NEAR EAST UNIVERSITY

Faculty of Engineering

Department of Computer Engineering

Storage Area Network

Graduation Project

COM 400

Student: Ayman Ghannam (20000915)

Supervisor: Mr. Jamal Fathi

Nicosia - 2005

ACKNOWLEDGMENTS

The work in this project was done under the supervision of Dr. Jamal Fathi, to whom I am grateful for his support, his interest in the progress of the project, and for his insightful and critical comments.

I am also wish to thank My best friend Mr. Murad hassan, he is Engineer in cyprus, who gave me his ever devotion and all valuable information which I really needed to complete my project.

I am also thankful to Mr. walid odtalla,. Al_kayed has helped me through many helpful and enjoyable discussions.

Also Thanks to all my friends which they support me in Cyprus.

Further I am thankful to Near East University academic staff and all those persons who helped me or encouraged me for the completion of my project. Thanks!

Finally, my thanks go to whom my love will never end, to my father and my mother, to my brothers and sisters, that helped me a lot and gave their lasting encouragement in my studies, so that I could be successful in my life time.

ABSTRACT

As we now appear to have safely navigated the sea that was the transition from one century to the next, the focus today is not on preventing or avoiding a potential disaster, but exploiting current technology. There is a storm on the storage horizon. Some may call it a SAN-storm that is approaching.

Storage Area Networks have lit up the storage world like nothing before it. SANs offer the ability to move data at astonishingly high speeds in a dedicated information management network. It is this dedicated network that provides the promise to alleviate the burden placed on the corporate network in this e-world.

Traditional networks, like LANs and WANs, which have long been the workhorses of information movement are becoming tired with the amount of load that is placed upon them, and usually slow down just when you want them to go faster. SANs offer the thoroughbred solution. More importantly, an IBM SAN solution offers the pedigree and bloodlines which have been proven in the most competitive of arenas.

TABLE OF CONTENTS

1. INTRODUCTION

1.1 Introduction to Storage Area Networks

Everyone working in the Information Technology industry is familiar with the continuous developments in technology, which constantly deliver improvements in performance, capacity, size, functionality and so on. A few of these developments have far reaching implications because they enable applications or functions which allow us fundamentally to rethink the way we do things and go about our everyday business. The advent of Storage Area Networks (SANs) is one such development. SANs can lead to a proverbial "paradigm shift" in the way we organize and use the IT infrastructure of an enterprise. In the chapter that follows, we show the market forces that have driven the need for a new storage infrastructure, coupled with the benefits that a SAN brings to the enterprise.

1.2 The need for a new storage infrastructure

The 1990's witnessed a major shift away from the traditional mainframe, host-centric model of computing to the client/server model. Today, many organizations have hundreds, even thousands, of distributed servers and client systems installed throughout the enterprise. Many of these systems are powerful computers, with more processing capability than many mainframe computers had only a few years ago. Storage, for the most part, is directly connected by a dedicated channel to the server it supports. Frequently the servers are interconnected using local area networks (LAN) and wide area networks (WAN), to communicate and exchange data. This is illustrated in Figure 1.1. The amount of disk storage capacity attached to such systems has grown exponentially in recent years. It is commonplace for a desktop Personal Computer today to have 5 or 10 Gigabytes, and single disk drives with up to 75 GB are available. There has been a move to disk arrays, comprising a number of disk drives. The arrays may be "just a bunch of disks" (JBOD), or various implementations of redundant arrays of independent disks (RAID). The capacity of such arrays may be measured in tens or hundreds of GBs, but I/O bandwidth has not kept pace with the rapid growth in processor speeds and disk capacities.

Distributed clients and servers are frequently chosen to meet specific application needs. They may, therefore, run different operating systems (such as Windows NT, UNIX of differing flavors, Novell Netware, VMS and so on), and different database software (for example, DB2, Oracle, Informix, SQL 4 Designing an IBM Storage Area Network Server). Consequently, they have different file systems and different data formats.

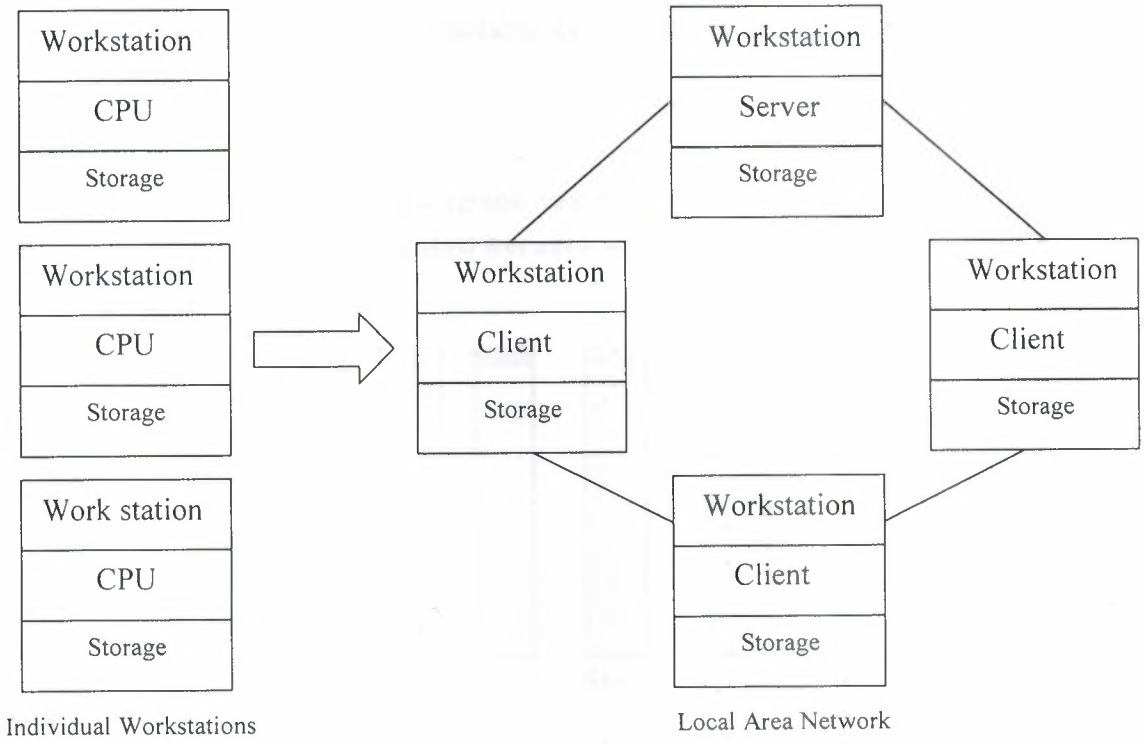

Figure 1.1 Servers are interconnected using Local Area Networks (LAN) and Wide Area Networks (WAN).

Typical distributed systems or client server infrastructure managing this multi-platform, multi-vendor, networked environment has become increasingly complex and costly. Multiple vendor's software tools, and appropriately-skilled human resources must be maintained to handle data and storage resource management on the many differing systems in the enterprise. Surveys published by industry analysts consistently show that management costs associated with distributed storage are much greater, up to 10 times

more, than the cost of managing consolidated or centralized storage. This includes costs of backup, recovery, space management, performance management and disaster recovery planning Disk storage is often purchased from the processor vendor as an integral feature, and it is difficult to establish if the price you pay per gigabyte (GB) is competitive, compared to the market price of disk storage. Disks and tapedrives, directlyattached to one client or server, cannot be used by other systems, leading to inefficient use of hardware resources. Organizations often find that they have to purchase more storage capacity, even though free capacity is available, but is attached to other platforms. This is illustrated in Figure 1.2.

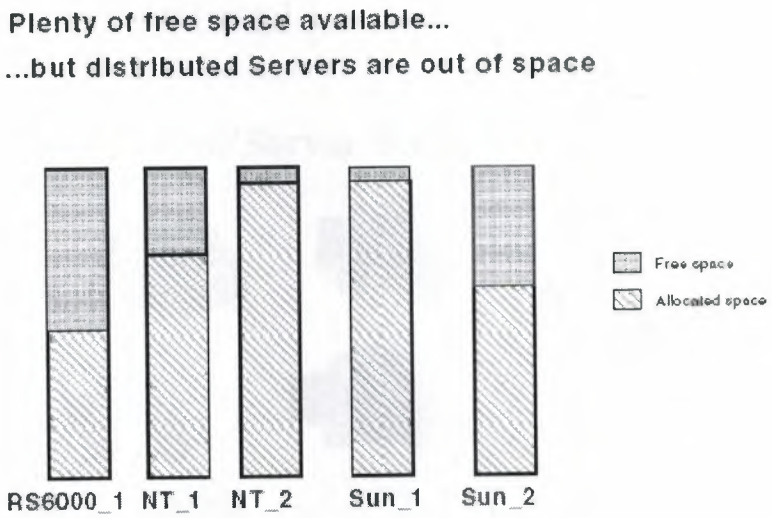

Figure 1.2. Inefficient Use of Available Disk to Individual Capacity Attached Servers

Additionally, it is difficult to scale capacity and performance to meet rapidly changing requirements, such as the explosive growth in e-business applications. Data stored on one system cannot readily be made available to other users, except by creating duplicate copies, and moving the copy to storage that is attached to another server. Movement of large files of data may result in significant degradation of performance of the LAN/WAN, causing conflicts with mission critical applications. Multiple copies of the same data may lead to inconsistencies between one copy and another. Data spread on multiple small systems is difficult to coordinate and share for enterprise-wide

applications, such as e-business, Enterprise Resource Planning (ERP), Data Warehouse, and Business Intelligence (BI). Backup and recovery operations across a LAN may also cause serious disruption to normal application traffic. Even using fast Gigabit Ethernet Transport, sustained throughput from a single server to tape is about 25 GB per hour. It would take approximately 12 hours to fully backup a relatively moderate departmental database of 300 GBs. This may exceed the available window of time in which this must be completed, and it may not be a practical solution if business operations span multiple time zones. It is increasingly evident to IT managers that these characteristics of client/server computing are too costly, and too inefficient. The islands of information resulting from the distributed model of computing do not match the needs of the ebusiness enterprise. We show this in Figure 1.3.

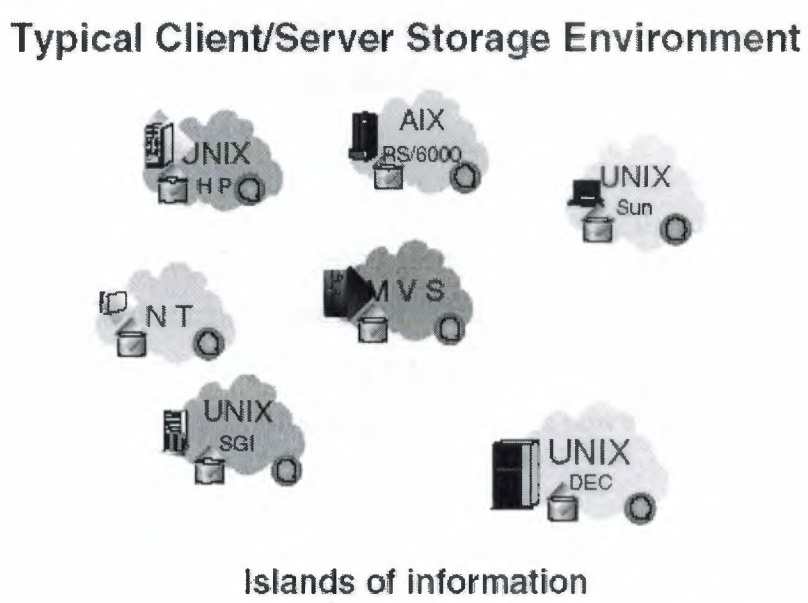

Figure 1.3 Distributed Computing Models Tends To Create Islands Of Information.

New ways must be found to control costs, to improve efficiency, and to properly align the storage infrastructure to meet the requirements of the business. One of the first steps to improved control of computing resources throughout the enterprise is improved

connectivity. In the topics that follow, we look at the advantages and disadvantages of the standard storage infrastructure of today.

1.3 The Small Computer Systems Interface legacy

The Small Computer Systems Interface (SCSI) is the conventional, server centric method of connecting peripheral devices (disks, tapes and printers) in the open client/server environment, as its name indicates, it was designed for the PC and small computer environment.

It is a b**Client/** standard storage infrastructure of today. Figure 1.3 Distributed computing model tends to create islands of information New ways must be found to control costs, to improve efficiency, and to properly align the storage infrastructure to meet the requirements of the business. One of the first steps to improved control of computing resources throughout the enterprise is improved connectivity. In the topics that follow, we look at the advantages and disadvantages of the standard storage infrastructure of today.

1.4 Storage network solutions

Today's enterprise IT planners need to link many users of multi-vendor, heterogeneous systems to multi-vendor shared storage resources, and they need to allow those users to access common data, wherever it is located in the enterprise. These requirements imply a network solution, and two types of network storage solutions are now available:

- Network attached storage (NAS)
- Storage Area Network (SAN)

1.4.1 What network attached storage is

NAS solutions utilize the LAN in front of the server, and transmit data over the LAN using messaging protocols, such as TCP/IP and Net BIOS. We illustrate this in Figure1.4

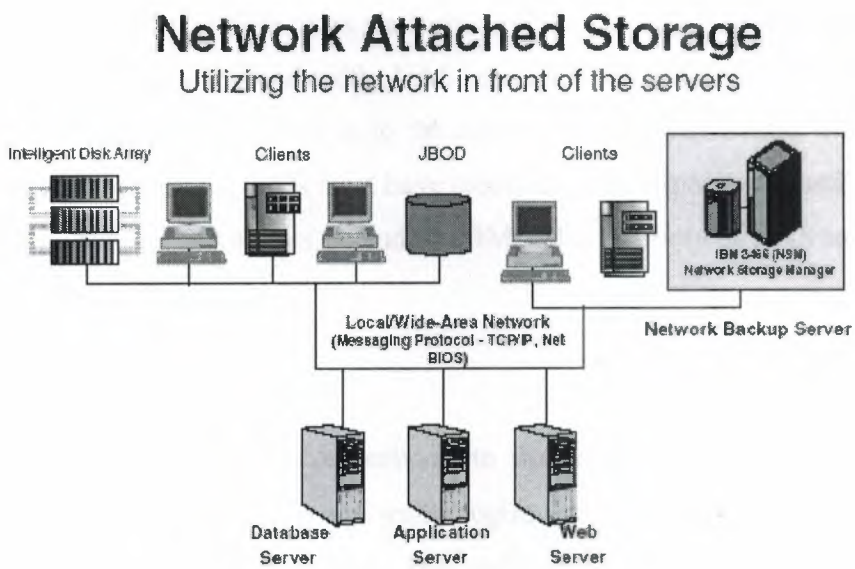

Figure 1.4. Network Attached Storage - Utilizing the Network In Front of The Servers.

Figure 1.4 Network attached storage - utilizing the network in front of the servers By making storage devices LAN addressable, the storage is freed from its direct attachment to a specific server. In principle, any user running any operating system can address the storage device by means of a common access protocol, for example, Network File System (NFS). In addition, a task, such as back-up to tape, can be performed across the LAN, enabling sharing of expensive hardware resources between multiple servers. Most storage devices cannot just attach to a LAN. NAS solutions are specialized file servers

which are designed for this type of attachment. NAS, therefore, offers a number of benefits, which address some of the limitations of parallel SCSI. However, by moving storage transactions, such as disk accesses, and tasks, such as backup and recovery of files, to the LAN, conflicts can occur with end user traffic on the network. LANs are tuned to favor short burst transmissions for rapid response to messaging requests, rather than large continuous data transmissions. Significant overhead can be imposed to move large blocks of data over the LAN, due to the small packet size used by messaging protocols. For instance, the maximum packet size for Ethernet is about 1500 bytes. A 10 MB file has to be segmented into more than 7000 individual packets, (each sent separately by the LAN access method), if it is to be read from a NAS device. Therefore, a NAS solution is best suited to handle cross platform direct access applications, not to deal with applications requiring high bandwidth. NAS solutions are relatively low cost, and straightforward to implement as they fit in to the existing LAN environment, which is a mature technology. However, the LAN must have plenty of spare capacity to justify NAS implementations. A number of vendors, including IBM, offer a variety of NAS solutions. These fall into two categories:

- File servers
- Backup/archive servers

However, it is not the purpose of this redbook to discuss these. NAS can be used separately or together with a SAN, as the technologies are complementary. In general terms, NAS offers lower cost solutions, but with more limited benefits, lower performance and less scalability, than Fibre Channel SANs.

1.4.2 What a Storage Area Network is

A SAN is a specialized, high speed network attaching servers and storage devices. It is sometimes called "the network behind the servers". A SAN allows "any to any" connection across the network, using interconnect elements such as routers, gateways, hubs and switches. It eliminates the traditional dedicated connection between a server and storage, and the concept that the server effectively "owns and manages" the storage devices. It also eliminates any restriction to the amount of data that a server can access, currently limited by the number of storage devices, which can be attached to the

individual server. Instead, a SAN introduces the flexibility of networking to enable one server or many heterogeneous servers to share a common storage "utility", which may comprise many storage devices, including disk, tape, and optical storage. And, the storage utility may be located far from the servers which use it. We show what the network behind the servers may look like, in Figure 1.5

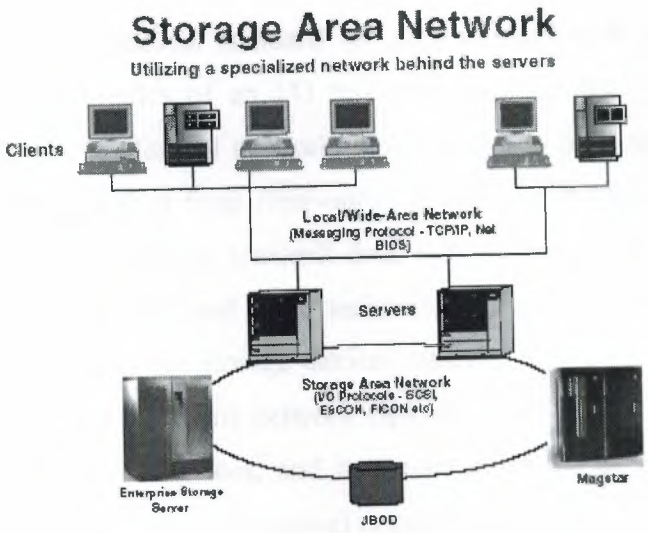

Figure 1.5. Storage Area Network - The Network behind the Servers.

A SAN differs from traditional networks, because it is constructed from storage interfaces. SAN solutions utilize a dedicated network behind the servers, based primarily (though, not necessarily) on Fibre Channel architecture. Fibre Channel provides a highly scalable bandwidth over long distances, and with the ability to provide full redundancy, including switched, parallel data paths to deliver high availability and high performance. Therefore, a SAN can bypass traditional network bottlenecks. It supports direct, high speed transfers between servers and storage devices in the following ways:

- **Server to storage.** This is the traditional method of interaction with storage devices. The SAN advantage is that the same storage device may be accessed serially or concurrently by multiple servers.
- **Server to server.** This is high speed, high volume communications between servers.

- **Storage to storage.** For example, a disk array could backup its data direct to tape across the SAN, without processor intervention. Or, a device could be mirrored remotely across the SAN. A SAN changes the server centric model of the typical open systems IT infrastructure, replacing it with a data centric infrastructure.

1.5 What Fibre Channel is

Fibre Channel is an open, technical standard for networking which incorporates the "channel transport" characteristics of an I/O bus, with the flexible connectivity and distance characteristics of a traditional network. Notice the European spelling of Fibre, which is intended to distinguish it from fiber-optics and fiber-optic cabling, which are physical hardware and media used to transmit data at high speed over long distances using light emitting diode (LED) and laser technology. Because of its channel-like qualities, hosts and applications see storage devices attached to the SAN as if they are locally attached storage. Because of its network characteristics it can support multiple protocols and a broad range of devices, and it can be managed as a network. Fibre Channel can use either optical fiber (for distance) or copper cable links (for short distance at low cost). Fibre Channel is a multi-layered network, based on a series of American National Standards Institute (ANSI) standards which define characteristics and functions for moving data across the network. These include definitions of physical interfaces, such as cabling, distances and signaling; data encoding and link controls; data delivery in terms of frames, flow control and classes of service; common services; and protocol interfaces.

Like other networks, information is sent in structured packets or frames, and data is serialized before transmission. But, unlike other networks, the Fibre Channel architecture includes a significant amount of hardware processing to deliver high performance. The speed currently achieved is 100 MB per second, (with the potential for 200 MB and 400 MB and higher data rates in the future). In all Fibre Channel topologies a single transmitter sends information to a single receiver. In most multi-user implementations this requires that routing information (source and target) must be provided. Transmission is defined in the Fibre Channel standards across three transport topologies:

- **Point to point:** a bi-directional, dedicated interconnection between two nodes, with full-duplex bandwidth (100 MB/second in each direction concurrently).
- **Arbitrated loop:** a uni-directional ring topology, similar to a token ring, supporting up to 126 interconnected nodes. Each node passes data to the next node in the loop, until the data reaches the target node. All nodes share the 100 MB/second Fibre Channel bandwidth. Devices must arbitrate for access to the loop. Therefore, with 100 active devices on a loop, the effective data rate for each is 1 MB/second, which is further reduced by the overhead of arbitration. A loop may also be connected to a Fibre Channel switch port, therefore, enabling attachment of the loop to a wider switched fabric environment. In this case, the loop may support up to 126 devices. Many fewer devices are normally attached in practice, because of arbitration overheads and shared bandwidth constraints. Due to fault isolation issues inherent with arbitrated loops, most FC-AL SANs have been implemented with a maximum of two servers, plus a number of peripheral storage devices. So FC-AL is suitable for small SAN configurations, or SANlets.
- **Switched fabric:** The term Fabric describes an intelligent switching infrastructure which delivers data from any source to any destination.The interconnection of up to 224 nodes is allowed, with each node able to utilize the full 100 MB/second duplex Fibre Channel bandwidth. Each logical connection receives dedicated bandwidth, so the overall bandwidth is multiplied by the number of connections (delivering a maximum of 200 MB/second x nnodes). The fabric itself is responsible for controlling the routing of information. It may be simply a single switch, or it may comprise multiple interconnected switches which function as a single logical entity. Complex fabrics must be managed by software which can exploit SAN management functions which are built into the fabric. Switched fabric is the basis for enterprise wide SANs.

A mix of these three topologies can be implemented to meet specific needs. Fibre Channel arbitrated loop (FC-AL) and switched fabric (FC-SW) are the two most commonly used topologies, satisfying differing requirements for scalability, distance, cost and performance. A fourth topology has been developed, known as slotted loop (FC-

SL); But, this appears to have limited application, specifically in aerospace, so it is not discussed in this book. Fibre Channel uses a serial data transport scheme, similar to other computer networks, streaming packets, (frames) of bits one behind the other in a single data line. To achieve the high data rate of 100 MB/second the transmission clock frequency is currently 1 Gigabit, or 1 bit per 0.94 nanoseconds. Serial transfer, of course, does not suffer from the problem of skew, so speed and distance is not restricted as with parallel data transfers as we show in Figure 1.6.

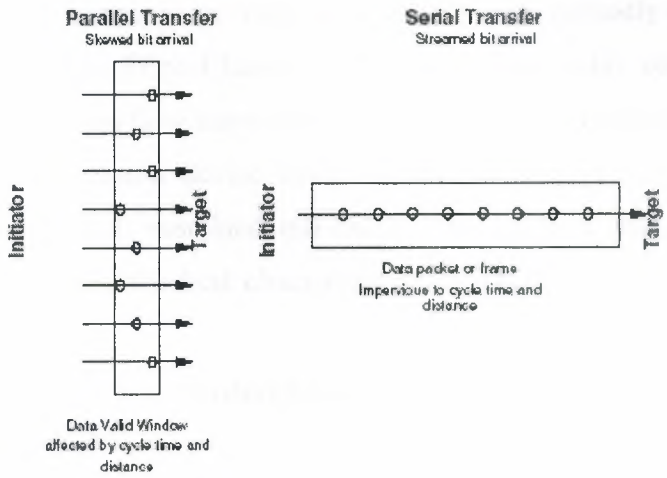

Figure 1.6. Parallel Data Transfers versus Serial Data Transfers

Serial transfer enables simpler cabling and connectors, and also routing of information through switched networks. Today, Fibre Channel can operate over distances of up to 10 km, link distances up to 90 km by implementing cascading, and longer with the introduction of repeaters. Just as LANs can be interlinked in WANs by using high speed gateways, so can campus SANs be interlinked to build enterprise wide SANs. Whatever the topology, information is sent between two nodes, which are the source (transmitter or initiator) and destination (receiver or target). A node is a device, such as a server (personal computer, workstation, or mainframe), or peripheral device, such as disk or tape drive, or video camera. Frames of information are passed between nodes, and the structure of the frame is defined by a protocol. Logically, a source and target node must utilize the same protocol, but each node may support several different protocols or data

types. Therefore, Fibre Channel architecture is extremely flexible in its potential application. Fibre Channel transport layers are protocol independent,enabling the transmission of multiple protocols. It is possible, therefore, to transport storage I/O protocols and commands, such as SCSI-3 Fibre Channel Protocol, (or FCP, the most common implementation today), ESCON, FICON, SSA, and HIPPI. Network packets may also be sent using messaging protocols, for instance, TCP/IP or Net BIOS, over the same physical interface using the same adapters, cables, switches and other infrastructure hardware. Theoretically then, multiple protocols can move concurrently over the same fabric. This capability is not in common use today, and, in any case, currently excludes concurrent FICON transport. Most Fibre Channel SAN installations today only use a single protocol. Using a credit based flow control methodology, Fibre Channel is able to deliver data as fast as the destination device buffer is able to receive it. And low transmission overheads enable high sustained utilization rates without loss of data. Therefore, Fibre Channel combines the best characteristics of traditional I/O channels with those of computer networks:

- High performance for large data transfers by using simple transport protocols and extensive hardware assists
- Serial data transmission
- A physical interface with a low error rate definition
- Reliable transmission of data with the ability to guarantee or confirm error free delivery of the data
- Packaging data in packets (frames in Fibre Channel terminology)
- Flexibility in terms of the types of information which can be transported in frames (such as data, video and audio)
- Use of existing device oriented command sets, such as SCSI and FCP
- A vast expansion in the number of devices which can be addressed when compared to I/O interfaces — a theoretical maximum of more than 16 million ports It is this high degree of flexibility, availability and scalability; the combination of multiple protocols at high speeds over long distances; and the broad acceptance of the Fibre Channel standards by vendors throughout the IT

industry, which makes the Fibre Channel architecture ideal for the development of enterprise SANs.

2. DRIVE FOR SAN INDUSTRY STANDARDIZATION

2.1 Overview

Given the strong drive towards SANs from users and vendors alike, one of the most critical success factors is the ability of systems and software from different vendors to operate together in a seamless way. Standards are the basis for the interoperability of devices and software from different vendors. A good benchmark is the level of standardization in today's LAN and WAN networks.

Standard interfaces for interoperability and management have been developed, and many vendors compete with products based on the implementation of these standards. Customers are free to mix and match components from multiple vendors to form a LAN or WAN solution. They are also free to choose from several different network management software vendors to manage their heterogeneous network.

The major vendors in the SAN industry recognize the need for standards, especially in the areas of interoperability interfaces and application programming interfaces (APIs), as these are the basis for wide acceptance of SANs. Standards will allow customers a greater breadth of choice, and will lead to the deployment of cross-platform, multi-vendor, enterprise-wide SAN solutions.

2.2 SAN industry associations and organizations

A number of industry associations, standards bodies and company groupings are involved in developing and publishing SAN standards. The major groups linked with SAN standards are shown in Figure 2.1. The roles of these associations and bodies fall into three categories:

- **Market development**— These associations are involved in market development, establishing requirements, conducting customer education, user conferences, and so on. The main organizations are the Storage Network Industry Association (SNIA); Fibre Channel Industry Association (merging the former Fibre Channel Association and the Fibre Channel Loop Community); and the SCSI Trade Association (SCSITA). Some of these organizations are also involved in the definition of defacto standards.

- **Defacto standards**— These organizations and bodies tend to be formed from two sources. They include working groups within the market development organizations, such as SNIA and FCIA. Others are partnerships between groups of companies in the industry, such as Jiro, Fibre Alliance, and the Open Standards Fabric Initiative (OSFI), which work as pressure groups towards defacto industry standards. They offer architectural definitions, write white papers, arrange technical conferences, and may reference implementations based on developments by their own partner companies. They may submit these specifications for formal standards acceptance and approval. The OSFI is a good example, comprising the five manufacturers of Fibre Channel switching products. In July 1999, they announced an initiative to accelerate the definition, finalization, and adoption of specific Fibre Channel standards that address switch interoperability.
- **Formal standards**— These are the formal standards organizations likeIETF, ANSI, and ISO, which are in place to review, obtain consensus, approve, and publish standards defined and submitted by the preceding two categories of organizations.

IBM and Tivoli Systems are heavily involved in most of these organizations, holding positions on boards of directors and technical councils and chairing projects in many key areas. We do this because it makes us aware of new work and emerging standards. The hardware and software management solutions we develop, therefore, can provide early and robust support for those standards that do emerge from the industry organizations into pervasive use. Secondly, IBM, as the innovation and technology leader in the storage industry, wants to drive reliability, availability, serviceability, and other functional features into standards. The standards organizations in which we participate are included in the following sections.

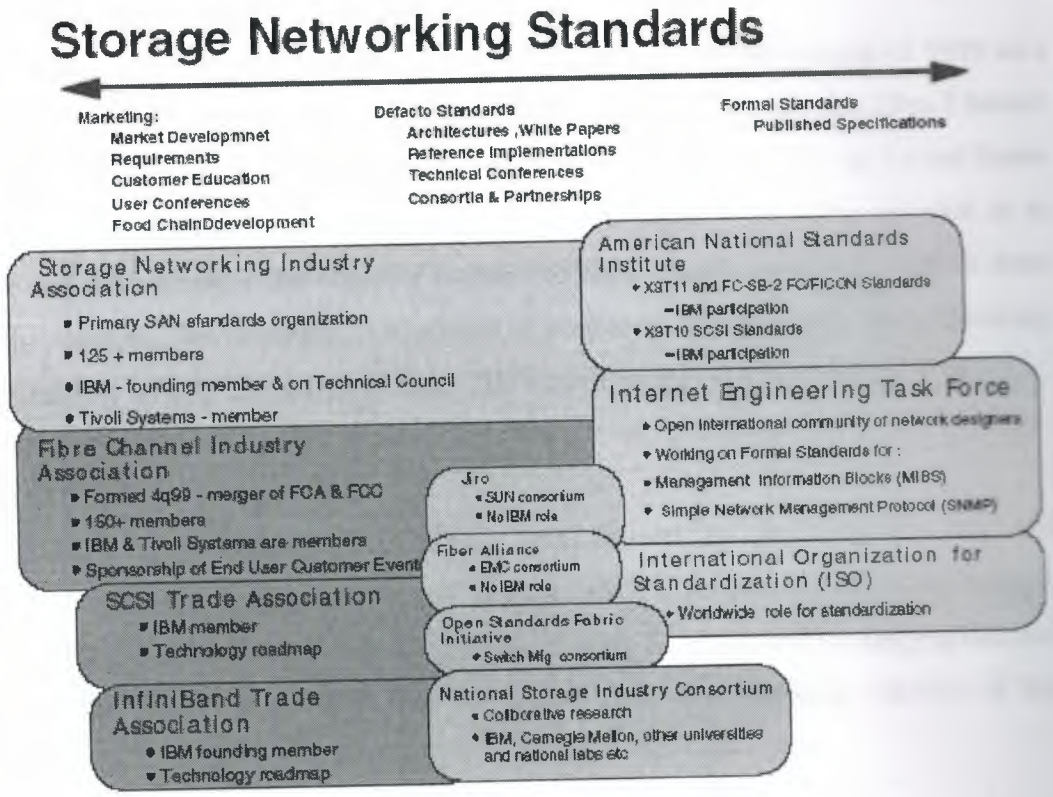

Figure 2.1 Groups Involved In Setting Storage Networking Standards.

2.2.1 Storage Networking Industry Association

Storage Networking Industry Association (SNIA) is an international computer industry forum of developers, integrators, and IT professionals who evolve and promote storage networking technology and solutions. SNIA was formed to ensure that storage networks become efficient, complete, and trusted solutions across the IT community. SNIA is accepted as the primary organization for the development of SAN standards, with over 125 companies as its members, including all the major server, storage, and fabric component vendors. SNIA also has a working group dedicated to the development of NAS standards. SNIA is committed to delivering architectures, education, and services that will propel storage networking solutions into a broader market. IBM is one of the founding members of SNIA, and has senior representatives participating on the board and in technical groups.

2.2.2 Fibre Channel Industry Association

The Fibre Channel Industry Association (FCIA) was formed in the autumn of 1999 as a result of a merger between the Fibre Channel Association (FCA) and the Fibre Channel Community (FCC). The FCIA currently has more than 150 members in the United States and through its affiliate organizations in Europe and Japan. The FCIA mission is to nurture and help develop the broadest market for fibre channel products. This is done through market development, education, standards monitoring and fostering interoperability among members' products. IBM is a principal member of the FCIA.

2.2.3 The SCSI Trade Association

The SCSI Trade Association (SCSITA) was formed to promote the use and understanding of small computer system interface (SCSI) parallel interface technology. The SCSITA provides a focal point for communicating SCSI benefits to the market, and influences the evolution of SCSI into the future. IBM is a founding member of the SCSITA.

2.2.4 InfiniBand (SM) Trade Association

The demands of the Internet and distributed computing are challenging the scalability, reliability, availability, and performance of servers. To meet this demand a balanced system architecture with equally good performance in the memory, processor, and input/output (I/O) subsystems is required. A number of leading companies have joined together to develop a new common I/O specification beyond the current PCI bus architecture, to deliver a channel based, switched fabric technology that the entire industry can adopt. InfiniBand™ Architecture represents a new approach to I/O technology and is based on the collective research, knowledge, and experience of the industry's leaders. IBM is a founding member of InfiniBand (SM) Trade Association.

2.2.5 National Storage Industry Consortium

The National Storage Industry Consortium membership consists of over fifty US corporations, universities, and national laboratories with common interests in the field of digital information storage. A number of projects are sponsored by NSIC, including

network attached storage devices (NASD), and network attached secure disks. The objective of the NASD project is to develop, explore, validate, and document the technologies required to enable the deployment and adoption of network attached devices, subsystems, and systems. IBM is a founding member of the NSIC.

2.2.6 Internet Engineering Task Force

The Internet Engineering Task Force (IETF) is a large, open international community of network designers, operators, vendors, and researchers concerned with the evolution of the Internet architecture, and the smooth operation of the Internet. It is responsible for the formal standards for the Management Information Blocks (MIB) and for Simple Network Management Protocol (SNMP) for SAN management.

2.2.7 American National Standards Institute

American National Standards Institute (ANSI) does not itself develop American national standards. It facilitates development by establishing consensus among qualified groups. IBM participates in numerous committees, including those for Fibre Channel and storage area networks.

2.3 SAN Software Management Standards

Traditionally, storage management has been the responsibility of the host server to which the storage resources are attached. With storage networks the focus has shifted away from individual server platforms, making storage management independent of the operating system, and offering the potential for greater flexibility by managing shared resources across the enterprise SAN infrastructure. Software is needed to configure, control, and monitor the SAN and all of its components in a consistent manner. Without good software tools, SANs cannot be implemented effectively.

The management challenges faced by SANs are very similar to those previously encountered by LANs and WANs. Single vendor proprietary management solutions will not satisfy customer requirements in a multi-vendor heterogeneous environment. The pressure is on the vendors to establish common methods and techniques. For instance, the

need for platform independence for management applications, to enable them to port between a variety of server platforms, has encouraged the use of Java.

The Storage Network Management Working Group (SNMWG) of SNIA is working to define and support open standards needed to address the increased management requirements imposed by SAN topologies, reliable transport of the data, as well as management of the data and resources (such as file access, backup, and volume management) are key to stable operation. SAN management requires a hierarchy of functions, from management of individual devices and components, to the network fabric, storage resources, data and applications. This is shown in Figure 2.2.

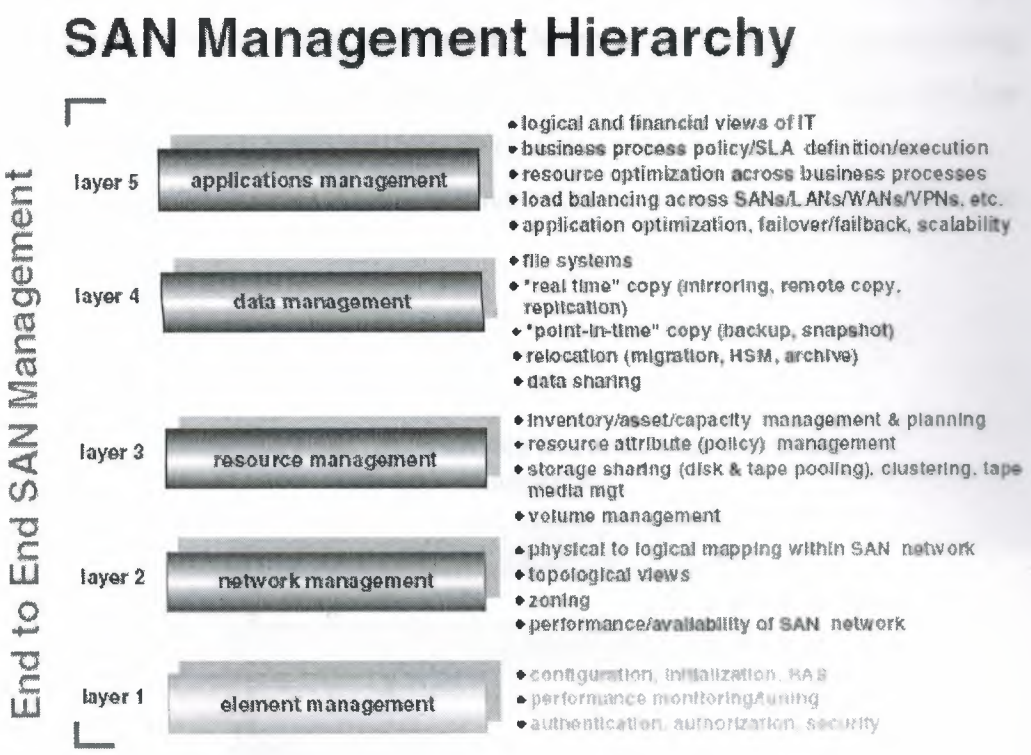

Figure 2.2 SAN Management Hierarchy.

These can be implemented separately, or potentially as a fully integrated solution to present a single interface to manage all SAN resources.

2.3.1 Application management

Application Management is concerned with the availability, performance, and recoverability of the applications that run your business. Failures in individual components are of little consequence if the application is unaffected. By the same measure, a fully functional infrastructure is of little use if it is configured incorrectly or if the data placement makes the application unusable. Enterprise application and systems management is at the top of the hierarchy and provides acomprehensive, organization-wide view of all network resources (fabric, storage, servers, applications). A flow of information regarding configuration, status, statistics, capacity utilization, performance, and so on, must be directed up the hierarchy from lower levels. A number of industry initiatives are directed at standardizing the storage specific information flow using a Common Information Model (CIM) sponsored by Microsoft, or application programming interfaces (API), such as those proposed by the Jiro initiative, sponsored by Sun Microsystems, and others by SNIA and SNMWG.

Figure 2.3 illustrates a common interface model for heterogeneous, multi-vendor SA management.

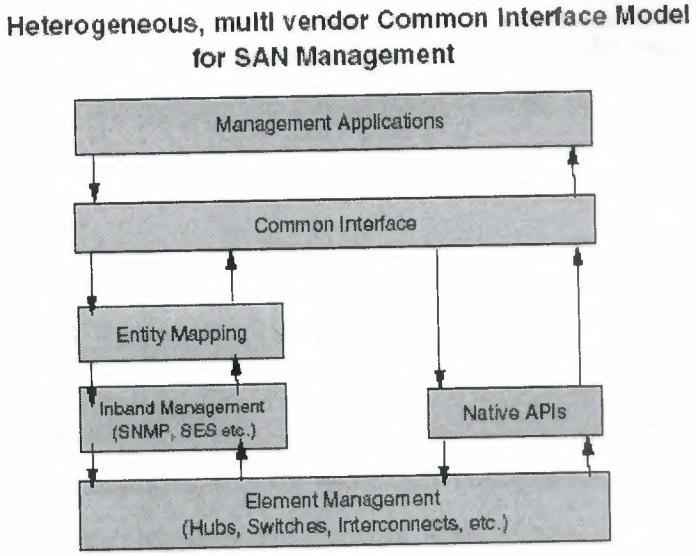

Figure 2.3 Common Interface Model For SAN Management

2.3.2 Data management

More than at any other time in history, digital data is fueling business. Data Management is concerned with Quality-of-Service (QoS) issues surrounding this data, such as:

- Ensuring data availability and accessibility for applications
- Ensuring proper performance of data for applications
- Ensuring recoverability of data

Data Management is carried out on mobile and remote storage, centralized host attached storage, network attached storage (NAS) and SAN attached storage (SAS). It incorporates backup and recovery, archive and recall, and disaster protection.

2.3.3 Resource management

Resource Management is concerned with the efficient utilization and consolidated, automated management of existing storage and fabric resources, as well as automating corrective actions where necessary. This requires the ability to manage all distributed storage resources, ideally

through a single management console, to provide a single view of enterprise resources. Without such a tool, storage administration is limited to individual servers. Typical enterprises today may have hundreds, or even thousands, of servers and storage subsystems. This makes impractical the manual consolidation of resource administration information, such as enterprise-wide disk utilization, or regarding the location of storage subsystems. SAN resource management addresses tasks, such as:

- Pooling of disk resources
- Space management
- Pooling and sharing of removable media resources
- Implementation of "just-in-time" storage

2.3.4 Network management

Every e-business depends on existing LAN and WAN connections in order to function. Because of their importance, sophisticated network management software has evolved. Now SANs are allowing us to bring the same physical connectivity concepts to storage.

And like LANs and WANs, SANs are vital to the operation of an e-business. Failures in the SAN can stop the operation of an enterprise. SANs can be viewed as both physical and logical entities.

SAN physical view

The physical view identifies the installed SAN components, and allows the physical SAN topology to be understood. A SAN environment typically consists of four major classes of components:

- End-user computers and clients
- Servers
- Storage devices and subsystems
- Interconnect components

End-user platforms and server systems are usually connected to traditional LAN and WAN networks. In addition, some end-user systems may be attached to the Fibre Channel network, and may access SAN storage devices directly. Storage subsystems are connected using the Fibre Channel network to servers, end-user platforms, and to each other. The Fibre Channel network is made up of various interconnect components, such as switches, hubs, and bridges, as shown in Figure 2.4.

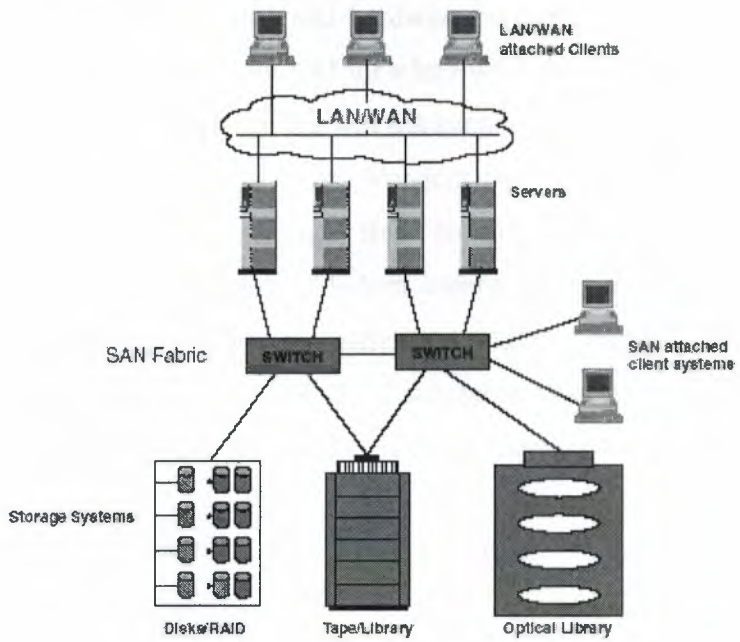

Figure 2.4 Typical SAN Environment

SAN logical view

The logical view identifies and understands the relationships between SAN entities. These relationships are not necessarily constrained by physical connectivity, and they play a fundamental role in the management of SANs.

For instance, a server and some storage devices may be classified as a logical entity. A logical entity group forms a private virtual network, or zone, within the SAN environment with a specific set of connected members.

Communication within each zone is restricted to its members. Network Management is concerned with the efficient management of the Fibre Channel SAN. This is especially in terms of physical connectivity mapping, fabric zoning, performance monitoring, error monitoring, and predictive capacity planning.

2.3.5 Element Management

The elements that make up the SAN infrastructure include intelligent disk subsystems, intelligent removable media subsystems, Fibre Channel switches, hubs and bridges, meta-data controllers, and out-board storage management controllers. The vendors of these components provide proprietary software tools to manage their individual elements, usually comprising software, firmware and hardware elements such as those shown in Figure 2.5. For instance, a management tool for a hub will provide information regarding its own configuration, status, and ports, but will not support other fabric components such as other hubs, switches, HBAs, and so on. Vendors that sell more than one element commonly provide a software package that consolidates the management and configuration of all of their elements. Modern enterprises, however, often purchase storage hardware from a number of different vendors. Fabric monitoring and management is an area where a great deal of standards work is being focused. Two management techniques are in use, inband and outband management.

2.3.5.1 Inband Management

Device communications to the network management facility is most commonly done directly across the Fibre Channel transport, using a protocol called SCSI Enclosure Services (SES). This is known as inband management. It is simple to implement, requires no LAN connections, and has inherent advantages, such as the ability for a switch to initiate a SAN topology map by means of SES queries to other fabric components. However, in the event of a failure of the Fibre Channel transport itself, the management information cannot be transmitted. Therefore, access to devices is lost, as is the ability to detect, isolate, and recover from network problems. This problem can be minimized by provision of redundant paths between devices in the fabric.

Inband management is evolving rapidly. Proposals exist for low level interfaces such as Return Node Identification (RNID) and Return Topology Identification (RTIN) to gather individual device and connection information, and for a Management Server that derives topology information. Inband management also allows attribute inquiries on storage devices and configuration changes for all elements of the SAN. Since inband management is performed over the SAN itself, administrators are not required to make additional TCP/IP connections.

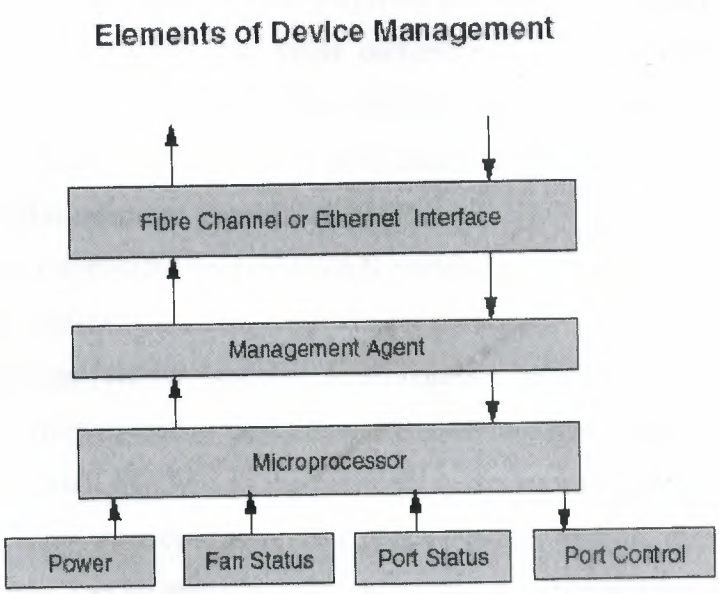

Figure 2.5 Device Management Elements

2.3.5.2 Outband Management

Outband management means that device management data are gathered over a TCP/IP connection such as Ethernet. Commands and queries can be sent using Simple Network Management Protocol (SNMP), Telnet (a text only command line interface), or a web browser Hyper Text Transfer Protocol (HTTP). Telnet and HTTP implementations are more suited to small networks.

Outband management does not rely on the Fibre Channel network. Its main advantage is that management commands and messages can be sent even if a loop or fabric link fails. Integrated SAN management facilities are more easily implemented, especially by using SNMP. However, unlike inband management, it cannot automatically provide SAN topology mapping.

(a)Outband developments

Two primary SNMP MIBs are being implemented for SAN fabric elements that allow outband monitoring. The ANSI Fibre Channel Fabric Element MIB provides significant operational and configuration information on individual devices. The emerging Fibre Channel Management MIB provides additional link table and switch zoning information that can be used to derive information about the physical and logical connections between individual devices. Even with these two MIBs, outband monitoring is incomplete. Most storage devices and some fabric devices don't support outband monitoring. In addition, many administrators simply don't attach their SAN elements to the TCP/IP network.

(b)Simple Network Management Protocol (SNMP)

This protocol is widely supported by LAN/WAN routers, gateways, hubs and switches, and is the predominant protocol used for multi vendor networks.

Device status information (vendor, machine serial number, port type and status, traffic, errors, and so on) can be provided to an enterprise SNMP manager. This usually runs on a UNIX or NT workstation attached to the network. A device can generate an alert by SNMP, in the event of an error condition. The device symbol, or icon, displayed on the SNMP manager console, can be made to turn red or yellow, and messages can be sent to the network operator.

(c)Management Information Base (MIB)

A management information base (MIB) organizes the statistics provided. The MIB runs on the SNMP management workstation, and also on the managed device. A number of industry standard MIBs have been defined for the LAN/WAN environment. Special MIBs for SANs are being built by the SNIA. When these are defined and adopted, multi-vendor SANs can be managed by common commands and queries.

Element management is concerned with providing a framework to centralize and automate the management of heterogeneous elements and to align this management with application or business policy.

2.4 SAN Status Today

SANs are in the same situation in which LANs and WANs were when these technologies began to emerge in the late 1980's. SAN technology is still relatively immature. Accepted industry standards are still under development in a number of key areas. However, vendors are working together in the standards organizations described, with the intention to rapidly improve this situation. For instance, in March 2000 Brocade Communications Systems announced that it would release elements of its Silkworm Fibre Channel interconnection protocol to the Technical Committee of the primary ANSI Fibre Channel standards group. Known as Fabric Shortest Path First (FSPF), this specifies a common method for routing and moving data among Fibre SANs are in the same situation in which LANs and WANs were when these technologies began to emerge in the late 1980's. SAN technology is still relatively immature. Accepted industry standards are still under development in a number of key areas. However, vendors are working together in the standards organizations described, with the intention to rapidly improve this situation. For instance, in March 2000 Brocade Communications Systems announced that it would release elements of its Silkworm Fibre Channel interconnection protocol to the Technical Committee of the primary ANSI Fibre Channel standards group. Known as Fabric Shortest Path First (FSPF), this specifies a common method for routing and moving data among Fibre Channel switches.

As a result the situation is fluid and changing quickly. What may be impractical today may be ready for prime time next week, next month, or next year. But, you can be confident that the industry standards initiatives will deliver effective cross platform solutions within the near term.

Many of the SAN solutions on the market today are restricted to specific applications. Interoperability is also often restricted, and currently available software management tools are limited in scope. But these considerations need not prevent you from actively planning and implementing SANs now.

They do mean that you need to take care in selecting solutions. You should try to ensure that your choices are not taking you in a direction which could be a dead end route, or locking you in to limited options for the future.

3. FIBRE CHANNEL BASICS

3.1 Overview

Fibre Channel (FC) is a technology standard that allows data to be transferred from one network node to another at very high speeds. Fibre Channel is simply the most reliable, highest performing solution for information storage, transfer, and retrieval available today. Current implementations transfer data at 100 MB/second, although, 200 MB/second and 400 MB/second data rates have already been tested.

This standard is backed by a consortium of industry vendors and has been accredited by the American National Standards Institute (ANSI). Many products are now on the market that take advantage of FC's high-speed, high-availability characteristics. In the topics that follow, we introduce Fibre Channel basic information to complement the solutions that we describe later in this redbook. We cover areas that are internal to Fibre Channel and show how data is moved and the medium upon which it travels.

3.2 SAN components

The industry considers Fibre Channel as the architecture on which most SAN implementations will be built, with FICON as the standard protocol for S/390 systems, and Fibre Channel Protocol (FCP) as the standard protocol for non-S/390 systems.

Based on this implementation, there are three main categories of SAN components:

- SAN servers
- SAN storage
- SAN interconnects

We show the typical SAN components that are likely to be encountered in Figure 3.1.

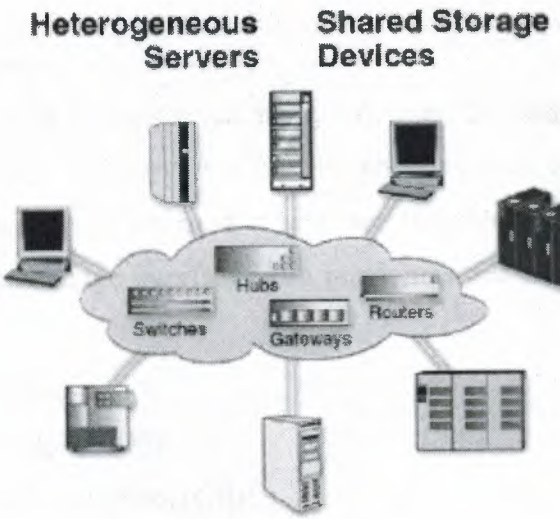

Figure 3.1. SAN Components

3.2.1 SAN servers

The server infrastructure is the underlying reason for all SAN solutions. This infrastructure includes a mix of server platforms, such as Windows NT, UNIX and its various flavors, and mainframes. With initiatives, such as server consolidation and e-business, the need for a SAN has become very strong.

Although most current SAN solutions are based on a homogeneous server platform, future implementations will take into account the heterogeneous nature of the IT world.

3.2.2 SAN storage

The storage infrastructure is the foundation on which information relies, and must support the business objectives and business model. In this environment, simply deploying more and faster storage devices is not enough; a new kind of infrastructure is needed, one that provides network availability, data accessibility, and system manageability. The SAN meets this challenge. It is a high-speed subnet that establishes a direct connection between storage resources and servers. The SAN liberates the storage device, so it is not on a particular server bus, and attaches it directly to the network. In other words, storage is externalized, and functionally distributed to the organization. The SAN also enables the centralization of storage devices and the clustering of servers, which makes for easier and less expensive administration.

3.2.3 SAN interconnects

The first element that must be considered in any SAN implementation is the connectivity of components of storage and servers using technologies such as Fibre Channel. The components listed here are typically used in LAN and WAN implementations. SANs, like LANs, interconnect the storage interfaces into many network configurations and across long distances.

- Cables and connectors
- Gigabit Link Model (GLM)
- Gigabit Interface Converters (GBIC)
- Media Interface Adapters (MIA)
- Adapters
- Extenders
- Multiplexers
- Hubs
- Routers
- Bridges
- Gateways
- Switches
- ESCON Directors
- FICON Directors.

3.3 Jargon terminology shift

Much of the terminology used for SAN has its origin in Internet Protocol (IP) network terminology. In some cases, companies in the industry use different terms that mean the same thing, and in some cases, the same terms are meaning different things. In this book we will attempt to define some of the terminology that is used and its changing nature among vendors.

3.4 Vendor standards and main vendors

This section gives an overview of the major SAN vendors in the industry:

- **Systems/storage SAN providers**

  IBM (Sequent), SUN, HP, EMC (DG Clariion), STK, HDS, Compaq, and Dell

- **Hub providers**

  Gadroon, Vixel and Emulex

- **Switch providers**

  Brocade, Ancor, McDATA, Vixel, STK/SND and Gadzoox

- **Gateway and Router providers**

  ATTO, Chaparrel Tech, CrossRoads Tech, Pathlight, Vicom

- **Host bus adapters (HBA) providers**

  Ancon, Compaq, Emulex, Genroco, Hewlett-Packard, Interphase, Jaycor Networks, Prisia, Qlogic and Sun Microsystems.

- **Software providers**

  IBM/Tivoli, Veritas, Legato, Computer Associates, DataDirect, Transoft (HP), Crosstor and Retrieve.

3.5 Physical characteristics

This section describes the components and technology associated with the physical aspects of Fibre Channel. We describe the supported cables and give an overview of the types of connectors that are generally available and are implemented in a SAN environment.

3.5.1 Cable

As with parallel SCSI and traditional networking, different types of cable are used for Fibre Channel configurations. Two types of cables are supported:

- Copper
- Fiber-optic

Fibre Channel can be run over optical or copper media, but fiber-optic enjoys a major advantage in noise immunity. It is for this reason that fiber-optic cabling is preferred. However, copper is also widely used and it is likely that in the short term a mixed environment will need to be tolerated and supported. Figure 3.2 shows fiber-optical data transmission.

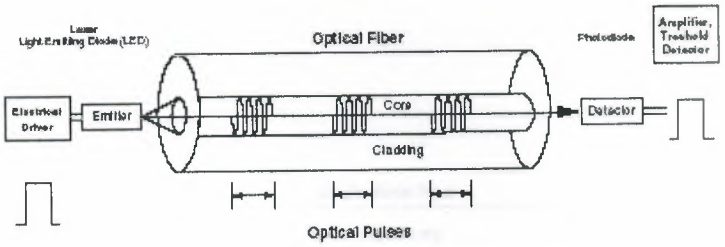

Figure 3.2 Fiber Optical Data Transmission

In addition to the noise immunity, fiber-optic cabling provides a number of distinct advantages over copper transmission lines that make it a very attractive medium for many applications.

At the forefronts of the advantages are:

- Greater distance capability than is generally possible with copper
- Insensitive to induced electro-magnetic interference (EMI)
- No emitted electro-magnetic radiation (RFI)
- No electrical connection between two ports
- Not susceptible to crosstalk
- Compact and lightweight cables and connectors

However, fiber-optic and optical links do have some drawbacks. Some of the considerations are:

- Optical links tend to be more expensive than copper links over short distances

- Optical connections don't lend themselves to backplane printed circuit wiring
- Optical connections may be affected by dirt and other contamination

Overall, optical fibers have provided a very high-performance transmission medium which has been refined and proven over many years.

Mixing fiber-optical and copper components in the same environment is supported, although not all products provide that flexibility and this should be taken into consideration when planning a SAN. Copper cables tend to be used for short distances, up to 30 meters, and can be identified by their DB-9, 9 pin, connector.

Normally fiber-optic cabling is referred to by mode or the frequencies of light waves that are carried by particular cable type. Fiber cables come in two distinct types, as shown in Figure 3.3.

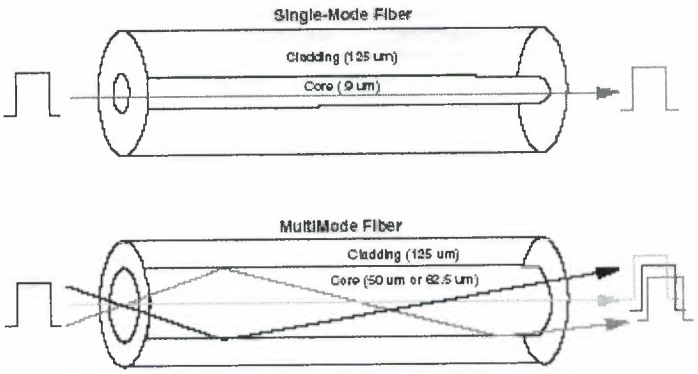

Figure 3.3. Multi-mode and single-mode propagation

- **Multi-mode fiber (MMF)** for short distances, up to 500m using FCP Multi-mode cabling is used with shortwave laser light and has either a 50 micron or a 62.5 micron core with a cladding of 125 micron. The 50 micron or 62.5 micron diameter is sufficiently large for injected light waves to be reflected off the core interior.
- **Single-mode fiber (SMF)** for long distances Single-mode is used to carry longwave laser light. With a much smaller 9 micron diameter core and a single-

mode light source, single-mode fiber supports much longer distances, currently up to 10 km at gigabit speed.

Fibre Channel architecture supports both short wave and long wave optical transmitter technologies, as follows:

- **Short wave laser** — this technology uses a wavelength of 780 nanometers and is only compatible with multi-mode fiber.
- **Long wave laser** — this technology uses a wavelength of 1300 nanometers. It is compatible with both single-mode and multi-mode fiber.

IBM will support the following distances for FCP as shown in Table 1.

Table 3.1. FCP distance.

Diameter (Microns)	Cladding (micron)	Mode	Laser type	Distance
9	125	Single mode	Longwave	=< 10 km
50	125	Multi mode	Shortwave	<= 500 m
62.5	125	Multimode	Shortwave	<= 175 m

Campus

A campus topology is nothing more than "cabling" buildings together, so that data can be transferred from a computer system in one building to storage devices, whether they are disk storage, or tape storage for backup, or other devices in another building. We show a campus topology in Figure 3.4.

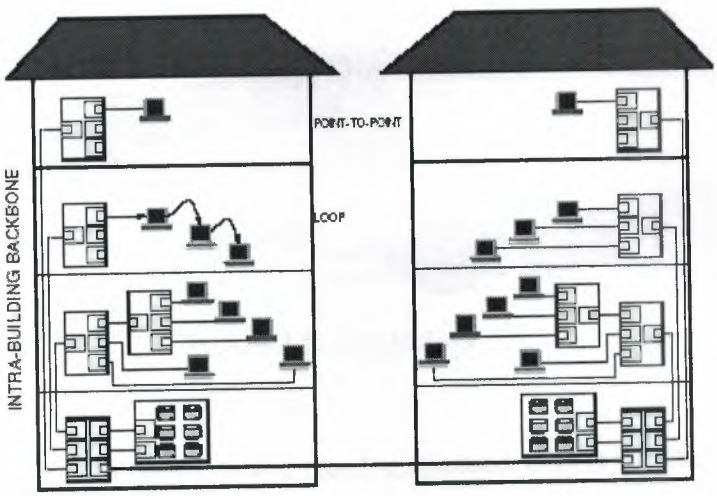

Figure 3.4. Campus Topology

3.5.2 Connectors

Three connector types are generally available. Fiber-optic connectors are usually provided using dual subscriber connectors (SC). Copper connections can be provided through standard DB-9 connectors or the more recentlydeveloped high speed serial direct connect (HSSDC) connectors. We show a selection of connectors in Figure 3.5.

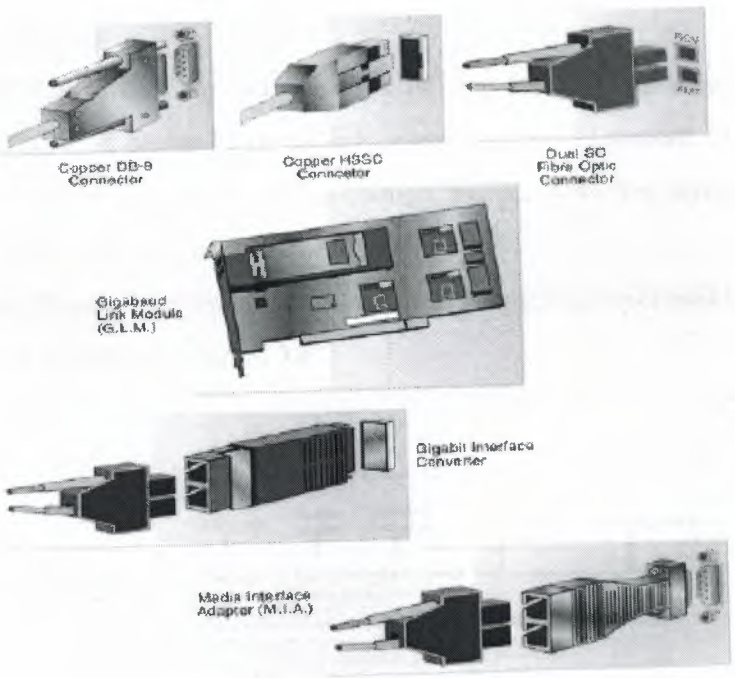

Figure 3.5. Connectors

Fibre Channel products may include a fixed, embedded copper or fiber-optic interface, or they may provide a media-independent interface. There are three media-independent interfaces available:

- **Gigabit Link Modules (GLMs)** — convert parallel signals to serial, and vice versa. GLMs include the serializer/de-serializer (SERDES) function and provide a 20-bit parallel interface to the Fibre Channel encoding and control logic. GLMs are primarily used to provide factory configurability, but may also be field exchanged or upgraded by users.
- **Gigabit Interface Converters (GBICs)** — provide a serial interface to the SERDES function. GBICs can be hot inserted or removed from installed devices. These are particularly useful in multiport devices, such as switches and hubs, where single ports can be reconfigured without affecting other ports.
- **Media Interface Adapters (MIAs)** — allow users to convert copper DB-9 connectors to multi-mode fibre optics. The power to support the optical transceivers is supplied by defined pins in the DB-9 interface.

3.6 Fibre Channel layers

Fibre Channel (FC) is broken up into a series of five layers. The concept of layers, starting with the ISO/OSI seven-layer model, allows the development of one layer to remain independent of the adjacent layers. Although, FC contains five layers, those layers follow the general principles stated in the ISO/OSI model.

The five layers are divided into two parts Physical and signaling layer and Upper layer

The five layers are illustrated in Figure 3.6.

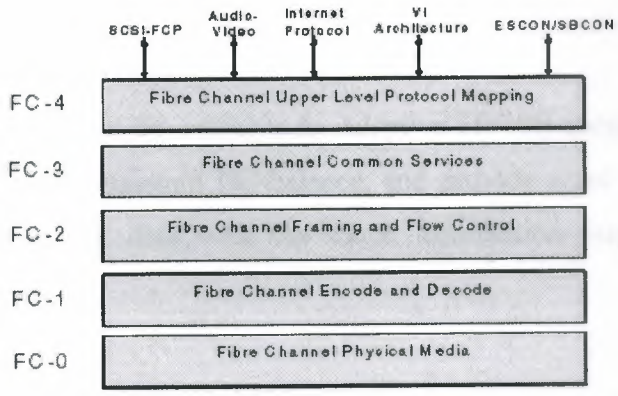

Figure 3.6. Fibre Channel layers

3.6.1 Physical and Signaling Layers

The physical and signaling layers include the three lowest layers: FC-0, FC-1, and FC-2.

3.6.1.1 Physical interface and media: FC-0

The lowest layer (FC-0) defines the physical link in the system, including the cabling, connectors, and electrical parameters for the system at a wide range of data rates. This level is designed for maximum flexibility, and allows the use of a large number of technologies to match the needs of the desired configuration.

A communication route between two nodes may be made up of links of different technologies. For example, in reaching its destination, a signal may start out on copper

wire and become converted to single-mode fibre for longer distances. This flexibility allows for specialized configurations depending on IT requirements.

Laser safety

Fibre Channel often uses lasers to transmit data, and can, therefore, present an optical health hazard. The FC-0 layer defines an open fibre control (OFC) system, and acts as a safety interlock for point-to-point fibre connections that use semiconductor laser diodes as the optical source. If the fibre connection is broken, the ports send a series of pulses until the physical connection is re-established and the necessary handshake procedures are followed.

3.6.1.2 Transmission protocol: FC-1

The second layer (FC-1) provides the methods for adaptive 8B/10B encoding to bind the maximum length of the code, maintain DC-balance, and provide word alignment. This layer is used to integrate the data with the clock information required by serial transmission technologies.

3.6.1.3 Framing and signaling protocol: FC-2

Reliable communications result from Fibre Channel's FC-2 framing and signaling protocol. FC-2 specifies a data transport mechanism that is independent of upper layer protocols. FC-2 is self-configuring and supports point-to-point, arbitrated loop, and switched environments. FC-2, which is the third layer of the FC-PH, provides the transport methods to determine:

- Topologies based on the presence or absence of a fabric
- Communication models
- Classes of service provided by the fabric and the nodes
- General fabric model
- Sequence and exchange identifiers
- Segmentation and reassembly

Data is transmitted in 4-byte ordered sets containing data and control characters. Ordered sets provide the availability to obtain bit and word synchronization, which also establishes word boundary alignment.

Together, FC-0, FC-1, and FC-2 form the Fibre Channel physical and signaling interface (FC-PH).

3.6.2 Upper layers

The Upper layer includes two layers: FC-3 and FC-4.

3.6.2.1 Common services: FC-3

FC-3 defines functions that span multiple ports on a single-node or fabric.

Functions that are currently supported include:

- **Hunt groups:** A hunt group is a set of associated N_Ports attached to a single node. This set is assigned an alias identifier that allows any frames containing the alias to be routed to any available N_Port within the set. This decreases latency in waiting for an N_Port to become available.
- **Striping:** Striping is used to multiply bandwidth, using multiple N_Ports in parallel to transmit a single information unit across multiple links.
- **Multicast:** Multicast delivers a single transmission to multiple destination ports. This includes the ability to broadcast to all nodes or a subset of nodes.

3.6.2.2 Upper layer protocol mapping (ULP): FC-4

The highest layer (FC-4) provides the application-specific protocols. Fibre Channel is equally adept at transporting both network and channel information and allows both protocol types to be concurrently transported over the same physical interface.

Through mapping rules, a specific FC-4 describes how ULP processes of the same FC-4 type interoperate. A channel example is sending SCSI commands to a disk drive, while a networking example is sending IP (Internet Protocol) packets between nodes.

3.7 The movement of data

To move data bits with integrity over a physical medium, there must be a mechanism to check that this has happened and integrity has not been compromised. This is provided by a reference clock which ensures that each bit is received as it was transmitted. In parallel topologies this can be accomplished by using a separate clock or strobe line. As data bits

are transmitted in parallel from the source, the strobe line alternates between high or low to signal the receiving end that a full byte has been sent. In the case of 16- and 32-bit wide parallel cable, it would indicate that multiple bytes have been sent.

The reflective differences in fiber-optic cabling mean that modal dispersion may occur. This may result in frames arriving at different times. This bit error rate (BER) is referred to as the jitter budget. No products are entirely jitter free, and this is an important consideration when selecting the components of a SAN.

As serial data transports only have two leads, transmit and receive, clocking is not possible using a separate line. Serial data must carry the reference timing which means that clocking is embedded in the bit stream.

Embedded clocking, though, can be accomplished by different means. Fibre Channel uses a byte-encoding scheme, which is covered in more detail in 3.7, "Data encoding" on page 56, and clock and data recovery (CDR) logic to recover the clock. From this, it determines the data bits that comprise bytes and words.

Gigabit speeds mean that maintaining valid signaling, and ultimately valid data recovery, is essential for data integrity. Fibre Channel standards allow for a single bit error to occur only once in a trillion bits (10-12). In the real IT world, this equates to a maximum of one bit error every 16 minutes, however actual occurrence is a lot less frequent than this.

3.8 Data encoding

In order to transfer data over a high-speed serial interface, the data is encoded prior to transmission and decoded upon reception. The encoding process ensures that sufficient clock information is present in the serial data stream to allow the receiver to synchronize to the embedded clock information and successfully recover the data at the required error rate. This 8b/10b encoding will find errors that a parity check cannot. A parity check will not find even numbers of bit errors, only odd numbers. The 8b/10b encoding logic will find almost all errors.

First developed by IBM, the 8b/10b encoding process will convert each 8-bit byte into two possible 10-bit characters.

This scheme is called 8b/10b encoding, because it refers to the number of data bits input to the encoder and the number of bits output from the encoder.

The format of the 8b/10b character is of the format Ann.m, where:

- A represents 'D' for data or 'K' for a special character
- nn is the decimal value of the lower 5 bits (EDCBA)
- '.' is a period
- m is the decimal value of the upper 3 bits (HGF)

We illustrate an encoding example in Figure 3.7.

In the encoding example the following occurs:

1. Hexadecimal representation x'59' is converted to binary: 01011001
2. Upper three bits are separated from the lower 5 bits: 010 11001
3. The order is reversed and each group is converted to decimal: 25 2
4. Letter notation D (for data) is assigned and becomes: D25.2

As we illustrate, the conversion of the 8-bit data bytes has resulted in two 10-bit results. The encoder needs to choose one of these results to use. This is achieved by monitoring the running disparity of the previously processedcharacter. For example, if the previous character had a positive disparity, then the next character issued should have an encoded value that represents

negative disparity.

You will notice that in our example the encoded value, when the running disparity is either positive or negative, is the same. This is legitimate. In some cases it (the encoded value) will differ, and in others it will be the same.

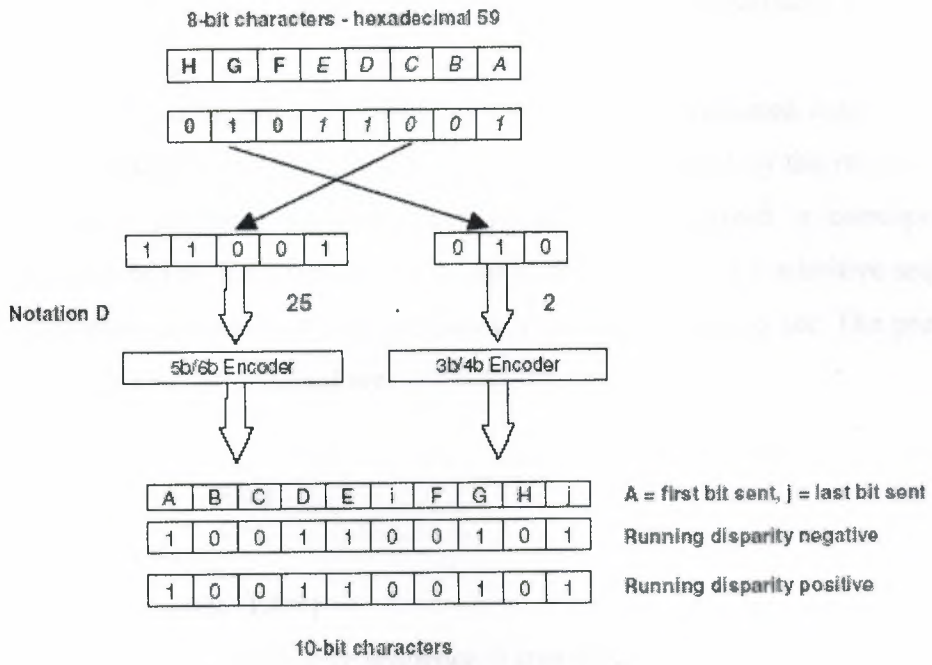

Figure 3.7. 8b/10b Encoding Logic

3.9 Ordered sets

Fibre Channel uses a command syntax, known as an ordered set, to move the data across the network. The ordered sets are four byte transmission words containing data and special characters which have a special meaning.

Ordered sets provide the availability to obtain bit and word synchronization, which also establishes word boundary alignment. An ordered set always begins with the special character K28.5. Three major types of ordered sets are defined by the signaling protocol.

The frame delimiters, the start-of-frame (SOF) and end-of-frame (EOF) ordered sets, establish the boundaries of a frame. They immediately precede or follow the contents of a Frame. There are 11 types of SOF and 8 types of EOF delimiters defined for the Fabric and N_Port Sequence control.

The two primitive signals: idle and receiver ready (R_RDY) are ordered sets designated by the standard to have a special meaning. An Idle is a primitive signal transmitted on the link to indicate an operational port facility ready for frame transmission and reception.

The R_RDY primitive signal indicates that the interface buffer is available for receiving further frames.

A primitive sequence is an ordered set that is transmitted and repeated continuously to indicate specific conditions within a port or conditions encountered by the receiver logic of a port. When a primitive sequence is received and recognized, a corresponding primitive sequence or Idle is transmitted in response. Recognition of a primitive sequence requires consecutive detection of three instances of the same ordered set. The primitive sequences supported by the standard are:

- Offline state (OLS)
- Not operational (NOS)
- Link reset (LR)
- Link reset response (LRR)

Offline (OLS): The offline primitive sequence is transmitted by a port to indicate one of the following conditions: The port is beginning the link initialization protocol, or the port has received and recognized the NOS protocol or the port is entering the offline status.

Not operational (NOS): The not operational primitive sequence is transmitted by a port in a point-to-point or fabric environment to indicate that the transmitting port has detected a link failure or is in an offline condition, waiting for the OLS sequence to be received.

Link reset (LR): The link reset primitive sequence is used to initiate a link reset.

Link reset response (LRR): Link reset response is transmitted by a port to indicate that it has recognized

a LR sequence and performed the appropriate link reset.

3.10 Frames

Frames are the basic building blocks of an FC connection. The frames contain the information to be transmitted, the address of the source and destination ports, and link control information. Frames are broadly categorized as Data frames and Link_control frames. When the frame is defined as a link control frame the length of the data field is zero bytes. If the frame is defined as a data frame, the data field may be any number of words between zero and 528 (0 and 2112 bytes). Data frames may be used as Link_Data

frames and Device_Data frames. Link control frames are classified as Acknowledge (ACK) and Link_Response (Busy and Reject) frames.

The primary function of the fabric is to receive the frames from the source port and route them to the destination port. It is the FC-2 layer's responsibility to break the data to be transmitted into frame size, and reassemble the frames. The frame structure is shown in Figure 3.8.

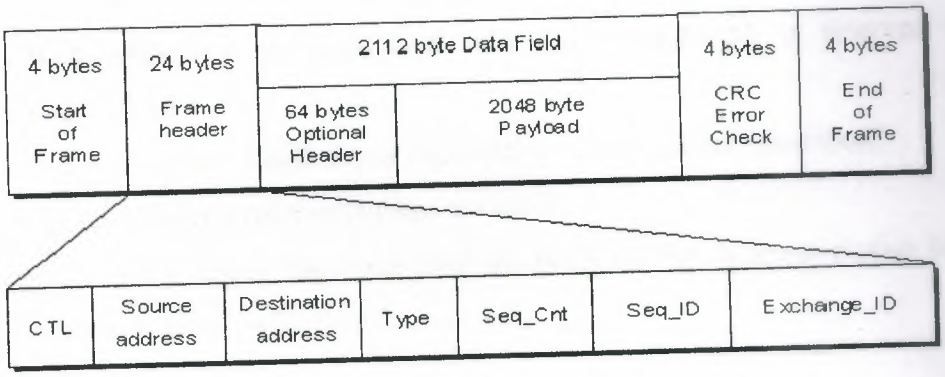

Figure 3.8. Frame Structure

Each frame begins and ends with a frame delimiter. The frame header immediately follows the SOF delimiter. The frame header is used to control link applications and control device protocol transfers, and to detect missing or out of order frames. An optional header may contain further link control information. A maximum 2112 byte long field contains the information to be transferred from a source N_Port to a destination N_Port. The 4 bytes cyclic redundancy check (CRC) precedes the EOF delimiter. The CRC is used to detect transmission errors.

3.11 Framing classes of service

Fibre Channel provides a logical system of communication called class of service that is allocated by various login protocols. Fibre Channel provides six different classes of service:

- Class 1: Acknowledged connection service
- Class 2: Acknowledged connectionless service
- Class 3: Unacknowledged connectionless service

- Class 4: Fractional bandwidth connection-oriented service
- Class 5: Reserved for future development
- Class 6: Uni-directional connection service

Each class of service has a specific set of delivery attributes involving characteristics, such as:

- Is a connection or circuit established?
- Is the in-order delivery of frames guaranteed?
- If a connection is established, how much bandwidth is reserved for that connection?
- Is confirmation of delivery or notification of non-delivery provided?
- Which flow control mechanisms are used?

The answers to the above questions form the basis for the different classes of service provided and are shown in Table 3.2.

Table 3.2 Classes of service

Attribute	Class 1	Class 2	Class 3	Class 4	Class 6
Connection or circuit established	Yes			Yes	Yes
In order frame delivery	Yes			Yes	Yes
Amount of link bandwidth	Full				
Confirmation of delivery	Yes	Yes		Yes	Yes
Support Multicast			Yes		Yes
Flow Control used: - End-to-End - Buffer-to-Buffer(R_RDY) - Virtual Circuit (virtual circuit_RDY)	Yes SOFc1 only No	Yes Yes No	No Yes No	Yes No Yes	Yes SOFc1 only No

Class 1: Acknowledged connection service

Class 1 provides true connection service. The result is circuit-switched, dedicated bandwidth connections.

An end-to-end path between the communicating devices is established through the switch. Fibre Channel Class 1 service provides an acknowledgment of receipt for guaranteed delivery. Class 1 also provides full-bandwidth, guaranteed delivery, and bandwidth for applications like image transfer and storage backup and recovery. Some applications use the guaranteed delivery feature to move data reliably and quickly without the overhead of a network protocol stack. Camp On is a Class 1 feature that enables a switch to monitor a busy port and queue that port for the next connection. As soon as the port is free, the switch makes the connection. This switch service speeds connect time, rather than sending a "busy" signal back to the originating N_Port and requiring the N_Port to retry to make the connection.

Stacked connect is a Class 1 feature that enables an originating N_Port to queue sequential connection requests with the switch. Again, this feature reduces overhead and makes the switch service more efficient.

Another form of Class 1 is called dedicated simplex service. Normally, Class 1 connections are bi-directional; However, in this service, communication is in one direction only. It is used to separate the transmit and receive switching. It permits one node to transfer to another node while simultaneously receiving from a third node. We show this in Figure 3.9.

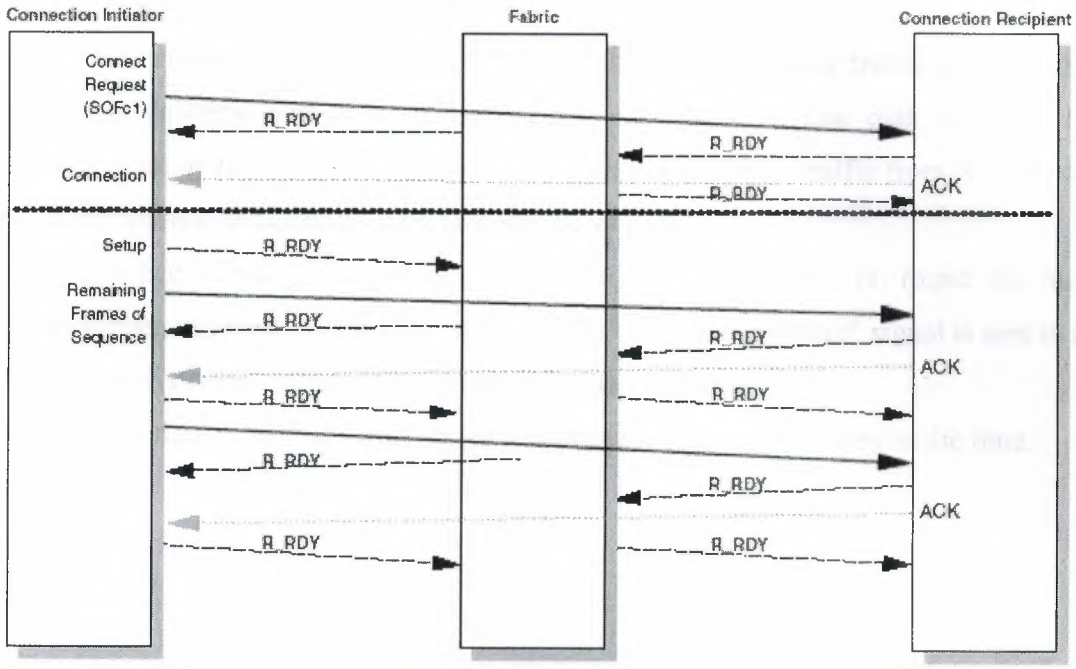

Figure 3.9. Class 1 flow control

Class 2: Acknowledged connectionless service

Class 2 is a connectionless service, independently switching each frame and providing guaranteed delivery with an acknowledgment of delivery. The path between two interconnected devices is not dedicated. The switch multiplexes traffic from N_Ports and NL_Ports without dedicating a path through the switch.

Class 2 credit-based flow control eliminates congestion that is found in many connectionless networks. If the destination port is congested, a "busy" signal is sent to the originating N_Port. The N_Port will then resend the message.

This way, no data is arbitrarily discarded just because the switch is busy at the time.

We show this in Figure 3.10.

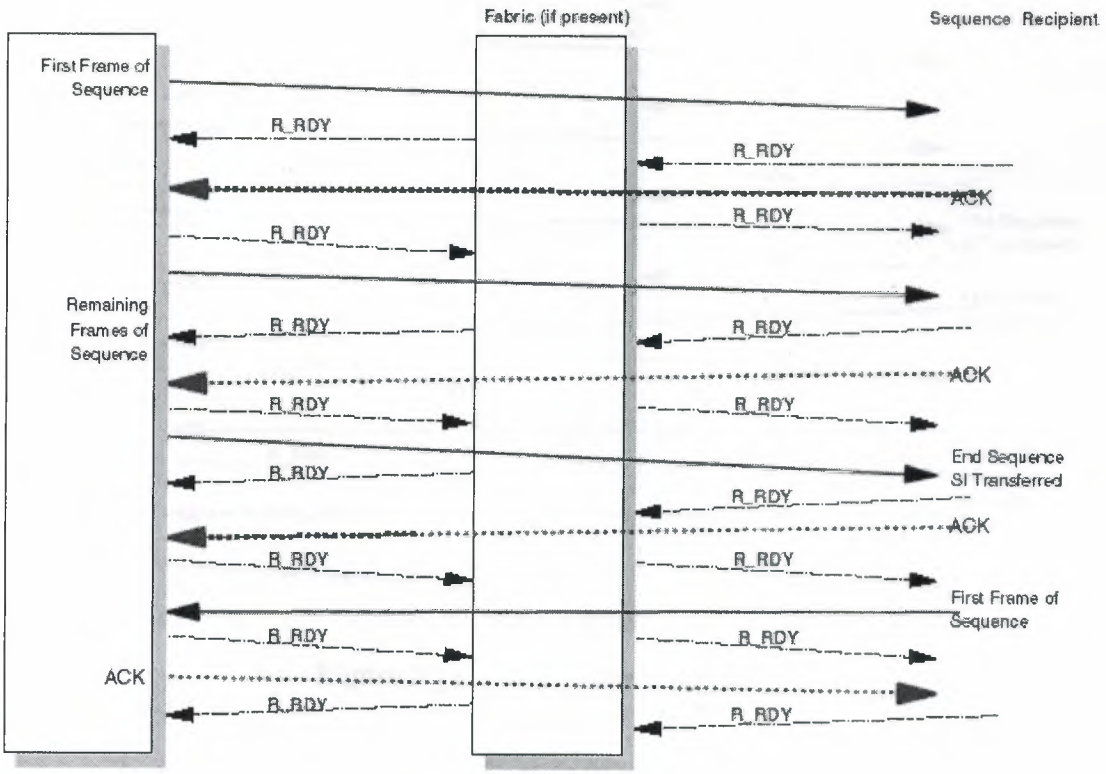

Figure 3.10 Class 2 Flow Control

Class 3: Unacknowledged connectionless service

Class 3 is a connectionless service, similar to Class 2, but no confirmation of receipt is given. This unacknowledged transfer is used for multicasts and broadcasts on networks, and for storage interfaces on Fibre Channel loops.

The loop establishes a logical point-to-point connection and reliably moves data to and from storage.

Class 3 arbitrated loop transfers are also used for IP networks. Some applications use logical point-to-point connections without using a network layer protocol, taking advantage of Fibre Channel's reliable data delivery. We show this in Figure 3.11.

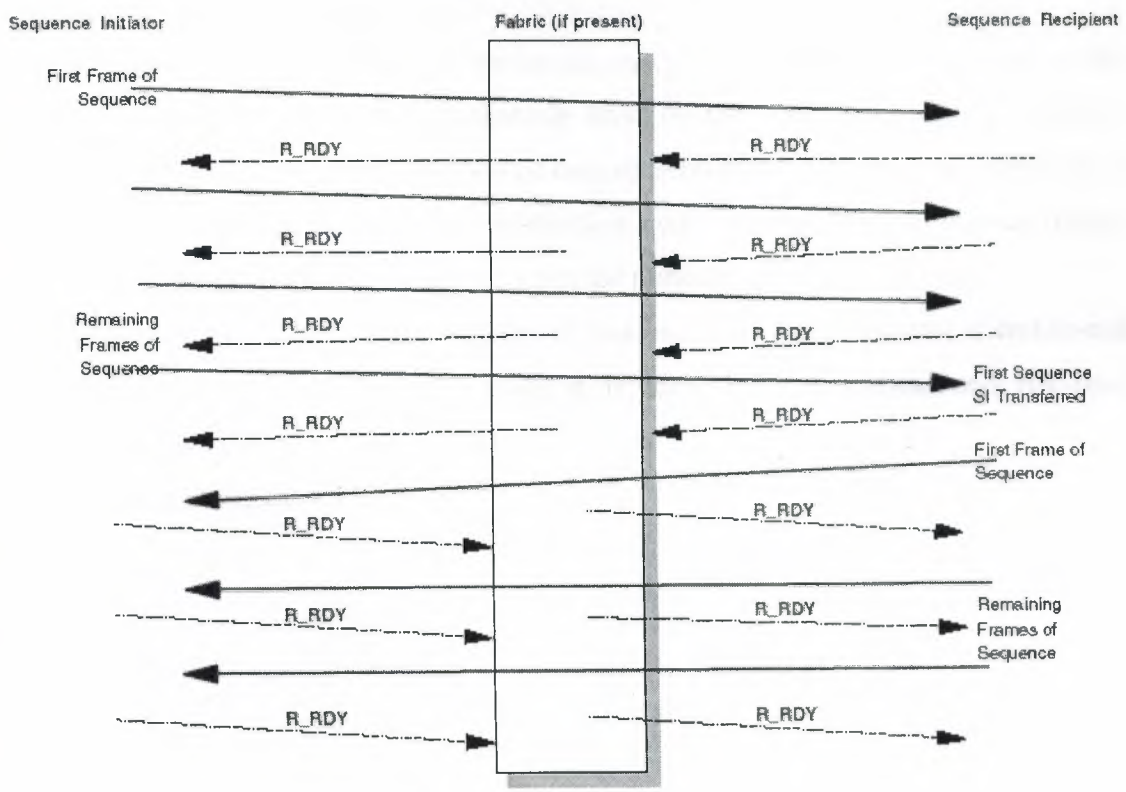

Figure 3.11. Class 3 flow control

Class 4: Fractional bandwidth acknowledged

Class 4 is a connection-oriented class of service which provides a virtual circuit. Virtual connections are established with bandwidth reservation for a predictable quality of service. A Class 4 connection is bi-directional, with one virtual circuit operational in each direction, and it supports a different set of quality of service parameters for each virtual circuit. These quality of service (QoS) parameters include guaranteed bandwidth and bounded end-to-end delay. A quality of service facilitator (QoSF) function is provided within the switch to manage and maintain the negotiated quality of service on each virtual circuit.

A node may reserve up to 256 concurrent Class 4 connections. Separate functions of Class 4 are the setup of the quality of service parameters and the connection itself.

When a Class 4 connection is active, the switch paces frames from the source node to the destination node. Pacing is the mechanism used by the switch to regulate available bandwidth per virtual circuit. This level of control permits congestion management for a switch and guarantees access to the destination node. The switch multiplexes frames belonging to different virtual circuits between the same or different node pairs.

Class 4 service provides in-order delivery of frames. Class 4 flow control is end-to-end and provides guaranteed delivery. Class 4 is ideal for time-critical and real-time applications like video.

We show this in Figure 3.12.

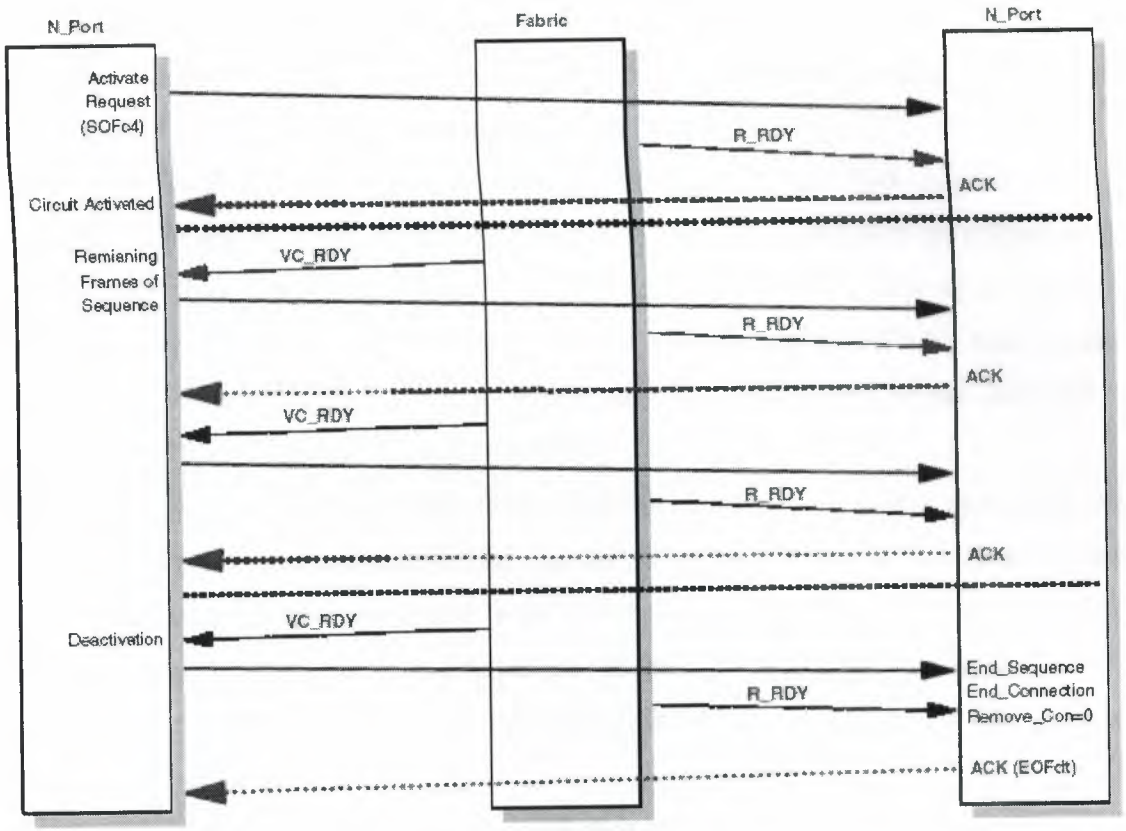

Figure 3.12. Class 4 Flow Control

Class 5: Still under development

Class 5 is still under development. This service allow for simultaneous (isochronous) data transfer to several participants and is especially applicable for audio and video servers in broadcast mode.

Class 6: Uni-directional connection service

Class 6 is similar to Class 1, providing uni-directional connection service. However, Class 6 also provides reliable multicast and pre-emption. Class 6 is ideal for video broadcast applications and real-time systems that move large quantities of data.

3.12 Naming and Addressing

In a Fibre Channel environment the unique identity of participants is maintained through a hierarchy of fixed names and assigned addresses identifiers.

In Fibre Channel terminology, a communicating device is a node. Each node has a fixed 64-bit Node_name assigned by the manufacturer. The node name will be unique if the manufacturer has registered a range of addresses with the IEEE, and so is normally referred to as a World-Wide Name. An N_Port within a parent (WWN) node is also assigned a unique 64-bit Port_Name, which aids the accessibility of the port and is known as the World-Wide Port Name (WWPN).

The WWN is a registered, unique 64-bit identifier assigned to nodes and ports. An example of a registration authority is the registration service support of the Media Access Control (MAC) address associated with the network interface card. In the IEEE understanding, a MAC address consists of 48 bits, 24 of which are assigned to a particular company through the registration process with the remaining 24 bits assigned by the user.

An example of the node and port name correlation is shown in Figure 3.13.

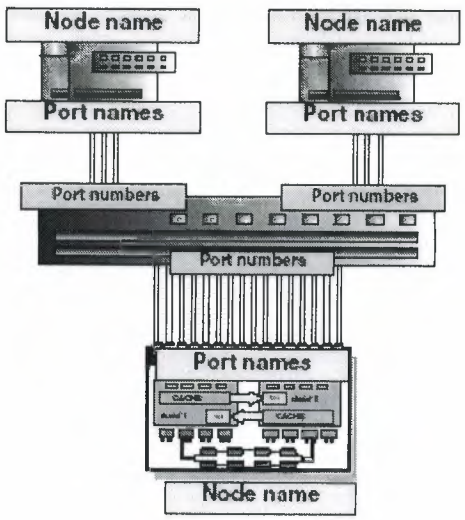

Figure 3.13. Nodes And Ports

For more information on the governing body and the WWN, go to:
standards.ieee.org/regauth/oui/index.html This naming convention allows each node and its associated N_Ports to be unique and accessible, even in a complex SANs.

The Fibre Channel naming convention allows either global or locally administered uniqueness to be assigned to a device. However, the administered name or WWN is not used for transporting frames across the network. In addition to a Fibre Channel WWN, a communicating device is dynamically assigned a 24-bit port address, or N_Port ID that is used for frame routing. As well as providing frame routing optimization, this 24-bit port address strategy removes the overhead of manual administration of addresses by allowing the topology to assign address.

In fabric environments, the switch is responsible for assigning a 24-bit address to each device as it logs on.

Allowing the topology to manage the assignment of addresses has the advantage that control of the addresses is now performed by the entity that is responsible for the routing of information. This means that address assignments can be made in a manner that results in the most efficient routing of frames within that topology. This approach mimics the behavior of the telephone system, where the telephone number (address) of a particular telephone is determined by where it is attached to the telephone system.

Fibre Channel ports

There is more than one kind of port, though, and its designation represents the use which is being made of it. We show some port designations in Figure 3.14.

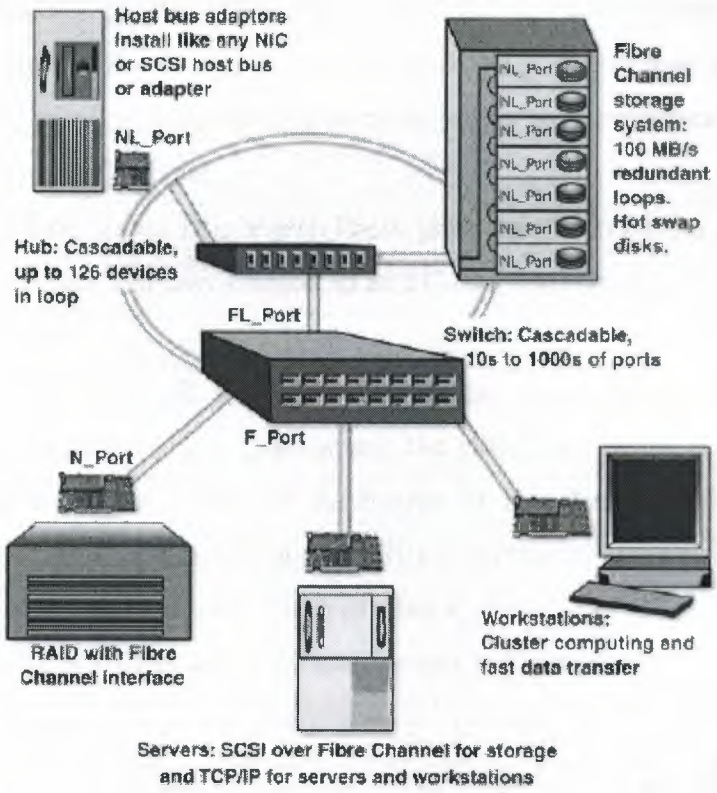

Figure 3.14. Fibre Channel Ports

There are six kinds of ports that we are concerned with in this redbook. They are:

- **Loop port (L_Port)** This is the basic port in a Fibre Channel arbitrated loop (FC-AL) topology. If an N_Port is operating on a loop it is referred to as an NL_Port. If a fabric port is on a loop it is known as an FL_Port. To draw the distinction, throughout this book we will always qualify L_Ports as either NL_Ports or FL_Ports.
- **Node ports (N_Port)** These ports are found in Fibre Channel nodes, which are defined to be the source or destination of information units (IU). I/O devices and

host systems interconnected in point-to-point or switched topologies use N_Ports for their connection. N_Ports can only attach to other N_Ports or to F_Ports.

- **Node-loop ports (NL_Port)** These ports are just like the N_Port described above, except that they connect to a Fibre Channel abritrated loop (FC-AL) topology. NL_Ports can only attach to other NL_Ports or to FL_Ports
- **Fabric ports (F_Port)** These ports are found in Fibre Channel switched fabrics. They are not the source or destination of IU's, but instead function only as a "middle-man" to relay the IUs from the sender to the receiver. F_Ports can only be attached to N_Ports.
- **Fabric-loop ports (FL_Port)** These ports are just like the F_Ports described above, except that they connect to an FC-AL topology. FL_Ports can only attach to NL_Ports.
- **Expansion ports (E_Port)** These ports are found in Fibre Channel switched fabrics and are used to interconnect the individual switch or routing elements. They are not the source or destination of IUs, but instead function like the F_Ports and FL_Ports to relay the IUs from one switch or routing elements to another. E_Ports can only attach to other E_Ports.

We show all these ports and how they interconnect in Figure 3.15.

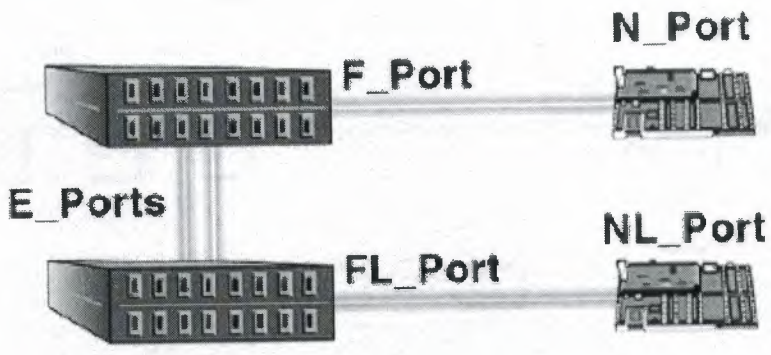

Figure 3.15. Port Interconnections

The Fibre Channel architecture specifies the link characteristics and protocol used between N_Ports, between N_Ports and F_Ports, an between NL_Ports and FL_Ports.

4. THE TECHNICAl TOPOLOGY OF A SAN

4.1 Overview

Fibre Channel provides three distinct and one hybrid interconnection topologies. By having more than one interconnection option available, a particular application can choose the topology that is best suited to its requirements. The three fibre channel topologies are:

- Point-to-point
- Arbitrated loop
- Switched — referred to as a fabric

The three topologies are shown in Figure 4.1.

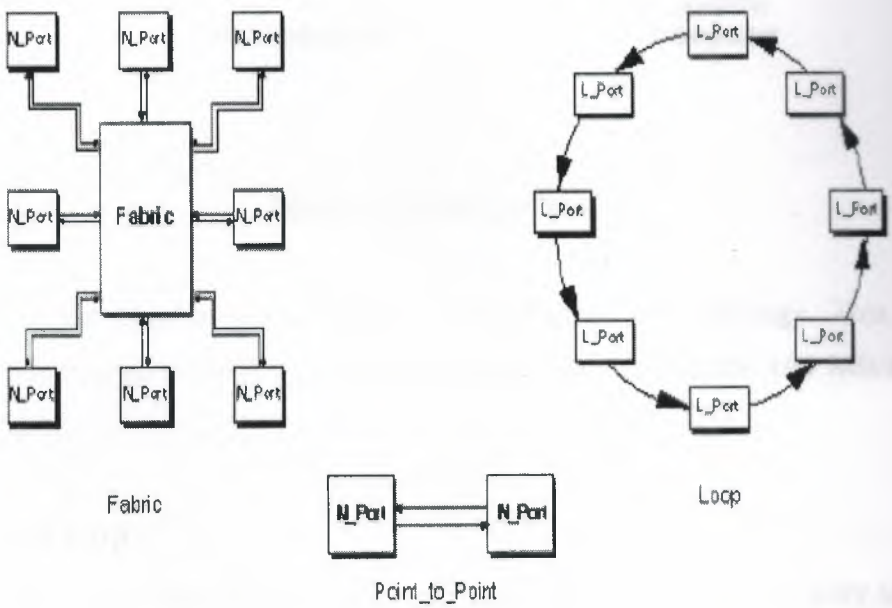

Figure 4.1. SAN Topologies.

4.2 Point-to-point

A point-to-pointconnection is the simplest topology. It is used when there are exactly two nodes, and future expansion is not predicted. There is no sharing of the media, which allows the devices to use the total bandwidth of the link.

A simple link initialization is needed before communications can begin. We illustrate a simple point-to-point connection in Figure 4.2.

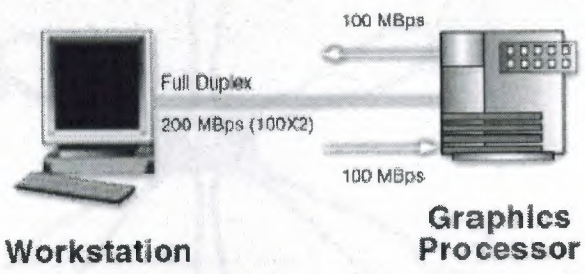

Figure 4.2 Point-To-Point

An extension of the point-to-point topology is the logical start topology. This is a collection of point-to-point topology links and both topologies provide 100 MB/s full duplex bandwidth.

4.3 Arbitrated loop

The second topology is Fibre Channel Arbitrated Loop (FC-AL). FC-AL is more useful for storage applications. It is a loop of up to 126 nodes (NL_Ports) that is managed as a shared bus. Traffic flows in one direction, carrying data frames and primitives around the loop with a total bandwidth of 100 MB/s. Using arbitration protocol, a single connection is established between a sender and a receiver, and a data frame is transferred around the loop. When the communication comes to an end between the two connected ports, the

loop becomes available for arbitration and a new connection may be established. Loops can be configured with hubs to make connection management easier. Up to 10 km distance is supported by the Fibre Channel standard for both of these configurations. However, latency on the arbitrated loop configuration is affected by the loop size.

A simple loop, configured using a hub, is shown in Figure 4.3.

Figure 4.3. Arbitrated loop

4.3.1 Loop protocols

To support the shared behavior of the arbitrated loop, a number of loop-specific protocols are used. These protocols are used to:

- Initialize the loop and assign addresses
- Arbitrate for access to the loop
- Open a loop circuit with another port in the loop
- Close a loop circuit when two ports have completed their current use of the loop

- Implement the access fairness mechanism to ensure that each port has an opportunity to access the loop

4.3.2 Loop initialization

Loop initialization is a necessary process for the introduction of new participants on to the loop. Whenever a loop port is powered on or initialized, it executes the loop initialization primitive (LIP) to perform loop initialization.

Optionally, loop initialization may build a positional map of all the ports on the loop. The positional map provides a count of the number of ports on the loop, their addresses and their position relative to the loop initialization master.

Following loop initialization, the loop enters a stable monitoring mode and begins with normal activity. An entire loop initialization sequence may take only a few milliseconds, depending on the number of NL_Ports attached to the loop. Loop initialization may be started by a number of causes. One of the most likely reasons for loop initialization is the introduction of a new device.

For instance, an active device may be moved from one hub port to another hub port, or a device that has been powered on could re-enter the loop.

A variety of ordered sets have been defined to take into account the conditions that an NL_Port may sense as it starts the initialization process. These ordered sets are sent continuously while a particular condition or state exists. As part of the initialization process, loop initialization primitive sequences (referred to collectively as LIPs) are issued. As an example, an NL_Port must issue at least three identical ordered sets to start initialization. An ordered set transmission word always begins with the special character K28.5. Once these identical ordered sets have been sent, and as each downstream device receives the LIP stream, devices enter a state known as open-init. This causes the suspension of any current operation and enables the device for the loop initialization procedure. LIPs are forwarded around the loop until all NL_Ports are in an open-init condition. At this point, the NL_Ports need to be managed. In contrast to a Token-Ring, the Arbitrated Loop has no permanent master to manage the topology. Therefore, loop initialization provides a selection process to determine which device will be the temporary loop master. After the loop master is chosen it assumes the responsibility for

directing or managing the rest of the initialization procedure. The loop master also has the responsibility for closing the loop and returning it to normal operation.

Selecting the loop master is carried out by a subroutine known as the Loop Initialization Select Master (LISM) procedure. A loop device can be considered for temporary master by continuously issuing LISM frames that contain a port type identifier and a 64-bit World-Wide Name. For FL_Ports the identifier is x'00' and for NL_Ports it is x'EF'.

When a downstream port receives a LISM frame from a upstream partner, the device will check the port type identifier. If the identifier indicates an NL_Port, the downstream device will compare the WWN in the LISM frame to its own.

The WWN with the lowest numeric value has priority. If the received frame's WWN indicates a higher priority, that is to say it has a lower numeric value, the device stops its LISM broadcast and starts transmitting the received LISM. Had the received frame been of a lower priority, the receiver would have thrown it away and continued broadcasting its own.

At some stage in proceedings, a node will receive its own LISM frame, which indicates that it has the highest priority, and succession to the throne of 'temporary loop master' has taken place. This node will then issue a special ordered set to indicate to the others that a temporary master has been selected.

4.3.3 Hub cascading

Since an arbitrated loop hub supplies a limited number of ports, building larger loops may require linking another hub. This is called hub cascading. A server with an FC-AL, shortwave, host bus adapter can connect to an FC-AL hub 500 meters away. Each port on the hub can connect to an FC-AL device up to 500 meters away. Cascaded hubs use one port on each hub for the hub-to-hub connection and this increases the potential distance between nodes in the loop by an additional 500 meters. In this topology the overall distance is 1500 meters. Both hubs can support other FC-AL devices at their physical locations. Stated distances assume a 50 micron multimode cable.

4.3.4 Loops

There are two different kinds of loops, the private and the public loop.

4.3.4.1 Private loop

The private loop does not connect with a fabric, only to other private nodes using attachment points called NL_Ports. A private loop is enclosed and known only to itself. In Figure 4.4 we show a private loop.

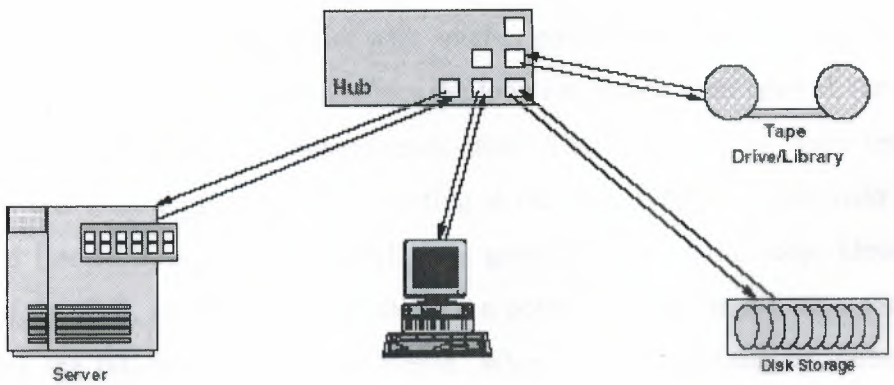

Figure 4.4. Private loop Implementation

4.3.4.2 Public loop

A public loop requires a fabric and has at least one FL_Port connection to a fabric. A public loop extends the reach of the loop topology by attaching the loop to a fabric. Figure 4.5 hows a public loop.

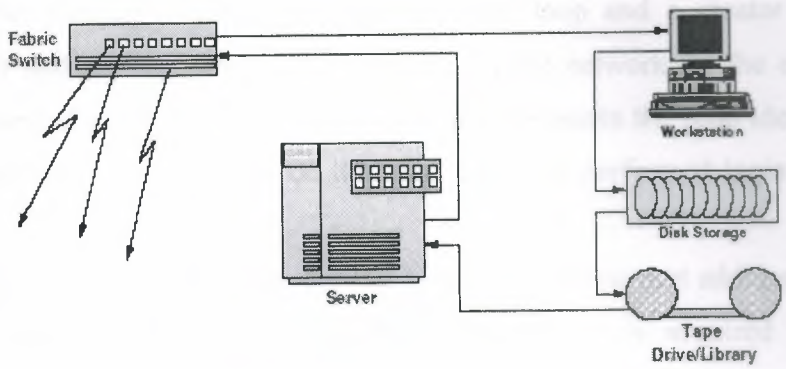

Figure 4.5 Public loop Implementation

4.3.5 Arbitration

When a loop port wants to gain access to the loop, it has to arbitrate. When the port wins arbitration, it can open a loop circuit with another port on the loop; a function similar to selecting a device on a bus interface. Once the loop circuit has been opened, the two ports can send and receive frames between each other. This is known as "loop tenancy". If more than one node on the loop is arbitrating at the same time, the node with the lower Arbitrated Loop Physical Address (AL_PA) gains control of the loop. Upon gaining control of the loop, the node then establishes a point-to-point transmission with another node using the full bandwidth of the media. When a node has finished transmitting its data, it is not required to give up control of the loop. This is a channel characteristic of Fibre Channel. However, there is a "fairness algorithm", which states that a device cannot regain control of the loop until the other nodes have had a chance to control the loop.

4.3.6 Loop addressing

An NL_Port, like a N_Port, has a 24-bit port address. If no switch connection exists, the two upper bytes of this port address are zeroes (x'00 00') and referred to as a private loop. The devices on the loop have no connection with the outside world. If the loop is attached to a fabric and an NL_Port supports a fabric login, the upper two bytes are assigned a positive value by the switch. We call this mode a public loop.

As fabric-capable NL_Ports are members of both a local loop and a greater fabric community, a 24-bit address is needed as an identifier in the network. In the case of public loop assignment, the value of the upper two bytes represents the loop identifier, and this will be common to all NL_Ports on the same loop that performed login to the fabric.

In both public and private arbitrated loops, the last byte of the 24-bit port address refers to the arbitrated loop physical address (AL_PA). The AL_PA is acquired during initialization of the loop and may, in the case of fabric-capable loop devices, be modified by the switch during login.

The total number of the AL_PAs available for arbitrated loop addressing is 127. This number is based on the requirements of 8b/10b running disparity between frames.

As a frame terminates with an end-of-frame character (EOF) this will force the current running disparity negative. In the Fibre Channel standard each transmission word between the end of one frame and the beginning of another frame should also leave the running disparity negative. If all 256 possible 8-bit bytes are sent to the 8b/10b encoder, 134 emerge with neutral disparity characters. Of these 134, seven are reserved for use by Fibre Channel. The 127 neutral disparity characters left have been assigned as AL_PAs. Put another way, the 127 AL_PA limit is simply the maximum number, minus reserved values, of neutral disparity addresses that can be assigned for use by the loop. This does not imply that we recommend this amount, or load, for a 100MB/s shared transport, but only that it is possible.

Arbitrated Loop will assign priority to AL_PAs, based on numeric value. The lower the numeric value, the higher the priority is. For example, an AL_PA of x'01' has a much better position to gain arbitration over devices that have a lower priority or higher numeric value. At the top of the hierarchy it is not unusual to find servers, but at the lower end you would expect to find disk arrays.

It is the arbitrated loop initialization that ensures each attached device is assigned a unique AL_PA. The possibility for address conflicts only arises when two separated loops are joined together without initialization.

4.3.7 Logins

There are three different types of login for Fibre Channel. These are:

- Fabric login
- Port login
- Process login

Port login

Port login is also known as PLOGI.

Port login is used to establish a session between two N_Ports (devices) and is necessary before any upper level commands or operations can be performed. During the port login, two N_Ports (devices) swap service parameters and make themselves known to each other.

Process login

Process login is also known as PRLI.

Process login is used to set up the environment between related processes on an originating N_Port and a responding N_Port. A group of related processes is collectively known as an image pair. The processes involved can be system processes, system images, such as mainframe logical partitions, control unit images, and FC-4 processes. Use of process login is optional from the perspective of Fibre Channel FC-2 layer, but may be required by a specific upper-level protocol as in the case of SCSI-FCP mapping.

We show Fibre Channel logins in Figure 4.6.

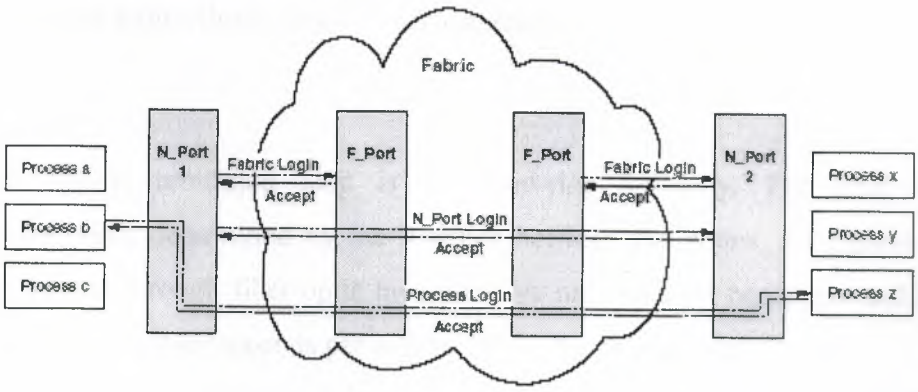

Figure 4.6 Fibre Channel logins

4.3.8 Closing a loop circuit

When two ports in a loop circuit complete their frame transmission, they may close the loop circuit to allow other ports to use the loop. The point at which the loop circuit is closed depends on the higher-level protocol, the operation in progress, and the design of the loop ports.

4.3.9 Supported devices

An arbitrated loop may support a variety of devices, including HBAs installed in the following servers:

- Individual Fibre Channel disk drives
- JBOD
- Fibre Channel RAID
- Native Fibre Channel tape sub-systems
- Fibre Channel to SCSI bridges

4.3.10 Broadcast

Arbitrated loop, in contrast to Ethernet, is a non-broadcast transport. When an NL_Port has successfully won the right to arbitration, it will open a target for frame transmission. Any subsequent loop devices in the path between the two will see the frames and forward them on to the next node in the loop.

It is this non-broadcast nature of arbitrated loop, by removing frame handling overhead from some of the loop, which enhances performance.

4.3.11 Distance

As stated before, arbitrated loop is a closed-ring topology. The total distance requirements being determined by the distance between the nodes. At gigabit speeds, signals propagate through fiber-optic media at five nanoseconds per meter and through copper media at four nanoseconds per meter. This is the delay factor.

Calculating the total propagation delay incurred by the loop's circumference is achieved by multiplying the length — both transmit and receive — of copper and fiber-optic

cabling deployed by the appropriate delay factor. For example, a single 10 km link to an NL_Port would cause a 50 microsecond (10 km x 5 nanoseconds delay factor) propagation delay in each direction and 100 microseconds in total. This equates to 1 MB/s of bandwidth used to satisfy the link.

4.3.12 Bandwidth

For optical interconnects for SANs, the bandwidth requirements are greatly influenced by the capabilities of:

- The system buses
- Network switches
- The interface adapters that interface with them
- Traffic locality

The exact bandwidth required is somewhat dependent on implementation, but are currently in the range of 100 to 1000 MB/s. Determining bandwidth requirements is difficult and there is no exact science that can take into account the unpredictability of sporadic bursts of data, for example. Planning bandwidth based on peak requirements could be wasteful. Designing for sustained bandwidth requirements, with the addition of safety margins, may be less wasteful.

4.4 Switched fabric

The third topology used in SAN implementations is Fibre Channel Switched Fabric (FC-SW). A Fibre Channel fabric is one or more fabric switches in a single, sometimes extended, configuration. Switched fabrics provide full 100MB/s bandwidth per port, compared to the shared bandwidth per port in Arbitrated Loop implementations.

If you add a new device into the arbitrated loop, you further divide the shared bandwidth. However, in a switched fabric, adding a new device or a new connection between existing ones actually increases the bandwidth. For example, an 8-port switch with three initiators and three targets can support three concurrent 100 MB/s conversations or a total 300 MB/s throughput (600 MB/s if full-duplex applications were available). A switched fabric configuration is shown in Figure 4.7.

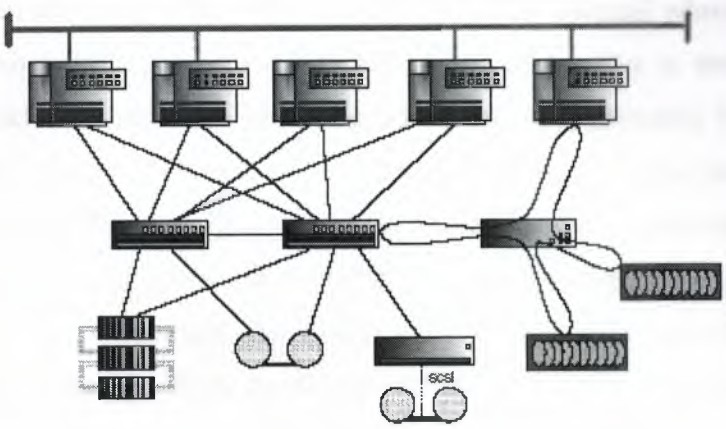

Figure 4.7 Sample Switched Fabric Configuration

4.4.1 Addressing

This ID is called the World Wide Name (WWN), This WWN is a 64-bit address and if two WWN addresses are put into the frame header, this leaves 16 bytes of data just for identifying destination and source address. So 64-bit addresses can impact routing performance.

Because of this there is another addressing scheme used in Fibre Channel networks. This scheme is used to address the ports in the switched fabric. Each port in the switched fabric has its own unique 24-bit address. With this 24-bit addressing scheme we get a smaller frame header and this can speed up the routing process. With this frame header and routing logic the Fibre Channel fabric is optimized for high-speed switching of frames.

With a 24-bit addressing scheme this allows for up to 16 million addresses, which is an address space larger than any practical SAN design in existence in today's world. Who knows what the future will bring? Maybe Fibre Channel addressing will have the same problems in the future as the internet does today, which is a lack of addresses. This 24-bit addressing has to be connected with the 64-bit addressing associated with World Wide Names. We explain this in the section that follows.

4.4.2 Name and addressing

The 24-bit address scheme also removes the overhead of manual administration of addresses by allowing the topology itself to assign addresses. This is not like WWN addressing, in which the addresses are assigned to the manufacturers by the IEEE standards committee, and are built in to the device at build time, similar to naming a child at birth. If the topology itself assigns the 24-bit addresses, then somebody has to be responsible for the addressing scheme from WWN addressing to port addressing.

In the switched fabric environment, the switch itself is responsible for assigning and maintaining the port addresses. When the device with its WWN is logging into the switch on a specific port, the switch will assign the port address to that port and the switch will also maintain the correlation between the port address and the WWN address of the device on that port. This function of the switch is implemented by using a Simple Name Server (SNS). The Simple Name Server is a component of the fabric operating system, which runs inside the switch. It is essentially a database of objects in which fabric-attached device registers its values.

Dynamic addressing also removes the potential element of human error in address maintenance, and provides more flexibility in additions, moves, and changes in the SAN.

4.4.2.1 Port address

A 24-bit port address consists of three parts:

- Domain (bits from 23 to 16)
- Area (bits from 15 to 08)
- Port or arbitrated loop physical address - AL_PA (bits from 07 to 00)

We show how the address is built up in Figure 4.8.

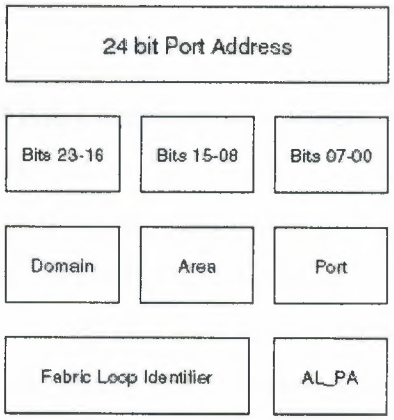

Figure 4.8. Fabric Port Address

We explain the significance of some of the bits that make up the port address in the following sections.

Domain

The most significant byte of the port address is the domain. This is the address of the switch itself. One byte allows up to 256 possible addresses. Because some of these are reserved (like the one for broadcast) there are only 239 addresses actually available. This means that you can have as many as 239 switches in your SAN environment. The domain number allows each switch to have a unique identifier if you have multiple interconnected switches in your environment.

Area

The area field provides 256 addresses. This part of the address is used to identify the individual FL_Ports supporting loops or it can be used as the identifier for a group of F_Ports; for example, a card with more ports on it. This means that each group of ports has a different area number, even if there is only one port in the group.

Port

The final part of the address provides 256 addresses for identifying attached N_Ports and NL_Ports. To arrive at the number of available addresses is a simple calculation based on:

Domain x Area x Ports

This means that there are 239 x 256 x 256 = 15,663,104 addresses available.

4.4.3 Fabric login

After the fabric capable Fibre Channel device is attached to a fabric switch, it will carry out a fabric login (FLOGI).

Similar to port login, FLOGI is an extended link service command that sets up a session between two participants. With FLOGI a session is created between an N_Port or NL_Port and the switch. An N_Port will send a FLOGI frame that contains its Node Name, its N_Port Name, and service parameters to a well-known address of 0xFFFFFE.

A public loop NL_Port first opens the destination AL_PA 0x00 before issuing the FLOGI request. In both cases the switch accepts the login and returns an accept (ACC) frame to the sender. If some of the service parameters requested by the N_Port or NL_Port are not supported, the switch will set the appropriate bits in the ACC frame to indicate this. When the N_Port logs in it uses a 24-bit port address of 0x000000. Because of this the fabric is allowed to assign the appropriate port address to that device, based on the Domain-Area-Port address format. The newly assigned address is contained in the ACC response frame.

When the NL_Port logs in a similar process starts, except that the least significant byte is used to assign AL_PA and the upper two bytes constitute a fabric loop identifier. Before an NL_Port logs in it will go through the LIP on the loop, which is started by the FL_Port, and from this process it has already derived an AL_PA. The switch then decides if it will accept this AL_PA for this device or not. If not a new AL_PA is assigned to the NL_Port, which then causes the start of another LIP. This ensures that the switch assigned AL_PA does not conflict with any previously selected AL_PAs on the loop.

After the N_Port or public NL_Port gets its fabric address from FLOGI, it needs to register with the SNS. This is done with port login (PLOGI) at the address 0xFFFFFC. The device may register values for all or just some database objects, but the most useful are its 24-bit port address, 64-bit Port Name (WWPN), 64-bit Node Name (WWN), class of service parameters, FC-4 protocols supported, and port type, such as N_Port or NL_Port.

4.4.4 Private devices on NL_Ports

It is easy to explain how the port to World Wide Name address resolution works when a single device from an N_Port is connected to an F_Port, or when a public NL_Port device is connected to FL_Port in the switch. The SNS will add an entry for the device World Wide Name and connects that with the port address which is selected from the selection of free port addresses for that switch. Problems may arise when a private Fibre Channel device is attached to the switch. Private Fibre Channel devices were designed to only to work in private loops.

When the arbitrated loop is connected to the FL_Port, this port obtains the highest priority address in the loop to which it is attached (0x00). Then the FL_Port performs a LIP. After this process is completed, the FL_Port registers all devices on the loop with the SNS. Devices on the arbitrated loop use only 8-bit addressing, but in the switched fabric, 24-bit addressing is used. When the FL_Port registers the devices on the loop to the SNS, it adds two most significant bytes to the existing 8-bit address.

The format of the address in the SNS table is 0xPPPPLL, where the PPPP is the two most significant bytes of the FL_Port address and the LL is the device ID on the arbitrated loop which is connected to this FL_Port. Modifying the private loop address in this fashion, all private devices can now talk to all public devices, and all public devices can talk to all private devices. Because we have stated that private devices can only talk to devices with private addresses, some form of translation must take place. We show an example of this in Figure 4.9.

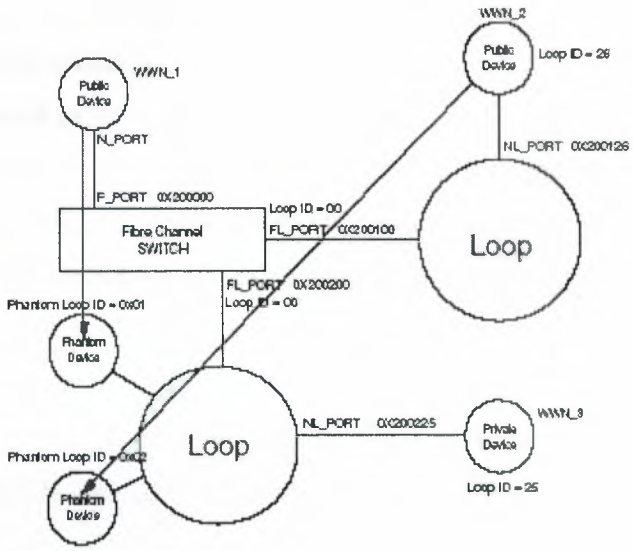

Figure 4.9 Arbitrated loop Address Translation

As you can see, we have three devices connected to the switch:

- Public device N_Port with WWN address WWN_1 on F_Port with the port address 0x200000
- Public device NL_Port with WWN address WWN_2 on FL_Port with the port address 0x200100. The device has AL_PA 0x26 on the loop which is attached on the FL_Port
- Private device NL_Port with WWN address WWN_3 on FL_Port with the port address 0x200200. The device has AL_PA 0x25 on the loop which is attached to the FL_Port

After all FLOGI and PLOGI functions are performed the SNS will have the entries shown in Table 4.3.

Table 4.3 SNS Entries

24 bit port address	WWN	FL_Port address
0x200000	WWN_1	n/a
0x200126	WWN_2	0x200100
0x200225	WWN_3	0x200200

We now explain some possible scenarios.

Public N_Port device accesses private NL_Port device

The communication from device to device starts with PLOGI to establish a session. When a public N_Port device wants to perform a PLOGI to a private NL_Port device, the FL_Port on which this private device exists will assign a "phantom" private address to the public device. This phantom address is known only inside this loop, and the switch keeps track of the assignments.

In our example, when the WWN_1 device wants to talk to the WWN_3 device, the following, shown in Table 4.4, is created in the switch.

Table 4.4 Phantom addresses

Switch port address	Phantom Loop Port ID
0x200000	0x01
0x200126	0x02

When the WWN_1 device enters into the loop it represents itself with AL_PA ID 0x01 (its phantom address). All private devices on that loop use this ID to talk to that public device. The switch itself acts as a proxy, and translates addresses in both directions.

Usually the number of phantom addresses is limited, and this number of phantom addresses decreases the number of devices allowed in the Arbitrated loop. For example, if the number of phantom addresses is 32 this limits the number of physical devices in the loop to 126 - 32 = 94.

Public N_Port device accesses public NL_Port device

If an N_Port public device wants to access an NL_Port public device, it simply performs a PLOGI with the whole 24-bit address.

Private NL_Port device accesses public N_Port or NL_Port device

When a private device needs to access a remote public device, it uses the public device's phantom address. When the FL_Port detects the use of a phantom AL_PA ID, it translates that to a switch port ID using its translation table similar to that shown in Table 4.4

4.4.5 QuickLoop

As we have already explained above, private devices can cooperate in the fabric using translative mode. However, if you have a private host (server), this is not possible. To solve this, switch vendors, including IBM, support a QuickLoop feature. The QuickLoop feature allows the whole switch or just a set of ports to operate as an arbitrated loop. In this mode, devices connected to the switch do not perform a fabric login, and the switch itself will emulate the loop for those devices. All public devices can still see all private devices on the QuickLoop in the translative mode.

4.4.6 Switching mechanism and performance

In a switched fabric, a "cut-through" switching mechanism is used. This is not unique to switched fabrics and it is also used in Ethernet switches. The function is to speed packet routing from port to port.

When a frame enters the switch, cut-through logic examines only the linklevel destination ID of the frame. Based on the destination ID, a routing decision is made, and the frame is switched to the appropriate port by internal routing logic contained in the switch. It is this cut-through which increases performance by reducing the time required to make a routing decision. The reason for this is that the destination ID resides in the first four bytes of the frame header, and this allows the cut-through to be accomplished quickly. A routing decision can be made at the instant the frame enters the switch.

An important criterion in selecting a switch is the number of frames that can be buffered on the port. During periods of high activity and frame movement, the switch may not be able to transmit a frame to its intended destination. This is true if two ports are sending data to the same destination. Given this situation, but depending on the class of service, the switch may sacrifice the frames it is not able to process. Not only does frame buffering reduce this likelihood, it also enhances performance.

Another great performance improvement can be realized in the way in which the 24-bit port address is built. Because the address is divided into domain, area and port, it is

possible to make the routing decision on a single byte. An example of this would be if the domain number of the destination address indicates that the frame is intended for a different switch, the routing process can forward the frame to the appropriate interconnection without the need to process the entire 24-bit address and the associated overhead.

4.4.7 Data path in switched fabric

A complex switched fabric can be created by interconnecting Fibre Channel switches. Switch to switch connections are performed by E_Port connections. This means that if you want to interconnect switches they need to support E_Ports. Switches may also support multiple E_Port connections to expand the bandwidth.

In such a configuration with interconnected switches, known as a meshed topology, multiple paths from one N_Port to another can exist. An example of a meshed topology is shown in Figure 4.10.

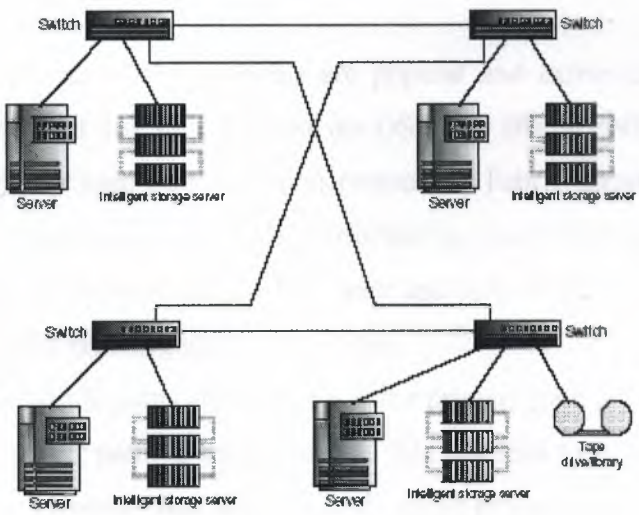

Figure 4.10 Meshed Topology Switched Fabric

4.4.7.1 Spanning tree

In case of failure, it is important to consider having an alternative path between source and destination available. This will allow the data still to reach its destination. However, having different paths available could lead to the delivery of frames being out of the order of transmission, due to a frame taking a different path and arriving earlier than one of its predecessors.

A solution to this, which can be incorporated into the meshed fabric, is called a spanning tree and is an IEEE 802.1 standard. This means that switches keep to certain paths as the spanning tree protocol will block certain paths to produce a simply connected active topology. Then the shortest path in terms of hops is used to deliver the frames and, most importantly, only one path is active at a time. This means that all associated frames go over the same path to the destination. The paths that are blocked can be held in reserve and used only if, for example, a primary path fails. The fact that one path is active at a time means that in the case of a meshed fabric, all frames will arrive in the expected order.

4.4.7.2 Path selection

For path selection, link state protocols are popular and extremely effective in today's networks. Examples of link state protocol are OSPF for IP and PNNI for ATM.

The most commonly used path selection protocol is Fabric Shortest Path First (FSPF). This type of path selection is usually performed at boot time and no configuration is needed. All paths are established at start time and only if the inter switch link (ISL) is broken or added will reconfiguration take place.

In the case that multiple paths are available if the primary path goes down, the traffic will be rerouted to another path. If the route fails this can lead to congestion of frames, and any new frames delivered over the new path could potentially arrive at the destination first. This will cause an out of sequence delivery.

One possible solution for this is to prevent the activation of the new route for a while, (this can be configured from milliseconds to a few seconds), so the congested frames are either delivered or rejected. Obviously, this can slow down the routing, so it should only

be used when the devices connected to the fabric are not in a position to, or cannot tolerate occasional out of order delivery. For instance, video can tolerate out of sequence delivery, but financial and commercial data cannot.

But today, Fibre Channel devices are much more sophisticated, and this is a feature that is not normally required. FSPF allows a fabric still to benefit from load balancing the delivery of frames by using multiple paths.

4.4.7.3 Route definition

Routes are usually dynamically defined. The fabric itself usually keeps only eight possible paths to the destination.

Static routes can also be defined. In the event that a static route fails, a dynamic route will take over. Once the static route becomes available, frames will return to utilizing that route.

If dynamic paths are used, FSPF path selection is used. This guarantees that only the shortest and fastest paths will be used for delivering the frames. We show an example of FSPF in Figure 4.11.

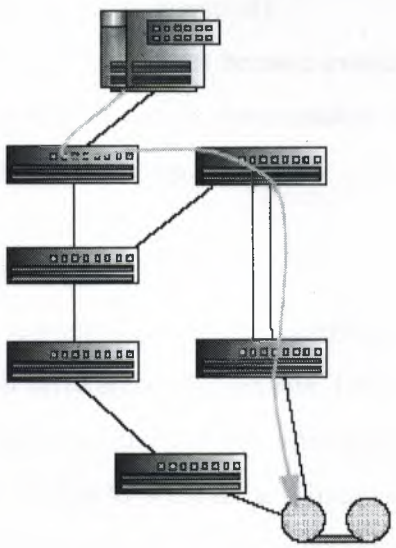

Figure 4.11 Fabric Shortest Path First

4.4.8 Adding new devices

Switched fabrics, by their very nature, are dynamic environments. They can handle topology changes as new devices are attached, or previously active devices are removed or taken offline. For these reasons it is important that notification of these types of events can be provided to participants (nodes) in the switched fabric.

Notification is provided by two functions:

- State Change Notification - SCN
- Registered State Change Notification - RSCN

These two functions are not obligatory, so each N_Port or NL_Port must register its interest in being notified of any topology changes, or if another device alters its state.

The original SCN service allowed an N_Port to send a notification change directly to another N_Port. This is not necessarily an optimum solution, as no other participants on the fabric will know about this change. RSCN offers a solution to this and will inform all registered devices about the change.

Perhaps the most important change that you would want to be notified about, is when an existing device goes offline. This information is very meaningful for participants which communicate with that device. For example, a server in the fabric environment would want to know if their resources are powered off

or removed, or as and when new resources became available for its use.

Changed notification provides the same functionality for the switched fabric as loop initialization provides for arbitrated loop.

4.4.9 Zoning

Zoning allows for finer segmentation of the switched fabric. Zoning can be used to instigate a barrier between different environments. Only the members of the same zone can communicate within that zone and all other attempts from outside are rejected.

For example, it may be desirable to separate a Windows NT environment from a UNIX environment. This is very useful because of the manner in which Windows attempts to claim all available storage for itself. Because not all storage devices are capable of

protecting their resources from any host seeking for available resources, it makes sound business sense to protect the environment in another manner.

Looking at zoning in this way, it could also be considered as a security feature and not just for separating environments. Zoning could also be used for test and maintenance purposes. For example, not many enterprises will mix their test and maintenance environments with their production environment. Within

a fabric, you could easily separate your test environment from your production bandwidth allocation on the same fabric using zoning.

We show an example of zoning in Figure 4.12.

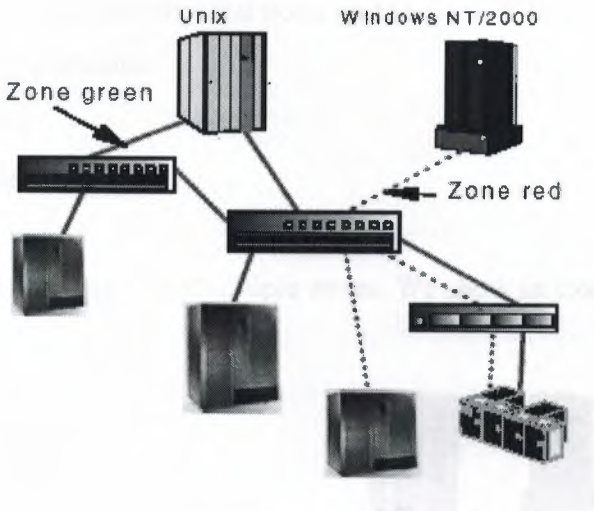

Figure 4.12 Zoning

Zoning also introduces the flexibility to manage a switched fabric to meet different user groups objectives.

4.4.10 Implementing zoning

Zoning can be implemented in two ways:

- Hardware zoning
- Software zoning

Hardware zoning

Hardware zoning is based on the physical fabric port number.

The members of a zone are physical ports on the fabric switch. It can be implemented in the following configurations:

- One to one
- One tomany
- Many to many

A single port can also belong to multiple zones. We show an example of hardware zoning in Figure 4.13.

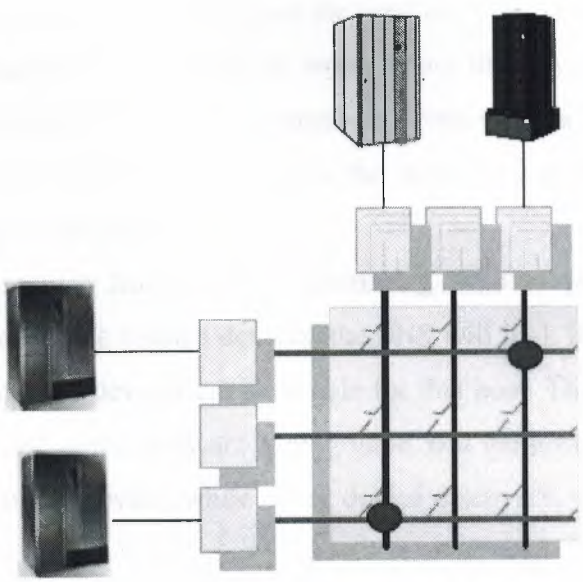

Figure 4.13 Hardware Zoning

One of the disadvantages of hardware zoning is that devices have to be connected to a specific port, and the whole zoning configuration could become unusable when the device is connected to a different port. In cases where the device connections are not permanent the use of software zoning is recommended.

The advantage of hardware zoning is that it can be implemented into a routing engine by filtering. As a result, this kind of zoning has a very low impact on the performance of the routing process.

Software zoning

Software zoning is implemented within the SNS running inside the fabric switch. When using software zoning the members of the zone can be defined with:

- Node WWN
- PortWWN

Usually zoning software also allows you to create symbolic names for the zone members and for the zones themselves.

The number of members possible in a zone is limited only by the amount of memory in the fabric switch. A member can belong to multiple zones. You can define multiple sets of zones for the fabric, but only one set can be active at any time. You can activate another zone set any time you want, without the need to power down the switch.

With software zoning there is no need to worry about the physical connections to the switch. If you use WWNs for the zone members, even when a device is connected to another physical port, it will still remain in the same zoning definition, because the device's WWN remains the same.

There is a potential security leak with software zoning. When a specific host logs into the fabric and asks for available storage devices, the SNS will look into the software zoning table to see which storage devices are allowable for that host. The host will only see the storage devices defined in the software zoning table. But the host can also make a direct connection to the storage device, while doing device discovery, without asking SNS for the informationit has.

4.4.11 LUN masking

Another approach to securing storage devices from hosts wishing to take over already assigned resources is logical unit number (LUN) masking. Every storage device offers its resources to the hosts by means of LUNs. For example, each partition in the storage server has its own LUN. If the host (server) wants to access the storage, it needs to request access to the LUN in the storage device. The purpose of LUN masking is to control access to the LUNs. The storage device itself accepts or rejects access requests from different hosts. The user defines which hosts can access which LUN by means of the storage device control program. Whenever the host accesses a particular LUN, the storage device will check its access list for that LUN, and it will allow or disallow access to the LUN.

4.4.12 Expanding the fabric

As the demand for the storage grows, a switched fabric can be expanded to service these needs. Not all storage requirements can be satisfied with fabrics alone. For some applications, the 100 MB/s per port and advanced services are overkill, and they amount to wasted bandwidth and unnecessary cost. When you design a storage network you need to consider the application's needs and not just rush to implement the latest technology available. SANs are often combinations of switched fabric and arbitrated loops.

4.4.12.1 Cascading

Expanding the fabric is called switch cascading. Cascading is basically interconnecting Fibre Channel switches. The cascading of switches provides the following benefits to a SAN environment:

- The fabric can be seamlessly extended. Additional switches can be added to the fabric, without powering down existing fabric.
- You can easily increase the distance between various SAN participants.
- By adding more switches to the fabric, you increase connectivity by providing more available ports.
- Cascading provides high resilience in the fabric.

- With Inter Switch Links (ISL) you can increase the bandwidth. The frames between the switches are delivered over all available data paths. So the more ISL you create, the faster the frame delivery will be, but careful consideration must be employed to ensure that a bottleneck is not introduced.
- When the fabric grows, the SNS is fully distributed across all the switches in fabric.
- With cascading, you also provide greater fault tolerance within the fabric.

4.4.12.2 Hops

As we stated in 4.3.2, the maximum number of switches allowed in the fabric is 239. The other limitation is that only seven hops are allowed between any source and destination. However, this is likely to change between vendors and over time.

We show a sample configuration that illustrates this in Figure 4.14.

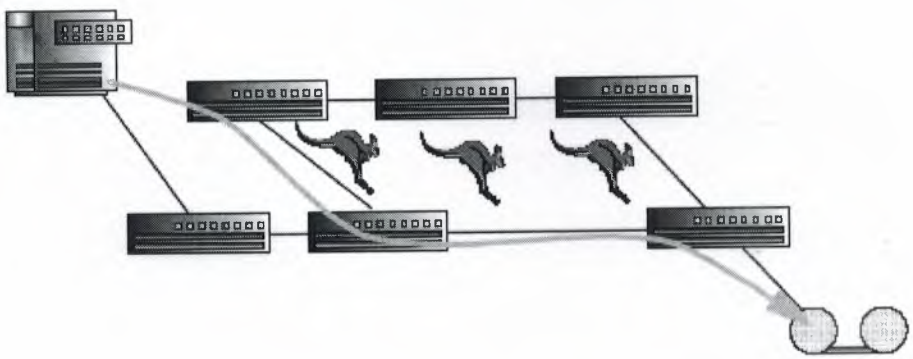

Figure 4.14 Cascading In Switched Fabric

The hop count limit is set by the fabric operating system and is used to derive a frame holdtime value for each switch. This holdtime value is the maximum amount of time that a frame can be held in a switch before it is dropped (Class 3) or the fabric is busy (F_BSY, Class 2) is returned. A frame would be held if its destination port is not available. The holdtime is derived from a formula using the error detect time-out value (E_D_TOV) and the resource allocation time-out value (R_A_TOV).

The value of seven hops is not 'hard-coded', and if manipulation of E_D_TOV or R_A_TOV was to take place, the reasonable limit of seven hops could be exceeded. However, be aware that this seven hop suggestion was not a limit that was arrived at without careful consideration of a number of factors. In the future the number of hops is likely to increase.

CONCLUSIONS

Information in this book was developed in conjunction with use of the equipment specified, and is limited in application to those specific hardware and software products and levels.

The advent of Storage Area Networks (SANs) is one such development. SANs can lead to a proverbial "paradigm shift" in the way we organize and use the IT infrastructure of an enterprise.

Standard interfaces for interoperability and management have been developed, and many vendors compete with products based on the implementation of these standards. Customers are free to mix and match components from multiple vendors to form a LAN or WAN solution.

Many products are now on the market that take advantage of FC's high-speed, high-availability characteristics. In the topics that follow, we introduce Fibre Channel basic information to complement the solutions that

When you design a storage network you need to consider the application's needs and not just rush to implement the latest technology available. SANs are often combinations of switched fabric and arbitrated loops

REFERNCES

[1] www.storage.ibm.com/ibmsan/index.htm IBM Enterprise SAN

[2] www.storage.ibm.com/hardsoft/products/fchub/fchub.htm IBM Fibre Channel Storage HUB

[3] www.pc.ibm.com/ww/netfinity/san IBM Storage Area Networks: Nefinity Servers

[4] www.storage.ibm.com/hardsoft/products/fcswitch/fcswitch.htm IBM SAN Fibre Channel Switch