The Complete Reference

Chapter 13

Architecture Overview

In 1993, the largest data warehouse application supported only 50GB of aggregate data. Although this appears trivial by today's standards, it drove the client/server and mainframe computing infrastructures to support larger and larger data capacities. Both mainframe and open server vendors responded by increasing the number of processors, channels, RAM memory, and bus capabilities. However, these technologies had reached a level of maturity in their development that limited dramatic improvements in many of these areas. Coupled with the momentum of online transactional workloads driven by Internet access, web applications, and more robust Windows-based client/server applications, the network began to play an increasing role in supporting the larger data-centric platforms. However, the distribution of work through the network only found the LAN to be challenged by its own data-centric traffic congestion.

In a different response to this challenge, both new and established vendors moved toward hybrid types of computing platforms, while methods appeared to handle the growing data-centric problem (see Chapter 8). These were parallel processing and high-end SMP computing platforms using high-performance relational database software that supported increased degrees of parallel processing. Vendors of systems and databases were thus caught in a dilemma, given that applications requiring data storage above 100GB capacities needed very costly high-end solutions.

Driven by the sheer value of access to data, end-user appetites for data-centric solutions continued to grow unabated. However, these requirements proved increasingly difficult for IT organizations as they struggled to supply, manage, and integrate hybrid solutions into their data centers. Obviously, one of the main areas experiencing dynamic growth within this maelstrom of datacentric activity was storage and storage-related products. This was an interesting dilemma for the storage vendors, given that the key to any enhancement of the data-centric application rollout revolved around a high-end I/O system. This drove storage vendors to the same traditional solutions as their mainframe and open-server brethren: enhance the existing infrastructure of disk density, bus performance (SCSI, PCI), and array scalability.

As the conservative approach from mainframe and open server vendors drove innovators offering alternative solutions to start their own companies and initiatives, a similar evolution began within storage companies. Existing storage vendors stuck to their conservative strategy to enhance existing technologies, spawning new initiatives in the storage industry.

Several initiatives studied applying a network concept to storage infrastructures that allowed processing nodes (for example, servers) to access the network for data. This evolved into creating a storage network from existing architectures and technologies, and interfacing this with existing I/O technologies of both server and storage products. Using the channel-oriented protocol of Fibre Channel, a network model of packet-based switching (FC uses frames in place of packets, however), and a specialized operating environment using the micro-kernel concept, the architecture of Storage Area Networks came into being.

This was such a major shift from the traditional architecture of directly connecting storage devices to a server or mainframe that an entire I/O architecture was turned upside down, prompting a major paradigm shift. This shift is not to be overlooked, or taken lightly, because with any major change in how things function, it must first be understood, analyzed, and evaluated before all its values can be seen. With Storage

Area Networks, we are in that mode. Not that the technology is not useful today. It can be very useful. However, the full value of Storage Area Networks as the next generation of I/O infrastructures still continues to evolve.

Creating a network for storage affects not only how we view storage systems and related products, but also how we can effectively use data within data-centric applications still in demand today. Here we are, ten years from the high-end data warehouse of 50GB, with today's applications supporting 500GB on average. This is only a ten-fold improvement. What can we achieve if we move our data-centric applications into an I/O network developed specifically for storage (perhaps a hundred-fold improvement)? Hopefully, ten years from now, this book will reflect the average database supporting 5 terabytes and being shared among all the servers in the data center.

Creating a Network for Storage

Network attached storage is very different from a Storage Area Network on many levels. First, SANs denote an entire infrastructure supporting storage systems. Secondly, (and maybe this should actually be first) SANs function on their own network developed specifically for shared I/O processing, enhanced storage devices, and scalability within a data center infrastructure. Thirdly, SANs operate on a protocol different than NAS (NAS integrates into traditional TCP/IP networks and associated network topologies) by using Fibre Channel. Lastly, SANs offer complete flexibility within their infrastructure by maintaining the intrinsic I/O communications protocols inherent with directly attached storage.

Whereas NAS remains a retrofit to existing computer networks, SANs offer an entire I/O infrastructure that breaks the storage boundaries of traditional storage I/O models and moves the support for data-centric applications into a new generation of computer processing models.

Several things happen when you create a storage network. The storage becomes accessible through a network model—for example, nodes logging in to the network can communicate with other nodes within the network. Nodes operating within the network offer a diverse amount of function depending on their device types—they can be storage devices, servers, routers, bridges, and even other networks. You can transfer data faster, with more throughput, and with increased flexibility. Managing the resources in a storage network can be performed from a centralized perspective across sharable domains.

This results in an inherent value to the Storage Area Network. Servers can now share the resources of storage systems such as disk arrays and devices. Conversely, storage arrays can be shared among the servers consolidating the number of devices required. This means increased access to centralized data by large numbers of applications, with support for larger storage capacities through increased ways of providing I/O operations.

Storage Area Networks are made up of four major parts. As we discussed with NAS, these parts cover the major areas of I/O operations, storage systems, and supported workloads. In SAN technology, however, there are more seemingly disparate parts that must be integrated to form an entire solution. With NAS, most of the products available are offered as bundled solutions. Not so with SANs. When considering SAN infrastructure, we must ponder more carefully the separate components that make up the infrastructure,

because each operates independently and interdependently with the Storage Area Network they participate in.

Given that consideration, we can discuss and consider the major components of a Storage Area Network and resulting architecture. SAN components include the following:

- **Network Part** SANs provide a separated network over and above what the client/server or mainframe networks utilize in connecting clients, terminals, and other devices, including NAS. This network, as mentioned, is based on the Fibre Channel protocol and standard.
- **Hardware Part** SANs depend on specific hardware devices just like other networks. Given that SANs provide a packet-switched network topology, they adhere to a standard layer of processing that the hardware devices operate within. Again, to avoid confusion between data communications packet topologies, the FC protocol relies on frames and an architecture more compatible with I/O channel operations.
- **Software Part** SANs operate within a separate set of software called a fabric that makes up the FC network. Additional software is required at the server connection where the average Windows or UNIX OS drivers must communicate within a SAN environment. The same thing is true for storage devices that must communicate with the SAN network to provide data to the servers.
- **Connectivity Part** Much of a SAN's value derives from its connectivity, which is made up of several hardware and software functions. Given the complete shift in I/O operations from bus level communications to network communications, many components must change or be modified to operate within this environment. Connectivity options within the SAN determine performance and workload applicability.

Figure 13-1 offers a glimpse into the SAN infrastructure.

The Network Part

SAN is a network and, as such, its main task is to provide communications among devices attached to it. It does so through the FC standard protocol, denoted by the FC standard and relative T-11 standards committee that maintains the protocol. The FC protocol offers a layered approach to communications similar to TCP/IP, but through a smaller set of processing layers. The FC layers are illustrated in Figure 13-2, showing functions FC-0 through FC-4. (More detail on FC layers can be found in Chapter 16.) However, what's important is that each layer is leveraged through different hardware components within the SAN.

What facilitates the network is the switch component that provides a set of circuits making up the communications paths for the devices attached. The communications paths are accommodated through the addressing associated with the devices' communications layers. Therefore, the FC switch device can provide an any-to-any connectivity matrix that allows communications from one device to another.

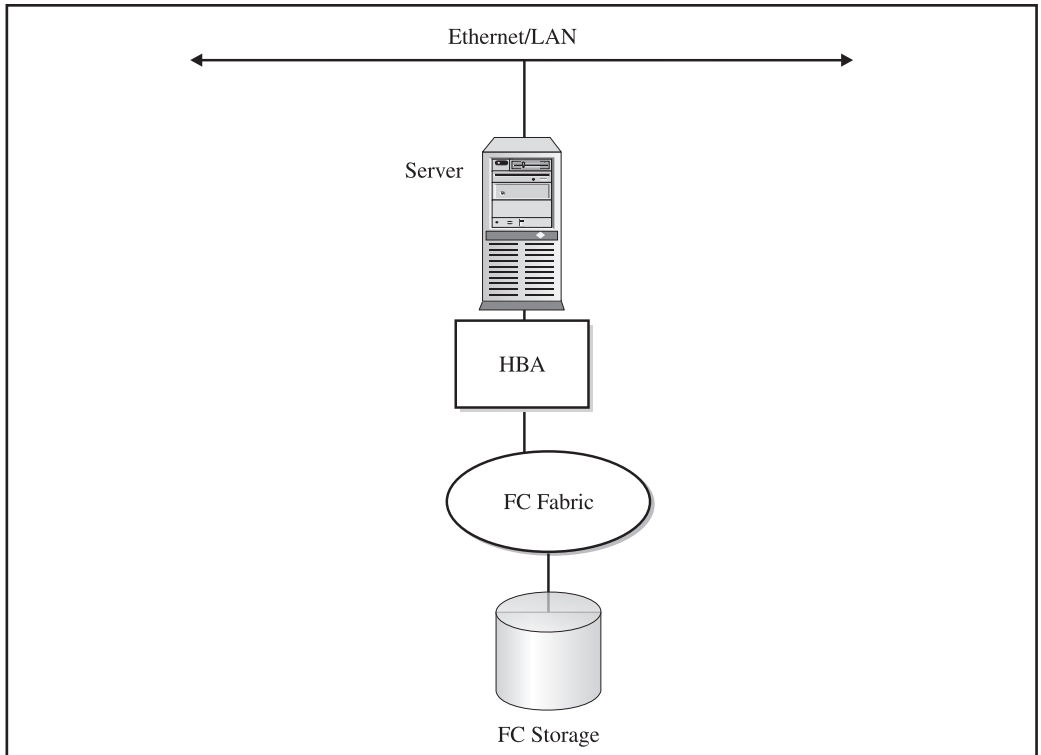

Figure 13-1. *SAN components overview*

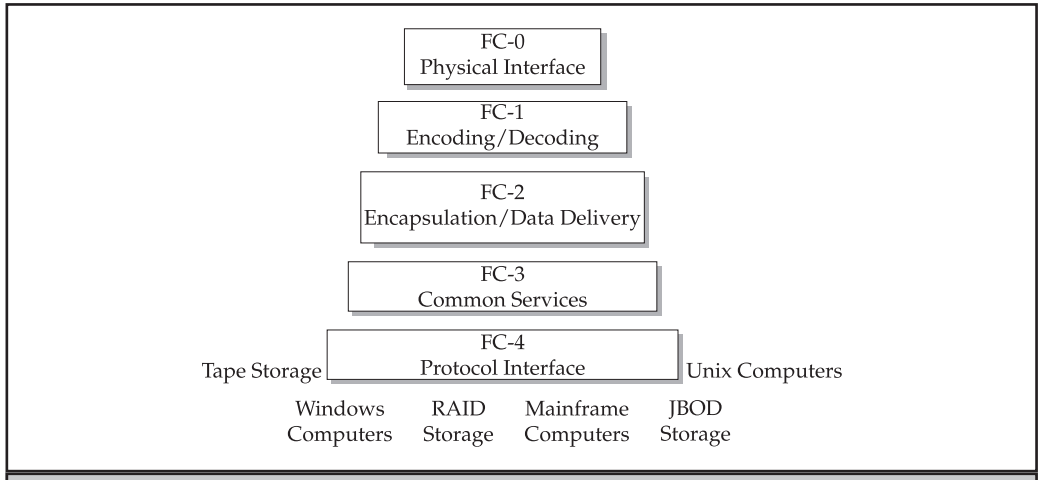

Figure 13-2. *FC layers and functions*

For example, a server may require a read from a particular disk. The server's I/O request provides the address for the requested device, the network fabric develops the connection, and that passes the read I/O to the storage device. Conversely, the storage responds with the same type of process, whereby the storage device requesting a connection with the server address returns the block of data from the read operation, the switch fabric makes the connection, and the I/O operation is completed.

It is important to note that one of the particular values of SANs, and specifically Fibre Channel, is its capability to provide a way for existing protocols to be encapsulated within the communications. This is especially valuable, as the SCSI commands for disk operations do not have to change. Consequently, disk and driver operations can operate within the FC protocol unchanged.

However, it is also important to note that in developing the SAN network, additional components are required beyond the FC switch for both the server and the storage devices. These are depicted in Figure 13-3, as we see the effects of our read I/O operation within the switch. These components—HBAs, FC switches, hubs, and routers—will be discussed in more detail in the section titled "The Hardware Part" in this chapter, as well as in Chapter 14.

Inside the FC network, the fabric software operates on a frame-based design. Similar to concepts used in TCP/IP packets, communications within the network are performed by transmitting FC frames throughout the fabric. The frames are composed of header, user data, and trailer information. FC layers determine the operations responsible for the frame development, frame addressing, and transmission (see Figure 13-3).

We would be remiss if we did not mention other ways the FC network can be configured. This is important because many of these configurations are still operating and some offer specific solutions to particular situations; they're also cost-effective. However, most can be attributed to implementations that occurred prior to the development of the FC Switched fabric solutions.

These solutions include the following:

- **Point-to-Point** This uses the FC to connect one device to another. Employed in initial FC implementations of FC disk arrays, it leveraged the increased bandwidth, but remained a direct attached solution. Also, tape devices could be used with an FC point-to-point configuration, increasing the number of drives supported from a single server.
- **Hub** This uses FC to connect in a loop fashion. Basically, the initial network fabric supported an arbitrated loop arrangement whereby, like SCSI bus configurations, the devices were configured in loop topologies which not only shared bandwidth but had to arbitrate for network communications, like a SCSI bus. It leveraged the speed of FC and the capability to place additional disk arrays within the network allowing connection of additional servers. FC-AL, however, provided latency issues that never fully facilitated the bandwidth and performance of FC-switched fabric implementations.

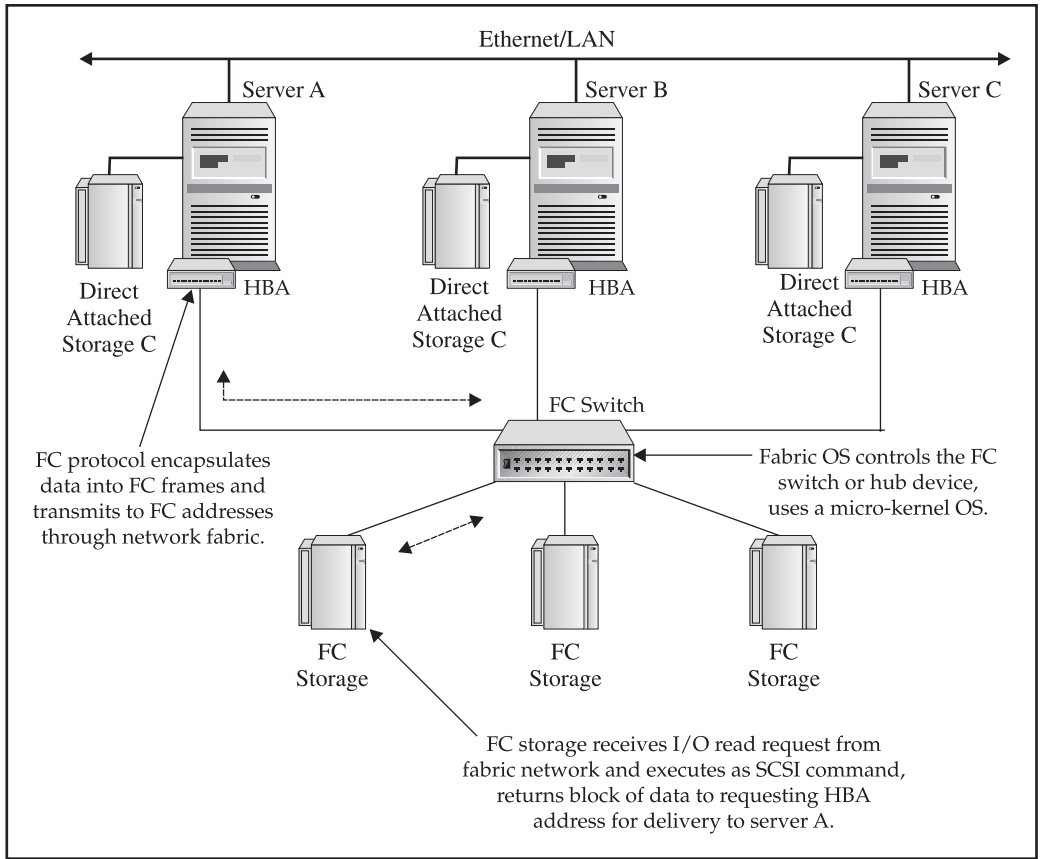

Figure 13-3. *A SAN devices overview and sample read I/O*

The Software Part

The central control of a SAN is contained within the Fabric Operating System functions (sometimes called the SAN OS, or just the fabric). Like micro-kernel components, SAN fabric utilizes the micro-kernel implementation as the basis for its operating system.

The Fabric OS runs within the FC switch. Figure 13-4 shows the typical makeup of the Fabric OS. As we can see, it truly is an effective implementation of a micro-kernel, supporting only the required functions of the fabric. This is good news for performance, but bad news whenever future enhancements and management functions are considered. It has similar limitations to the NAS bundled solution—a micro-kernel conundrum shared among storage networking solutions.

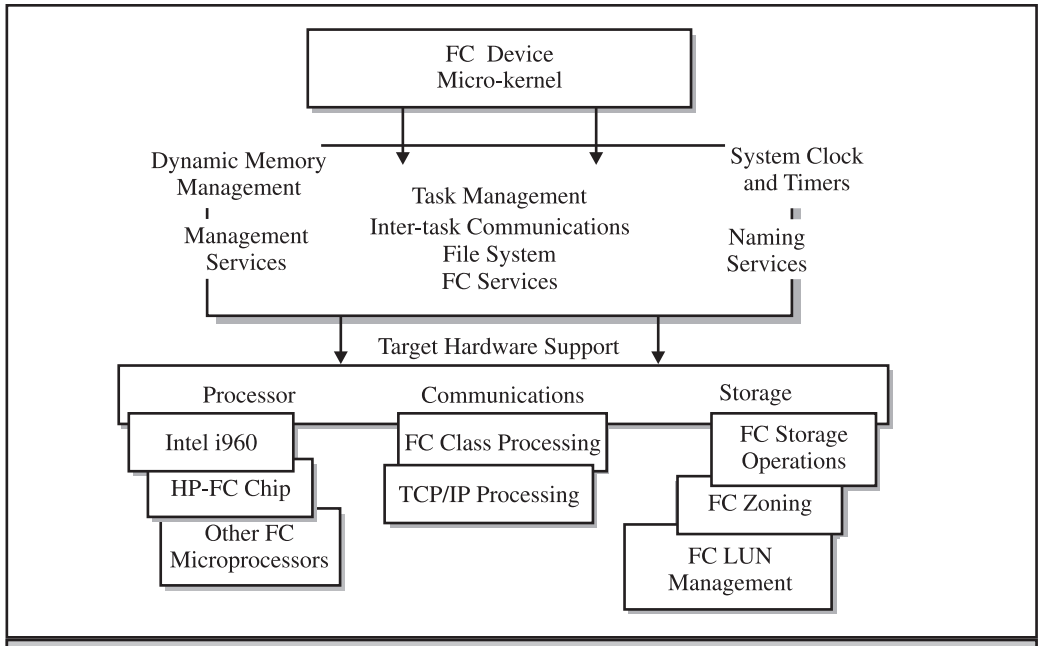

Figure 13-4. *A typical fabric OS micro-kernel*

Fabric OS Services

The Fabric OS offers a common set of services provided by any network. However, it also provides services specific to Fibre Channel and I/O operations. These services are summarized in Figure 13-5 and described next:

- **OS Services** Required functions configured with micro-kernel, such as task, memory, and file management, are included within these services. These form the basic OS functions that provide the core-level processing specific to supporting a fabric.
- **FC Services** These functions interact closely with OS services to facilitate the FC layer processing of frame management within the switch. The FC services provide both switched fabric and arbitrated loop modes of processing, as well as frame processing and protocol routing for the complex interconnections that occur within the switch.
- **Simple Name Services** The FC network works on the basis of a naming convention that identifies devices (such as HBA port and storage adapter port) by their names as they are attached to the fabric. These services provide a rudimentary database (for example, a file that supports the SAN naming conventions, the status of device connections, and so forth) and functions to identify new devices as they are attached.

- **Alias Services** The fabric provides functions that allow particular devices to broadcast frames to a set of aliases as noted within the naming conventions (for instance, a generic set of names that one device can broadcast to).
- **Management Services** FC switches that provide startup and configuration utilities allowing administrators to set up and configure the switch. These utilities range from direct attached access through a PC attached to a serial interface of the switch to web browser-based access if the switch is configured with an Ethernet interface. The management functions over and above configuration and setup consist of status and activity information stored and accessed through a Simple Network Management Protocol (SNMP) database. This database referred to as a MIB (Management Information Base) is specific to SNMP and requires this protocol to extract or store any information. Additional information and discussion of SAN management can be found in Part VI.

Fabric APIs and Applications

FC-switched fabric systems operate similar to other software operating environments by providing a set of basic services to the supporting hardware (refer to the previous bullet points within the OS services discussion). However, like any base operating environment, it must have interfaces to the outside world for access from administrators and other software. These access points will come from application programming interfaces, or APIs. This software interface will provide third-party software vendors, or any systems administrators brave enough to try, the specification for writing

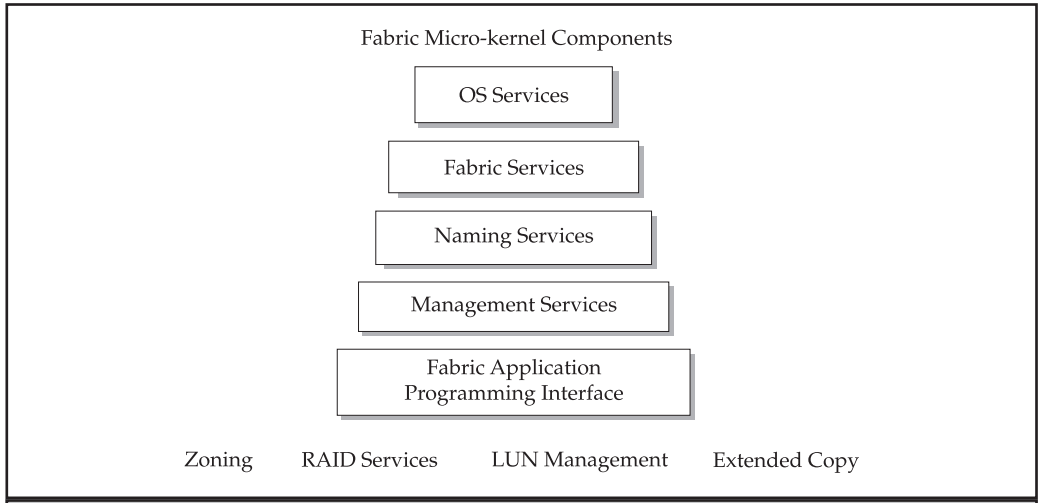

Figure 13-5. *Fabric OS micro-kernel components*

a program to work with the fabric OS. We provide an overview of both API and fabric applications below.

- **Application Programming Interfaces** FC switch vendors all include various levels of Application Programming Interfaces created specifically to allow other software vendors and complementary hardware vendors to develop applications that enhance the operation of the switch. These interfaces are generally so complex that it is beyond most IT administrators to leverage, nor is it recommended by the switch vendors to facilitate local API utilization by customers. Regardless of complexity, this does provide an important feature which enables software vendors—mostly management vendors—to supply applications that interface with the switch and therefore enhance its operation.
- **Fabric Applications** FC fabric supports specific applications that adhere to the vendor's API structure. Applications generally provided by the switch vendor are characterized by their close operation of the switch through facilities such as zoning, hardware enclosure services, and advanced configuration features. Third-party tools, offered as products from software vendors, will interface with the switch to offer mainly management services such as backup and recovery, device and component monitoring, and storage services.

Server OS and HBA Drivers

Dealing with I/Os generated from servers requires that the server operating systems be SAN-compliant. This generally means that the OS is able to recognize the FC Host Bus Adapter driver and related functions necessary to link to SAN attached disks. Although mostly transparent for disk operations, this aspect of the software dependencies and functions supporting the SAN configuration is becoming increasingly complex as file and database management becomes more compliant with SAN network functions.

Most system vendors offering operating systems support SANs. However, it should be noted that further diligence is necessary to identify specific OS release levels that support a specific SAN implementation. Given the variety of hardware and network devices supported through operating system compatible drivers, the support for particular SAN components requires detailed analysis to ensure the attached servers support the SAN infrastructure.

All HBA vendors provide drivers to support various SAN configurations and topologies. These drivers are matched to a specific level of OS. Given most SANs are available as a packaged solution, it is critical that the driver software provides sufficient flexibility to support not only the FC switch configuration, but also the storage arrays and other devices attached, such as routers/bridges, tapes, and other server HBAs. Refer to the OS macro guide to cross reference the SAN support for a particular configuration. As with the server OS consideration, more diligence is necessary to identify specific driver release levels that support a specific SAN implementation and set of hardware components and fabric.

Storage Intelligence

An additional area of software to note is the functional area inherent with particular storage systems. Storage arrays are all delivered with particular implementations of FC, SCSI, and added utility applications. These software implementations must be compatible with both the server OS and fabric OS, as well as the HBA drivers.

Other Networks As a final note on software, we must recognize the effects SAN fabrics are having on TCP/IP and their integration into other enterprise networks. We noted previously that the integration of NAS devices with FC storage components is now available. This configuration does not require a FC switch to support Ethernet attachment. However, it does not develop into a storage area network despite the fact that FC storage can be accessed through TCP/IP stacks and compatible file protocols supported by the NAS device. The capability for a fully functional SAN to participate within the IP network remains problematic.

This also provides a method for switch-to-switch connectivity, thus allowing remote connectivity from one SAN configuration to another. This should be given due consideration, given its ability to execute a block I/O operation through an IP network, something which impacts security, performance, and loading. We examine this in more detail when discussing SAN Connectivity in Chapter 16.

The Hardware Part

The most striking part of a storage area network is its diversity of components. NAS, as we have discussed in Part III, is almost polarized in the opposite direction given its bundled and simplistic structure. The SAN, on the other hand, must have a central network of new devices as well as supporting network appendages attached to the storage and servers it supports. This is most apparent in the required hardware of switches, HBAs, and sometimes bridges and routers. In support of our discussion on the SAN, the following overview will further familiarize you with these components.

The Fastest Storage Possible: 100MB/sec

The hardware part of the SAN is made up of components that allow a complete storage network to be enabled. The minimum devices necessary are an FC switch, an FC-enabled server, and FC-enabled storage systems. Although the FC switch could be replaced by a FC hub to facilitate the storage network, we will address it separately (in the section titled "The Unusual Configuration Club, Optical, NAS, and Mainframe Channels" later in the chapter), given its legacy position and its enhancement of switches to handle both FC fabric and arbitrated loop operations.

Figure 13-6 illustrates the minimum components necessary to configure a simple SAN. The FC switch centers the network as it connects the server and storage array. The FC server is connected through an FC Host Bus Adapter (HBA). The FC HBA provides the necessary FC protocol processing and interfaces with the server's operating system. The FC storage array is connected through an integrated FC port attachment that injects the necessary FC protocol communications into the storage controller's mechanisms.

This forms the simplest SAN configuration. However, it is the one that facilitates the FC SAN architecture and provides the basis for additional SAN configurations no matter how complex. In addition to the basic devices, Figure 13-6 also shows the type of connections required for implementation into an existing data center.

The Usual Configuration Gang, RAID, Disk Array, and Tape

Figure 13-7 takes our simple SAN configuration and places into a data center setting. As we can see, the SAN has been enhanced to include additional servers, two storage arrays both supporting RAID levels, and a tape device connected through a FC bridge. This characterizes a sample configuration that supports all the necessary data center operations including database, backup and recovery, and shared access.

The Unusual Configuration Club, Optical, NAS, and Mainframe Channels

Figure 13-8 takes our simple SAN configuration and provides some interesting twists. Although these are unusual implementation scenarios, they are important to point out because they show the increasing levels of flexibility SANs have. Note that there is FC optical storage involved, as well as connection to file storage systems through an IP

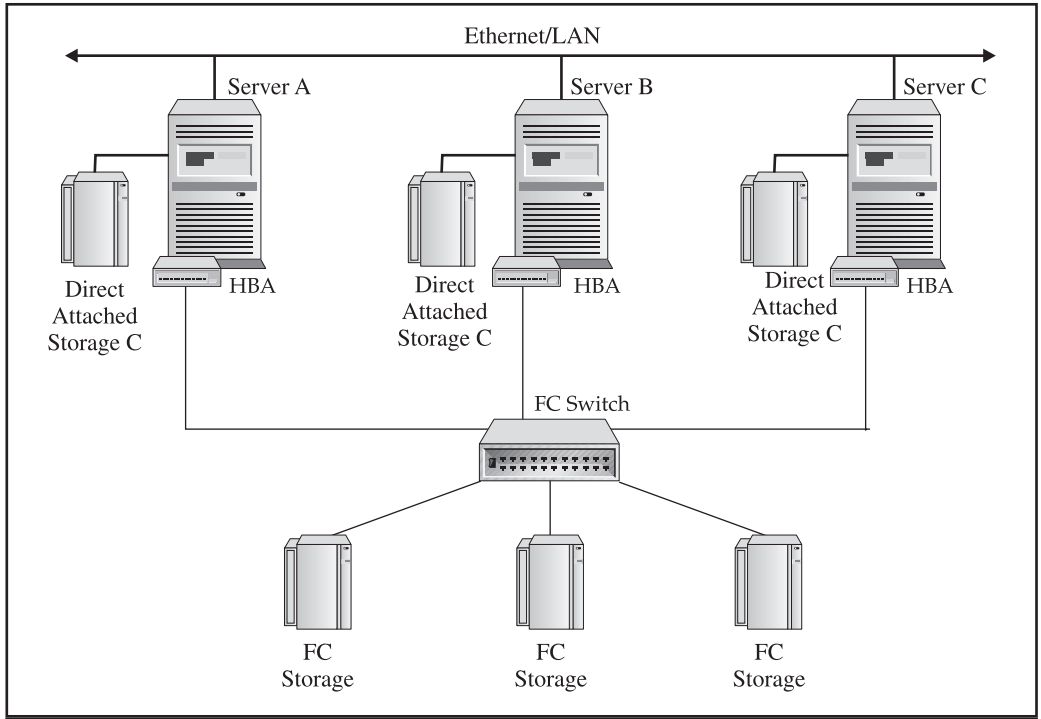

Figure 13-6. *A simple SAN architecture*

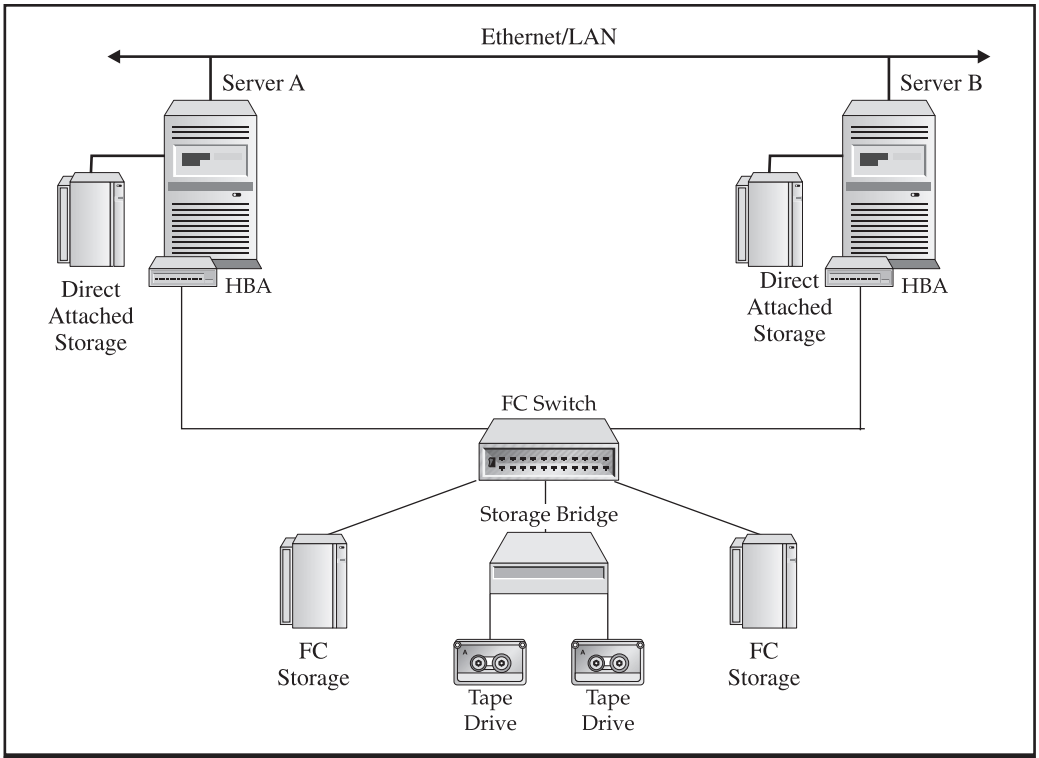

Figure 13-7. *A simple SAN configuration*

connection, and, finally, the connection to alternative servers such as mainframes supporting FC connectivity, like IBM's FICON.

The Connectivity Part

If we further analyze our simple SAN implementation, we find that switch connections come in two forms: the physical connection and the logical port configuration. The physical connection supports both optical and copper connections. Obviously, most switch products operate at optimum using the optical connections; however, the capability to support several connectivity schemes can be important when integrating into existing cabling structures.

The logical connection requires understanding the type of port configuration the switch provides. Ports can be configured as particular types of connections. These are defined according to topology and class of processing. In some switches, these are performed automatically per port, as different types of processing configurations are defined for the fabric—for instance, Classes 1, 4, and 6 for connection types of circuits, and Classes 2 and 3 for connectionless circuits or frame switching (that is, packet-switched networks).

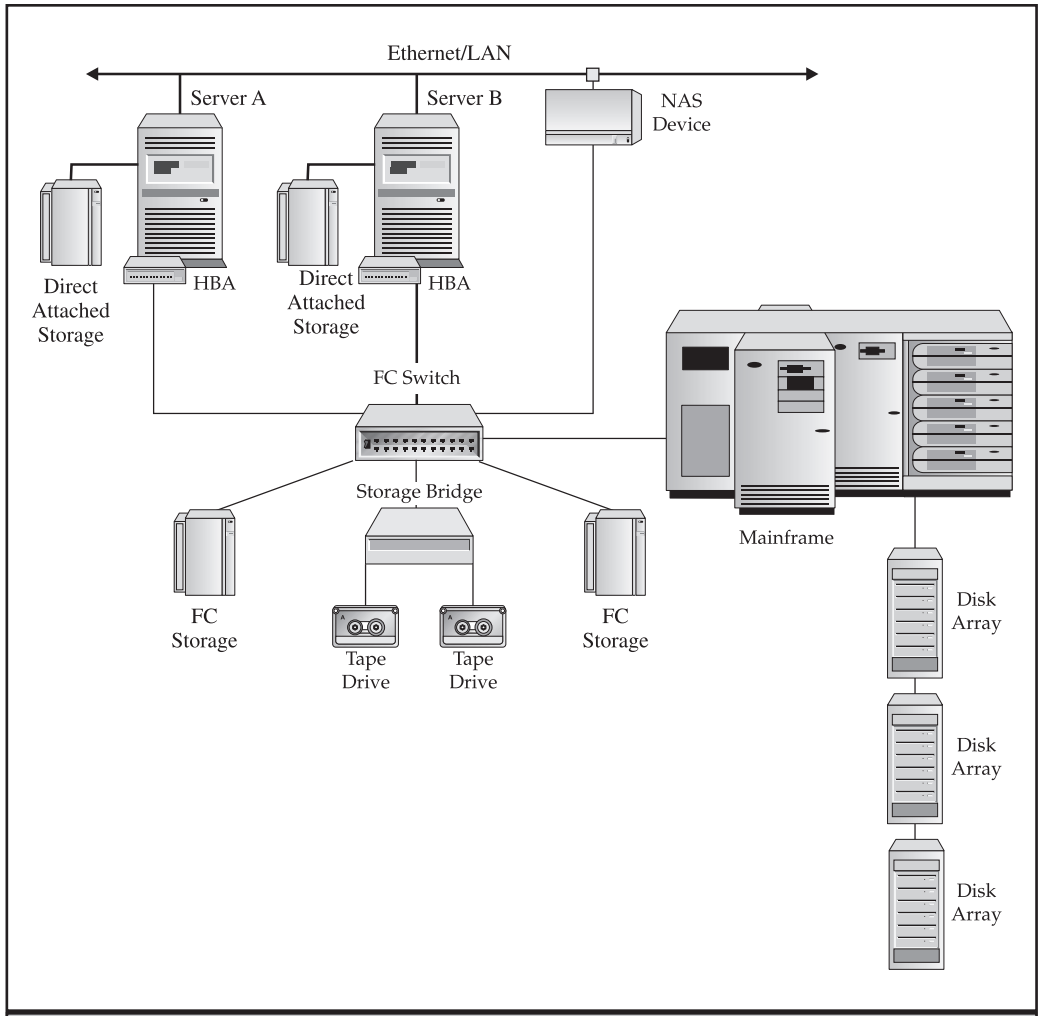

Figure 13-8. *A specialized/advanced SAN configuration*

Connecting the Server: The Host Bus Adapters

Connecting the server requires the use of a Host Bus Adapter (HBA). This is a component similar to the Network Interface Card (NIC) discussed previously in Chapter 8 as network fundamentals in storage connectivity options, and in Chapter 12. It essentially performs the same function as the NIC in terms of providing a physical connection to the FC network, using either copper or optical connectors. In addition, it provides the software functions that translate the FC protocol communications and interfaces these commands with the server operating system. This is provided (similar to NICs) with a set of drivers and is configured according to the type of FC topology the SAN is operating with.

Connecting the Storage: The SCSI/RAID Operations

One of the most effective architectural design strategies within the FC protocol and FC fabric elements is its ability to frame higher-level protocols within communications processing. This means that the SCSI commands that communicate disk and tape operations can continue to operate. These commands are packaged within the Fibre Channel frame and transmitted. Upon delivery, the FC frame is translated and the SCSI commands are reassembled for delivery and execution on the target device.

This is not only effective for server operations, which have historically interfaced with SCSI bus technologies in direct-attached disks, but also for the disk and related controllers themselves. As the SCSI commands are developed from I/O operation activities within the server OS, the I/O operations at the disk controller perform as if they continued to be directly connected. However, this begins to become complex as the reference to logical units (LUN) within a bus are tracked and configured for an FC network. This then becomes a process that grows more challenging when operating with LUNs that are dynamically modified or hidden by layer 2 or 3 processing—say, in the case of a bridge or router.

SAN Configurations

SANs evolve within data centers from simple configurations using single switch or hub operations with a limited number of disk devices. These are often limited traffic projects that provide an entry level of operations and proof of concept for targeted applications. However, they soon grow to configurations with multiple FC switches that are interconnected to optimize the reliability or availability of the applications. The following three levels of configurations best characterize the types of configurations IT will become familiar with.

Entry Level

An entry-level configuration will basically be a proof of concept or beta installation to prove the reliability of the SAN project. As stated earlier, these are often limited projects handling limited user traffic and designed to support a targeted single application. The entry-level SAN configuration can also allow IT personnel to become familiar and comfortable with the new types of hardware, give them some experience with the new software fabric configurations and operation, and let them experience real traffic within the SAN.

Figure 13-9 shows an example configuration of an entry-level SAN. Note this configuration contains a single FC switch, two storage arrays, and a tape device. There are three Windows-based servers attached to the switch which complete the SAN network. In this example, anywhere from three to five relational database systems provide production test beds for application testing. Each database has its aggregate data stored across the storage arrays. The storage arrays themselves are configured for RAID level 5 and level 1. This allows several databases to share storage in a single array; in this case, array A using RAID level 5, while array B is configured for RAID level 1.

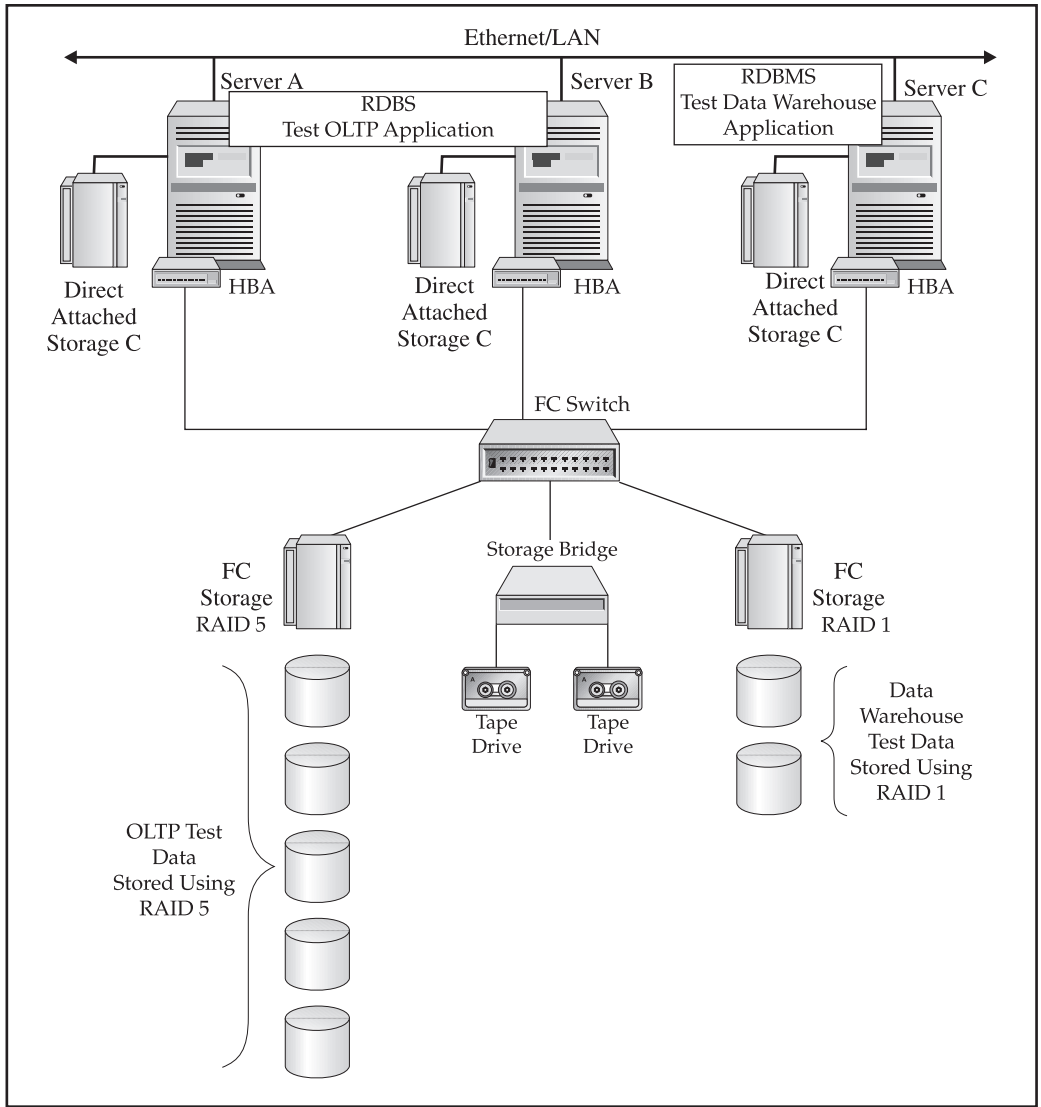

Figure 13-9. *An entry-level SAN configuration*

This configuration provides an entry-level SAN that provides a test bed for three database applications; some are for typical OLTP workloads while the third is for a data warehouse application. The systems are set to allow applications and DBAs to test new query applications against test databases. The SAN configuration replaces the six servers used to support database testing and the six storage arrays directly attached to their respective servers.

Consequently, even for an entry configuration, the architecture of the SAN can consolidate a considerable amount of hardware and software.

Mid-Range: Production Workloads, Multiple Switches

If we move up to supporting production workloads, we must consider all the necessary reliability and availability configuration requirements. Figure 13-10 shows an example SAN that supports production applications with ten servers using Windows OSs. The servers are attached to three FC switches configured in what's called a *cascading arrangement*. The multiple storage arrays support both RAID 5 and RAID 0 in our example set of applications. Our production configuration also has the addition of backup with the attachment of a tape library.

Our production workload supports several database applications that utilize five of the ten servers. These transactional OLTP-type requirements utilize the storage arrays configured for RAID 5. Our other workloads are a combination of an intranet for internal users of the company's internal web application. This application utilizes the storage arrays configured for RAID 0, which has the data stripped but has no failover capability.

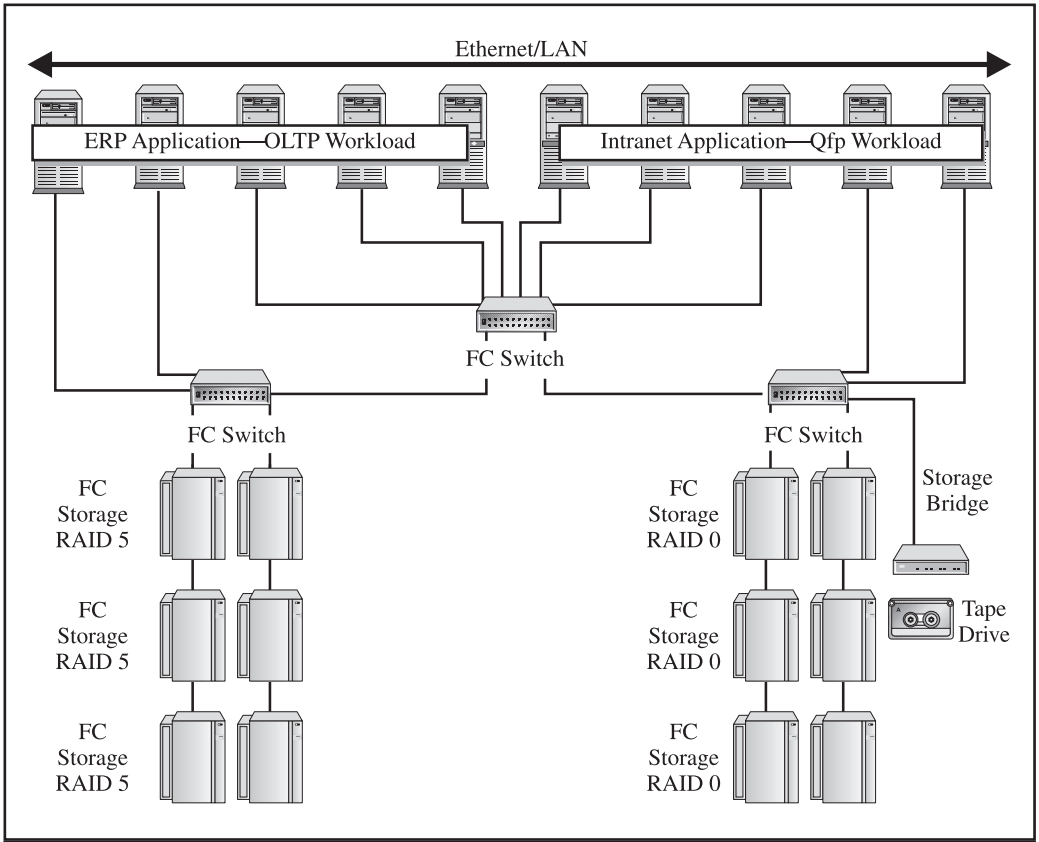

Figure 13-10. *A mid-range SAN configuration*

Enterprise: Director-Level with Duplicate Paths and Redundancy

Moving up to the support of an enterprise level and the support of multiple applications, we have an example enterprise configuration in Figure 13-11. This much larger configuration shows 20 servers, consisting of both UNIX- and Windows-based OSs, attached to six FC switches. Storage arrays support 2 terabytes of both file- and database-oriented data models, given that the workloads range from OLTP database applications to web servers, and data-centric datamart applications. A datamart, incidentally, is similar to a data warehouse with limited subject scope; in some industry sectors, however, datamarts have proliferated faster given their singular and more simplistic database design. This allows for faster design and implementation activities, which makes them demanding in their storage capacity, access and data sourcing requirements from larger data warehouses, and operational databases.

The storage arrays are configured in RAID levels 5, 1, and 0. As with any production workloads, appropriate failover has been configured with a more complex cascading switch configuration. There is a new twist, however—a remote connection within the switch that allows for a disaster recovery operation to be in place. The usual configuration of a tape library to facilitate the backup strategies of various workloads and applications is also in place.

As demonstrated here, SAN configurations have an inherent value to storage network architecture. Even though the architecture reflects a new model of processing, SAN configurations provide the most effective and efficient method of server consolidation, storage consolidation, improved data access, and scalability of data storage.

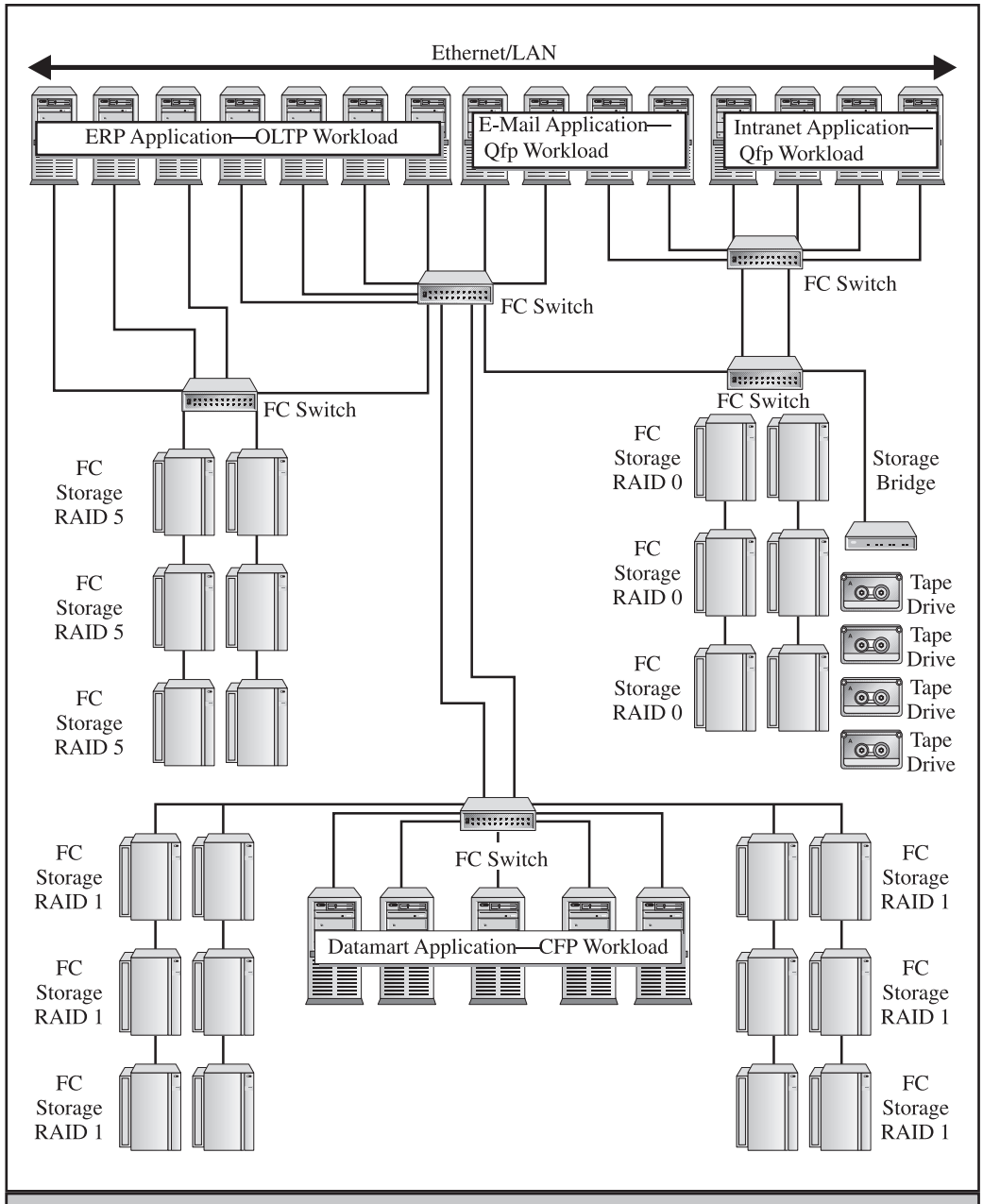

Figure 13-11. *An enterprise SAN configuration*