

White Paper

Switched Storage (Switched Bunch of Disks or SBOD) Technologies in
Real-Time Mission Critical Environments

By Jason Mancebo

Overview

This paper discusses the benefits of a switched (SBOD) storage device versus a single or dual fibre channel arbitrated loop (JBOD) storage device. It reviews switched FC-AL storage as well as port bypass based arbitrated loop technologies, presents an overview and discussion of real-world issues in both technologies and reviews the reliability and performance of these technologies.

1 Progression of Technology

When computer systems first emerged as reliable tools in the video and film marketplace, SCSI (Small Computer Storage Interface) was the de facto (and only) choice. First starting at less than 20 then up to 30 then 40 megabytes per second (MB/s), SCSI was a true solid performer and continues to be so with its current Low Voltage Differential or LVD 160 MB/s and LVD 320 MB/s technologies.

Unfortunately, SCSI is limited to relatively short distances and it is limited to direct attached storage solutions. With the advent of fibre channel (FC), users opted for the more robust, faster (at the time), and flexible in distance and topology FC technology.

While FC is more robust and flexible, at the time it was invented, point-to-point FC (or fabric) was cost prohibitive and the Fibre Channel Arbitrated Loop (FC-AL) was invented. Today most FC storage systems use FC-AL for their drive connectivity, and maintain the loop topology to this day.

This second storage wave introduced many benefits to the demanding real-time and high performance storage environments. Distance limitations were nearly eliminated and in doing so, pockets of desk-side storage islands that were isolated and could not be accessed by hosts other than the local system were moved to centralized machine rooms. Storage efficiencies increased and with the introduction of fibre channel hubs and switches from manufactures like Brocade and others, a complete and functional storage area network (SAN) infrastructure was born. The FC storage network now enables multiple video, film or computer systems to utilize all centrally located and managed storage resources as if local and centrally located and managed resources have reduced the cost and increased the reliability resulting in much more efficient solutions to users' storage problems. While current storage networks are much more advanced and offer much better solutions than did their predecessors, the issues of the arbitrated loop on the connection to disk (back-end) still present themselves as serious detractors in high availability storage to the mission critical demands of the entertainment and digital video marketplace. (See figure 1)

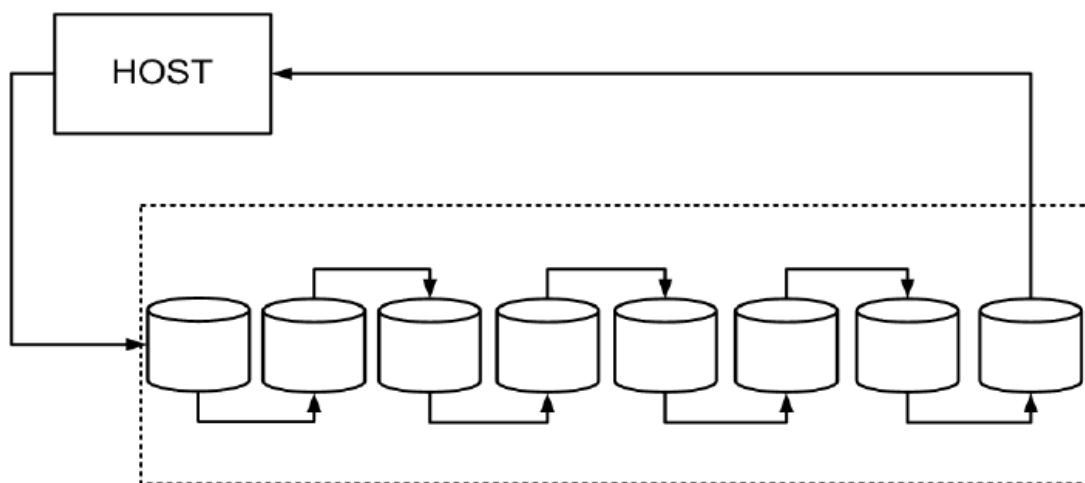


Figure 1 – JBOD Signal Path

These critical FC-AL issues are as follows:

1. Loop Initialization Procedure (LIP) on errors
2. Bandwidth limitations & Degradation (when using high drive counts on loops)
3. Bandwidth limitations with multiple hosts on a single or multiple loops
4. Reliability, failure prevention and fault isolation

All of these issues are attributed to each disk drive being on an arbitrated loop data path. The key to avoiding and solving these important issues is to eliminate the FCAL loops and establish point to point connections between the disk drives and system controller. This ensures individual disk drives have no effect on one another

but maintain one to one relationships with their initiator, effectively eliminating the data path.

2 Next Generation Brings a Sea Change – SBOD

Removing the problems that are inherent to FC-AL was one of the last hurdles preventing delivery of the most desired, most robust and highest performance storage network to systems with real-time demands. By constructing a non-blocking switched drive connection rather than a loop connection with all devices on a data path, it is possible to eliminate the critical FC-AL issues and develop a much more secure, robust and valuable storage solution. The SBOD or **Switched Bunch Of Disks** integrates the disks of a JBOD with a fibre channel loop switch to provide a full point-to-point storage topology. (See fig. 2)

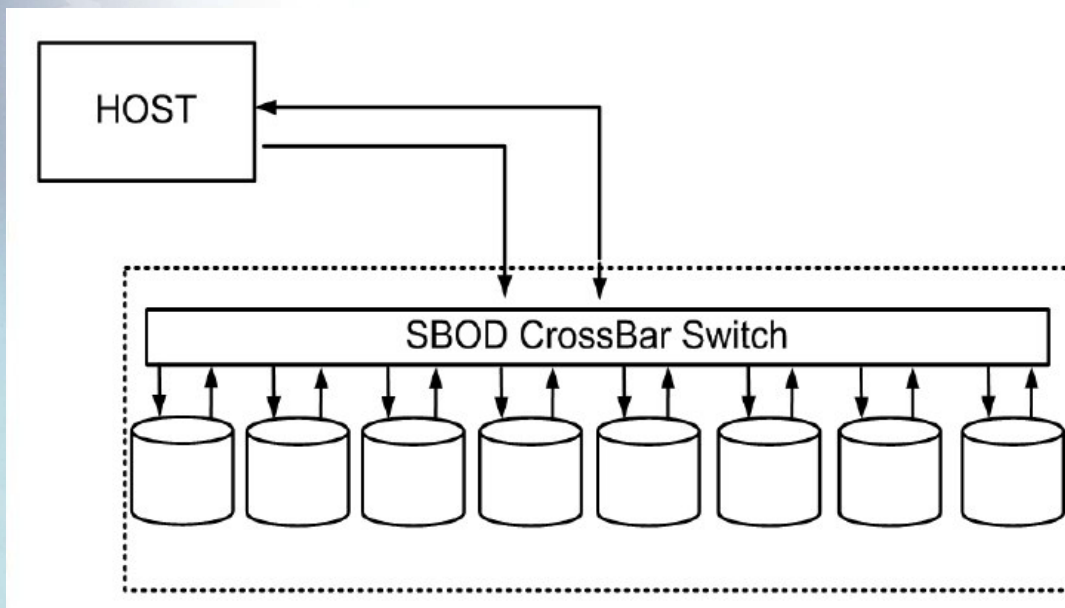


Figure 2 – SBOD Data Path

By the Numbers – Addressing and Solving the Critical Issues

A) LIP Resets

The characteristics of FC-AL can be limiting in an environment requiring real-time data availability, and a limited discussion of important performance characteristics is in order.

Due to the arbitration scheme employed by FC-AL, a serious error in communication or removal of a device requires the loop to be suspended and reset with a loop initialization process (LIP). No data conversations can occur until the physical loop is re-established and all the devices accessing the loop initialize, obtain address, and begin arbitrating. This remains the case for all planned maintenance or unplanned failures. During the LIP, all data is inaccessible. This data inaccessibility is inconsistent with the requirements of a real-time application's need for deterministic and constant data (See Figure 3)

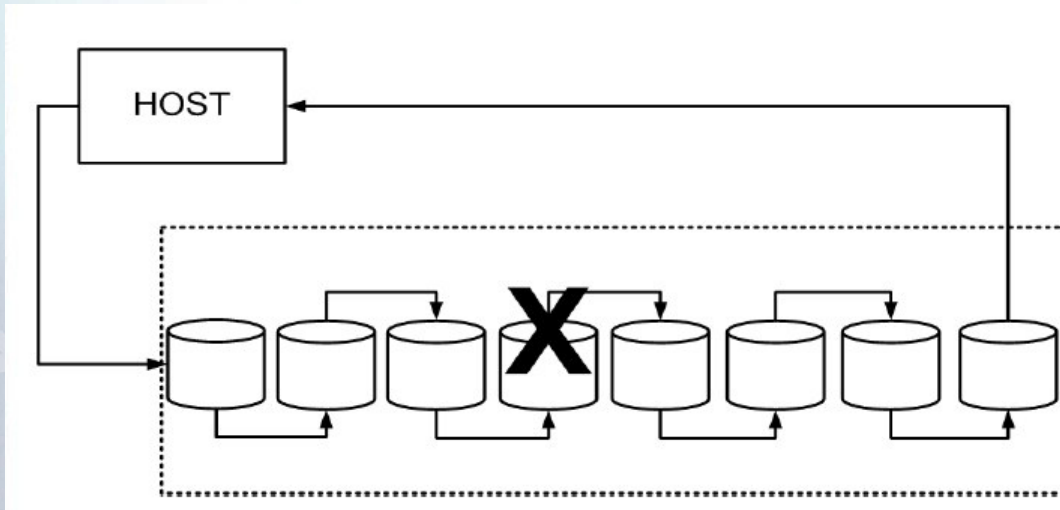


Figure 3 – LIP Reset in Data Path on JBOD

As indicated in the above diagram, any latency caused by an error or other failure on any drive in the loop causes the entire loop to cease data I/O and begin a LIP reset sequence. No data I/O can occur during a LIP reset and recovery. If such an event does occur, the application using the storage device experiences excessive latency. This latency takes the form of data starvation, in the case of playback, or buffer overflow, in the case of recording (if the duration of the LIP event exceeds the duration capacity of the application to use cache in memory and or disk). As an SBOD device does not involve all drives on interconnected loops, but each individual drive in each I/O, an error event on a single drive does not affect the entire system, but is an isolated drive only event (see figure 4).

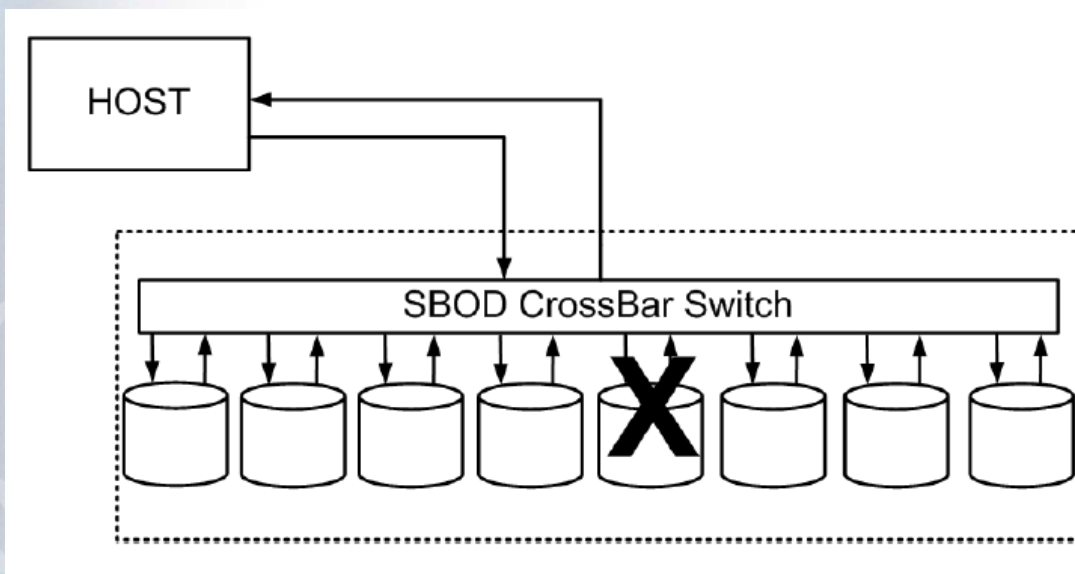


Figure 4 - Failed Drive in Switched Storage Device

Managing the event becomes the preferred method of maintaining real-time data integrity and availability in mission critical applications. This greatly extends the availability and decreases or eliminates excessive latency to ensure performance is maintained. The failure rates of these systems are also dramatically reduced.

B) Bandwidth Limitations

For the purposes of this discussion, all bandwidth rates and numbers are described in half-duplex mode. While greater bandwidth is possible in full-duplex mode, the inherent nature of a disk drive only allows it to read or write at any one time. The maximum aggregate bandwidth of current FC-AL interfaces is fixed at 2 Gb/sec or 200 MB/sec. Many quality JBOD systems are designed with the potential to utilize up to four segmented loops because the maximum theoretical bandwidth for the system is fixed at 800MB/sec. Although there are 8 SFP connectors, only four are host ports. The others are output ports for daisy-chaining additional expansion enclosures (see figure 5)

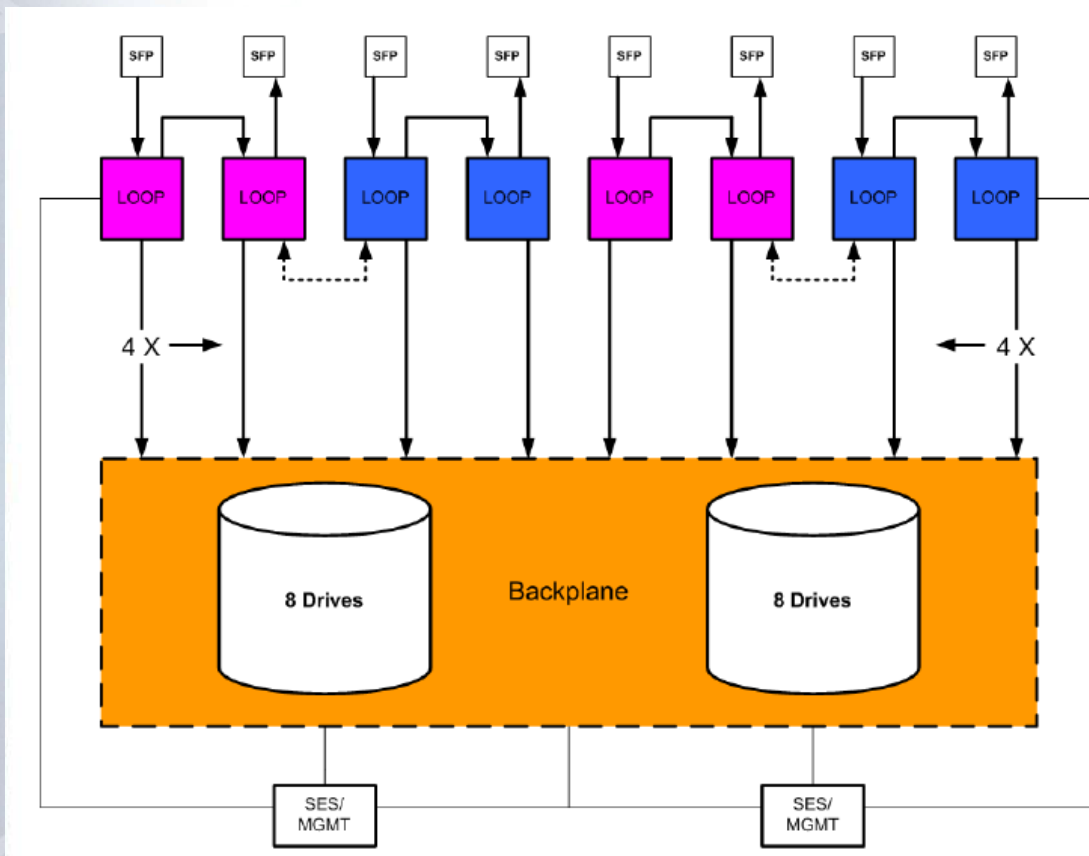


Figure 5 JBOD Loop Diagram

By removing the loops and implementing point to point connections with the disks through the switch, eight host connections are enabled and their combined bandwidth potential raises to 1600 MB/sec or 1.6 GByte/sec. (See figure 6)

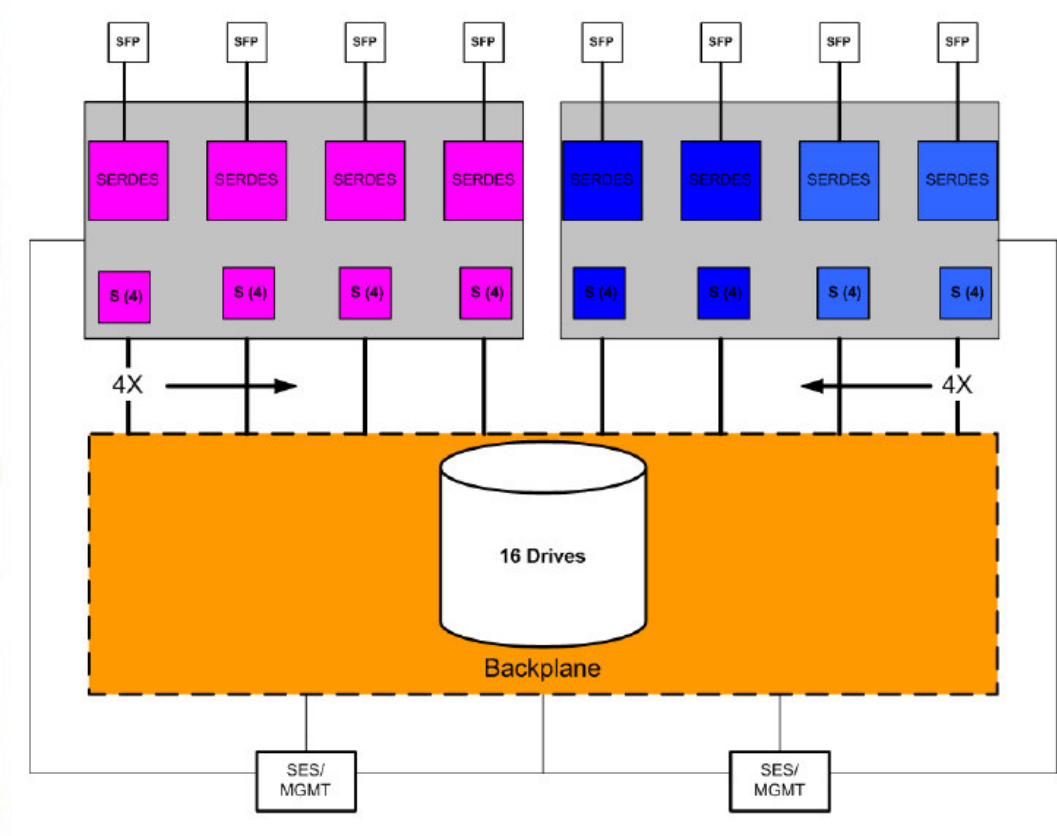


Figure 6 SBOD Interconnect Diagram

With current FC drives comfortably delivering approximately 70 MB/sec in sustained sequential read I/O, only 12 drives are necessary to over-saturate the a four loop JBOD system (12*70= 840 MB/sec). Addition of further drives can only increase capacity in the system, not performance and in fact, will negatively impact latency and performance on FC-AL loops. In order to increase the bandwidth beyond this level, the addition of another enclosure with additional loops is necessary. While this is certainly technically possible, it further increases the cost and decreases the efficiency of the enclosure. In order to use the potential of all sixteen drives (16*70MB/sec= 1.12GB/sec), one would need to purchase two enclosures and four FC host bus adapters (2 port) in addition to the 16 drives. (See figure 7)

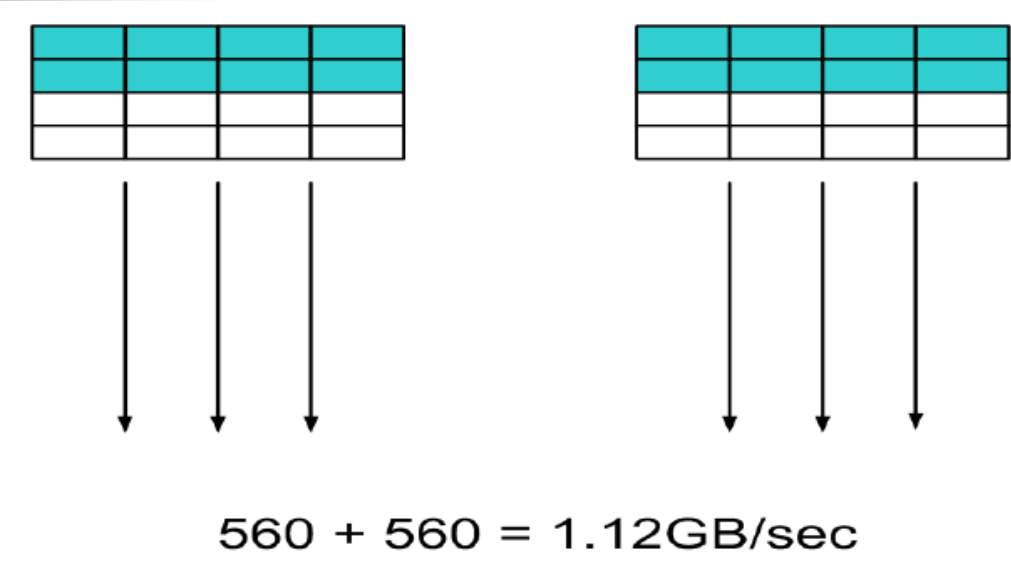


Figure 7 – 2 Enclosures with Total of 16 Disks (shaded)

Conversely, in a switched storage system, the maximum bandwidth limitations imposed by a common FC-AL are removed. In these systems, maximum bandwidth is 1.6 GB/sec. Using an SBOD system for this discussion and to deliver the full capacity of 16 drives, only one enclosure is needed as the bandwidth provided is greater than the drives could sustain ($16 \times 70 = 1.12 \text{ GB/sec} < 1.6 \text{ GB/sec}$ capacity). (See figure 8)

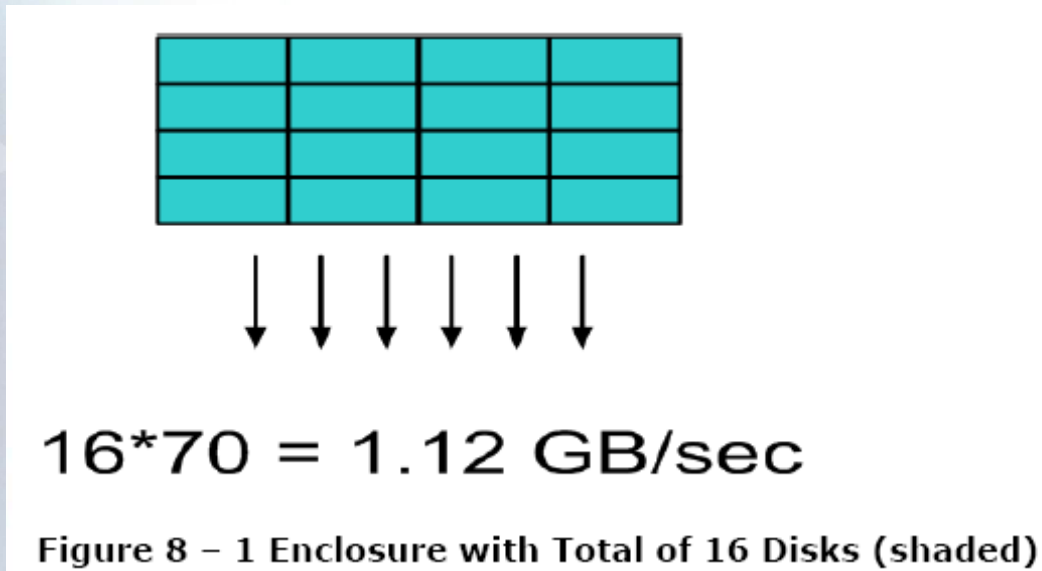


Figure 8 – 1 Enclosure with Total of 16 Disks (shaded)

At the JBOD level above, the price to deliver the full performance capabilities of the 16 drives includes:

- 2 Enclosures
- 4 Dual Port Host Bus Adapters
- 16 FC Drives

To deliver the same performance with SBOD using the same assumptions, the price includes:

- 1 Enclosure
- 4 Dual Port Host Bus Adapters
- 16 FC Drives

As indicated above, moving to the SBOD technology reduces not only the cost of purchase, maintenance and operation of an extra enclosure but it also decreases the rack space requirements by up to 50%. Ignoring price performance and discussing strictly comparative performance, the SBOD system again shows its strength. In JBOD systems using FC-AL, each node (host or disk) on the loop is responsible for its own address on the loop and must arbitrate (negotiate) for control in order to communicate within the storage system. Although 126 nodes can be attached in a loop, only one conversation can occur at any point in time. Each node adds latency and will reduce overall bandwidth. Each loop is limited to 2 Gb or 200 MB/sec of bandwidth. SBOD uses fibre channel loop switches to remove shared total aggregate bandwidth limitations of 2 Gb/sec. Rather than 2 Gb/sec maximum aggregate bandwidth, any port on the switch can provide full-speed access to all other ports, thus allowing 2 Gb/sec bandwidth per switched connection or disk and removing the need for all disks to share bandwidth. As enclosures are cascaded to connect to further drives the latency in an SBOD solution only increases by that incurred by the SBOD device not the 16 drives in that enclosure. If large drive configurations are required on the loop then the link between enclosures will limit the overall systems performance to 200MB/s. The SBOD technology allows a unique feature 'dual path trunking', this automatically provides twice the bandwidth between cascaded enclosures. By increasing the bandwidth available to each port, and by removing the necessity for common arbitration overhead, SBOD provides progressive bandwidth gains. In the

following comparison, a typical sequential read I/O pattern is demonstrated in both JBOD and SBOD systems. While this comparison is 2 loops, not 4, it is indicative of the performance gains seen. (See figure 9)

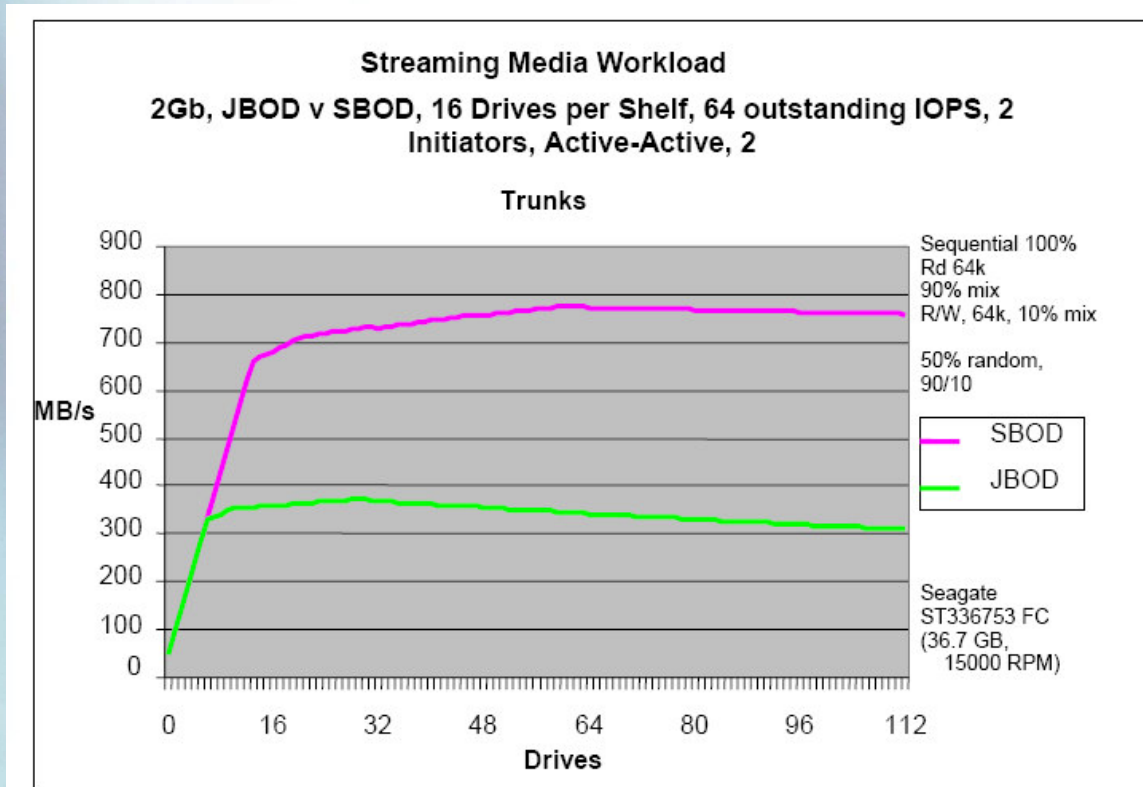


Figure 9 – JBOD v. SBOD Bandwidth

As indicated in figure 9 above, the SBOD system is capable of delivering much more performance. It does, in fact, *double* the performance at all drive count levels (at and above 16). As demonstrated, SBOD does not suffer from the typical performance degradation experienced with large count JBOD systems. This is due to the vastly reduced latency in the SBOD system which is especially prevalent when compared to a JBOD system with more than approximately 32 drives on a FCAL.

C) Performance with Multiple Hosts Attached

Another issue that is critical is that of multiple host systems. On a dual loop JBOD system, each of the drive’s loops is completely isolated and can only be seen by hosts in that particular loop. If there are multiple hosts wishing to be able to read or write to drives on multiple loops, a switch must be included, dramatically adding cost to the storage system. With the current popularity of shared storage and shared file-systems, this is an issue which will increasingly need to be addressed. With the SBOD solution, all drives are available to all hosts with equal performance. The addition of the external switch is not necessary, and configuration and costs of the system are greatly reduced.

D) Reliability / Failure Prevention

In looking at current storage systems used by digital media professionals today we find that the fundamental factors to failures lead to at minimum, excessive latency and at worst, catastrophic failures that are both mechanical and electrical in nature.

As an example, an intermittent or poorly performing optical interconnect SFP is quite common, the cause can be extremely difficult to isolate and diagnose as its effects can appear on data transfers to any devices on the loop. Excessive delays in getting the system back on line often cause schedule delays and result in lost revenue.

The de-facto solution to allow data to bypass a failed or missing drive in an FC-AL JBOD system is the use of a port bypass circuit or PBC. Unfortunately, this is not a very intuitive device. It is binary in nature and its only capabilities are to direct data to the drive it is connected to, or allow data to bypass said drive. With SBOD technology, the link to each device is via a mechanism that is aware of the data flowing through it, each device is effectively on its own isolated link rather than the overall loop. This allows an autonomous monitoring facility to check for errors at each connection point and take the appropriate action to manage and report them. This provides heightened preventative monitoring of all devices within the system on a "per device" basis. Configuration of these facilities is available through an extended SES (SCSI Enclosure Services) facility (SES pages 80-86). The technology includes the following autonomous functionalities:

Intelligent Device Monitoring

"Per device monitoring provides extensive and critically needed improvements over simple PBCs. In a storage system where a new drive is being introduced for the purposes of scaling the amount of storage or to replace a defective drive, SBOD architectural advantages extend far beyond the simple signal detection or drive presence offered by a JBOD. SBOD's ensure a device meets the basic requirements of the FC-AL protocol and in advanced SBOD implementations a new drive can be quarantined to ensure correct operation before being allowed admission to an on-line storage pool. This ensures a faulty drive does not bring down an entire loop or SAN and does not prevent access to any other data. Once allowed admission to on-line storage, the intelligent monitoring continues. The state of all the links is continually monitored and if a problem is detected the offending device can then be autonomously detected and removed from on-line storage. This implementation improves access to critical storage resources and critical data, and boosts reliability to levels non-existent before the introduction of SBOD technology.

Trend Monitoring

SBODs provide access to several metrics and diagnostics tools that can be used to help monitor trends over time or during an active troubleshooting session. SBODs in particular are in a unique position to monitor the primary source of failures – the hard drives.

Traffic Monitor

The amount of traffic, in percent of frames or switching per period, can be tracked in order to help monitor bottlenecks and help in load leveling or to detect problems. A drive may show no symptoms of communication problems, but is actually experiencing severe access problems. Monitoring traffic over time will help pinpoint this failure mode" (Advancing Storage Reliability – T. Hammond-Doel, - 2003)

Also available for monitoring are:

- Word Error and CRC Error Counters (providing details of communication errors detected at each and every link)
- CRC Error Source ALPA (Allowing the source of errors to be tracked)
- Relative Frequency Check (Providing isolation of drives with faulty oscillators)
- Ordered Set Detector (Used to trap protocol errors in the FC system)

The migration to SBOD and managed, monitored systems from unintelligent JBOD systems is truly of benefit to the critical needs of the customers seeking real-time performance and the unique mission critical nature of the systems used.

3 Conclusion

While, on the surface, switched back end storage systems offer obvious benefits in the areas of increased bandwidth and price performance, added benefits are clearly provided in the reliability and stability demanded by digital media professionals with critical real-time and data accessibility needs. Demonstrated in our discussion was the elimination of LIP's in nominal operations as well as data error events, the extended SES monitoring and management facilities as well as fewer components to potentially fail. While increasing bandwidth, density, price performance, up-time, accessibility and reliability, this unique technology can also provide a significant reduction in the total cost of ownership of a storage systems and network. It provides the changes needed in the storage systems to allow operators, engineers and owners to focus on their individual business critical activities whilst allowing the robust infrastructure they purchased to provide them with more benefits than just simple data storage.

Sources

Advancing Storage Reliability – T. Hammond-Doel, (2003)
Protected SANs for Broadcast Environments - C. Jason Mancebo (2001)