Chapter 4

Decoupling the Storage Component: Creating a Network for Storage

Chapter 3 summarized the architecture, configurations, and benefits as well as caveats of putting storage on an existing network using NAS storage devices. NAS can use existing network resources and excels in addressing large-scale data access read-only applications. However, it is limited by its file-oriented I/O architecture that limits this storage networking solution when addressing heavy online transactional processing (OLTP) applications.

An alternative to NAS is the other cornerstone of storage networking, the Storage Area Network (SAN). SANs, like their counterparts in NAS, allow storage to be connected to a network and provide access to multiple clients and servers. The fundamental differences are the methods in which they accomplish this. SANs require their own network to operate, which provides a significant increase to throughput. SANs also provide direct I/O access from the devices connected to the network, thus providing yet another fundamental shift in distributing the I/O workload among applications.

This chapter will discuss the fundamental concepts of the Storage Area Network as well as providing an insight into the evolution of distributing I/O operations within and throughout processing configurations. Because SANs fundamentally shift the I/O architecture of processing configurations by moving direct storage operations to a network, it's important to understand a holistic picture of how these complex technologies developed. Within this chapter is a brief discussion of the evolutionary ancestors of the SAN, most of which have not been rendered obsolete or extinct. Many of the solutions may be configurations installed within your data center.

The Data-centric World

We live in a data-centric world that consumes information at an amazingly fast pace. The information we process as individuals all starts out as data stored somewhere and, more importantly, is a requirement driven by an application that generates the information that we are presented with. Where did all these applications come from and why do they require such vast amounts of data? Certainly these are questions that are beyond the scope of this book; however, they remain the fundamental driving force behind the innovation and advancement of I/O processing.

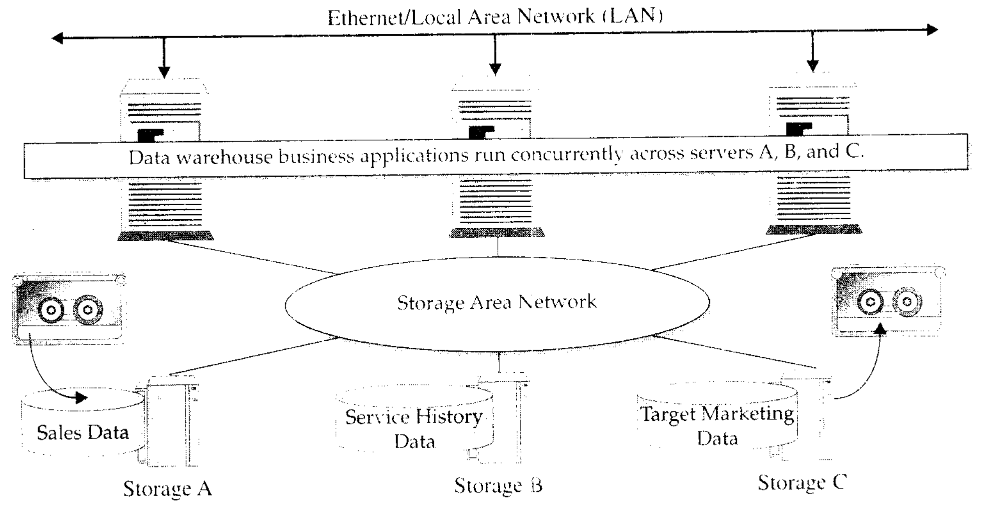
The support of the increasing data-centric nature of OLTP and data analysis application systems evolved through experiences in large database applications using centralized mainframe configurations. Even as some high-end processor configurations became distributed by connecting systems within a tightly coupled network, the ability to handle high I/O transactional rates and large-scale databases exceeded their limitations. The advancement and proliferation of relational database solutions exacerbated the problem with its exponential storage resource appetite when compared to traditional processing of the time. As this relational database phenomenon filtered into the client/server processing configurations, this increased not only the number of server configurations, but their complexities and challenges as well. It also provided another target to support the ever increasing population of data-centric applications.

Many of today's applications rely on some form of relational database products that, as discussed in Chapter 3, were designed to work directly with disk storage remote to the server. These new software functions encapsulated within the FC fabric (some of them user accessible and some micro-code enabled) drive the next logical caveat: management.

Management of the SAN remains problematic even though new products and solutions continue to appear. However, this scenario of rapid change, due to the newness of the technology, largely renders only management solutions, which in the end will be merely transient given that SANs will continue to outgrow their simple configurations, add significant new functions from software vendors, and become integrated with other network and server technologies.

On the other hand, SANs provides the next generation of scalable solutions for large-scale enterprise applications. The SAN architecture supports environments that require scalable solutions for large user populations that access datacentric (that is, database-driven) information that has read/write volatility. This has proven itself within the enterprise data centers where SANs are the solution of choice given their capability to handle large amounts of data, using applications such as OLTP and Data Warehousing. In addition, customers who are dealing with large volumes of unstructured data (such as audio, video, and static images) are migrating to SANs to support these applications.

In discussing both the NAS and SAN architectures in summary fashion, a more thorough picture of the components that make them up will come to light. How these technologies operate, as well as how to apply their value to solutions within data center settings, will be discussed in greater detail in Parts V and VI.


Ethernet/Local Area Network (LAN)
Data warehouse business applications run concurrently across servers A, B, and C.
Storage Area Network
Sales Data
Service History Data
Target Marketing Data
Storage A
Storage B
Storage C

Data operations for copy, update, backup, and recovery can move to and from offline media directly to target media sources reducing the impact and downtime associated with these activities. Very Large Databases (VLDB) are shared with SAN attached servers.

Figure 4-10. *SANs for a data warehouse environment*

The Caveats for SAN

The caveats for SAN are several. The main issues of complexity and cost center around the heterogeneous nature of the SAN itself. Because SANs are configured with several discrete components (see Figure 4-5), the complexity of configuration and implementation becomes a challenge.

Cost becomes an issue as the SAN devices remain fairly new to storage infrastructure markets and therefore have not reached commodity-pricing status. It's also important to note the rapid change due to the relative immaturity of the products, and the reliance on other areas of technologies such as Operating Systems, storage devices, and management software. The associated cost in terms of user knowledge and training in new technologies that make up the SAN—such as switches, fabrics, Fibre Channel, and HBAs—contribute to this caveat.

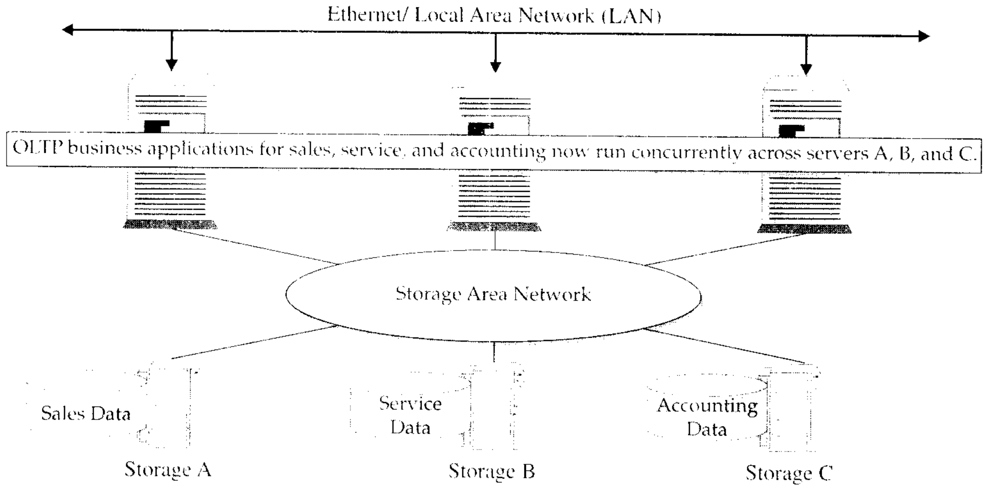
However, an often-overlooked challenge is the esoteric nature of the logical extensions of a server's software that deals with storage. File systems, volume managers, and physical I/O operations all play necessary roles in performing storage operations. Within the SAN, these operations become more logical and have to coexist with other servers that share the fabric network and devices connected. I/O functions must extend into the fabric itself and operate through Fibre Channel software, storage software, and device micro-code, all being

SANs for Data Access

As more paths to data become available, so does the theoretical limit of users accessing that data. SANs have the capability to increase the number of paths to the data by way of the number of actual I/O paths, or through the actual transfer of data to the application and, subsequently, the user. If we add up the FC's capability to operate at gigabit speeds, the SANs' capacity to support a large user base becomes apparent. An example is depicted in Figure 4-9.

SANs for Data Size

The best example for dealing with size is data warehouses. As described in Chapter 1, these datacentric applications have extreme data storage requirements in terms of the number of bytes of online storage. What makes these applications particularly difficult to configure is the complexity involved in processing the data and the movement of large subsets of data to source and process complex transactions. Figure 4-10 shows the use of SANs to enhance a particular data warehouse application in terms of the size of databases involved. This also pertains to SANs data movement activities, which are accomplished while maintaining a service level.


Ethernet/ Local Area Network (LAN)
OLTP business applications for sales, service, and accounting now run concurrently across servers A, B, and C.
Storage Area Network
Sales Data
Service Data
Accounting Data
Storage A
Storage B
Storage C

Business application data is consolidated into single production databases for sales, service, and accounting applications, respectively. Data is protected via mirroring or RAID technologies that provide Fault Tolerance.

Figure 4-9. *SANs for an OLTP database environment*

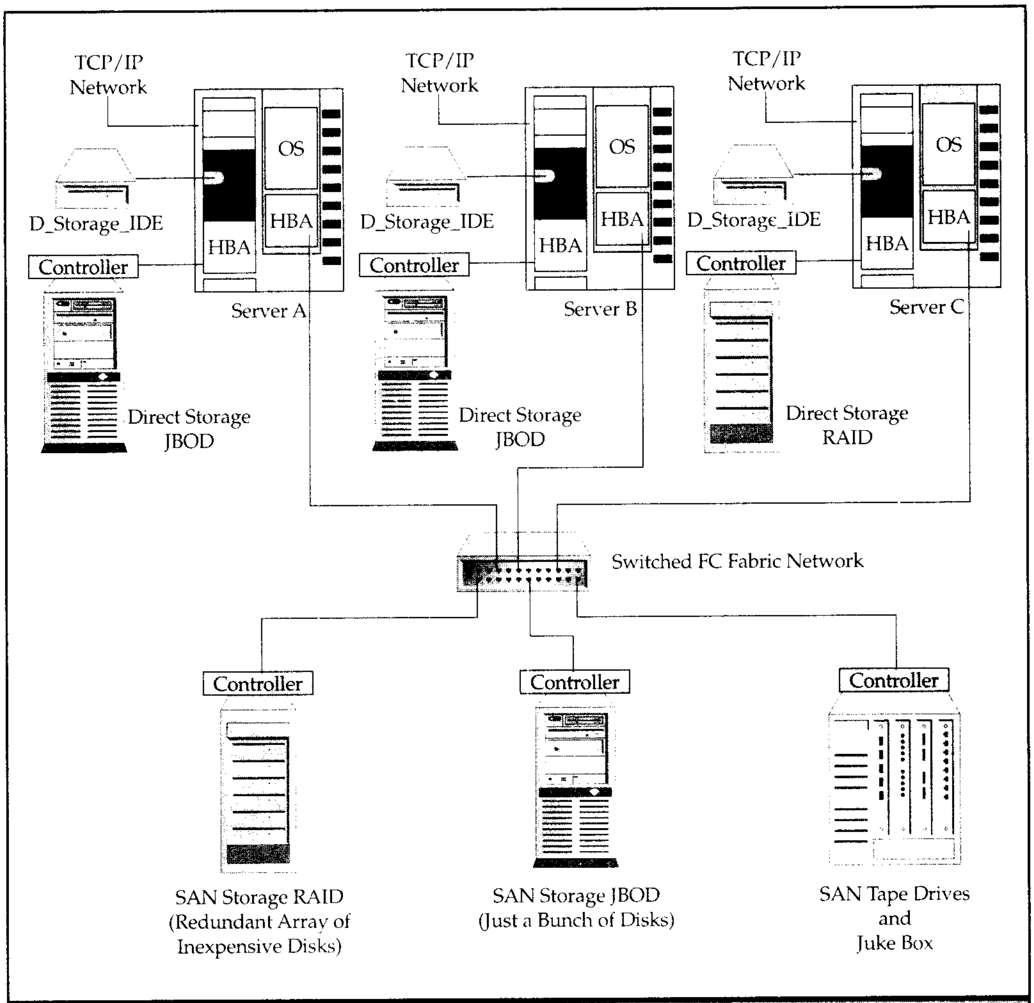


Figure 4-8. *Current SAN configurations using fabric architectures*

storage network drivers that allow the server to communicate with the switch and ultimately log in and communicate with storage devices.

Finally, storage devices used with the fabric must be FC-compliant devices (that is, they must speak FC to communicate with the network). As with Ethernet networks, early users of SANs required the use of bridges and routers to allow traditional storage devices (such as tape drives) to participate within the fabric. These devices translated FC protocols to SCSI bus level protocols, basically breaking frames down into bus segments so a SCSI-level device (such as a tape drive) could be connected to the FC fabric network.

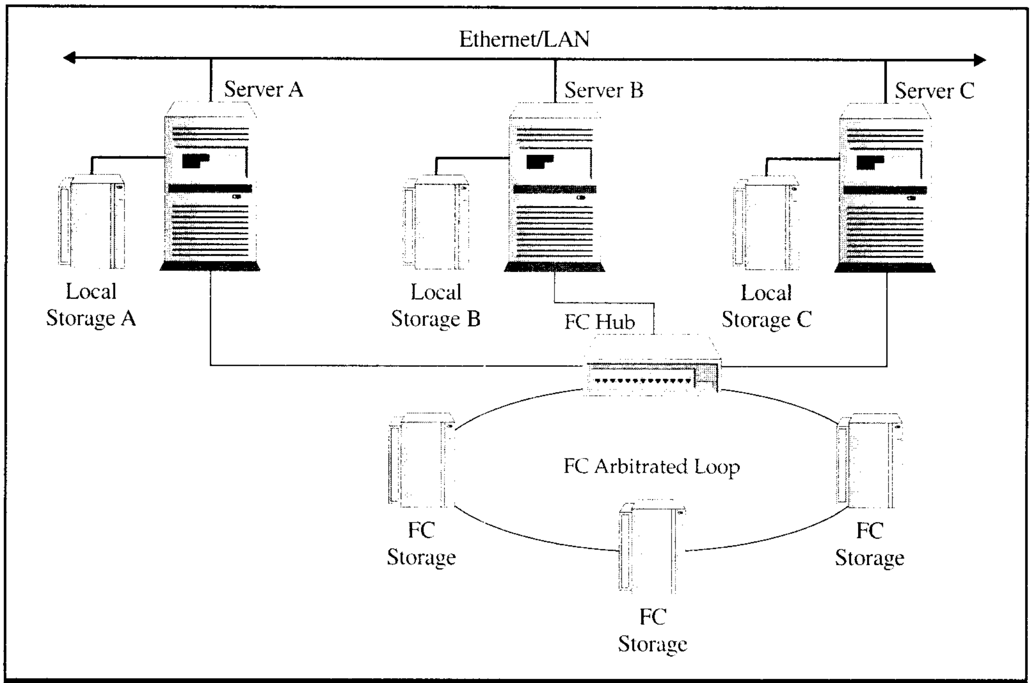


Figure 4-7. *Early SAN configurations using HUB architectures*

The most current state of SAN configuration is the switched fabric network, as shown in Figure 4-8. This configuration allows "any-to-any" device communications to take place within the network. Each device participating within the network gets full use of 100 MBps bandwidth without sacrificing its performance as other devices are added. Minimal overhead to moving FC frames through the network is accomplished through an FC network switch.

Switched Network allowed a significant increase in the number of devices connected with the network. This allowed for servers and storage devices to be connected in a network fashion and communicate bidirectionally in an "any-to-any" manner.

An Operational Overview of SAN

SANs are constructed with many new components from the storage network. The foundation is the FC switch, which provides the physical connection that allows "any-to-any" communications within the fabric. SAN switches provide the hardware and software foundations needed to facilitate the network—the hardware itself being composed of ports that permit the connection of FC-based devices, such as storage arrays and servers.

To participate in the storage network servers requires a special host adapter, similar to a network adapter known as an FC Host Bus Adapter (HBA). The HBAs supply the

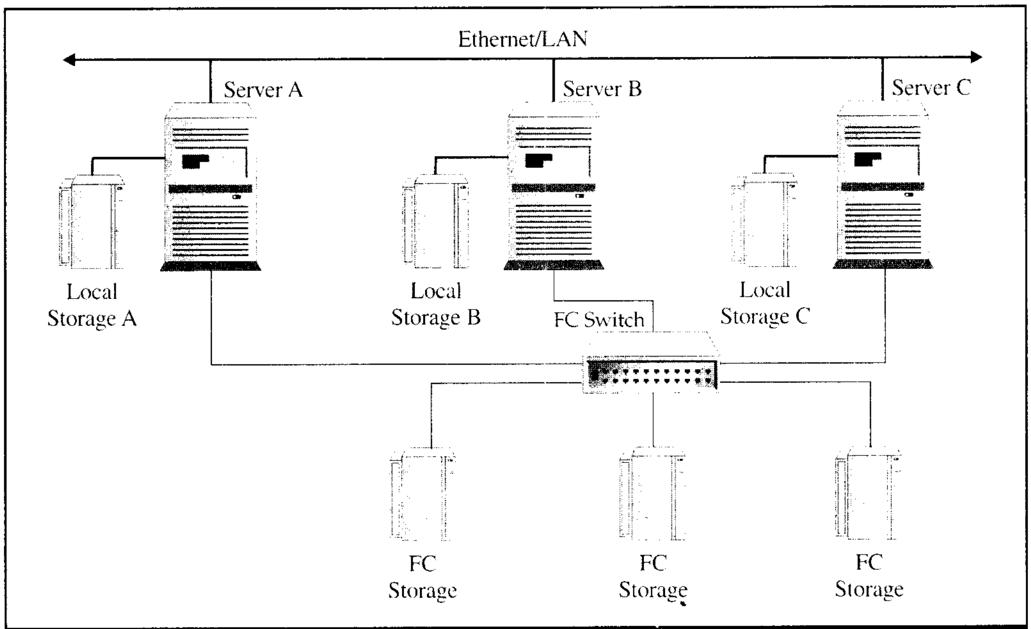


Figure 4-6. *A storage area network*

SAN Evolution and Development

The first stage of SAN evolution was the development of Fibre Channel (FC)-based storage devices used with direct connections, or a point-to-point configuration. These storage devices used FC addressing to process requests to and from the connected devices, presumably a storage array attached to a server. Although this enhanced the bandwidth of transferred data, FC devices had the capacity to handle 100 MBps, and were limited to the two devices within the point-to-point configuration. Although this worked well and used the existing device communications of SCSI commands, it proved restrictive in regards to the growth of large storage infrastructures.

As shown in Figure 4-7, the second stage employed arbitrated loop architectures, letting more devices participate through the use of FC Hubs. This allowed storage devices to continue operating within an arbitration scheme to share bandwidth, but allowed additional devices to participate within a network fashion. Although this worked better than point-to-point solutions, the overhead necessary to mediate the Hub, adding additional devices within the loops required sharing the bandwidth and thus sacrificed much of the high-speed efficiencies gained with FC devices.

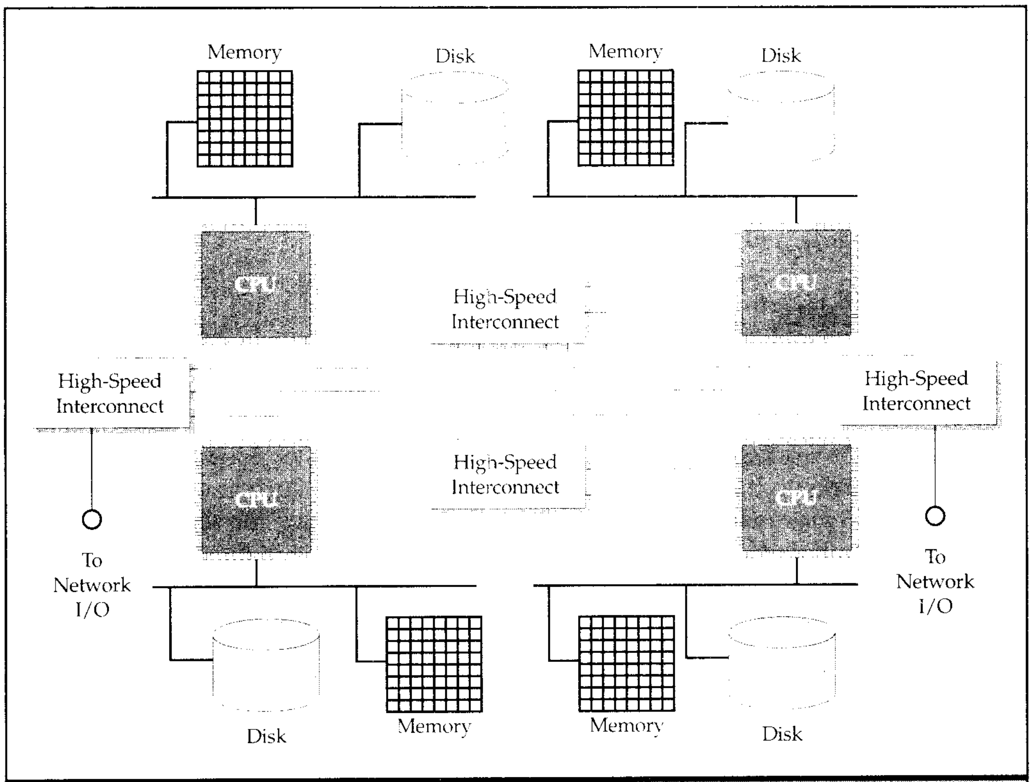


Figure 4-5. *"Shared Nothing" Architecture*

Channel, and the capability for multiple application and storage processing nodes to communicate and share resources.

The SAN Idea

Direct connect storage in the client/server storage model proved to have limitations in bandwidth, the maximum number of devices available on a bus, and concurrency of traffic. Attaching application servers and Fibre Channel storage devices to a central location creates a network of server and storage devices. This configuration allows the application servers to access the storage devices in a network fashion. In other words, the devices, both servers and storage, must have an address on the network, have the capability to log in, and have methods available to communicate with other devices within the network.

As depicted in Figure 4-6, servers and storage devices participate in the network. All devices log in to the network, or fabric OS, which is done by sending messages through Fibre Channel protocol frames.

A High-Speed Interconnect—Switched Fabric Network

FC operates on a serial link design and uses a packet type of approach for encapsulation of the user data. FC transmits and receives these data packets through the node participants within the fabric. Figure 4-4 shows a simplified view of the FC fabric transmission from server node to storage node. The packets shipped by FC are called frames and are made up of header information, including addresses, the user data at an incredible 2,048 bytes, 2k bytes per frame, and ERC information.

"Shared Nothing" Architecture

As we discussed earlier in Chapter 2, these systems form the foundation for Massively Parallel Processing (MPP) systems. Each node is connected to a high-speed interconnect and communicates with all nodes within the system (see Figure 4-5). These self-contained computer nodes work together, or in parallel, on a particular workload. These systems generally have nodes that specialize in particular operations, such as database query parsing and preprocessing for input services. Other nodes share the search for data within a database that is distributed among nodes specializing in data access and ownership. The sophistication of these machines sometimes outweighs their effectiveness, given that they require a multi-image operating system, for example an OS on each node, sophisticated database, and storage functions to partition the data throughout the configuration, and finally, the speed, latency, and throughput of the interconnect. In these systems, both workload input processing and data acquisition can be performed in parallel, providing significant throughput increases.

Each of these seemingly disparate technologies evolved separately: fabrics coming from developments in the network industry, Fibre Channel resulting from work on scientific device interconnects, and "shared nothing" architectures arising from parallel processing advancements and developments in the VLDB technologies. Directing these technologies toward storage formed an entirely new network that provides all the benefits of a fabric, the enhanced performance of frame-level protocols within Fibre

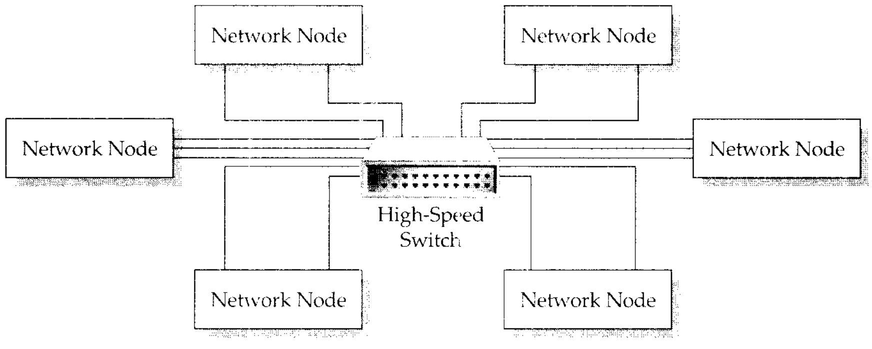


Figure 4-4. *Switched Fabric Network*

(Additional discussions of the SMP and MPP models of processing and I/O systems can be found in Part II.)

The processing advancements driven by SMP and MPP technologies began to set the foundation for advanced storage architectures. Storage provided the focal point to integrate several innovations such as an enhanced I/O channel protocol, a high-speed interconnection, and "shared nothing" architecture to support the ever increasing need for data access resulting from exponential increases in data size.

An Enhanced I/O Protocol—Fibre Channel

Fibre Channel is a layered connectivity standard, as illustrated in Figure 4-3. It demonstrates the channel characteristics of an I/O bus, the flexibility of a network, and the scalability potential of MIMD computer architectures.

Note *MIMD stands for Multi-Instruction Multi-Data. It defines a computer design taxonomy in which multiple instructions can be executed against multiple data sources. Although synonymous with MPP computer configurations, the MIMD operations can also be supported by SMP configurations where application tasks are processing in parallel by multiple CPUs operating on multiple data sources. Because MPP configurations are made up of discrete computers, their I/O scalability is much greater than SMP configurations that must share I/O operations across a single bus configuration.*

In implementation, fibre channel uses a serial connectivity scheme that allows for the highest-level bandwidth of any connectivity solution available today, 10gigbit. This architecture allows for implementations to reach as high as 200MB/sec burst rate for I/O operations, with aggregate rates depending on workload and network latency considerations. Regardless of the latency issues, this is a tremendous enhancement to throughput over traditional bus connectivity.

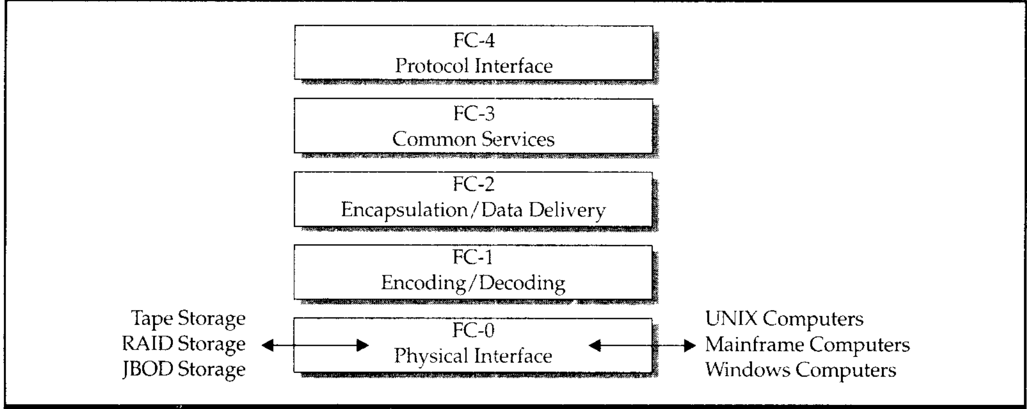


Figure 4-3. *Fibre Channel Protocol*

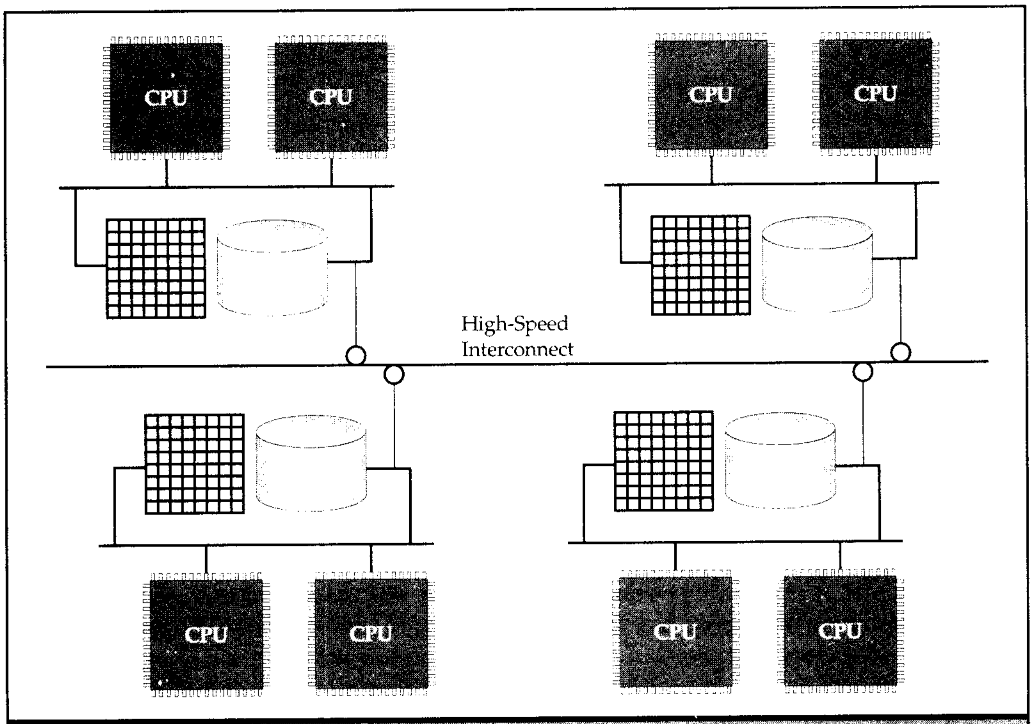


Figure 4-2. MPP processing

Even with the potential that MPP architectures demonstrated, the complexities have proved far greater and limited the proliferation of these machines as general-purpose servers. The complicated enhancements to operating systems communications to enable the coordination of multiple single-image operating systems has proved to be problematic and costly. In addition, the overhead required as increased computing nodes are added increases in a non-linear fashion and consequently limits the effectiveness and throughput. The size, operational complexities, and management challenges of MPP configurations have limited their usage to specialized applications.

Distributing I/O Processing

As advancements were being made in SMP and MPP architectures, storage technologies remained tightly coupled with their computer counterparts. They continued to lag behind processing advancements and followed an I/O execution model of direct attachment to the server regardless of whether it was connected to SMP configurations or MPP processing nodes. However, some advancement was made with shared I/O as evidenced in particular MPP configurations. Sometimes considered the predecessor to the SAN, the "shared nothing" model can be viewed as a logical extension to the MPP parallel environments.

processing capabilities by providing additional CPU components. The machines that tightly integrated these features with the operating system (MVS) were known by their design terms of dyadic, triadic, and quadratic, depending on the number of CPUs available on the machine. This further enhanced the processing of IBM mainframe configurations with their available loosely clustered systems of MVS-JES2 and little-known tightly coupled clustered systems of MVS_JES3 (see the Note in this section).

This type of architecture has become widely available through all major system vendors as the SMP designs provided an entry into increased processing power necessary to handle larger applications. However, as with IBM mainframe configurations that are integrated with MVS features and functions, all SMP systems are closely integrated with operating systems features. This is necessary to handle the additional activities of processing with more than a single CPU, sharing memory resources and spaces, and coordination within a shared I/O bus. Although they provide additional processing capacities for large-scale applications, they also bring their own limitations. Two are very apparent in Figure 4-1: the sharing of system RAM and the scalability of the I/O system.

Note *Multiple Virtual Systems (MVS) is a proprietary operating system offered by IBM and used exclusively with its mainframe computer systems. Available since the middle 1970s, MVS's longevity can partly be attributed to its modular architecture. This allows software subsystems to specialize in particular areas of operation. MVS was early in its move to separate I/O functions into a separate subsystem. In addition, it offers proprietary subsystems of Job Entry Subsystems (providing two alternatives, JES2 and JES3) that provide, among many other functions, the inter-systems communications necessary for a cluster of mainframes to share processing workloads. IBM installations are moving to zOS, an enhanced version of MVS that supports POSIX compliance and open systems standard functionality such as TCP/IP access and web software services.*

Massive Parallel Processing Systems (MPP)

Another advancement from the traditional client/server model was the innovation of Massively Parallel Processing Systems (MPP). These systems enabled multiple computer systems, each specializing in a particular aspect of an application process communicating through a high-speed link. The ability to apply parallel tasks to complex applications provided the processing throughput necessary to complete what was once considered impossible. The MPP architecture evolved into two categories of machines: one for process intensive applications and one for database intensive applications. Each machine category was differentiated by its high-speed link architecture, its integrated database functionality, and configuration flexibility.

The links could be network or switched based (or in some cases a hybrid of both). This link provided a communications network that enabled each node to work on individual processing tasks, while controlling its own systems resources, as illustrated in Figure 4-2. This delineates the node structure within the MPP system whereby each computing node does not share computer resources such as CPU, RAM, or local I/O, thus acquiring the name "Shared Nothing" configurations.

devices. In addition to this evolution of processing, combining large user populations with Very Large Data Bases (VLDB) requires a greater sophistication of I/O and storage functionality than what is available on traditional, albeit large-scale, mainframe, client/server storage systems, or NAS.

The effects of the high end OLTP and data-centric data warehouse applications accelerated the advancement of I/O architectures within high-end server systems and mainframes. These also spawned the development of increasingly complex distributed processing configurations to support the growth of I/O operations driven by business requirements that called for increasing amounts of data to be available online. These advancements have taken the form of Symmetric Multiprocessing Systems (SMP) and Massive Parallel Processing (MPP) systems. These architectures advanced the I/O capabilities and functionality of traditional computer systems, and as a lasting effect set a foundation for the development of storage area networking.

Distributing Computer Processing

The enhancement to traditional systems architectures by computer designers in meeting the requirements of data-centric applications was to distribute the I/O processing. The efforts in research and design moved I/O further away from the computer elements, (CPU, RAM, and bus connectivity components), which in turn provided an opportunity to increase the number of I/O paths to storage devices.

Symmetric Multiprocessing Systems (SMP)

Symmetric Multiprocessing Systems (SMP) became one of the first alternative designs to offer the use of more than one CPU component, as indicated in Figure 4-1. This first became available with IBM mainframes to increase the throughput scalability and

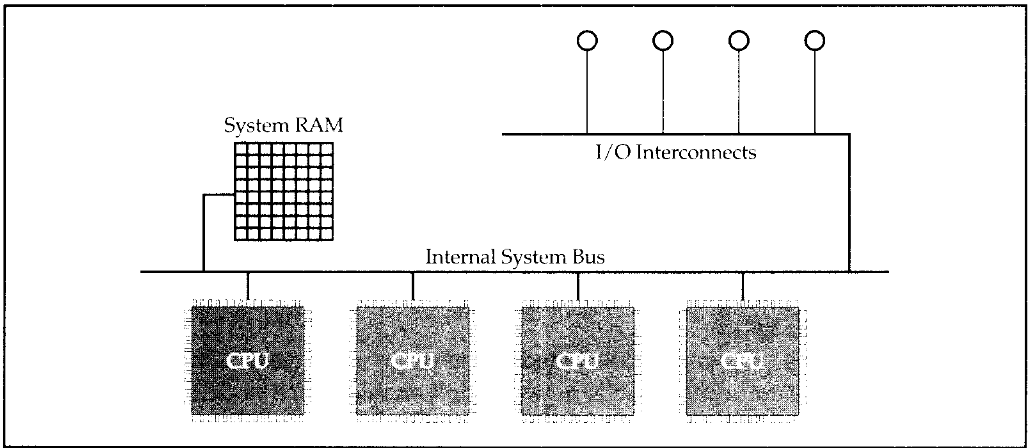


Figure 4-1. *SMP processing*