Advances in Technology—
Redefining Large Shared Storage Systems

By Todd Roth

With the adoption of high definition (HD) and its commensurate requirement for increased storage capacity and bandwidth, the need for scalability has increased. Because of its deterministic performance and dual-ported redundancy, Fiber Channel is the leading technology for storage connection and storage area network (SAN) development (see Annex 2 SAN Storage Interfaces). By applying the latest advances in Fiber Channel technology, larger, more capable, and more redundant systems can be built to meet this need.

By definition, a broadcast video server must base all channel capacity calculations on worst-case system loading. Therefore, a 12-channel server must support 12 simultaneous independent channels, a 120-channel server, 120 simultaneous channels, etc. Whether the same medium is playing back in all channels, or different media in each channel, performance must be constant, quality-of-service (QoS) guaranteed. If all drives are functioning, or one or two drives fail, performance must be constant and QoS-guaranteed. To have a defined deterministic performance and reliability in the broadcast environment, the most rudimentary calculation has always been Channel capacity = Storage bandwidth/channel data rate.

Other factors governing system size include:

1. Maximum (Fiber Channel) FC link bandwidth—limits storage bandwidth.
2. Available FC ports—limits total frames—channel capacity.
3. Drive enclosure/drive arrangement—limits possible drive configurations/drives per FC link.
4. Host memory—limits total storage.
5. FC adapter address space—limits total drives and frames.
6. Bandwidth reserved for media file interchange—reduces total available bandwidth.

The same calculations that determine channel capacity by reserving bandwidth for on-air channels also allow the use of unreserved bandwidth for non-realtime operations such as file transfer. By integrating and managing file transfer protocol (FTP) and common internet file system (CIFS) support, large servers can support file transfer in formats such as material exchange format (MXF), audio video interleave (AVI), and Quicktime, without affecting on-air performance. Archive, near-line backup, off-site mirror, and third-party media interchange are intrinsically supported in such architecture without the need for any additional data bridging or gateway devices.

Obviously, managing the interdependence of these governing factors is a complex challenge. This overwhelming myriad of potential solutions can be addressed by limiting system designs to the most efficient implementation choices.

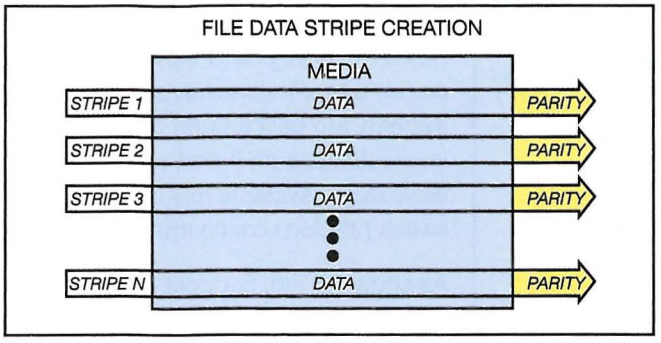

Figure 1. Data striping.

Data Striping and Bandwidth Aggregation

Some storage area network (SAN) designs are developed around the concept of the RAID set and its associated "aggregate" bandwidth. The aggregate storage system bandwidth is created by striping data across individual drives, allowing the individual capacity and bandwidth of each component drive to add to the total sum.

The process involves breaking files down into data blocks and calculating parity. Depending on the data protection strategy, either single redundant array of inexpensive disks (RAID 3) or multiple error-correcting codes (ECC) parities are calculated. Contiguous data and associated parity blocks are organized into data stripes such that the total number of blocks is equal the number of drives in the RAID set. Calculating optimal block size to optimize data transfer performance is necessary; too small is inefficient, too large unwieldy (Fig. 1).

The data stripes are then written across all the drives constituting the RAID set. This allows each drive to contribute in an additive manner to not only the capacity, but also the bandwidth of the aggregate logical volume (Fig. 2).

Non-interruptive expansion is limited to increasing the storage capacity by adding RAID sets. RAID set bandwidth can only be expanded by increasing the number of member drives in each set, a process which requires adding drives to the array and re-striping data.

Host Overview

Each SAN-based host consists of fiber channel interfaces, baseband SDI interfaces, and a gigabit Ethernet connection. For realtime server applications, host performance is best measured by maximum sustained throughput (Mbits/sec). Current and potential platform configurations are shown in Fig. 3.

System Bandwidth Distribution

System bandwidth, aggregated in the SAN, is distributed and shared across all connected host servers. A system designed to have a 4 Gbit/sec aggregate bandwidth is not simply created by connecting four 1 Gbit/sec systems together; each piece of media on the SAN has 4 Gbits/sec shared access (Fig. 4).

Reliability, Resiliency, and Redundancy

SAN-based server system reliability is achieved through support of multiple redundancy and resiliency mechanisms. System resiliency is achieved by eliminating single points of failure and can be approached by either

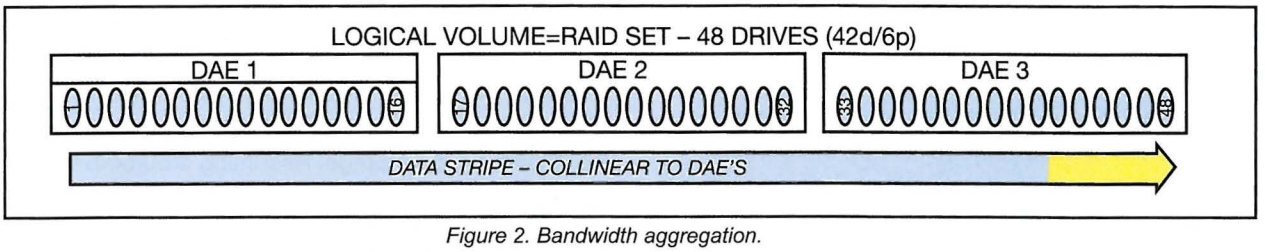

Figure 2. Bandwidth aggregation.

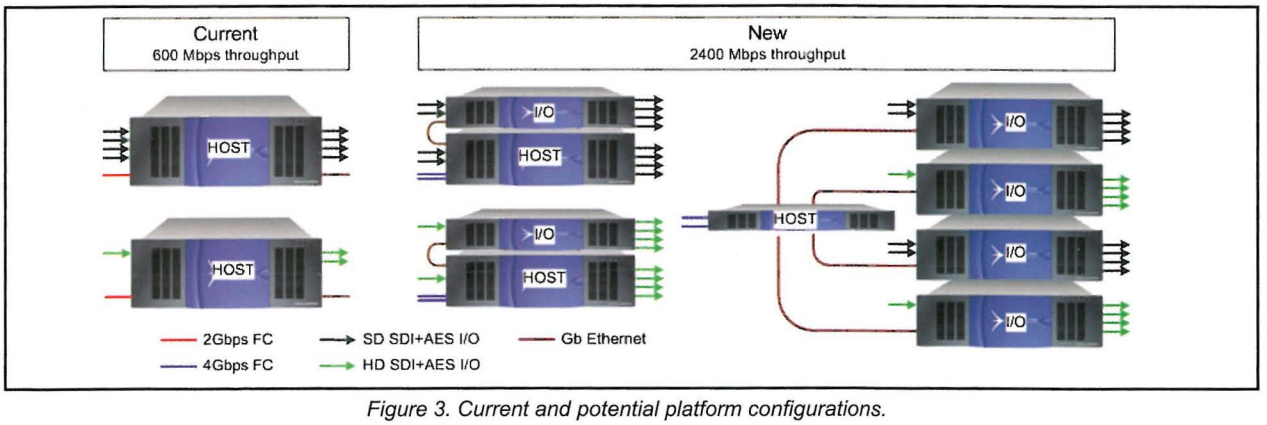

Figure 3. Current and potential platform configurations.

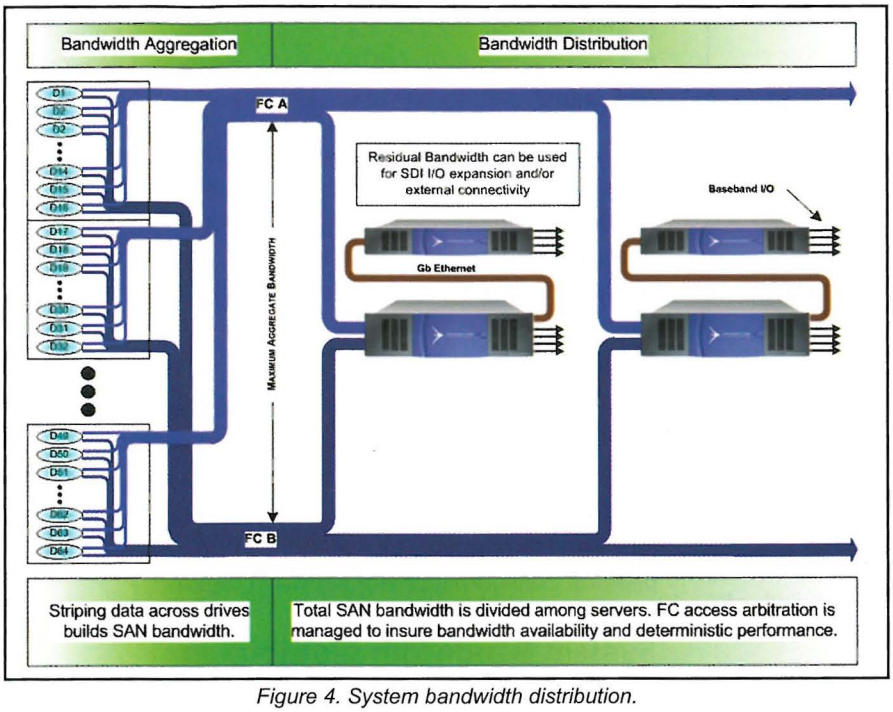

Figure 4. System bandwidth distribution.

SAN Configurations

Nexio FC SAN-based servers rely on two basic architectures. Smaller systems are built around direct host connections to an 8-port storage array, while larger systems rely on a dual-switch (32 port) configuration.

As shown in Fig. 5, direct-connect servers reach a maximum aggregate bandwidth of about 2200 Mbits/sec, while switched systems can reach close to 6000 Mbits/sec. Whereas the aggregate bandwidth of 4 Gbit/sec fiber channel start at nearly 4000 Mbits/sec and can reach 8000 Mbits/sec.

Direct Connect—Maximum 2200 Mbit/sec Bandwidth

Direct-connect systems are designed around the storage bandwidth available by striping enough drives to saturate the aggregation of the two available fiber channel links. Using just-a-bunch-of-disks (JBOD) storage chassis, expansion is performed via dual 4-port hubs built into each drive enclosure. Aside from bandwidth saturation, further limitations are incurred due to the availability of host connection ports and fiber channel address space (Fig. 6).

Dual Switched—Maximum 6000 Mbit/sec Bandwidth

Switched systems are designed around the dual-switched concept. This eliminates exposure of the entire system to a single point of failure at the switch. Even though switches are highly reliable, this system is designed so that a switch failure only affects half of the hosts and has no impact on the storage. Assuming a 50/50 balance of host and storage ports, and subtracting 4 ports for inter-switch connection, 14 ports are available for host connection (Fig. 7).

N+1 or mirroring strategies. In systems with only single-host fiber channel connections, the addition of a single "reserve" host (N+1) protects against most failures, while costly full mirroring is required to eliminate all single points of failure. With the advent of dual-ported FC hosts, the less expensive N+1 strategy eliminates all single points of failure. This mitigates the cost versus protection tradeoff associated with choosing between N+1, or mirroring as a system design approach.

Data protection is achieved through RAID. Smaller RAID sets can be protected against single data errors with a single parity drive, while larger RAID sets are protected against dual data errors by multiple parity drives in an ECC manner. Full dual-path fiber channel allows the development of drive-only (RAID set) mirroring for additional protection against entire chassis failures. Taken a step further, enhanced Ethernet connectivity allows for provision of off-site RAID set mirrors.

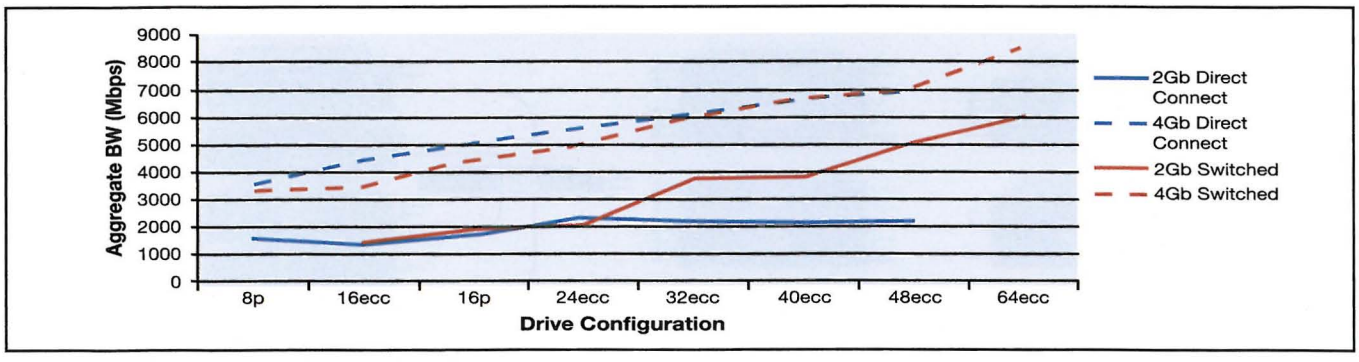

Figure 5. Data Bandwidth -vs- Drive Configuration.

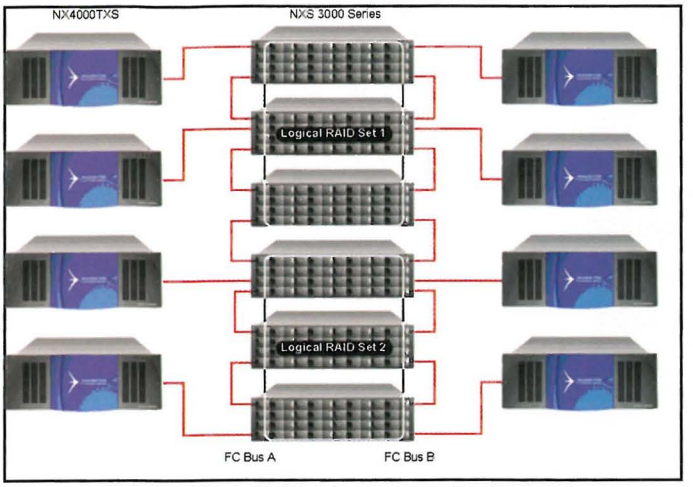

Figure 6. Direct-connect system.

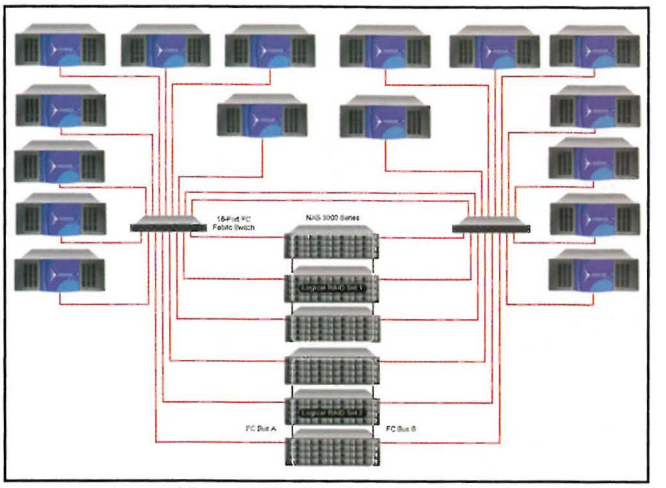

Figure 7. Dual-switched system.

Switches allow for aggregation of more than 2 fiber channel links. Using switched 2 Gbit/sec fiber channel, maximum storage performance can be achieved by striping data across a 64-drive RAID set (organized in 4 chassis of 16 drives attached to 8 switch ports). Storage bandwidth is limited by practical considerations such as number of switch ports, address space, and performance considerations of larger RAID sets. Striped sets beyond 64 incur diminishing returns on performance and unacceptable latencies due to increasing buffer sizes.

Technological Improvement Overview

Several technical improvements and innovations have allowed a revisit of the infrastructure design and scalability of some SAN-based servers. Since the bulk of these improved technologies focus around fiber channel advances, the storage infrastructure of a SAN-based server system potentially reaps the largest benefit (Table 1).

- Transition from 2 to 4 Gbits/sec FC–Moving from a 2 to 4 Gbit/sec FC infrastructure requires the transition or all storage system components, including adapters, switches, drives, and storage enclosures. The aggregate bandwidth of a 4 Gbit/sec SAN could conceivably double 2 Gbit/sec performance, however, practical considerations will impose limitations.
- Support for dual-ported FC HBAs–Dual-ported host adapters allow improved reliability by providing a second, redundant path to each drive. Two independent paths from host to storage, eliminates the loss of system capacity associated with an FC link or switch failure.
- Support for increased FC address space–Along with the new generation of HBAs, comes an increase in FC address space from 128 to 2048 devices. This allows for the support of more drives, increasing maximum storage. This is necessary for dual-ported

Technical Driver	Advantage	Impact	Timeframe
Transition from 2 to 4 Gbits/sec FC	Greater storage bandwidth	More channels	Now under test
Support for dual-ported FC host adapters (HBA)	Fully dual-path fiber channel	Better redundancy	Now under test
Support for increased HBA address space	Increases maximum FC connections	More storage	Now under test
Higher capacity "stackable" FC switches	Better port scalability at lower cost	Up to 64 port switches	Now under test
Transition from 10K to 15K rpm drives	Greater drive bandwidth	Fewer drives to saturate a FC link	Now under test
Introduction of switched BOD drive arrays	More ports/higher bandwidth	Larger nonswitched systems	Now under test
Increase in host memory	Ability to address more storage	Increased media storage capability	Done

Table 1. Summary of Technical Advances and Benefits.

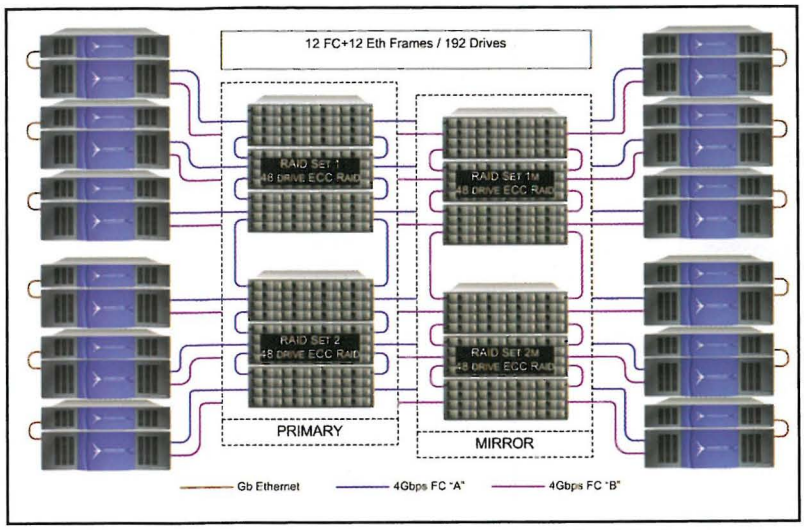

Figure 8. Extension of 4 Gbit/sec direct-connected architecture.

HBA support, since each drive appears as multiple addresses (2 ports x 2 paths = 4 addresses).

- "Stackable" 16-port switches–Stackable switches improve over existing switches by allowing expansion via dedicated inter-switch links (ISL), where connecting older switches requires using 2 switch ports (leaving 28 ports available in a dual-switched configuration). Four dedicated ISL ports per switch allow 4 switches to be cascaded and all 64 ports to be available. This increase in port count allows for the creation of larger, more redundant systems.
- Transition from 10 to 15K rpm disk drives–A faster spindle speed improves the drives data bandwidth. Since our SAN bandwidth is based upon the aggregation of individual drive bandwidths, this transition allows fewer drives to saturate an FC link. Where 10 10K drives saturated a 2 Gbit/sec link, around 14 15K drives should saturate a 4 Gbit/sec link. This should allow maximum (link saturation) bandwidth in a 1 x 16 operating mode.
- Support for SBOD array–The transition from Just-Bunch-Of-Disks (JBOD) to Switched Bunch-Of-Disks (SBOD) increases the port count from 4 shared (via internal dual hubs) to 6 switched (via dual internal switches) per storage enclosure. This increases additive connectivity and allows for triple-port trunking; additional ports for host connection and bandwidth aggregation across 3 instead of 2 FC links. SBODs increase the channel capacity of non-switched systems.
- Increased host memory space–While not a fiber channel improvement in itself, increased host memory capacity from 2 to 4 Gbits/sec is necessary to address more storage and create larger SANs.

Storage system bandwidth improvements not only increase servers I/O channel capacity but also residual bandwidth available for Ethernet interconnectivity. Along with SDI I/O, each host also provides a gigabit Ethernet port that can be used for file-oriented activities such as remote mirroring, archival or third-party file interchange. Server content can be accessed via FTP or CIFS.

Additional host bandwidth and processing power also allows connection, via gigabit Ethernet, of a second "expansion" chassis to each primary fiber channel connected frame, doubling the channel capacity of each host connection. The introduction of hosts that serve as bridges between dual-port fiber channel and multiple gigabit Ethernet ports allow the bypass of limitations based on maximum host limitations of the FC SAN.

Although evolutionary, these technical drivers make it possible to substantially scale up maximum system size, while leveraging experience gained from previous generation designs. While taken together, these advancements allow the option for larger, more scalable, more redundant generations of current SAN architectures.

Potential System Configurations

With the application of the technical improvements, new system designs arise. Based both on direct-connected as well as switched architectures, larger systems with more capability can be developed. Combining Ethernet connectivity with FC SAN performance, new configurations arise that go far beyond existing limitations.

Larger direct-connected systems can be built around more capable SBOD drive arrays, taking advantage of greater aggregate bandwidth, and improved connectivity.

Direct-Connect

This design is the extension to 4 Gbits/sec of existing direct-connected architecture. Additional ports on switched BOD (SBODS) allow for more hosts, while the use of dual-ported HBAs allow for the support of mirrored storage. Ethernet-based I/O expansion further expands channel capacity (Fig. 8).

Switched

With the addition of inter-switch-links (ISL), "stackable" FC-switched SAN designs can grow to support more than 2 switches. Four 16-port switches can create a 64-port

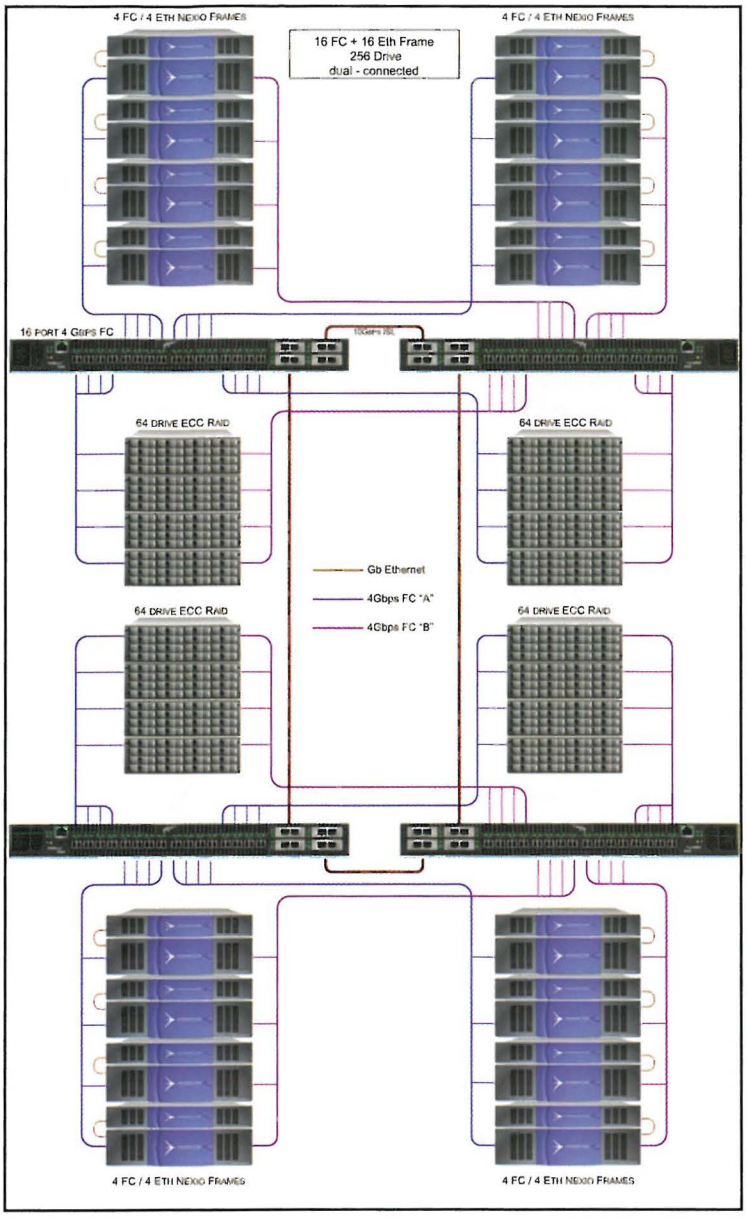

Figure 9. A 4-switch system.

	Channels \SD/HD	Capacity
Direct Connected	96/48	12.1 Tbyte
FC Switched	256/128	68.4 Tbyte
Dual-path FC Switched	128/64	68.4 Tbyte

Table 2. New Targeted System Capacities

switched SAN infrastructure with full dual-path redundancy. With the anticipated introduction of faster, higher performance drives, optimal drive configuration will need to be determined. Exactly how many drives will be needed to saturate 4, 6, or even 8 switch ports, and the resulting aggregate bandwidth has yet to be determined.

Dual-port FC host adapters allows for complete redundancy of the host-to-drive connection layer and overcomes a significant point of failure that has been recognized in dual-switched system design. This comes at the cost of increased host port capacity requirement on the switching fabric.

Again assuming a 50/50 balance of host-to-drive ports, and since ISL links are used to inter-switch connection, a 4-switch system has 32 host ports, providing connection for 16 dual-ported hosts (Fig. 9).

Conclusion

While the introduction of improved technology is a catalyst for the evolutionary design and development of higher capacity server architectures, applying these technical drivers to an existing framework, allows the proposal of future systems that both make technical sense and are constructible (Table 2).

The technical drivers outlined in this paper show that the current path allows for improvement in reliability and resiliency, as well as scalability and capacity, while lowering cost. By leveraging the information technology's industry cost curve, lower per-channel and per-gigabyte of storage costs can be achieved while increasing performance.

The system designs proposed in this paper describe design goals that try to balance benefits and tradeoffs. Some implementations will obviously precede others, some limitations will be overcome, and new limitations will arise.

Annex

1. Brief Glossary of Storage Terms

DAS – direct attached storage—A method of connecting one or more drives to a single host. The host maintains file system directory data and accesses drives via low-level block read/write commands.

DAS Interfaces:

ATA – traditional parallel desktop drive connection, supports one or two drives.

SATA – newer serial attached ATA, supports 1 drive per link.

SCSI – small computer storage interface—traditional parallel enterprise drive connection, supports up to 16 drives per bus. SAS—new serial attached SCSI, supports many drives per link.

SAN – storage area network—A method of connecting many drives to many hosts. Each host maintains file system directory data and accesses drives via low-level block read/write commands.

SAN Interfaces:

FC – Fiber Channel—dual ported network architecture for drive attachment.

iSCSI – internet SCSI—adaptation of SCSI over Ethernet

NAS – network attached storage—a method of connecting many drives to many hosts via Ethernet. File system directories are store in the NAS device, hosts access data via file

commands. NAS systems typically combine DAS storage with an embedded server.

Miscellaneous

ISL – inter switch link—high-performance data link for moving traffic between switches.

2. SAN Storage Interfaces: FC -vs.- iSCSI -vs.- FCoE

Fiber channel has evolved through four generations (1 Gbit/sec hub-based, 1 Gbit/sec switched, 2 Gbits/sec switched and now 4 Gbit/sec switched) as a storage-interconnection optimized architecture. A more recent SAN contender, iSCSI maps the SCSI disk access protocol over Ethernet, taking advantage of existing LAN components and other off-the-shelf networking components. Lacking the ubiquitous nature of Ethernet, fiber channel is the only true end-to-end storage network; FC host adapters connect to FC switches to FC drive enclosures to FC drives.

While not as scalable as Ethernet iSCSI based SANs, fiber trades off scalability for deterministic performance. Fiber Channel's low-overhead state machine-based media access control (MAC) layer allows for sub 1 ms latencies, whereas the collision-based MAC of Ethernet and processing overhead of TCP/IP leads to non-deterministic latencies that can range from 50 ms to over 100 ms. iSCSI SANs basically trade determinism for IP-based scalability and interconnect ability. Deterministic storage performance is critical to deterministic server performance.

Further, the dual-ported nature of fiber channel allows for systems to be designed with fully redundant data paths all the way from the host to the drive. This redundancy can only be partially achieved in iSCSI-based SANs, since no native iSCSI drives exist, and the only dual-ported drives are fiber channel.

The newest SAN contender, Fiber Channel-over-Ethernet (FCoE) has recently been proposed with expected implementation in 2009. Targeting 10 Gbit/sec Ethernet, FCoE is designed to allow FC SAN to run over Ethernet networks. Unlike iSCSI, FCoE bypasses the TCP/IP layer and adds flow-control mechanisms aimed at reducing the packet loss problems that plague current Ethernet implementations.

Published in the IBC 2006 Conference Proceedings, Amsterdam, The Netherlands, September 7-11, 2006.

The Author

Todd Roth received a BSE degree from Arizona State University in 1990. He has been leading the technical development of broadcast video servers since the introduction of the VR30 line in 1993, the industries first PC-based, videodisk recorder, through the current Nexio, the industries first software-codec-based HD/SD video server.

Innovations by Roth and his team led to the introduction of the industries first shared storage, multi-channel videodisk recorder in 1994. Roth led the team that pioneered the development of Fiber Channel-based shared storage and software RAID, introduced by the VR300 line in 1996.

In 2000, Roth was granted a patent for the invention of the "Shared Video Data Storage System," which led to Leitch's 2001 Emmy for "Pioneering Developments in Shared Video-Data Storage Technology." In July of 2002 Roth was granted a patent for the invention of the "Method and apparatus for synchronized multiple format storage."