3 Storage Networking Building Blocks

A STORAGE NETWORK may be composed of a wide variety of hardware and software products. Software products invariably involve management of some kind: management of logical volumes, management of disk resources, management of tape subsystems and backup, and management of the network transport status. Hardware products include the storage end systems and their interfaces to the network as well as the switches and bridge products that provide connectivity. Storage network interfaces typically include legacy SCSI, Fibre Channel, and Gigabit Ethernet. Just as first-generation Fibre Channel SANs accommodated legacy SCSI disk and tape devices, IP-based storage networks must incorporate both Fibre Channel and legacy SCSI devices. The challenge for vendors is to make this complexity disappear, and find inventive ways to integrate new and old technologies with minimal user involvement.

The following sections discuss the major building blocks of storage networks and highlight the technical aspects that differentiate them.

3.1 Storage Networking Terminology

Storage networking end systems include storage devices and the interfaces that bring them into the network. RAID arrays, just a bunch of disks (JBODs), tape subsystems, optical storage systems, and host adapter cards installed in servers are all end systems. The interconnect products that shape these end systems into a coherent network are discussed in the following sections on Fibre Channel, Gigabit Ethernet, and NAS. New SAN products such as virtualization devices and SAN appliances that front-end storage may, depending on implementation, also be considered end systems from the standpoint of storage targets, or interconnects from the standpoint of hosts.

3.1.1 RAID

RAID is both a generic term for intelligent storage arrays and a set of methods for the placement of data on multiple disks. Depending on the methods used, RAID can both enhance storage performance and enable data integrity. Because logic is required to distribute and retrieve data safely from multiple disk resources, the RAID function is performed by an intelligent controller. The controller may be implemented in either hardware or software, although optimal performance is through hardware, in the form of application-specific integrated circuits (ASICs) or dedicated microprocessors. A RAID array embeds the controller function in the array enclosure, with the controller standing between the external interface to the host and the internal configuration of disks. RAID arrays may include eight to ten internal disks for departmental applications or many more for data center requirements.

The performance problem that RAID solves stems from the ability of host systems to deliver data much faster than storage systems can absorb it. When a server is connected to a single disk drive, reads or writes of multiple data blocks are limited by the buffering capability, seek time, and rotation speed of the disk. While the disk is busy processing one or more blocks of data, the host must wait for acknowledgment before sending or receiving more. Throughput can be significantly increased by distributing the stream of data block traffic across several disks in an array, a technique called *striping*. In a write operation, for example, the host can avoid swamping the buffering capacity of any individual drive by subdividing the data blocks into several concurrent transfers sent to multiple targets. This simplified RAID is called *level 0*. If the total latency of an individual disk restricted its bandwidth to 10 to 15 MBps, then eight disks in an array could saturate a gigabit link that provided approximately 100 MBps of effective throughput.

Although boosting performance, RAID 0 does not provide data integrity. If a single disk in the RAID set fails, data cannot be reconstructed from the survivors. Other RAID techniques address this problem by either writing parity data on each drive in the array or by dedicating a single drive for parity information. RAID 3, for example, writes byte-level parity to a dedicated drive, and RAID 4 writes block-level parity to a dedicated drive. In either case, the dedicated parity drive contains the information required to reconstruct data if a disk failure occurs. RAID 3 and RAID 4 are less commonly used as stand-alone solutions because the parity drive itself poses a performance bottleneck problem. RAID 5 is the preferred method for striped data because it distributes block-level parity information across each drive in the

array. If an individual disk fails, its data can be reconstructed from the parity information contained on the other drives, and parity operations are spread out among the disks.

The RAID striping algorithms range from simple to complex and thus imply much higher logic at the RAID 5 level. More logic implies more expense. RAID 5 controllers must not only provide the intelligence to distribute data and parity information across multiple drives, but must also be able to reconstruct data automatically in case of a disk failure. Typically, additional drives are provisioned in standby mode (spare), available for duty if a primary disk fails. The additional manipulation of data provided by RAID 5 also implies latency, and vendors of these products compete on the basis of performance and optimized controller logic.

Storage applications may also require full data redundancy. If, for example, the RAID controller failed or a break occurred along the cable plant, none of the RAID striping methods would ensure data availability. RAID level 1 achieves full data integrity by trading the performance advantage of striping for the integrity of data replication. This is accomplished by mirroring. In disk mirroring, every write operation to a primary disk is repeated on a secondary or mirrored disk. If the primary disk fails, a host can switch over to the backup disk. Mirroring on its own is subject to two performance hits: once for the latency of disk buffering, seek time, and spindle speed, and once for the additional logic required to write to two targets simultaneously. The input/output (I/O) does not complete until both writes are successful. Mirroring is also an expensive data integrity solution because the investment in storage capacity is doubled with every new installation. Mirroring, however, is the only solution that offers a near-absolute guarantee that data will be readily available in the event of disk failure. In addition, because data is written to two separate disk targets, the targets themselves may be separated by distance. Mirroring thus provides a ready solution for disaster recovery applications, provided sufficient latency and bandwidth are available between primary and remote sites. Storage replication over distance also makes mirroring a prime application for IP-based storage solutions.

RAID implementations may be combined to provide both striping throughput and data redundancy via mirroring. RAID 0 + 1, for example, simply replicates a RAID 0 disk set to create two separate copies of data on two separate, high-performance arrays. Just as mirroring doubled the cost of the capacity of a single drive, RAID 0 + 1 doubles the cost of a RAID striping array. As shown in Figure 3–1, data blocks are striped over disks in each array, providing an exact copy of data that can be written or read at high

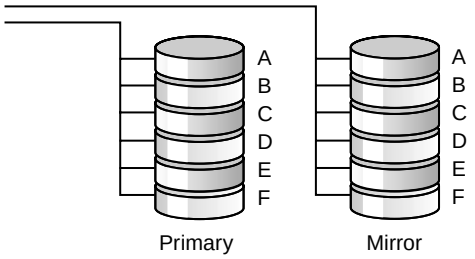

Figure 3–1 RAID 0 + 1 striping plus mirroring. Blocks A, B, C, D, E, and F of a single data file.

speed. Some implementations may combine RAID 5 with RAID 1 for an even higher level of data integrity and availability.

Because RAID implementations treat the drives in an array as a single resource, homogeneity of access is imposed on each individual drive. The performance of the disk set is thus determined by the lowest common denominator. The disk with the slowest access time will dictate the performance of all other drives. RAID-based storage arrays are thus typically populated with the same drive type per set, including unused hot spares that can be brought on-line if a drive fails. As newer, lower cost, and higher capacity drives are constantly introduced into the market, however, a customer may replace a failed disk with a much higher performing unit. The performance and capacity advantage of the newer unit will not be realized because the operational parameters of the original drive will throttle the new disk.

The main components of a RAID subsystem include the interface (parallel SCSI, Fibre Channel, or Gigabit Ethernet), the RAID controller logic, the backplane that accommodates the disks, the disks themselves, and the enclosure, including power supplies and fans. At the high end of the RAID food chain, RAID arrays may provide redundant power supplies and fans, diagnostic and "phone home" features for servicing, and high-performance logic to optimize RAID operations. At the low end, RAID may be implemented via software on a host system with data striped to unrelated disks (JBOD). Software RAID places additional burdens on the server or host because CPU cycles must be devoted to striping across multiple targets. The RAID function may also be provided by an HBA or interface card, which off-loads the host CPU and manages striping or mirroring functions to disk targets. This is advantageous from the standpoint of the host, but, as with software RAID, places additional transactions on the storage network for the parity operations. Despite higher cost, a RAID subsystem offers optimal performance for

both RAID functions and storage network traffic compared with software or adapter card implementations.

A RAID enclosure hides the access method between the RAID controller and the disks it manages. The external interface between the host systems and the array may be parallel SCSI, Fibre Channel, or Gigabit Ethernet. The internal interface between the RAID controller and disks may be parallel SCSI or Fibre Channel, and at the very low end of the spectrum, even Integrated Drive Electronics (IDE). This separation between the internal workings of the array and the external interface provides flexibility in designing low-, medium-, and high-end RAID systems that target different markets. It also facilitates the introduction of new external interfaces such as IP over Gigabit Ethernet, which although not trivial from an engineering standpoint at least do not require redesign of the entire subsystem including the back-end disks.

In Figure 3–2, the basic architecture of RAID subsystems is shown with the variations of external interfaces and internal disk configurations. If Fibre Channel disks are used, the internal Fibre Channel disk interface may be based on a shared loop or switched fabric (discussed later).

RAID systems are a powerful component of storage networks. They offer the performance and data integrity features required for mission-critical applications, and the flexibility (given sufficient budget) for resilient data replication and disaster recovery strategies. They also provide, depending on vendor implementation, the ability to scale to terabytes of data in a single, highly available storage resource. In a storage network based on Fibre Channel or Gigabit Ethernet, the resources of a RAID array may be shared by multiple servers, thus facilitating data availability and reduction of storage management costs through storage consolidation.

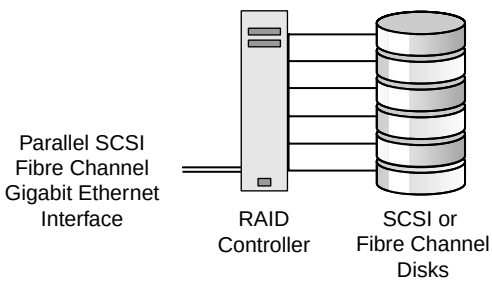

Figure 3–2 Basic architecture of a RAID subsystem.

3.1.2 JBODs

A JBOD is an enclosure with multiple disk drives installed in a common backplane. Unlike a RAID array, a JBOD has no front-end logic to manage the distribution of data over the disks. Instead, the disks are addressed individually, either as separate storage resources or as part of a host-based software or an adapter card RAID set. JBODs may be used for direct-attached storage based on parallel SCSI cabling, or on a storage network with, typically, a Fibre Channel interface.

The advantage of a JBOD is its lower cost vis-à-vis a RAID array, and the consolidation of multiple disks into a single enclosure that share power supplies and fans. JBODs are often marketed for installation in 19-inch racks and thus provide an economical and space-saving means to deploy storage. As disk drives with ever-higher capacity are brought to market, it is possible to build JBOD configurations with hundreds of gigabytes of storage.

Because a JBOD has no intelligence and no independent interface to a storage network, the interface type of the individual drives determines the type of connectivity to the SAN. An IP-based storage network using Gigabit Ethernet as a transport would therefore require Gigabit Ethernet/IP interfaces on the individual JBOD disks, or a bridge device to translate between Gigabit Ethernet and IP to Fibre Channel or parallel SCSI. Over time, disk drive manufacturers will determine the type of interface required by the market.

As shown in Figure 3–3, a JBOD built with SCSI disks is an enclosed SCSI daisy chain and offers a parallel SCSI connection to the host. A JBOD built with Fibre Channel disks may provide one or two Fibre Channel interfaces to the host and internally is composed of shared loop segments. In either configuration, a central issue is the vulnerability of the JBOD to individual disk failure. Without the appropriate bypass capability, the failure of a single drive could disable the entire JBOD.

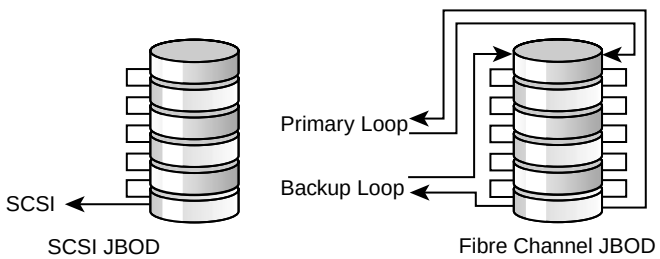

Figure 3–3 JBOD disk configurations.

Management of a JBOD enclosure is normally limited to simple power and fan status. In-band management may be provided by the SCSI Enclosure Services protocol, which can be used in both parallel SCSI and Fibre Channel environments. Some vendor offerings also allow the JBOD to be divided into separate groups of disks via a hardware switch or jumpers. As shown in Figure 3–4, a single Fibre Channel JBOD may appear as two separate resources to the host.

How the individual disk drives within a JBOD are used for data storage is determined by the host server or workstation, or by RAID intelligence on an HBA. Windows Disk Administrator, for example, can be used to create individual volumes from individual JBOD disks, or can assign groups of JBOD disks as a volume composed of a striped software RAID set. Software RAID will increase performance in reads and writes to the JBOD, but will also give exclusive ownership of the striped set to a single server. Without volume-sharing middleware, multiple servers cannot simultaneously manage the organization of striped data on a JBOD without data corruption. The symptom of unsanctioned sharing is the triggering of endless check disk sequences as a host struggles with unexpected reorganization of data on the disks. Generally, software RAID on JBODs offers higher performance and redundancy for dedicated server-to-storage relationships, but does not lend itself to server clustering or serverless tape backup across the SAN.

One means of leveraging JBODs for shared storage is the use of storage virtualization appliances that sit between the host systems and the JBOD targets. The virtualization appliance manages the placement of data to multiple JBODs or RAID arrays, while presenting the illusion of a single storage resource to each host. This makes it possible to dispense with software RAID on the host because this function is now assumed by the appliance. Essentially, storage virtualization fulfills the same function of an intelligent RAID controller, except that the virtualization appliance and storage arrays now sit

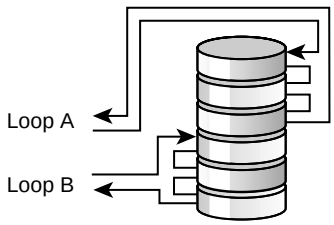

Figure 3–4 Dividing a Fibre Channel JBOD backplane into separate loops.

in separate enclosures across the storage network. As with Dorothy in *The Wizard of Oz,* however, one really should pay attention to the little man behind the curtain. Although presenting a simplified view of storage resources to the host systems, storage virtualization appliances must necessarily assume the complexity of managing data placement and be able to recover automatically from failures or disruptions. This is not a trivial task.

3.1.3 Tape Subsystems

Data storage on media of different kinds divides into a hierarchy of cost per megabyte, performance, and capacity. Generally, higher cost per megabyte brings greater performance and, ironically, lower capacity. At the high end, solid-state storage offers the performance of memory access, but with limited capacity. Spinning media such as disk drives in RAID configurations can support gigabit access speeds, with the capacity of more than a terabyte of data. Tape subsystems may only support a fifth of that speed, but large libraries can store multiple terabytes of data. At the low-performance end of the scale, optical media libraries offer nearly unlimited storage capacity. The two most commonly used solutions—disk and tape—are deployed for a sequential storage strategy, with normal data transactions based on disk, followed by periodic backup of that data to tape. Hierarchical storage management applications may be used to rationalize this process and to determine the frequency of and appropriate migration of data from disk to tape, and sometimes from tape to optical storage.

Securing a backup copy of data as a safeguard against disk or system failure is a universal problem. No institution or enterprise is likely to survive a loss of mission-critical information. In addition, a company may be obliged to keep reliable copies of its data according to government or commercial regulations. Financial institutions, for example, must keep long-term records of their transactions, which may require both tape and optical storage archiving.

If spinning disk media could provide affordable, scalable, and highly reliable long-term storage, mirroring would suffice. Although this solution is strongly supported by disk manufacturers, mirroring for ever-increasing amounts of data has proved too costly to implement. Tape, by contrast, has proved to be economical, scalable, and highly reliable, and despite the advances made by disk technology, tape is likely to endure as the primary tool for the archival preservation of data.

In its pure SCSI incarnation, tape provides a secure copy of data, but its performance is constrained by the typical topology in which it is deployed. For optimal performance in traditional environments, a SCSI tape device can

be attached to a server/storage SCSI daisy chain. In this configuration, the tape device (like the storage arrays) becomes captive to an individual server. Each server would thus require its own tape unit for backup, multiplying the cost of tape systems and management throughout the network. Alternately, a dedicated tape backup server and SCSI-attached tape library can provide centralized backup over the LAN. This facilitates resource sharing, but places large block data transfer on the same LAN that is used for user traffic. The bandwidth of the LAN itself may create additional problems. The backup window, or period of time in which a nondisruptive backup could occur, may not be sufficient for the amount of data that requires duplication to tape.

The conflict between backup requirements and the constraints imposed by LAN-based backup is resolved by storage networking. By placing tape subsystems on a storage network, they become shared resources for multiple servers and can now move backup traffic independently of the LAN. This simultaneously reduces costs, simplifies administration, and provides greater bandwidth for backup streams. LAN-free backup on a storage network has also enabled new backup solutions. Even on a SAN, the backup traffic moves from disk storage to server and from server to tape. The server is in the backup path because it is responsible for reads from disk and for writes to tape. However, because servers, storage, and tape subsystems are now peers on the storage network, data paths can be enabled directly between disk storage and tape resources. Server-free backup is predicated on intelligent backup agents on the SAN that can perform the server's read/write functions for tape. A third-party copy (extended copy) agent may be embedded in the tape library, in the SAN switch, or in a SAN-to-SCSI bridge used to connect a SCSI tape library. Because the third-party copy agent assumes the task of reading from disk and writing to tape, server CPU cycles are freed for user transactions. LAN-free and server-free backup solutions for IP-based SANs are discussed further in Chapter 13.

The internal design of a tape subsystem is vendor specific, but typically includes an external interface for accessing the subsystem, controller logic for formatting data for tape placement, one or more tape drives that perform the write and read functions, robotics for manipulating tape cartridges and feeding the drives, and slots to hold the cartridges while not in use. Vendors may promote a variety of tape technologies, including advanced intelligent tape, linear tape-open, and digital linear tape, which are differentiated by performance and capacity.

The external interface to a tape library may be legacy SCSI, Fibre Channel, or Gigabit Ethernet. Although each tape drive within a library may only

support 10 to 15 MBps of throughput, multiple drives can leverage the bandwidth provided by a gigabit interface. Theoretically this would allow multiple servers (or third-party copy agents) to back up simultaneously to a single library over the SAN, although the library controller must support this feature. Another potential performance enhancement is provided by the application of RAID striping algorithms to tape, or tape RAID. As with disk RAID, tape RAID implies additional complexity and logic, and therefore additional expense.

The initiative for IP-based SCSI solutions for tape was launched by SpectraLogic Corporation in the spring of 2001. Tape, like storage arrays, relies on block SCSI data for moving large volumes of data efficiently. SCSI over IP on Gigabit Ethernet infrastructures provides tape vendors with much greater flexibility in deploying their solutions. The Gigabit Ethernet network may be a dedicated SAN or a virtual segment (VLAN) of an enterprise network. Backups may thus occur wherever sufficient bandwidth has been allocated, and familiar IP and Ethernet management tools can be leveraged to monitor backup traffic. Because third-party copy is infrastructure neutral, serverless backup can also be used for IP-based tape subsystems.

3.1.4 SCSI Over IP-to-Parallel SCSI Bridges

Like first-generation Fibre Channel SANs, IP-based SANs must accommodate legacy devices, including SCSI disk arrays and SCSI tape subsystems. SCSI tape libraries, in particular, represent a substantial investment, and few information technology (IT) administrators have the luxury of discarding a valuable resource simply because interface technology has improved. The common denominator between the IP SAN and the legacy tape device is the SCSI protocol. The legacy tape device, however, supports parallel SCSI, or the SCSI-2 protocol. The IP SAN supports serial SCSI, or the SCSI-3 protocol. The function of a bridge is to translate between the two SCSI variants, and to make the SCSI-2 tape or storage subsystem appear to be a bona fide IP-addressable device.

An IP storage-to-SCSI bridge may provide multiple parallel SCSI ports to accommodate legacy units, and one or more Gigabit Ethernet ports to front the SAN. Just as a Fibre Channel-to-SCSI bridge must assign a Fibre Channel address to each legacy SCSI device, an IP storage-to-SCSI bridge must proxy IP addresses. The specific type of serial SCSI-3 supported by the Gigabit Ethernet ports on the bridge is vendor dependent, but may be iSCSI or iFCP.

SAN bridges provide a valuable function, both for preserving the customer's investment in expensive subsystems and for making those subsystems

participants in a shared storage network. Although bridges are normally used to bring tape subsystems into a SAN, legacy SCSI disk arrays can also be supported. In some vendor implementations, even SCSI hosts (for example, servers with SCSI adapter cards) can be accommodated. This enables an IP storage network to be constructed with SCSI end systems only, using IP and Gigabit Ethernet as the SAN infrastructure. Customers with large SCSI installations could thus enjoy the benefits of shared storage networking without investing in new host adapters, storage, or tape.

3.1.5 Host Adapters

Host adapter cards provide the interface between the server or workstation internal bus and the external storage network. Interface cards are available for different bus architectures and may offer different physical connections for network interface. The adapter card vendor also supplies a device driver that allows the card to be recognized by the operating system. The device driver software may also perform protocol translation or other functions if these are not already executed by onboard logic.

Whether Fibre Channel or Gigabit Ethernet, the HBA or network interface card (NIC) must provide reliable gigabit communication at the physical and data link levels. As discussed later, Gigabit Ethernet has taken the physical layer and data encoding standards from Fibre Channel. Above the data encoding level, however, Gigabit Ethernet must appear as standard Ethernet to provide seamless integration to operating systems and applications.

For storage networking, two additional components may appear on the host Gigabit Ethernet interface card. To support storage data transfer efficiently, a storage NIC must incorporate an upper protocol layer for serial SCSI-3. This may be an iSCSI interface or an FCP interface. The purpose of this protocol interface is to deliver SCSI data to the operating system with high performance and low processor overhead. Fibre Channel FCP has solved this SCSI delivery issue, whereas the iSCSI initiative is reengineering an entirely new solution.

The second component that may be embedded on a storage NIC is additional logic to off-load TCP/IP processing from the host CPU. At gigabit speeds, TCP overhead may completely consume the resources of a host system. This would be unacceptable for servers in a storage network, which must simultaneously service both disk and user transactions. TCP off-load engines (for those lacking enough acronyms, TOEs) may be provided by software routines or, more efficiently, in ASICs or dedicated processors onboard the NIC.

Figure 3–5 depicts the basic functions of a Fibre Channel HBA, whereas Figure 3–6 shows the basic components of a storage NIC.

Both adapters provide transmit and receive connections to the storage network, which may be via fiber-optic or copper cabling. Both adapter types provide a clock and data recovery (CDR) logic to retrieve gigabit signaling from the inbound bit stream. Both provide serializing/deserializing logic to convert serial bits into parallel data characters for inbound streams, and convert parallel into serial for outbound streams. The mechanism for data encoding is provided by an 8b/10b encoding scheme originally developed by IBM and discussed in further detail later. For transmission on gigabit links, the data encoding method also uses special formatting of data and commands known as *ordered sets*.

Above the ordered set logic, a storage NIC will include a LAN controller chip, auxiliary logic and memory, an optional TOE, and hardware-based or software drivers for the serial SCSI-3 protocol. All of this functionality is made physically accessible to the host platform through the PCI, S-bus, or other bus interface and is logically accessible through the host device driver supplied by the manufacturer.

One of the often-cited benefits of IP-based storage networking is the ability to leverage familiar hardware and management software to deploy and maintain a SAN. In the example given here of a storage network adapter, there are clearly common components to ordinary Ethernet NICs. It is unlikely, however, that off-the-shelf Gigabit Ethernet NICs will be suited to storage applications. Without embedded logic to speed serial SCSI processing

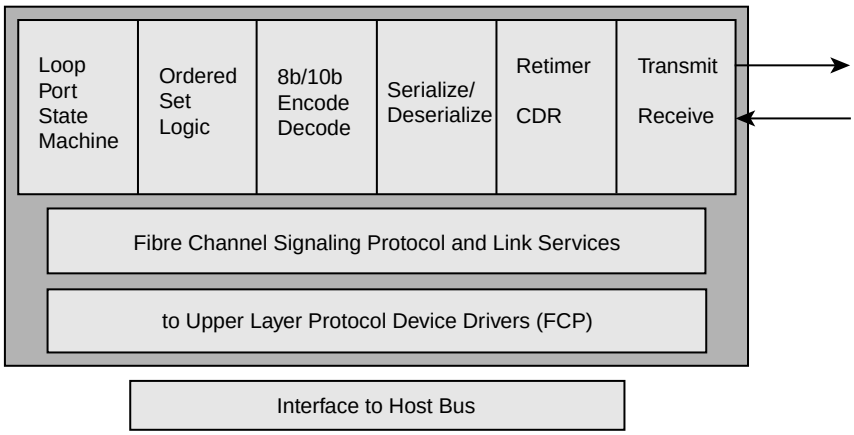

Figure 3–5 A Fibre Channel HBA.

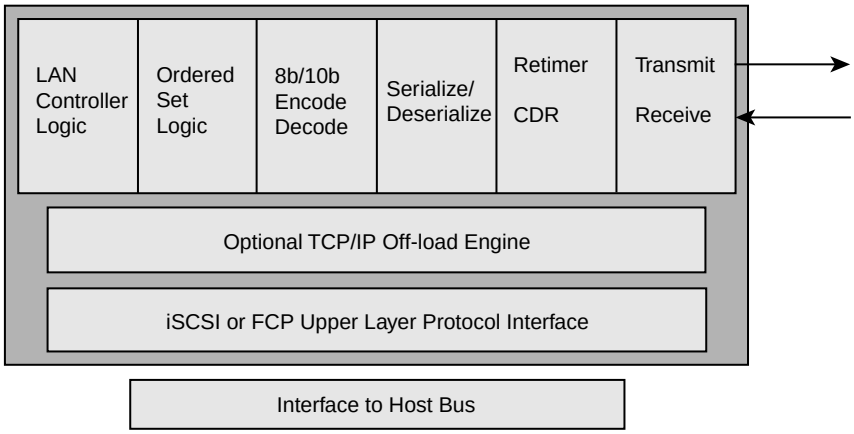

Figure 3–6 A storage over IP NIC.

and to off-load TCP overhead, a standard Gigabit Ethernet NIC will not provide the performance required for storage applications. From a management and support standpoint, however, the distinction between a specialized storage NIC and a standard NIC are minimal when compared with Fibre Channel HBAs.

3.2 Legacy SCSI Cabling

SCSI cabling is a parallel wiring plant analogous to common personal computer parallel printer cabling. In the SCSI cable, each bit of a data byte is given its own wire. A SCSI cable may have 8 (narrow SCSI) or 16 (wide SCSI) data lines, plus control lines to manage device selection and data transmission. The effective transmission rate is governed by the number of data lines provided and the clock rate at which bytes of data are sent concurrently. Although new parallel SCSI designs such as high-voltage differential and low-voltage differential (LVD) have enabled higher throughput, the overall distance limitation of SCSI cabling is 25 m, and device support is limited to 15 devices on a SCSI chain.

In a parallel SCSI configuration, the sending party places the 8 or 16 bits of data on the cable plant and toggles a control line to indicate transmission. Small differences in propagation delay may occur as the data bits are sent, so it is essential that the receiving device be able to capture all bits accurately within a certain window of time. This window is known as *skew*. The greater the differences in propagation delay along the length of the cable, the wider

the window must be for all data bits to be captured. Parallel transmission thus favors shorter cable lengths to minimize skew and to avoid data corruption. To achieve an effective bandwidth greater than serial gigabit transports, the SCSI cable must be less than 12 m long.

Distance limitations and limited device population have relegated SCSI technology to the middle and low end of the storage market, where resource sharing and high availability are not required. Additionally, adds, moves, or changes in a SCSI configuration may necessitate downtime and disruption to users. This is less likely to be the case in gigabit storage networking.

3.3 Network-Attached Storage

Like the acronym SAN, NAS is largely a marketing term that has, through repeated use, gained a technical definition. The primary distinction between NAS and SAN rests on the difference between data files and data blocks. NAS transports files; SANs transport blocks. NAS uses file-oriented delivery protocols such as NFS and CIFS, whereas SANs use block-oriented delivery protocols such as SCSI-3. Because data blocks are the raw material from which files are formed, NAS also has a block component. These blocks are addressed on a per-file basis, using meta-data (directory information) to determine which file to use. The block access methods of a NAS device, however, are typically hidden in the NAS enclosure. To the outside world, the NAS device is a server of files and directories.

NAS accomplishes the central goal of storage networking: the sharing of storage resources through the separation of servers and storage over a common network. Like SAN-based storage, NAS overcomes the limitations of parallel SCSI and enables a more flexible deployment of shared storage. In redundant configurations, NAS can also provide highly available, nondisruptive storage access.

As shown in Figure 3–7, a NAS architecture includes disk arrays for data placement, a NAS processor, and an external interface to the user network. This architecture has a number of implications. Although the block data transport of a SAN normally occurs over a dedicated storage network (or VLANs within a Gigabit Ethernet network), the file transport of a NAS device assumes direct connectivity to the user network. NAS performance is thus partially determined by the bandwidth available on the messaging network. In addition, although the NAS processor is optimized for file transport, it is essentially a thin server sitting on a LAN "front-ending" storage

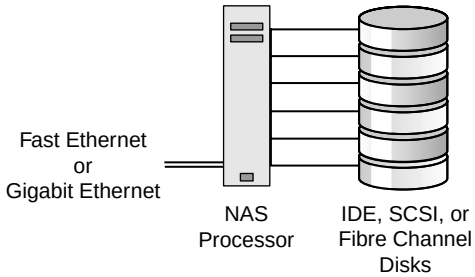

Figure 3–7 NAS architecture.

arrays. To avoid the definition of NAS as "network-attached server" or simply "file server," vendors of NAS products have attempted to make the thin server component so thin that it is invisible from the user's perspective. The term *NAS* is therefore always linked in marketing literature to the concept of "appliance"—a device that is simply plugged into the network and requires little administration. Finally, the connection between NAS intelligence and its disks is immaterial from a user's perspective. Although SAN storage is predicated on high-performance gigabit interfaces for storage (either directly or via bridge products), a NAS device may rely on parallel SCSI, IDE, or Fibre Channel for storage connectivity.

The challenge of NAS vendors is to make their products more appliance-like and to reduce the overhead of NFS/CIFS protocols over TCP/IP. Although TCP/IP imposes its own latency on file transactions, NFS and CIFS also engender latency as file transfer sessions are established and torn down, and as files are found through directory lookup. Chip-based TOEs offer some relief, although the main initiative from some NAS vendors for dealing with latency is support of the Direct Access File System (DAFS) over the VI protocol. DAFS relies on remote direct memory access (RDMA) techniques to move data directly to systems memory. If a NAS controller can place file data directly into the memory of the client efficiently, transport latency is offset by the higher performance of file data placement.

The convergence of NAS and SAN is accelerated by the development of IP storage networks. With 10Gb backbones and gigabit interfaces to end devices, both file and block data can share a common network infrastructure. This promotes the deployment of shared storage solutions on the basis of user application requirements, without the artificial limits imposed by incompatible network topologies.

3.4 Fibre Channel

As the first network architecture to implement storage networking applications successfully, Fibre Channel has faced substantial challenges in product development, standardization, interoperability, and market acceptance. It has also achieved technological breakthroughs in the areas of gigabit transport and upper layer protocol support for SCSI-3, which ironically have made it vulnerable to competing storage network technologies. As the pioneer of storage networking, Fibre Channel has had to create its own vocabulary, and this, in turn, has made it difficult for customers to understand, deploy, and support. The basic lexicon of Fibre Channel is reviewed here.

3.4.1 Fibre Channel Layers

Fibre Channel standards are developed in the National Committee of Industrial Technology Standards (NCITS) T11 standards body, which has defined a multilayer architecture for the transport of block data over a network infrastructure. As shown in Table 3–1, Fibre Channel layers are numbered from FC-0 to FC-4.

The upper layer, FC-4, establishes the interface between the Fibre Channel transport and upper level applications and operating system. For storage applications, FC-4 is responsible for mapping the SCSI-3 protocol for transactions between host initiators and storage targets. The FC-3 layer is still in standards development, and includes facilities for data encryption and

TABLE 3–1 FIBRE CHANNEL LAYERED ARCHITECTURE

Fibre Channel Layer	Layer Title	Comments
FC-4	Upper layer protocol interface	SCSI-3, IP, VI, and so on
FC-3	Common services	Under development
FC-2	Data delivery	Framing, flow control, service class
FC-1	Ordered sets/byte encoding	8b/10b encoding, link controls
FC-0	Physical interface	Optical/electrical, cable plant

compression. The FC-2 layer defines how blocks of data handed down by the upper level application are segmented into sequences of frames for handoff to the transport layers. This layer also includes class-of-service implementations and flow control mechanisms to facilitate transaction integrity. The lower two layers, FC-1 and FC-0, focus on the actual transport of data across the network. FC-1 provides facilities for encoding and decoding data for shipment at gigabit speeds, and defines the command structure for accessing the media. FC-0 establishes standards for different media types, allowable lengths, and signaling.

Collectively, the Fibre Channel layers fall within the first four layers of the OSI model: physical, data link, network, and transport. Fibre Channel assumes a single unpartitioned network and homogeneous address space for the network fabric. Although theoretically this address space can be quite large (15.5 million addresses in a switched fabric), a single network space has implications for large Fibre Channel SANs. Without network segmentation, the entire fabric is potentially vulnerable to disruption in the event of failures.

3.4.2 FC-0: Fibre Channel Physical Layer

As the first successful serial gigabit transport, Fibre Channel has defined the basic principles and methods required for data integrity over high-speed serial links. At the physical layer, these include standards for gigabit signaling, supported cable types, allowable cable distances, and physical interfaces. Because Gigabit Ethernet has borrowed heavily from the Fibre Channel physical layer standards, it is useful to understand what they provide.

Unlike SCSI parallel cabling, a serial network cabling scheme does not have a separate control line to signal the rate of data transmission so that the recipient can accurately capture data. In a serial implementation, this clock signaling must be embedded in the bit stream itself. Fibre Channel uses an FC-1-defined byte-encoding scheme and CDR circuitry to recover the clock signal from the serial bit stream. The physical layer standards dictate that data integrity for gigabit transmission must be no less than 10^{-12} bit error rate, or a maximum of 1 bit error every 16 minutes over 1Gb media. To meet or exceed this rigorous standard, the physical interfaces and cabling must minimize the amount of jitter or timing deviation that may occur along the physical transport.

Deviations from the original clock signaling, or jitter, may be the result of natural propagation delays through fiber-optic or copper cabling as well as unnatural transients from poorly designed interfaces, laser optics, circuit

boards, or power supplies. Jitter may be measured and represented in graphical form by an *eye diagram* on a test scope, as illustrated in Figure 3–8. The crossover points or intersections forming the eye represent signaling transitions to high or low voltages. Ideally, all transitions should occur at precisely the same interval. If this were the case, the eye would be perfectly formed and the CDR circuitry could recover all data bits with no bit errors whatsoever. In reality, some deviation will always be present. If the jitter is too extreme, the CDR will miss one or several bits, resulting in the corruption of data. This will in turn trigger recovery routines at the higher layers.

If jitter reduction is essential for Fibre Channel's 1.0625-Gbps clock rate, it is even more essential for Gigabit Ethernet's faster 1.25 Gbps. The faster the clock, the greater the statistical occurrence of bit errors over the same time span. A faulty transceiver, substandard fiber-optic cabling or connectors, exceeding cable distance guidelines, improperly shielded copper components, or simply bad product design can introduce system instability at the physical layer.

For cable plant, Fibre Channel accommodates both copper and fiber-optic cabling. Copper cabling is typically twin axial as opposed to shielded twisted pair, and is deployed for intracabinet and intercabinet usage. Intracabinet copper cabling may be used within an enclosed 19-inch rack for connecting storage devices or HBAs to Fibre Channel hubs or switches. The maximum length of intracabinet copper is 13 m. Intercabinet copper cabling may be used externally to 19-inch racks, to a maximum of 30 m. Both varieties are problematic because any copper cable plant is succeptable to electromagnetic interference (EMI) and may create ground loop problems between devices.

For both Fibre Channel and Gigabit Ethernet, fiber-optic cabling is the preferred cable plant because of its immunity to EMI. Fiber-optic cable types

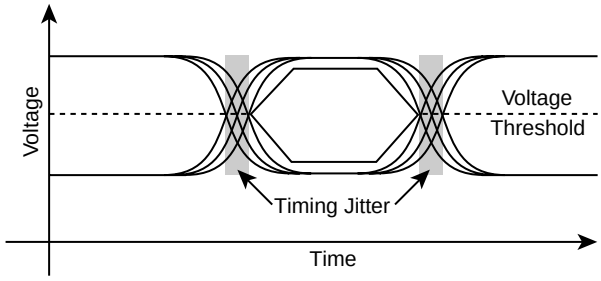

Figure 3–8 An eye diagram showing timing deviations in a gigabit stream.

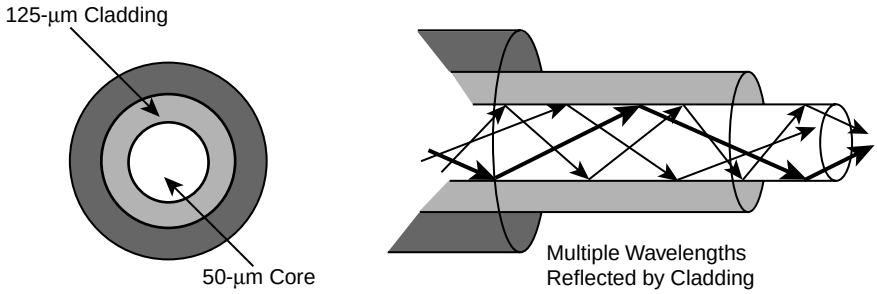

Figure 3–9 Multimode fiber-optic cable.

are distinguished by "mode," or by the frequencies of light waves that the optical cable supports.

Multimode cabling is used with shortwave laser light and has either a 50-μm or 62.5-μm core with 125-μm cladding. The reflective cladding around the core restricts light to the core. As shown in Figure 3–9, a shortwave laser beam is composed of hundreds of light modes that reflect off the core at different angles. This dispersion effect reduces the total distance at which the original signal can be reclaimed. In Fibre Channel configurations, multimode fiber supports 175 m with 62.5-μm/125-μm cable, and supports 500 m with 50-μm/125-μm cable.

Single-mode fiber is constructed with a 9-μm core and 125-μm cladding. Single mode is used to carry long-wave laser light, which has little of the dispersion effect of multimode lasers because the diameter of the core is matched to the wavelength of the light. With a much smaller diameter core and a single-mode light source, single-mode fiber supports much longer distances, currently as much as 10 km at gigabit speeds.

At either end of the cable plant, transceivers or adapters are used to bring the gigabit bit stream onto the circuit boards of HBAs or controller cards. Gigabit interface converters, or GBICs, connect the cabling to the device interface. Small-form factor GBICs are steadily replacing the older SC connectors, because they enable higher port density for switches. Optical transceivers may be permanently mounted onto the HBA, storage, or switch port, or may be removable to facilitate changes in media type or to service a failed unit.

3.4.3 FC-1: Fibre Channel Link Controls and Data Encoding

Suppose that the cable plant, transceivers, and interfaces all provided a stable physical layer transport for gigabit transmission. Turning bits of serial data

into intelligible bytes is still an issue. If raw data bytes were dropped serially onto a gigabit transport, it would be impossible to tell where one byte ended and another began. Sending a stream of hex 'FF' bytes, for example, would create a sustained direct current (DC) voltage on the link, making it impossible to recover the embedded clock signaling needed to establish byte boundaries.

Fibre Channel standards have addressed this problem by using a byte encoding algorithm first developed by IBM. The 8b/10b encoding method converts each 8-bit data byte into two possible 10-bit characters. To avoid sustained DC states, each of the two 10-bit characters will have no more than six total ones or zeros. Of all the possible 10-bit characters that can be generated by standard 8-bit data bytes, about half will have an equal number of ones and zeros. The 8b/10b encoding scheme thus ensures a healthy mix of ones and zeros that allows recovery of the embedded clock signaling and thus recovery of data.

Because the 8b/10b encoder generates two different 10-bit characters for each byte, which one should be used for data transmission? This selection is made based on the *running disparity* of the character stream (Figure 3–10). If a 10-bit character has more ones than zeros, it has *positive disparity*. If it has more zeros than ones, it has *negative disparity*. An equal number of ones and zeros results in *neutral disparity*. The concept of running disparity is key to maintaining a more consistent distribution of ones and zeros in the bit

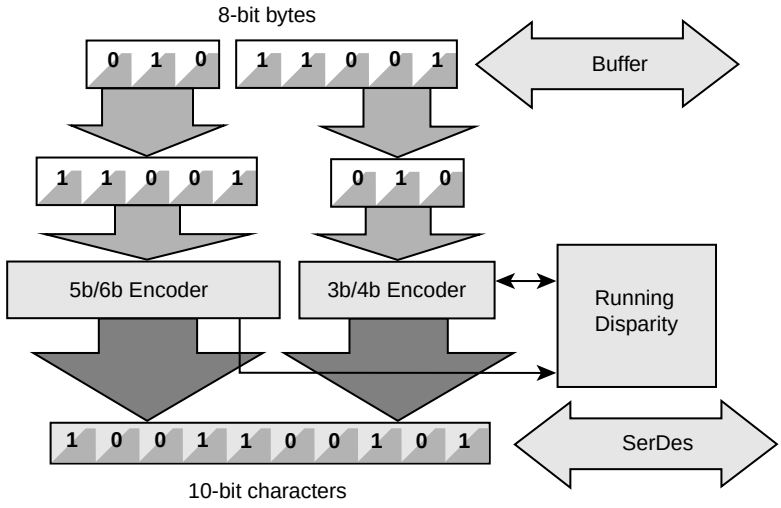

Figure 3–10 8b/10b encoding logic.

stream. A 10-bit data character with positive disparity should be followed by a character with neutral disparity (which leaves the running disparity positive), or by a character with negative disparity (which would leave the running disparity negative). This alternation between positive and negative disparity patterns ensures that no sequential combination of 10-bit characters will result in persistent ones or zeros bit states.

For all the 10-bit characters that can be generated by standard data bytes, none have more than four ones or zeros in sequence. Running disparity maintains this minimal occurrence for data characters. Some nonstandard 10-bit combinations, however, result in five ones or zeros in sequence. These characters are reserved as special characters and are inserted into the character stream as a means to establish boundaries between 10-bit characters. In Fibre Channel standards, the presence of a special "K28.5" character is monitored by the CDR circuitry. As soon as five ones or zeros in sequence are detected, the CDR can begin buffering sets of 10 bit streams that can then be converted accurately to valid 8-bit data bytes.

Fibre Channel standards for the FC-1 layer leverage the 8b/10b encoding method to create a command syntax known as *ordered sets*. The 8b/10b encoding scheme and running disparity ensure that the embedded gigabit signaling can be recovered and that data bytes can be successfully retrieved. Ordered sets are composed of four 10-bit characters, or 40 bits that constitute a transmission word. The ordered set leads with the special K28.5 character to indicate that the transmission word is a link-layer command or a signal of a change in state. The three data characters following the K28.5 character define the function of the ordered set; for example, start of frame (SOF), end of frame (EOF), initialization, and class of service.

Gigabit Ethernet has borrowed the ordered set command and signaling structure from Fibre Channel, but as we see later, uses fewer commands. Fibre Channel ordered sets are divided into frame delimiters, primitive signals, and primitive sequences. Frame delimiters mark the frame boundaries and may include frame sequencing information for multiframe transmissions. Primitive signals include the IDLE ordered set, which is used to maintain CDR when no user data is present on the link. Primitive sequences are ordered sets that must occur at least three times on the link before any action is taken (for example, a loop initialization or LIP primitive). Fibre Channel standards define more than 20 ordered sets for frame delimiting, more than 10 for primitive signals, and more than 15 for primitive sequences. Because only a single instance of a primitive signal is required to initiate an action, the CDR mechanism for gigabit transmission must be very precise.

3.4.4 FC-2: Fibre Channel Framing, Flow Control, and Class of Service

The data bytes that were encoded by FC-1 for reliable transmission on the physical media were handed down by FC-2 as a series of frames. The FC-2 layer receives blocks from the upper layer protocol (for example, FCP) and subdivides those into sequences of frames that can be reassembled on the other end. Frames are grouped into *sequences* of related frames. A database record, for example, may be written to disk as a single sequence of frames. The sequence is the smallest unit of error recovery in Fibre Channel. If a transmission word within a frame is corrupted and cannot be recovered, the entire sequence of frames must be retransmitted. At gigabit speeds, it is more efficient simply to retransmit an entire sequence of frames than to buffer and provide recovery constantly at the frame level. In the hierarchy of frame delivery at FC-2, multiple sequences of frames can occur within a single *exchange*. The exchange binding between two communicating devices maximizes utilization of the link between them and avoids constant setup and teardown of logical connections.

Fibre Channel framing allows for a variable-length frame with a payload of 0 to 2,112 bytes. Because the Fibre Channel maximal frame size does not map directly to Ethernet framing, issues can arise when Fibre Channel is tunneled over IP/Ethernet. The basic format of the Fibre Channel frame is shown in Table 3–2. The ordered sets used for the SOF and EOF delimiters indicate where the frame falls within a sequence of frames, as well as the class of service required. The header field contains the destination and source Fibre Channel addresses as well as payload length. The cyclic redundancy check (CRC) is calculated before the data is run through the 8b/10b encoder, with the 4-byte CRC itself later encoded along with the rest of the frame contents. At the receiving end, the CRC is recalculated and compared against the frame's CRC to ensure data integrity.

The SOF delimiter establishes the class of service that will be used for frame transmission, whereas the EOF delimiter may indicate when that class of service may be terminated. Class of service is used to guarantee bandwidth

TABLE 3–2 FIBRE CHANNEL FRAME FORMAT

SOF	Header	Data Field	CRC	EOF
1 word	6 words	0–2112 bytes	4 bytes	1 word

or to require acknowledgment of frame receipt for secure data transport. Storage applications may require different classes of service, but the vast majority of Fibre Channel transactions are performed with class 3 datagram service.

Class 1 service defines a dedicated connection between two devices (for example, a file server and a disk array) with acknowledgment of frame delivery. Class 1 service can be assumed in a point-to-point connection between two devices because there are no other participants to impose bandwidth demands. Class 1 service through a Fibre Channel fabric, however, requires the fabric switches to establish dedicated data paths between the communicating pair. A 16-port switch, for example, could only support 8 concurrent class 1 sessions. Consequently, class 1 is almost never deployed in SAN applications.

Class 2 service avoids the issue of connection-oriented, dedicated bandwidth, but provides acknowledgment of frame delivery. Frame acknowledgment imposes its own overhead, however, and so impacts the efficiency of link utilization. Like class 1, class 2 service is fully defined in standards but is infrequently used.

Ironically, although storage network applications revolve around mission-critical applications that require the highest degree of data integrity, the most commonly used class of service in Fibre Channel is both connectionless and unacknowledged in terms of frame delivery. Class 3 service in Fibre Channel is analogous to datagram service such as UDP/IP in LAN environments. Frames are streamed from initiator to target with no acknowledgment of receipt. In the early days of Fibre Channel adoption, customers balked at the idea of committing their mission-critical data to a datagram type of service. In practice, however, class 3 gained respectability simply because it worked. As a connectionless protocol, class 3 facilitates the efficient utilization of fabric resources because bandwidth is not hoarded by communicators as in class 1. And by eliminating acknowledgments, class 3 service imposes minimal protocol overhead on the link.

The ability of a datagram class of service to transmit and receive data reliably is predicated on a highly stable and properly provisioned infrastructure. The 10^{-12} bit error rate mandated by Fibre Channel standards and the thoughtful allocation of bandwidth for storage network resources enables the use of class 3 service for a wide variety of storage applications. This has significant implications for storage networks based on Gigabit Ethernet, which shares the link integrity requirements of Fibre Channel. For contained switch environments such as data centers, a datagram type of service is viable

for stable, high-performance data transfer. This is not the case for potentially congested or lossy implementations, such as wide area switched networks.

Other Fibre Channel classes of service include class 4 for virtual circuits and class 6 for acknowledged multicast applications. As with many other Fibre Channel features that may be supported in fabrics, these are still immature in terms of product implementation and have lacked the engineering focus that Gigabit Ethernet has enjoyed.

Class 3 service requires a flow control mechanism to ensure that a target is not flooded with frames and forced to discard them. Fibre Channel flow control is based on a system of credits, with each credit representing an available frame buffer in the receiving device. If, for example, a disk array has 20 frame buffers, a server could stream 20 frames of a sequence in a single burst before waiting for additional credits to be issued by the array. As the array absorbs the 20 frames, the first in sequence are passed to the FC-2 frame reassembly logic for reconstruction into data blocks for FC-4. As individual frames move up the assembly line, buffers are freed for additional inbound frames. The array issues a credit for each newly emptied buffer, allowing the server to send additional frames.

This frame-pacing algorithm based on credits prevents frame loss and reduces the frequency of sequence retransmission over the fabric. For class 3 service in a fabric, the credit relationship is not end to end between storage devices and servers, but between each device and the switch port to which it is attached. Providing adequate buffers on switch ports is essential for minimizing frame discards. For Gigabit Ethernet SANs, port buffering is also an issue, although the link-level flow control is implemented differently.

3.4.5 FC-3: Common Services

The FC-3 layer has been a placeholder in Fibre Channel standards as the more basic functions of the other layers have been developed. Because FC-3 sits between the FC-4 upper layer protocol and the FC-2 framing layer, FC-3 would contain services that would be performed immediately prior to handoff to the lower layers. This would include services such as encryption and authentication, although there are currently no such services in Fibre Channel implementations. Arguably, there has not been a lot of incentive to develop such facilities for Fibre Channel, because Fibre Channel SANs presuppose a private, secure environment. Putting storage traffic over metropolitan and wide area networks (MANs and WANs), however, may raise security concerns. This is another area that highlights the advantages of

storage traffic over IP, because encryption and authentication tools are readily available to safeguard sensitive storage data.

3.4.6 FC-4: Fibre Channel Upper Layer Protocol

The purpose of engineering a highly reliable physical plant, a rigorous byte encoding scheme, link-layer controls, efficient framing and sequence transmission, viable classes of service, and flow control is, of course, to service the upper layer applications behind which sit end users who are constantly creating and accessing stored data. Although the FC-4 layer standards include support for VI, IP, and other protocols, the most well-developed and most widely used FC-4 protocol is serial SCSI-3 (FCP).

The central task of FCP is to make Fibre Channel end devices appear as standard SCSI entities to the operating system. For host systems, the FCP function is embedded in the Fibre Channel HBA and the device driver supplied by the manufacturer. This allows Windows Disk Administrator, for example, to see Fibre Channel disks as SCSI-addressable storage resources. The operating system does not need to distinguish between storage resources that are direct-SCSI attached, ATA/IDE attached, or SAN attached. If, alternately, Fibre Channel as a storage networking solution had required changes to Windows, Solaris, or UNIX operating systems, it is doubtful that it ever would have been deployed. This is because of the much longer development and test cycles required for operating system revisions and the reluctance of customers to introduce additional complexity into their server environments. Just as FC-0 and FC-1 enabled reliable gigabit transmission of data at the physical and link levels, FCP has enabled a reliable protocol interface to the operating system and supported applications.

As shown in Figure 3–11, the upper layer protocol interface supports standard SCSI mapping for the operating system while maintaining Fibre Channel device address mapping for data destinations in the form of logical unit numbers (LUNs) on the target disks. The Fibre Channel frame header holds the 3-byte Fibre Channel network address, with identifying

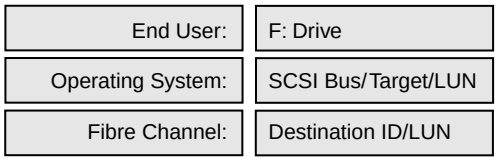

Figure 3–11 Perspectives on the Fibre Channel storage target.

LUN information contained in the frame payload. This tiered mapping is, thankfully, transparent to the end user, whose primary interface is through the drive designation assignable through the operating system's file system/volume management interface.

From the standpoint of the operating system, FCP translates standard SCSI commands into the appropriate SCSI-3 equivalents required for block data transfer over a serial network infrastructure. A SCSI I/O launched by the operating system to read blocks of data from disk, for example, would initiate an FCP exchange between the host and target using command frames known as *information units* (IUs). Within the exchange session, groups of frame comprising one or more sequences would be used to transport data from target to host. SCSI commands and responses between the operating system and FCP are implemented through the lower layers as serial SCSI FCP functions, as shown in Table 3–3.

Device drivers for HBAs must translate between conventional SCSI addressing and Fibre Channel device addresses. As a legacy from parallel SCSI, storage devices are identified by a bus/target/LUN triad. The bus is a SCSI chain hung from a specific SCSI port or SCSI adapter card. Multiple SCSI ports on a server require multiple bus designations. The target is a storage device, such as a disk. The logical unit identified by an LUN may represent a logical division of the disk—for example, a disk with two partitions that are accessible from the operating system as drives E: and F:. The device

TABLE 3–3 FCP EQUIVALENTS TO STANDARD SCSI FUNCTIONS (FROM AMERICAN NATIONAL STANDARDS INSTITUTE [ANSI] T10 FCP-2)

SCSI Function	FCP Equivalent
I/O operation	Exchange (concurrent sequences)
Protocol service request and response	Sequence (related frames)
Send SCSI command request	Unsolicited command IU (FCP_CMND)
Data delivery request	Data descriptor IU (FCP_XFER_RDY)
Data delivery action	Solicited data IU (FCP_DATA)
Send command complete response	Command status IU (FCP_ RSP)
REQ/ACK for command complete	Confirmation IU (FCP_CONF)

driver of the HBA or storage adapter card must translate this bus/target/LUN designation into a network-addressable identifier so that data can be passed to the appropriate storage target on the SAN. How this is implemented in IP storage environments is examined in the following chapters.

3.4.7 Fibre Channel Topologies

The three topologies supported by Fibre Channel are significant for IP-based storage networks both in terms of legacy support of Fibre Channel SAN segments and for understanding common features that can assist IP storage solutions for faster time-to-market. Different generations of Fibre Channel end devices may be optimized for specific topology protocols. Accommodating these devices via IP storage switch ports will encourage the transition from Fibre Channel SANs to IP-based SANs. In addition, the stability and demonstrated interoperability of Fibre Channel end devices and the FCP protocol enables IP-based storage networks to leverage the intellectual effort that has been vested in these technologies to date.

As shown in Figure 3–12, Fibre Channel supports direct point-to-point connections between two devices (typically a server and a single storage array); a shared, arbitrated loop topology; and a switched fabric. Gigabit Ethernet can support an analogous point-to-point connection as well as switched fabric, but in practical implementation has no shared media option. Fibre Channel point to point was commonly deployed for first-generation solutions, but because it only supports two devices it does not quite qualify as a storage network. Fibre Channel arbitrated loop is similar in concept to

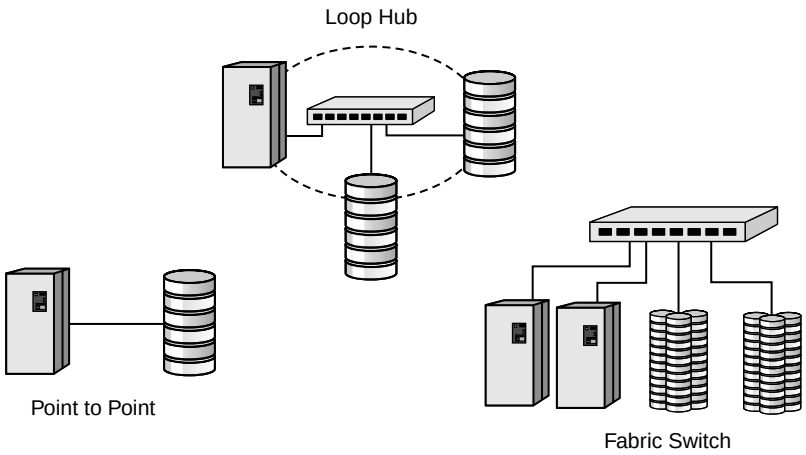

Figure 3–12 Fibre Channel topologies for point to point, loop, and fabric.

Token Ring. Multiple devices (as many as 126 end nodes) can share a common medium, but must arbitrate for access to it before beginning transmission. A Fibre Channel fabric is one or more Fibre Channel switches in a single network. Each device has dedicated bandwidth (200 MBps full duplex for 1Gb switches), and a device population of as many as 15.5 million devices is supported. This large number is strictly theoretical, because in practice it has been difficult for Fibre Channel fabrics to support even a few hundred devices.

Loop and fabric devices may be supported on a single network. A loop hub with six devices, for example, can be attached to a fabric switch port. Each of the devices registers its presence with the fabric switch so that it can communicate to resources on other switch ports. Such devices are referred to as *public loop devices*. They can also communicate with each other on the same loop segment without switch intervention. However, they each must arbitrate for access to their shared loop before any data transaction can occur.

One caveat for loop devices on fabrics is the result of the evolution of Fibre Channel device drivers. Not all loop-capable HBAs or storage devices can support fabric attachment. Such devices are known as *private loop devices*. To support these older loop devices, the fabric switch must provide proxy registration for them so that they become visible to the rest of the network and accessible as storage resources. There are no specific Fibre Channel standards covering this private loop proxy feature, and consequently every switch vendor's implementation is proprietary.

Switched fabrics pose significant issues, many of which are still unresolved. Fabric switches provide a number of services to facilitate device discovery and to adjust for changes in the network infrastructure. Devices register their presence on the switch via a simple name server (SNS), which is essentially a small database with fields for the device's network address, unique WWN, upper layer protocol support, and so on. When a server attaches to the fabric, it queries the SNS to discover its potential disk targets. This relieves the server from polling 15.5 million possible addresses to discover and establish sessions with storage resources. The SNS table in a standalone switch may be fairly small, with only 10 to 30 entries. When multiple switches are connected into a single fabric, however, they must exchange SNS information so that a server anywhere on the network can discover storage. The larger the fabric, the more difficult it becomes to update the collective SNS data and to ensure reliable device discovery.

Another fabric issue for large fabrics is the ability to track changes in the network. Fabric switches provide a registered state change notification (SCN)

entity that is responsible for alerting hosts to changes in the availability of resources. If, for example, a server has a session with a target array, it can be proactively notified if the array goes off-line or if the path to it through the fabric is broken. Because a Fibre Channel fabric is one large infrastructure, marginal components that trigger repeated SCNs can be disruptive to the entire network.

Management of Fibre Channel fabrics has evolved over time, but has been hindered by the challenges that a new technology faces. Out-of-band management with more familiar Simple Network Management Protocol (SNMP) protocols has enabled device and configuration management, although these are not as mature as the management platforms used by LAN and WAN networking. In-band management over the Fibre Channel links eliminates the need to have a parallel 10/100 Ethernet network for SNMP management, but is vulnerable to link failures. In-band management in Ethernet and WAN networks is predicated on redundant links. If both data and management traffic ride on the same network links, the failure of a single link would simply reroute traffic to available links in the meshed network. For these environments, provisioning redundant links is relatively inexpensive and simplifies network design and management. Achieving this level of redundancy in Fibre Channel networks is both expensive and awkward to implement. Without redundant links for in-band management, however, the loss of a data path may also be the loss of management traffic. Just when management information is needed the most, it would be unavailable.

Probably the most publicized issue for large Fibre Channel fabrics has been the lack of interoperability between vendor switch products. The standard for switch-to-switch connectivity (NCITS T11 FC-SW2) defines the connectivity and routing protocol for fabric switches. Switches are joined to each other via expansion ports, or E_Ports, and share routing information through the Fabric Shortest Path First (FSPF) protocol, a variant of the more commonly used Open Shortest Path First (OSPF) protocol for LAN and WAN networks. Although FSPF itself has not presented an overwhelming engineering challenge, competitive interests among Fibre Channel switch vendors have retarded its implementation. Until recently, dominant vendors have been unwilling to accelerate interoperability, fearing that openess would result in loss of market share. This, in turn, has led to absurd fabric designs simply to achieve higher port counts that could easily be accommodated through vendor interoperability. Some SAN designs have attempted to provide a high port count by deploying stacks of 10 or more 16-port switches, with nearly half the switch ports sacrificed for interswitch links.

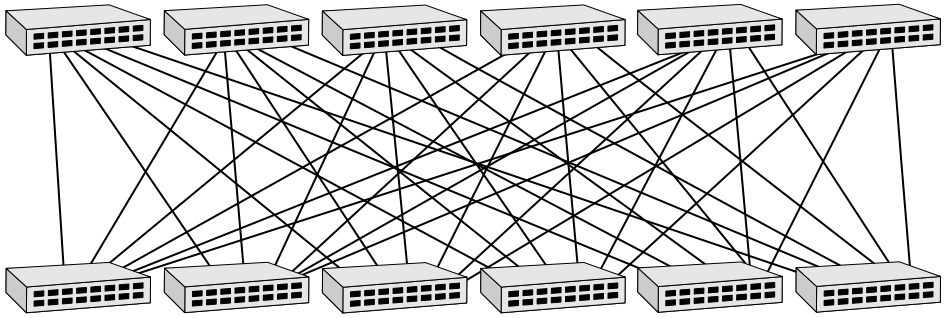

Figure 3–13 An actual vendor example of a higher port count fabric.

The result is a conglomeration of cabling that still results in a blocking architecture if more than one switch-to-switch transaction is started. With switch interoperability, it would be possible to combine high-port-count director-class switches with departmental 16-port switches for a more efficient deployment (Figure 3–13).

One inherent issue for large fabrics is the fact that a fabric is a single network. OSPF in LANs and WANs allows for the subdividing of networks into nondisruptive areas. A disruptive occurrence within a single area does not propagate throughout the entire network. FSPF does not provide this facility. Consequently, as a Fibre Channel fabric grows in population (and importance), it becomes increasingly vulnerable to outages.

Taken collectively, the issues associated with Fibre Channel fabrics are not insurmountable, but will require significant engineering resources to overcome them. The Fibre Channel fabrics that join Fibre Channel end devices have not achieved the level of stability and interoperability already attained by Fibre Channel HBAs, storage arrays, and tape subsystems. This makes the Fibre Channel fabric itself a prime candidate for replacement, which is the stated goal of IP storage solutions.

3.5 Gigabit Ethernet

Gigabit Ethernet owes its existence to the technical innovations of Fibre Channel transport and the historical momentum of Ethernet and IP networking. From Fibre Channel, Gigabit Ethernet has taken the breakthrough technology of gigabit physical specifications, fiber optics, CDR, 8b/10b data encoding, and ordered sets for link commands and delimiters. From Ethernet, it has inherited mainstream status and seamless integration to a vast

installed base of operating systems and network infrastructures. Although Fibre Channel has had to struggle for credibility as an emergent technology, Gigabit Ethernet's credibility was established before it was even implemented. Today, 10 Gigabit Ethernet and higher speeds are assumed to be the logical evolution of the technology and of future enterprise networks. In the process, Ethernet is shedding some of its characteristic attributes such as collision detection and shared topology, but is retaining its name.

3.5.1 Gigabit Ethernet Layers

The reference model for Gigabit Ethernet is defined in the Institute of Electrical and Electronics Engineers (IEEE) 802.3z standard. Like 10/100 Ethernet, Gigabit Ethernet is a physical and data link technology, corresponding to the lower two OSI layers, as shown in Table 3–4.

The Gigabit Ethernet physical layer contains both media-dependent and media-independent components. This allows the gigabit media-independent interface to be implemented in silicon and still interface with a variety of network cabling, including long- and shortwave optical fiber and shielded copper. The reconciliation sublayer passes signaling primitives between upper and lower layers, including transmit and receive status as well as carrier sense and collision detection. In practice, Gigabit Ethernet switching relies on dedicated, full duplex links and does not need a collision detection method. Carrier sense multiple access with collision detection (CSMA/CD) is

TABLE 3–4 GIGABIT ETHERNET PHYSICAL AND DATA LINK LAYERS

OSI Reference Layer	Gigabit Ethernet Layer
	Media access control (MAC) client sublayer
Data link layer	MAC control (optional)
	MAC
	Reconciliation
	Gigabit media independent interface
Physical Layer	Media-dependent PHY group
	Medium-dependent interface
	Medium

incorporated into the standard to provide backward compatibility with standard and Fast Ethernet.

Unlike Fibre Channel, Gigabit Ethernet's 8b/10b encoding occurs at the physical layer via sublayers in the media-dependent physical (PHY) group. As shown in Figure 3–14, Fibre Channel layers FC-0 and FC-1 are brought into the lower layer physical interface, whereas traditional 802.3 Ethernet provides MAC and logical link control (LLC; or its offspring, MAC client) to support the upper layer protocols.

To facilitate its integration into conventional Ethernet networks and wide area transports, Gigabit Ethernet uses standard Ethernet framing as shown in Figure 3–15. The preamble and SOF delimiter are followed by the destination (DA) and source (SA) MAC addresses of the communicating devices. Creative use of bytes within the length/type field enable enhanced functionality such as VLAN tagging, as discussed later. The data field may contain as much as 1,500 bytes of user data, with pad bytes if required. The CRC is part of the frame check sequence. Optional frame padding is provided by the extension field, although this is only required for gigabit half-duplex transmissions.

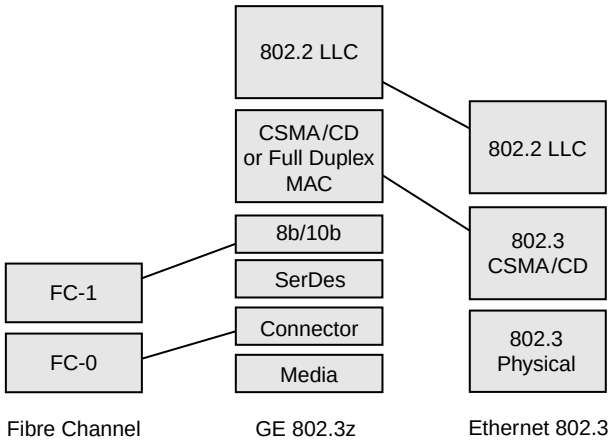

Figure 3–14 Gigabit Ethernet/Ethernet and Fibre Channel.

Preamble	SOF Delimeter	DA	SA	Length/Type	Data	(pad)	CRC	Ext

Figure 3–15 Standard Ethernet frame format.

IP over Ethernet is inserted into the data field and provides the network layer routing information to move user data from one network segment to another. TCP/IP provides higher level session control for traffic pacing and the ability to recover from packet loss. Although IP can be carried in other frame formats, link-layer enchancements for Ethernet offer additional reliability and performance capability unmatched by other transports, including Fibre Channel. These include VLANs, QoS, link-layer flow control, and trunking. Collectively, these functions provide a set of powerful tools for constructing storage networks based on IP and Gigabit Ethernet.

3.5.2 802.1Q VLAN Tagging

LANs in a switched Ethernet infrastructure enable the sharing of network resources such as large Gigabit Ethernet switches while segregating traffic from designated groups of devices. Members of a VLAN can communicate among themselves but lack visibility to the rest of the network. Sensitive information (for example, financial or human resource) can thus be isolated from other users, although the traffic is running through a common infrastructure. VLAN tagging was standardized in 1998 through the IEEE 802.1Q committee. An analogous capability is provided in Fibre Channel through a technique called *zoning*. Standards for zoning are still under construction, but only relate to the exchange of zone information by Fibre Channel switches. How zones are actually implemented within a switch is still proprietary. Consequently, there is no direct equivalent to 802.1Q's more open and flexible format.

VLAN tagging is accomplished by manipulating the length/type field in the Ethernet frame. To indicate that the frame is tagged, a unique 2-byte descriptor of hex "81–00" is inserted into the field. This tag type field is followed by a 2-byte tag control information field, as shown in Table 3–5, which carries the VLAN identifier and user priority bits as described later.

TABLE 3–5 IEEE 802.1Q VLAN TAG FIELDS

802.1Q Tag Type Field	Tag Control Information Field		
	User Priority	Canonical format indicator bit (CFI)	VLAN Identifier
81-00			
16 bits	3 bits	1 bit	12 bits

The 12-bit VLAN identifier allows for as many as 4,096 VLANs to be assigned on a single switched infrastructure—far more than the number of zones typically offered by Fibre Channel switch vendors.

From a performance standpoint, VLAN tagging is a highly efficient means of segregating network participants into communicating groups without incurring the overhead of MAC address filtering. Intervening switches use the logical VLAN identifier, rather than the MAC address, to route traffic properly from switch to switch, and this in turn simplifies the switch decision process. As long as the appropriate switch port is associated with the proper VLAN identifier, no examination of the MAC address is required. Final filtering against the MAC address occurs at the end point.

All major Gigabit Ethernet switch vendors support the 802.1Q standard. This makes it a very useful feature not only for data paths that must cross switch boundaries, but for heterogeneous switched networks as well. For IP storage network applications, 802.1Q facilitates separation of storage traffic from user messaging traffic as well as segregation of different types of storage traffic (for example, on-line transaction processing) from tape backup. Compared with Fibre Channel zoning, 802.1Q VLANs offer more flexibility and lack the complexity of vendor-specific implementations.

3.5.3 802.1p/Q Frame Prioritization

The 802.1Q VLAN tag control information field allocates 3 bits for user priority. The definition for these User Priority bits is provided by IEEE 802.1p/Q, and enables individual frames to be marked for priority delivery. The QoS supported by 802.1p/Q allows for eight levels of priority assignment. This ensures that mission-critical traffic will receive preferential treatment in potentially congested conditions across multiswitch networks, and thus minimizes frame loss resulting from transient bottlenecks.

For storage network applications, the ability to prioritize transactions in an IP-based SAN is a tremendous asset. Storage networks normally support a wide variety of applications, not all of which require high priority. Updating an on-line customer order or a financial transaction between banks, for example, rates a much higher priority for business operations than a tape backup stream. The class of service provided by 802.1p/Q allows storage administrators to select the applications that should receive higher priority transport and assign them to one of the eight available priority levels. In a multiswitch network, class of service ensures that prioritized frames will have preference across interswitch links. Except for a few proprietary port-based implementations, Fibre Channel currently does not support frame

prioritization and thus cannot distinguish between mission-critical and less essential storage applications.

3.5.4 802.3x Flow Control

Flow control at the data link level helps to minimize frame loss and avoids latency resulting from error recovery at the higher layer protocols. In Fibre Channel, flow control for class 3 service is provided by a buffer credit scheme. As buffers are available to receive more frames, the target device issues receiver readys (R_RDYs) to the initiator, one per available buffer. In Gigabit Ethernet, link-layer flow control is provided by the IEEE 802.3x standard. The 802.3x implementation uses a MAC control PAUSE frame to hold off the sending party if congestion is detected. If, for example, receive buffers on a switch port are approaching saturation, the switch can issue a PAUSE frame to the transmitting device so that the receive buffers have time to empty. Typically, the PAUSE frame is issued when a certain high-water mark is reached, but before the switch buffers are completely full.

Because the PAUSE frame is a type of MAC control frame, the frame structure is slightly different from the conventional data frame. Like VLAN tagging, the length/type field is used to indicate the special nature of the frame, in this case hex "88–08" to indicate a MAC control frame. As shown in Table 3–6, this indicator is followed by an opcode of hex "00–01" to define further the MAC control frame as a PAUSE frame. The amount of time that a transmitting device should cease issuing frames is specified by the opcode parameter field. pause_time cannot be specified in fixed units such as microseconds, because this would prove too inflexible for backward compatibility and future Ethernet transmission rates. Instead, pause_time is specified in pause_quanta, with one pause_quanta equal to 512 bit_times for the link speed being used. The timer value can be between 0 and 65,535 pause_quanta, or a maximum of approximately 33 msec at Gigabit Ethernet's 1.25-Gbps transmission rate. If the device that issued the PAUSE frame empties its buffers before the stated pause_time has elapsed, it can issue

TABLE 3–6 IEEE 802.3X PAUSE FRAME FORMAT

Length/Type	MAC Control Opcode	Parameters
88-08	00-01	Pause_time
16 bits	16 bits	16 bits

another PAUSE frame with pause_time set to zero. This signals the transmitting device that frame transmission can resume.

Because PAUSE frames may be used between any devices and the switch ports to which they are attached, and because Gigabit Ethernet only allows one device per port, there is no need to personalize the PAUSE frame with the recipient's MAC address. Instead, a universal, well-known address of 04-80-C2-00-00-01 is used in the destination address field. When a switch port receives the PAUSE frame with this address, it processes the frame but does not forward it to the network.

The 802.3x flow control provided by Gigabit Ethernet switches creates new opportunities for high-performance storage traffic over IP. Fibre Channel class 3 service has already demonstrated the viability of a connectionless, unacknowledged class of service, providing there is a flow control mechanism to pace frame transmission. In Fibre Channel fabrics using class 3, as with 802.3x in Ethernet, the flow control conversation occurs between the switch port and its attached device. As the switch port buffers fill, it stops sending R_RDYs until additional buffers are freed. In Gigabit Ethernet, this function is performed with PAUSE frames, with the same practical result. In either case, buffer overruns and the consequent loss of frames are avoided, and this is accomplished with minimal impact on performance.

The reliability provided by the gigabit infrastructure through data link flow control enables streamlined protocols to be run at the upper layer. For IP storage, the equivalent to Fibre Channel class 3 is UDP. UDP is connectionless and unacknowledged, and thus is unsuited to very congested environments such as the Internet. For contained data center storage applications, however, 802.3x flow control and storage over UDP/IP can offer a reliable and extremely high-performance solution without incurring the protocol overhead of TCP/IP.

3.5.5 802.3ad Link Aggregation

Link aggregation, or trunking, provides higher bandwidth for switched networks by provisioning multiple connections between switches or between a switch and an end device such as a server. Link aggregation also facilitates scaling the network over time, because additional links to a trunked group can be added incrementally as bandwidth requirements increase. In Figure 3–16, two Gigabit Ethernet switches share three aggregated links for a total available bandwidth of 7.5 Gbps full duplex.

Originally, the 802.3ad standards initiative was promoted as a means to provide higher bandwidth for standard 10/100-Mbps Ethernet networks.

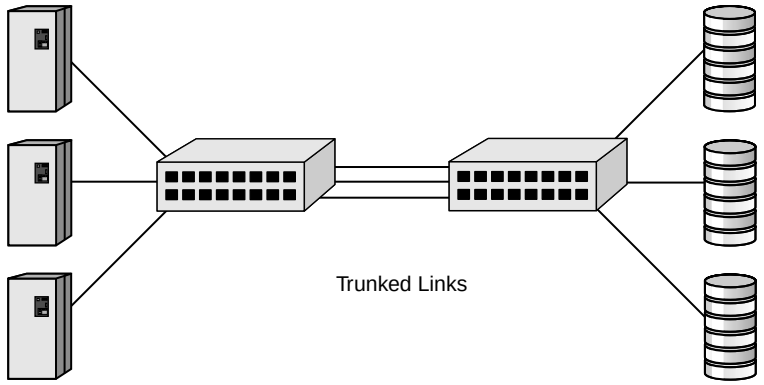

Figure 3–16 Link aggregation between two Gigabit Ethernet switches.

Link aggregation was a means of satisfying higher bandwidth requirements while Gigabit Ethernet was still being developed. As with memory, CPUs, and storage, however, whatever performance or capacity is reached at any given point in time is never sufficient for the increasing demands of users and applications. Consequently, bundled Ethernet links have been replaced with bundled Gigabit Ethernet links, which at some point will be superseded by bundled 10 Gigabit and higher Ethernet links. Replicators, for example, will no doubt require bundled 100 Gigabit Ethernet links.

Link aggregation must resolve several issues to avoid creating more problems than it fixes. In normal bridge environments, the spanning tree algorithm would, on encountering multiple links between two devices, simply disable the redundant links and only allow a single data path. This would prevent duplication of frames and potential out-of-order delivery. Link aggregation must therefore make multiple links between two devices appear as a single path, while simultaneously providing a mechanism to avoid frame duplication and ensure in-order frame delivery. This could be accomplished by manipulating MAC addresses (for example, assigning the same MAC address to every trunked link) or by inserting a link aggregation intelligence between the MAC client and MAC layers. The status of link availability, current load, and conversations through the trunk require monitoring to ensure that frames are not lost or inadvertantly reordered.

In-order delivery of frames is guaranteed if a conversation between two end devices is maintained across a single link in the trunk. Although this is not as efficient link utilization as simply shipping each frame over any available connection, it avoids the extra logic required to monitor frame

ordering and to reassemble them before delivery to the recipient. At the same time, additional transactions by other devices benefit from the availability of the aggregated interswitch links, and switch-to-switch bottlenecks are avoided.

Link aggregation as specified in 802.3ad is almost mandatory for IP-based storage networks, particularly when multiple Gigabit Ethernet switches are used to build the SAN backbone. Along with 802.1p/Q prioritization, link aggregation can ensure that mission-critical storage traffic has an available path through the network and that multiple instances of mission-critical transactions can occur simultaneously. This requirement will be satisfied temporarily by the arrival of 10Gb uplinks between switches, but these will inevitably be "trunked" to provide even higher bandwidth over time.

3.5.6 Gigabit Ethernet Physical Layer Considerations

Gigabit Ethernet has borrowed so heavily from the Fibre Channel physical layer that there are relatively few differences between them. Gigabit Ethernet has a slightly higher transmission rate of 1.25 Gbps, compared with Fibre Channel's 1.0625 Gbps. For storage applications, Gigabit Ethernet's faster clock can drive approximately 15 MBps more bandwidth over interswitch links.

Transceivers for Gigabit Ethernet applications may also vary, although some GBICs can interoperate with both Fibre Channel and Gigabit Ethernet transmission speeds. Gigabit Ethernet has introduced support for new cable types for gigabit transport, including category 5 unshielded twisted pair. As shown in Table 3–7, cable distances are comparable with Fibre

TABLE 3–7: GIGABIT ETHERNET CABLE SPECIFICATIONS

Cable Type	Diameter	Laser Type	Maximum Distance (m)
1000BASE-T	CAT-5 UTP	N/A	100
1000BASE-CX	STP	N/A	25
1000BASE-LX	10 μm	Long wave	5,000
1000BASE-SX	50 μm	Short wave	500
1000BASE-LX	50 μm	Long wave	550

Channel, with the exception of long-wave, single-mode cabling (10 km for Fibre Channel).

3.6 Assumptions for IP-Based SANs

Because it is probable that legacy SCSI, Fibre Channel, and IP-based SANs will coexist for some time, the adoption rate of IP storage solutions will depend on the ability of vendors to supply stable, interoperable, and high-performance products that can accommodate a variety of storage network interfaces.

As the first-generation storage network infrastructure, Fibre Channel has set expectations in terms of storage performance and network flexibility. Although management and interoperability of fabrics may be problematic, stability and performance have at least been achieved for Fibre Channel interfaces on end devices. The onus is therefore on IP storage vendors to accommodate these devices and to ensure the same level of reliability and interoperability for native IP storage interfaces.

Block storage data over IP and Gigabit Ethernet must provide performance equal to or greater than other storage solutions. This is facilitated by functionality inherent in Gigabit Ethernet, including faster transmission speeds (going to 10Gb), link-layer flow control, and link aggregation.

IP storage must also provide enhancements for storage transport unavailable by other means. The more flexible capabilities of VLANs over Gigabit Ethernet and traffic priorization give storage administrators new tools for securing mission-critical transactions in the SAN and for sharing the SAN infrastructure between disparate storage applications.

In terms of management, IP-based SANs benefit from the much wider deployment of sophisticated transport management platforms for enterprise data networks. The merger of storage and networking, however, creates unique requirements beyond network transport management. The integration of network management with storage management is still required to simplify the configuration and management of the SAN, and to reduce administrative overhead.

In terms of interoperability, the greater stability of Fibre Channel end devices and demonstrated interoperability between Gigabit Ethernet switches presents opportunities to combine the best from both worlds to facilitate IP SANs. Native IP storage devices, however, also have to demonstrate both standards compliance and interoperability to achieve acceptance in the market.

3.7 Chapter Summary

Storage Networking Terminology

- RAID provides performance and data redundancy.
- RAID 0 stripes data blocks over multiple disks.
- RAID 1 provides data duplication (mirroring) between two disks.
- RAID 5 provides striping of data blocks and distributed parity for data reconstruction in the event of failure.
- RAID levels may be combined—for example, RAID 0 + 1.
- JBODs are more economical than RAID but have no inherent redundancy or mirroring capability.
- To maximize performance, software RAID on host systems can be applied against JBOD targets.
- Tape subsystems may be SCSI, Fibre Channel, or Gigabit Ethernet interfaces.
- Bridge products can bring legacy SCSI devices into an IP-based SAN.
- SAN host adapters may include Fibre Channel HBAs and IP storage NICs.
- IP storage NICs may provide optimized logic for TCP off-loading.

Legacy SCSI Cabling

- SCSI cabling provides parallel wires for simultaneous transfer of data bits.
- The maximal SCSI cabling is 25 m.
- The maximal device population for SCSI cabling is 15 devices on a string.
- Skew refers to the window of time required to capture all data bits in a parallel transmission.

Network-Attached Storage

- NAS serves files; SANs provide data blocks.
- The NAS architecture is comprised of a thin server and attached storage.

- NAS storage may be ATA, SCSI, or Fibre Channel.
- NAS uses NFS or CIFS over IP for file access.
- NAS products are typically marketed as appliances, requiring little configuration or management.
- NAS and IP-based SAN traffic may share a common network infrastructure.

Fibre Channel

- Fibre Channel is a standards-based, layered architecture.
- FC-0 defines gigabit physical layer specifications.
- FC-1 provides data encoding and link-level controls.
- FC-2 defines segmentation and reassembly of data via frames, flow control, and classes of service.
- FC-3 is being developed for common services such as encryption.
- FC-4 is the upper layer protocol interface between Fibre Channel and IP, SCSI-3, and other protocols.
- The most commonly used FC-4 protocol is FCP for serial SCSI-3.
- FC-0 and FC-1 provide the foundation layers for Gigabit Ethernet.
- The 8b/10b data encoding algorithm coverts 8-bit bytes into 10-bit characters.
- Encoding is required to prevent sustained DC states on the gigabit link.
- A proportional representation of ones and zeros is maintained via running disparity.
- Ordered sets are 10-bit characters used for frame delimitation, signaling, and link change notification.
- The maximal payload for a Fibre Channel frame is 2,112 bytes.
- The most commonly used class of service for Fibre Channel is class 3, which is connectionless and requires no acknowledgment of frame receipt.

- Fibre Channel has no standardized encryption or authentication methods.
- FCP is responsible for mapping SCSI devices at the operating system level to Fibre Channel-attached storage resources.
- Fibre Channel topologies include point to point, arbitrated loop, and fabric.
- Fibre Channel fabrics use a subset of OSPF called FSPF for fabric routing.
- The Fibre Channel fabric appears as one integral network.
- Interoperability and management issues for fabric switches have retarded the deployment of Fibre Channel SANs.

Gigabit Ethernet

- Gigabit Ethernet is a data link transport that borrows from both Fibre Channel and conventional 802.3 Ethernet.
- Ethernet framing is used to transport TCP/IP data over Gigabit Ethernet networks.
- 802.1Q VLAN tagging allows segregation of devices on the SAN.
- 802.1p/Q frame prioritization enables mission-critical traffic to be assigned one of eight levels of priority for SAN transport.
- 802.3x flow control provides reliable transport of storage data over connectionless protocols such as UDP/IP.
- 802.3ad link aggregation allows scalability of IP-based SANs with no loss in performance.
- Gigabit Ethernet's transmission rate of 1.25 Gbps provides slightly better performance than Fibre Channel.
- Gigabit Ethernet cabling includes category 5 unshielded twisted pair copper cabling as well as standard multimode and single-mode fiber-optic cabling.

Assumptions for IP-Based SANs

- Storage networks based on IP and Gigabit Ethernet can leverage new functionality for class of service, VLANs, flow control, and trunking provided by Ethernet standards.

- An optimal storage over IP solution accommodates legacy SCSI, Fibre Channel, and native IP storage devices.
- Management of storage networks requires the integration of transport management and storage management.
- Interoperability is a key driver for market adoption.