Electronically Switched Architecture for NetApp Enterprise Storage Configurations

Kris Neff | August 2004 | TR 3265

TECHNICAL REPORT

Network Appliance, a pioneer and industry leader in data storage technology, helps organizations understand and meet complex technical challenges with advanced storage solutions and global data management strategies.

Abstract

Network Appliance constantly searches for innovations to improve its portfolio of hardware and software products. The recently introduced electronically switched hub (ESH) module, for use in NetApp® disk shelf products, is a great example of continuous improvement at the hardware level. By employing leading-edge embedded Fibre Channel switch technology, NetApp is able to offer improved functionality, serviceability, and performance to our customers. This paper describes the following improvements in more detail.

Table of Contents

1. Increased Link Speeds

When combined with the DS14mk2 FC disk shelf, storage configurations run at 2Gb link speeds. The performance of an SBOD (Switched Bunch Of Disks) is much improved over a JBOD (Just a Bunch Of Disks). SBOD performance increases are dependent on both the technology and the number of drives in a FC-AL loop. As we are introducing new technology, system performance gains from the technology are most interesting (the relationship between number of drives and performance is well documented and understood). The performance difference between 1Gb and 2Gb link speeds is most readily apparent during degraded mode (RAID reconstruction) operations. During degraded mode, loop traffic greatly increases from movement of (additional) reconstruction data. Depending on the nature of existing workloads, this can result in slower response times to end-users. Link speed increases are one way to offset the impact–larger “pipe” connections result in a system that is more capable under heavy traffic conditions.

2. Increased Capability and Performance

From a purely technical viewpoint, fibre channel loops support 126 devices. From a practical position, traditional FC-AL daisy-chain topologies (e.g. loop resiliency circuits) require limits on the number of devices for performance reasons. The performance impact is directly attributable to loop overhead traffic. In the past, the NetApp recommendation for LRC topologies is 56 devices per loop.

Advanced FC-AL topologies allow the “cost” of loop overheads to be minimized, thereby increasing the number of supported disk drives. This is true for system configurations that include ESHs. ESH topology is a hub and spoke arrangement with local neighborhoods surrounding each device. Loop overhead is minimized by the fact that traffic no longer flows through each disk drive. The hubs are capable of local communication to the disk drives and then more efficiently conveying this information to the filer.

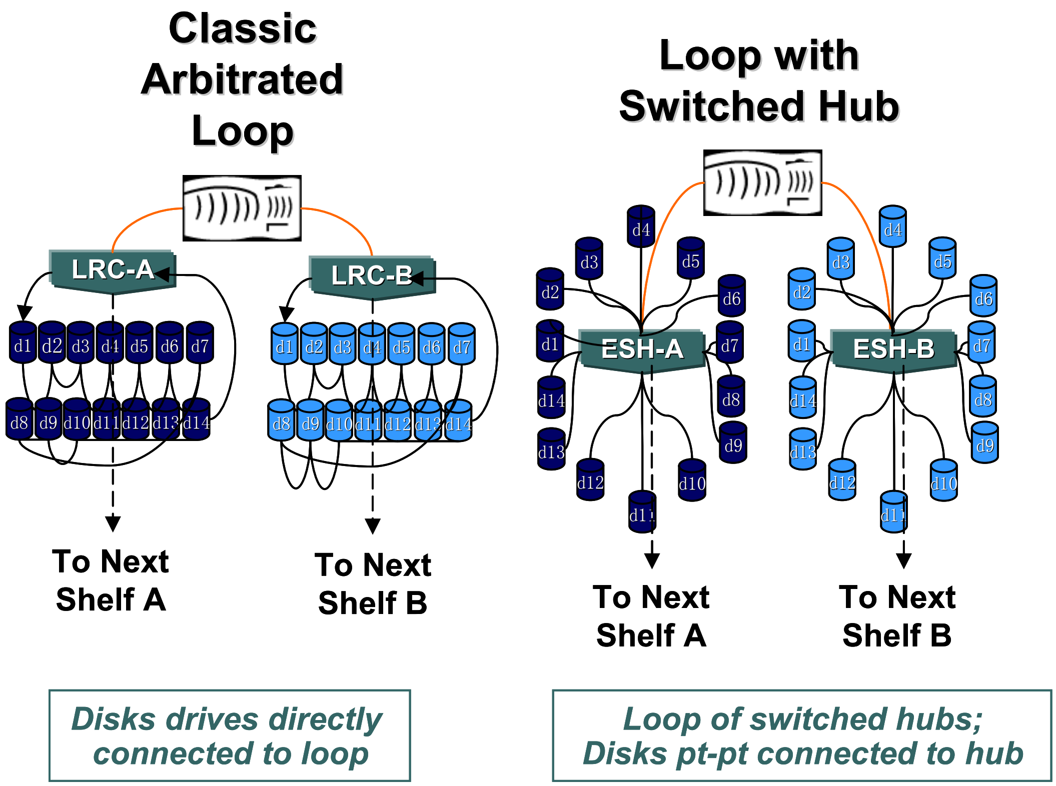

Figure 1: Loop Topology Differences

The practical implications from more densely packed loops are:

Economics and simplicity – reduced number of host bus adapters in the filer

Better configurability and upgradeability – fewer PCI expansion slots required for disk connects

With the introduction of the ESH product, the NetApp recommendation for loop devices increases to 84 devices.

3. Serviceability Improvements

SBODs introduce new advanced diagnostic tools:

Intelligent per device monitoring

Intelligent per device monitoring provides extensive improvement over simple JBODs. SBODs diagnostics and monitoring tools allow for direct allocation of failing components in a SAN environment.

Storage systems typically see new disk drives as either a replacement or to increase the amount of useable capacity. During these events, SBOD architectural advantages extend far beyond the simple signal detection or drive presence offered by a JBOD. SBOD ensures that the device meets the basic requirements of the FC-AL protocol. Additionally, in advanced SBOD implementations, a drive can be placed in its own private storage segment to ensure correct operation before being allowed admission to on-line storage. This powerful feature ensures that a rogue drive does not bring down an entire SBOD – or even worse – an entire SAN.

Once a device has been qualified and allowed admission to on-line storage, the intelligent monitoring continues. The state of the SAN is continually monitored and if a problem is detected (for example, the device identifies the SBOD as being in the incorrect state for a set period of time), the SBOD identifies and removes the offending device from on-line storage.

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 SAN 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 showing no other symptoms of a problem, such as Ordered Set (OS) or Cyclic Redundancy Check (CRC) errors, but appears to be completely operational may actually be experiencing severe access problems. Monitoring traffic over time will help pinpoint this failure mode.

Word Error and CRC Error Counters

All words on each link are continuously monitored and errors in any 10B word are examined for compliance to specifications. A count of detected word errors is maintained. All frame data is also monitored and a calculation of the CRC is performed on-the-fly allowing tracking of frame errors. Typically, word errors occurring in a frame will also generate a CRC error.

Monitoring the number of word or CRC errors over time provides visibility into the general health of a storage system. Trending errors over time, however, provides a means of predicting failures before they occur. Predicting and preventing a hard failure before it occurs extends system uptime.

CRC Error Source ALPA

The CRC of each frame is monitored and calculated. If a CRC error is detected, the Source ALPA of the frame is recorded. In a switching environment this provides another powerful tool for maintaining and advancing SAN reliability. In a JBOD, even if CRC errors can be detected, finding the source of the error can be elusive and may appear on every port of a SAN. With SBODs, a CRC error is usually confined to a single port. With the port and the CRC source ALPA known, identifying the exact location of an error happens much faster.

Relative Frequency Check

SBODs typically connect 10 - 16 drives to 1 - 4 initiators or RAID controllers. This leads to the SBOD being in the middle of a multitude of clock domains. Over time clocks tend to drift and eventually one or more devices may drift out of tolerance. SBODs employ a clock check feature comparing the relative frequency of all attached devices to the clock connected to the SBOD. A drifting clock can be detected or – if several clocks appear to be drifting – the clock of the SBOD itself can be determined to be drifting.

Ordered Set Detector

SBODs also offer the ability to inspect the integrity and contents of the signals and the data traveling through the switch. The information is stored in Ordered Set detectors on each port. This information can be used to help troubleshoot the SAN and to quickly identify problem areas or failure modes.

4. Advanced Storage Reliability

JBODs are used most commonly in storage systems. JBODs provide dual channel access to a group of disks typically using FC-AL protocols. Figure 2 highlights the reliability exposure inherent in JBODs designs today, where a single drive can cause loss of access to all drives in a shelf. JBODs typically use a port bypass circuit (PBC) to automatically bypass a drive when the drive is no longer present or has no signal, i.e., usually a binary decision. SBODs technology increase reliability and recovery requirements of today's 24x7x365 data centers.

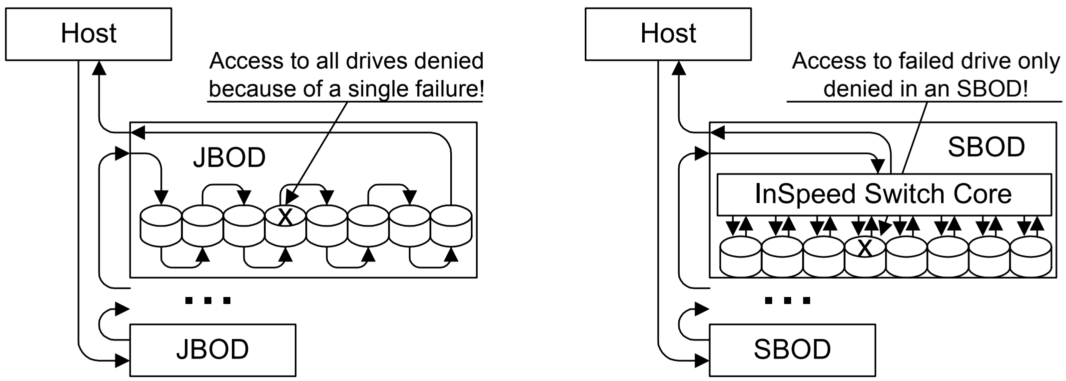

Figure 2 - JBOD versus SBOD

Network Appliance, Inc.
495 East Java Drive
Sunnyvale, CA 94089
www.netapp.com