be used to maximize IOPS especially if the transaction size is in the 2KB to 16KB range. If performance is important but not critical then SAS HDD's may be used at the next level. Finally, data warehouse or documentation may reside on large capacity HDD's such as SATA. Figure 4 gives an example of a Tiered Storage deployment where SSD resides in the server, SAS drives have the most immediate access externally, and SATA drives are cascaded.

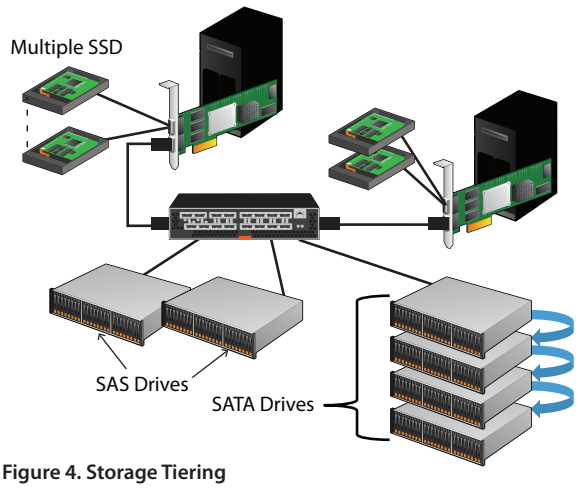


Figure 4. Storage Tiering

Distributed Applications certainly require network infrastructure to connect computing resources across a campus or across the globe. There are requirements for fail-over, high availability, and data backup/restore. However, these applications have requirements for a high transaction rate that may be best served by SAS storage as part of a larger infrastructure. Emerging applications provide local and remote replication services for backup and disaster recovery. Examples of these services are Cluster and Standby Continuous Replication(2) (CCR/SCR) where local repositories can be managed as external storage attached to a server (see Figure 5). In this case, a port (or physical connection) on the HBA, Switch, or Disk Drive could fail and a back-up path is available. Yet, this also keeps the passive node storage separate from the active node storage.

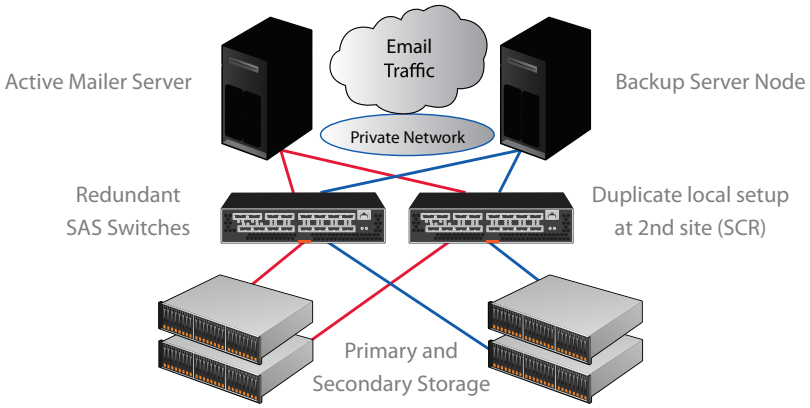


Figure 5. Local/Continuous Replication

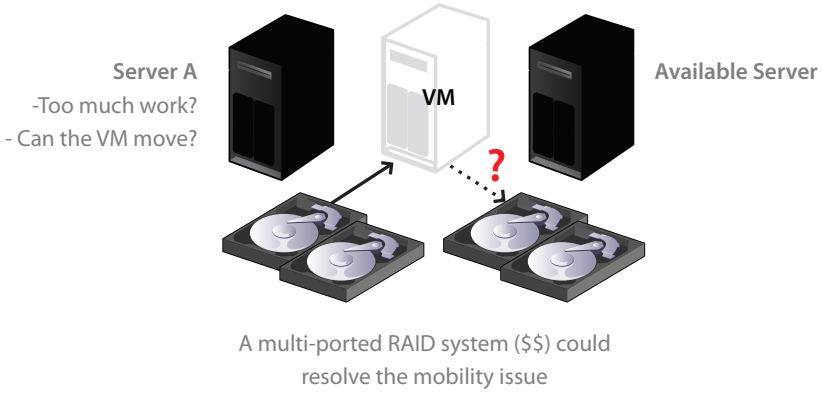


Figure 6. Old DAS Model

An example of Virtual Server deployment is in the large, homogeneous racks of servers + storage that are built for economy. These racks may be further bundled with other racks to create a "pod". While many pods may make up a Mega Datacenter, it's the racks and pods that can benefit from externally connected SAS storage as the most efficient configuration because the storage can be shared amongst the servers. Refer to the contrasting images presented in Figures 6 and 7. Emerging applications in the form of virtual server mobility help keep the OpEx in line with the CapEx by simply shifting workloads to another physical server without reconfiguring hardware.

Virtual Server mobility is assisted by the concept of exposing all storage to a mobility manager via a SAS Switch. By placing metadata on the disk drives indicating which virtual server image

is stored there, the mobility manager can easily tie a virtual server image, an its applications, to a virtual machine at the physical server level.

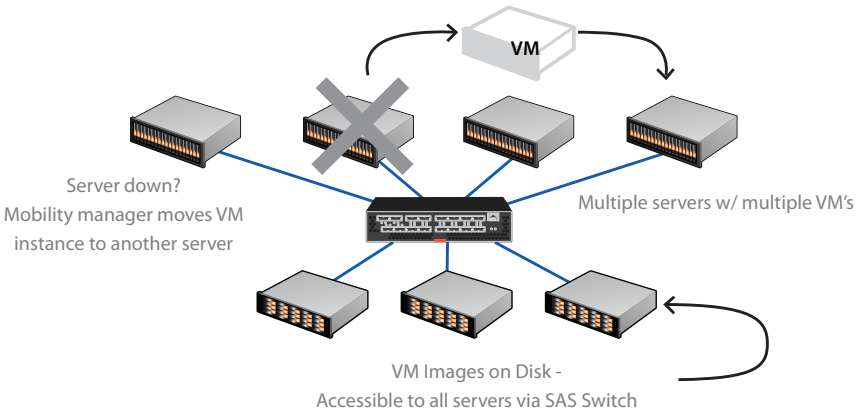
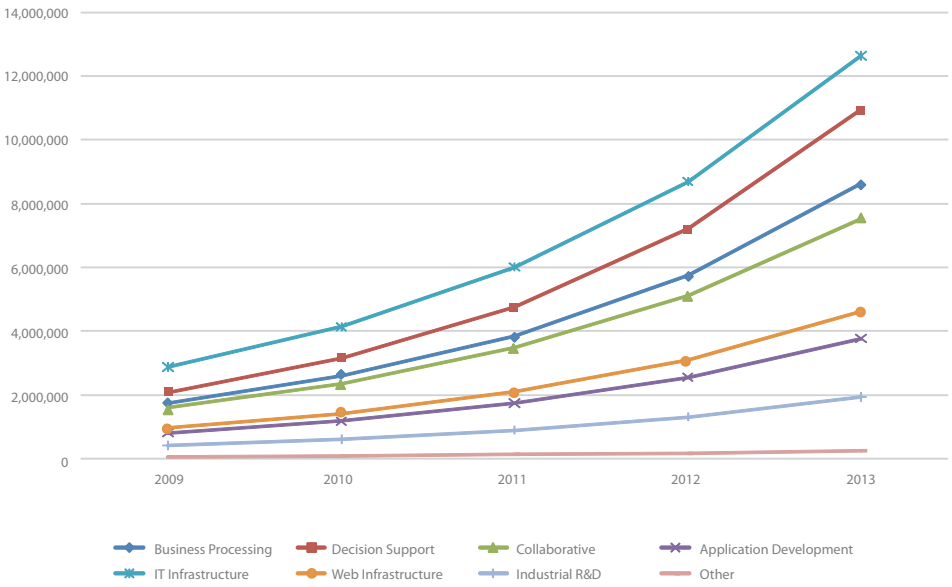


Figure 7. VM Mobility Vs. External Storage

IT Requirements

Enterprise Datacenters operate with a core set of philosophies that every solution must address. There is a continual growth in data (see Figure 8) and required services that must be managed within highly scrutinized CapEx and OpEx budgets.



Scalable/Sharable. Simplicity of storage expansion, re-provisioning, or varying deployment models. In the Switched SAS model, one may add storage without disruption. Broadcast events notify members in the topology of a change and the new storage is automatically "discovered". The storage may be further segregated (zoned) to a specific server, or set of servers, by out-of-band management utilities. This simple philosophy of conveying presence and taking action applies to re-assigning storage to a new function or deploying different numbers of servers and storage units. Simply plug in the components and the topology is understood by all entities.

The SAS standard has proven its flexibility in accommodating legacy SAS technology. Expecting this trend to continue allows for future storage expansion even with new, faster technology, thus preserving current investments. All SAS components have options to work at less than their maximum link rate while supporting legacy protocol. Therefore, a 12Gb/s SAS HBA could be deployed in a new server and connect to a 3, 6, or 12Gb/s SAS Switch. The 12Gb/s SAS Switch could connect to 3, 6, or 12Gb/s JBOD's. The JBOD's could connect to 3, 6, or 12Gb/s storage devices.

Ease of scaling can be applied to modest setups or more complicated setups that extend into dozens of servers and hundreds of disk drives. By layering SAS Switches and cascading storage units, the potential for access to thousands of storage devices exists (see Figure 9). Future SAS roadmaps are driving towards larger Switch port-counts such that many configurations avoid layering the Switches.

High Availability. Figure 10 shows a Blade deployment scenario. Each Blade Server has a custom form-factor HBA to fit best within the server blade. Note how multiple SAS Switches can be used to complete a separate IO path (best accomplished using dual-port SAS drives). If any connection fails, the workload can be re-routed. Also, this example uses RAID controllers in the external storage unit to further secure the data.

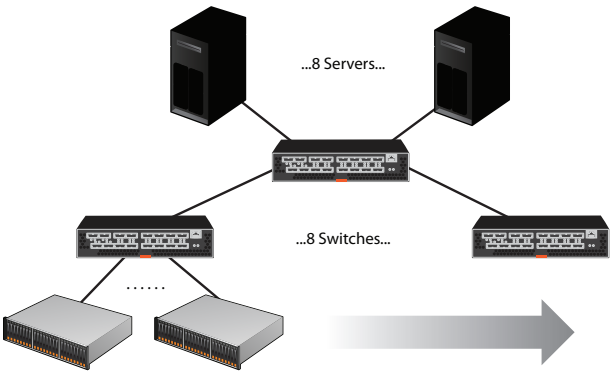


Figure 9. 1,536 Drive Configuration

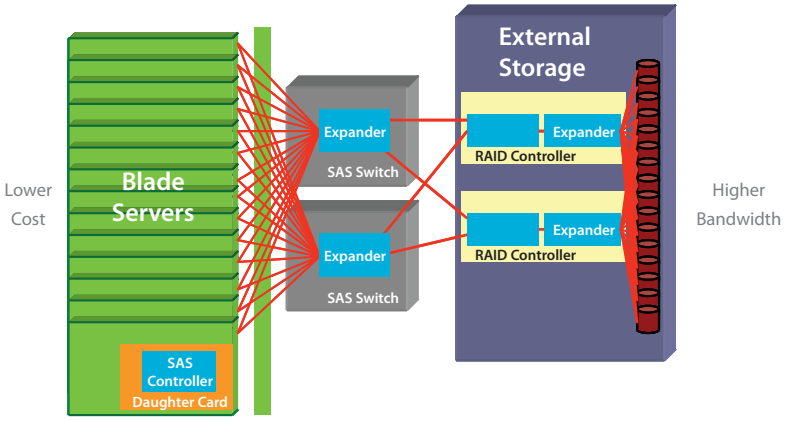


Figure 10. Blade Chassis w/ SAS Switch

Manageability. The SAS infrastructure provides low-touch management. Along with the automated discovery for adding/removing storage, management tasks such as software/firmware updates and new installations are designed to be "low-touch". Well behaved software at the host level informs the operating system that a device is "busy" when a link is dropped for a short period of time. Firmware should be designed to minimize re-discovery times, or not drop links at all, in case of an upgrade. Installations can be replicated by pre-assigning connection attributes since there is no device or address specific knowledge involved in the connection (e.g. deploy the same zoning configuration for identical server/storage racks).

Configuration of zoning (Figure 3), as well as a centralized point of contact to view/manage the topology, is through the SAS Switch. By providing a network interface, TCP/IP, and Telnet services, one can manage the Switch out-of-band. Also, the SAS Switch is a logical focal point for physical storage maintenance. By having the Switch in place, layers of storage units may be reduced or eliminated thus making replacement and servicing easier by not tracing cable hops through a challenging set of connections.

Theoretical Performance

Performance aspects of a SAS Switch are focused on throughput and latency. IOPS isn't considered in this discussion as the disk drive (or SSD) and HBA have a profound effect on IOPS capability. Latency is Switch/Expander centric while throughput is a function of PCI-Express bandwidth in the host, the SAS link rate, disk drive capability, and the Switch's ability to route the traffic.

Switch Bandwidth. The concurrent bits of data that a 16-port SAS Switch is intended to support is as follows:

- 12Gb/s (full duplex) x 16 ports x 4 phys per port = 768Gb/s

- A nominal SAS configuration puts four SAS physical links (phys) into one port, hence, one cable

Using a 16-port SAS Switch as an example, we examine the best scenario where there are eight upstream ports to the host servers and eight downstream ports to JBOD storage. Throughput is discussed in terms of half-duplex, despite the full-duplex capability of SAS, to accommodate practical expectations of how servers and storage integrate within a SAS infrastructure

Throughput per port:

- 6Gb/s x 4 phys = 24Gb/s per port
- (24Gb/s per port) / (8b/10b encoding) = 2.4MB/s per port
- 2.4MB/s per port x 88.33% (to accommodate arbitration delays and additional framing) = 2160MB/s per port

SAS Switch Throughput in a practical application:

- 2160MB/s per port x 8 port pairs = 17,280MB/s per SAS Switch

Requires one link from the initiator to the SAS Switch, and another from the SAS Switch to the next SAS Switch or Expander (i.e. JBOD), to complete a connection from server to storage

The time delay of completing a physical connection impacts overall latency within an Expander-based SAS topology. Times of 100ns or less are to be expected for each expander in the connection path. Referring back to Figure 4, there are two levels of latency. Noting that each JBOD has one expander, the SAS drives each have two "hops", one for the Switch and one for the JBOD. Contrast this to the last JBOD in the stack of SATA drives where there are five hops to get to a disk drive. Fanning out storage at the highest level is a requirement for maximizing performance.

Actual Performance

Test Setup:

- Intel-based white-box servers
- Single Intel Xeon® 5520 microprocessor per server
- LSISAS9200-8e® 6Gb/s SAS HBA "Initiator"
- Astek® 6Gb/s SAS JBOD
- Seagate ST9146803SS®, Firmware v0006
- One LSISAS6160® 6Gb/s SAS Switch (all configurations)

- IOmeter 2006 in client/server configuration
- IO Queue Depth of 8
- Disk drives limited to range of 1000 LBA thus avoiding Seek operations
- See Figure 7 for similar layout

Test Cases (Sequential Read and Write 256KB Block Size):

1) 8 Servers w/ 1 HBA (x4 SAS) per server, 8 JBOD, 16 HDD per JBOD

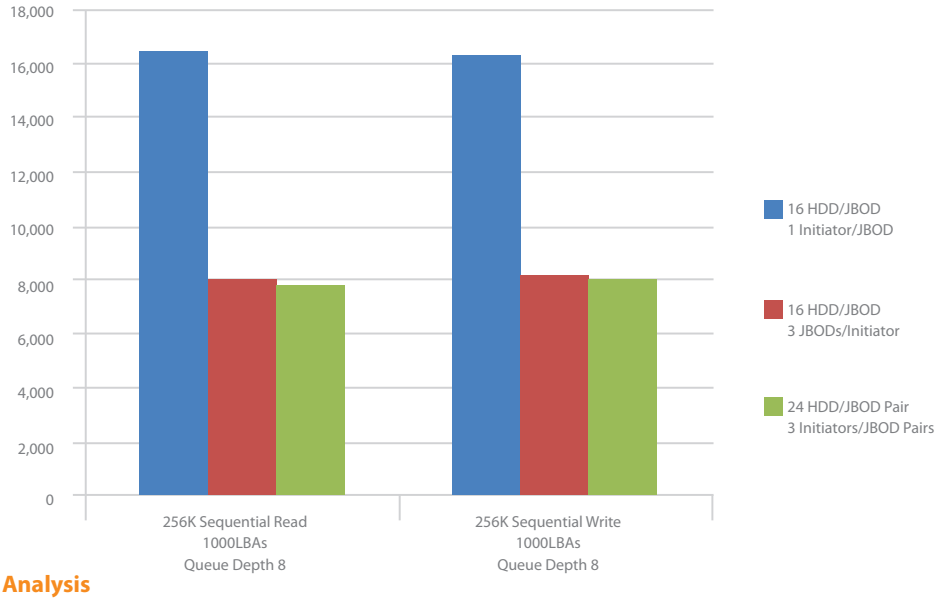
- 1:1 Server to JBOD ratio – best performing solution

2) 4 Servers w/ 1 HBA (x4 SAS) per server, 12 JBOD, 16 HDD per JBOD

- Less processing capacity, more storage capacity

3) 6 Servers w/ 2 HBA (x4 SAS) per server, 4 JBOD pairs (cascaded), 24 HDD per JBOD pair

- Note how cascading the JBOD's adds an additional hop w/ a minor performance impact



Analysis

With optimal conditions set, the Read and Write tests measure approximately 16,400MB/s or 95% of the goal. As server-to-storage ratios were changed, the expected throughput was cut in half. This is due to the fact that either the server side or storage side was limited to four ports for each of these experiments.

Summary

The old adage, "I need more storage, I'll buy another server," need not apply any longer. The sharable nature at the core of SAS Expander Devices allows creation of new products such as the SAS Switch that is capable of managing traffic connections in a broad topology. New connector technology and further expansion of Expander connection management capabilities offer a roadmap to greater possibilities in the future.

Furthermore, SAS is served by new techniques for data management and replication. These new applications have a two-fold purpose: First, to supply new ways to provide high availability and reliability; Second, to challenge the status quo in terms of new levels of transaction processing and throughput.

SAS protocol and the SAS Switch provide functionality for simplifying the storage environment and reducing operating expense by providing:

- Simplified cabling that reduces/eliminates hops that add performance delays and physical layout challenges
- Equal performance across all drives and servers
- Single configuration, reporting, and control point via a SAS Switch
- Stateless servers: any-to-any topology, enables virtualization mobility
- Access control via SAS Zoning

SAS is an excellent infrastructure for making affordable, manageable, and compact storage deployments an attractive alternative to network based storage designs.

List of references

(1) Working Draft Project American National T10/2124-D Standard: Information Technology – SAS Protocol Layer (SPL)

(2) Concept from Microsoft® Exchange Server 2007® at technet.microsoft.com

(3)SCSI Trade Association (www.scsita.org): Advanced Connectivity Solutions Unleash SAS Potential – SCSI Trade Association White Paper (Author: Harry Mason, Contributor: Jay Neer)

LSI
Storage. Networking. Accelerated.

For more information and sales office locations, please visit the LSI website at: www.lsi.com

North American Headquarters
San Jose, CA
T: +1.866.574.5741 (within U.S.)
T: +1.408.954.3108 (outside U.S.)

LSI Europe Ltd.
European Headquarters
United Kingdom
T: [+44] 1344.413200

LSI KK Headquarters
Tokyo, Japan
T: [+81] 3.5463.7165