Rivier College
Department of Math and Computer Science

Fibre Channel Technology for Storage Area Networks

David Norman
CS553A Introduction to Networking Technology
02Dec2001

Contents

Figures

Introduction

The problem.

Today's applications are rapidly overwhelming the capacity of networks and of storage space. In e-commerce, huge databases support electronic cataloging and ordering while large numbers of customers attempt to simultaneously access the information. As corporations grow and enter the international business environment, enterprise systems maintain corporate information across not only states but countries. To maintain and make available to all users that large amount of information reliably and in a timely manner is challenging to say the least. More and more feature films are incorporating digital effects. Video editing software, Computer Aided Drafting and photo-realistic rendering software are utilized to either modify a film or even create one from scratch. Even a few seconds worth of a film requires hundreds of megabytes of storage space. When teams of 20 animators/digital artists are trying to work on their own piece of a film, the burden on the storage and the network facilities are tremendous. Web sites that serve up streaming audio and or video are consuming more resources as the demand for these services go up. In addition to simply supporting bandwidth and storage increases, corporations now want to be able to safeguard their data. This typically entails making backups of data (to tape) and saving data off the corporate premises. This is an extremely small sample of the applications that are challenging the storage and networking architectures.

Traditionally, these applications have been supported by file servers with either large internal disks or disk farms directly attached to the server The disks are typically connected to the server via SCSI (Small Computer System Interconnect). The SCSI standard defines a high throughput parallel interface that is used to connect up to 7 peripherals (including the host adapter card itself) to the computer. Examples of these peripherals are scanners, CD (Compact Disk) players/recorders, digitizers, tape drives and as previously stated hard disks. This architecture has several limitations. The server can only access data on devices directly attached to it. If a server or any part of its SCSI hardware fails, access to its data is cut off. Also, SCSI supports a finite number of devices, therefore the amount of data a server can access is limited. If more storage space is needed, but there is no more room on the SCSI bus, expansion is no longer possible. SCSI, due to its parallel structure, has distance limitations as well. This requires that the storage be near the servers. These limitations are the driving force behind a new paradigm for data storage and access.

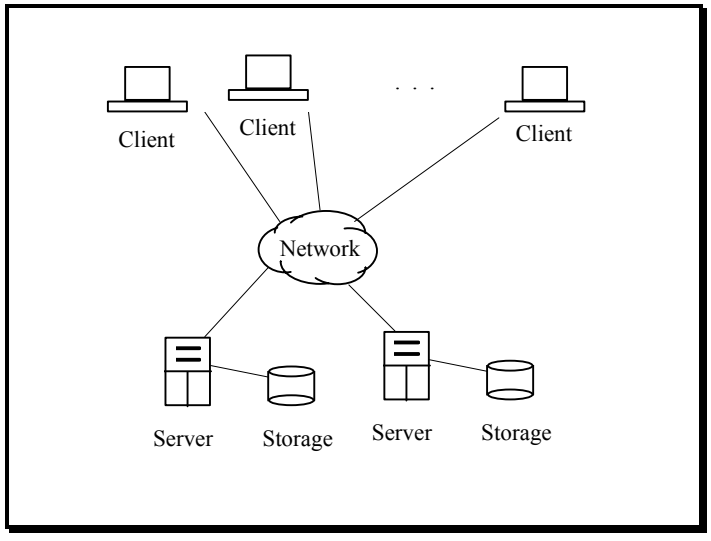

Figure 1: Traditional storage architecture with storage devices directly attached to servers.

What is a Storage Area Network?

Enter the Storage Area Network (SAN). A SAN is consists of a network that sits between the servers and the storage devices. Contrast this with the traditional architecture of direct attached storage and you can immediately see the difference. A SAN allows multiple servers to access any storage device. This helps to increase fault tolerance. If a particular server goes down, it does not take down a block of storage. A SAN has greater range than SCSI, that is, the storage devices do not need to be co-located with the servers. This is attributed to the architecture of the network that sits between the servers and the storage. The predominant technology used to implement SANs is Fibre Channel.

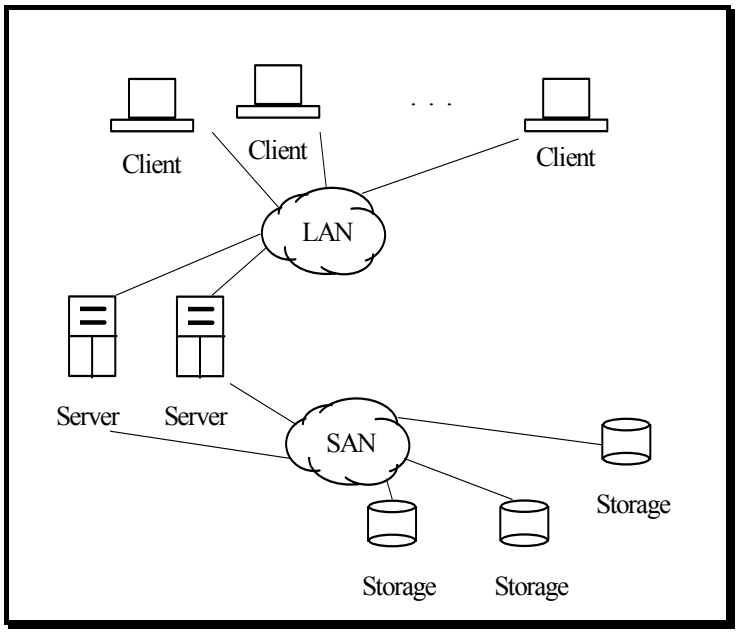

Figure 2 : Storage Area Network architecture with storage devices attached to servers through a network.

What is Fibre Channel?

Fibre Channel (FC) is a high speed serial interface for connecting computers and storage systems. FC is a mature technology having been approved and continually moderated by the American National Standards Institute (ANSI) T11X3 committee. New additions to the Fibre Channel specifications are continually being added. There is strong industry support for Fibre Channel products. Vendors such as McData, Vixel, Emulex, Qlogic and StorageTek develop Host Bus Adapters, FC-AL Hubs, Fabrics and management software. They work together with solution providers such as Compaq, Dell, HP, IBM and Sun to create entire solutions.

Why use Fibre Channel for SAN?

Fibre Channel is the predominant technology for implementing SANs today because it does the best job at meeting the requirements of today's applications. FC is fast, currently it supports speed of up to 1Gbps with 2, 4 & 10Gbps in the works. FC, being a network architecture, allows storage devices to be accessed by all servers on the SAN thus improving reliability. It supports several different topologies. The topologies have varying levels of cost to capabilities tradeoffs, thus allowing corporations to start with a small, reliable setup and scale up as needed. It also supports distances up to 10mi via fiber optic cable. This supports the capability of off-site data storage for disaster recovery and high speed local area networking between buildings on a campus or in the vicinity. Fibre Channel is a proven and fielded technology with many companies manufacturing FC components for SANs.

Introduction to Fibre Channel Technology

Fibre Channel technology is a networking technology which is designed to facilitate high speed data transfer between computer systems and storage devices. It supports several common transport protocols including IP and SCSI. The support for multiple protocols allows FC to merge high-speed I/O with networking functionality in a single package. To understand FC, we must discuss the concept of a network versus a channel.

Many networks currently in use are connectionless. Data along with addressing information are bundled into packets and transmitted over a shared medium. This methodology is relatively straight forward to implement and works relatively well. These networks are flexible and can handle both changes in configuration and also varying load. However, it requires a large amount of overhead to get the packets to go to the proper destination with some level of reliability. Reliability is a major concern as evidenced by the current concentration of work in the Quality Of Service (QoS) area. Also, as more users are added to the network and the resources are used up, the timeliness and reliability of correct packet delivery goes down.

The concerns that apply to networking are not exactly the same as the concerns that apply to hardware I/O. Specifically the kind of I/O involved with storage devices but also with network hardware as well. The primary concern here is performance. This is where the concept of a channel is appropriate. An I/O channel's purpose is to move data from one end of the link to the other with the least latency. Therefore, an I/O channel is typically hardware intensive with minimal software overhead. The tradeoff here is minimal error correction. In order to achieve this level of performance, channels operate in a very clearly defined domain. Reliability is achieved by minimizing errors by utilizing a rigorous and simple protocol. Therefore, channels are not as flexible in configuration as networks.

Fibre Channel is designed to combine the best features of both networks and channels. FC maintains the speed and low overhead of a channel while adding the flexibility (through connectivity) and the longer distances that are characteristic of a network.

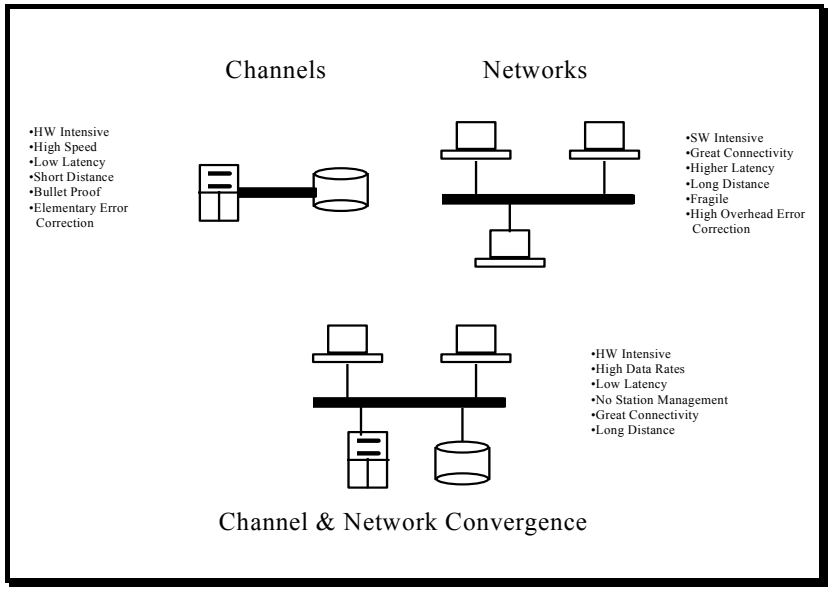

Figure 3: Fibre Channel is designed for Channel & Network Convergence

FC incorporates channel type features such as point-to-point dedicated connections and hardware intensive processing for buffer control, error handling and encoding/decoding. FC incorporates networking type features such as circuit and packet switching and use of data structures to pass control information in the data stream. Other features include full-duplex transmission, support for both optical and copper as a transport medium and the ability to serve as a transport for other protocols, both channel and network.

The Fibre Channel Architecture.

The generic Fibre Channel network is composed of one or more bi-directional point-to-point channels. The links support 1Gbps (or 100MBps) data rates in each direction. The transport media may be fiber optic cable, copper twisted pair or coax cable. The links in the FC network are between communication ports known as N_ports. N_port stands for Node Port where a node is a device on the FC network. The links may be point-to-point between N_ports or the may be set up as a Fabric. A Fabric consists of several N_Ports connected to a switch. Note: Ports on the switch are called F_ports. Finally, the ports may be "daisy chained" to form a ring. This is called an Arbitrated Loop (FC-AL). In this configuration the ports are referred to as L_ports. No switch is necessary for FC-AL. These basic layouts may be combined in different ways to create more complex topologies.

FC is typically realized in one of 3 topologies: Point-To-Point, Loop or Fabric. The Point-To-Point connection is the simplest type of connection. It can exist by itself or as a subset in a Fabric or Loop topology. An example of a point-to-point topology would be a FC tape unit connected directly to a server.

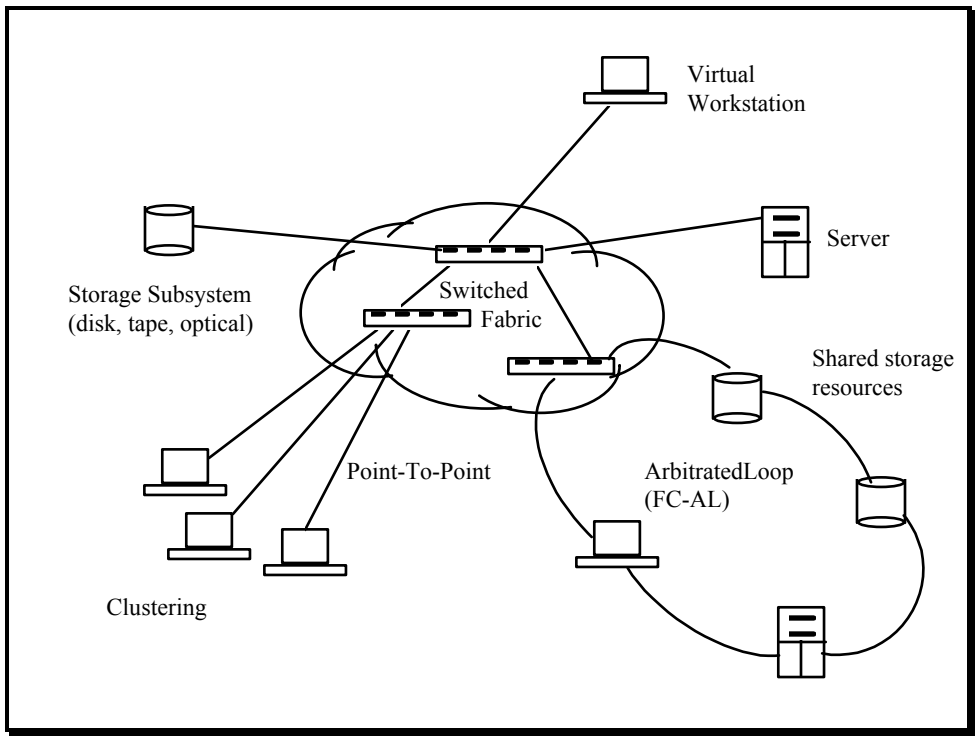

Figure 4: Fibre Channel Topologies.

The Arbitrated Loop (FC-AL) topology in its basic form is when L_ports are connected together. The primary advantages are that Loops are easy to setup and maintain and are relatively inexpensive. Additionally, several Loops may be connected via switch to share the load. Disadvantages inherent in a Loop are that Devices on a Loop must share the full bandwidth. The more devices, the less bandwidth is available to each. Also, they are subject to a failure similar to a series circuit. If one node on the loop goes

down, the entire loop is out. A means of mitigating the effect of a node taking out the entire loop is the use of a FC-AL hub. Circuitry in the hub allows bypassing of failed nodes. An extra side benefit of the hub is that it simplifies wiring. All the wires come into a central point instead of going from node to node.

The Fabric topology, in some ways can be considered the "ideal" FC topology. The Fabric was originally designed to be a generic interface between each node and the physical layer. In theory, it would not matter to the N_port whether it was connected to a loop, hub, switch or a storage device. It would simply work. In current context however, the definition of Fabric has come to mean actual FC switch hardware. Advantages of Fabric topology are high performance, excellent connectivity and redundancy. Also, several different media types may be tied together with a Fabric. On the other hand, Fabrics are relatively costly and configuring and maintaining them is not a simple task.

Fibre Channel uses a multi layer protocol architecture along the lines of the 7 Layer OSI Model. There are 5 layers. They are FC-0: Physical layer, FC-1: Encode/Decode layer, FC-2: Framing Protocol/Flow Control, FC-3: Common Services and FC-4: Upper Level Protocol Support. Additionally, there is another layer, which although is not typically considered part of the basic architecture is so important as to warrant mention. This is the FC-AL (Arbitrated Loop) layer.

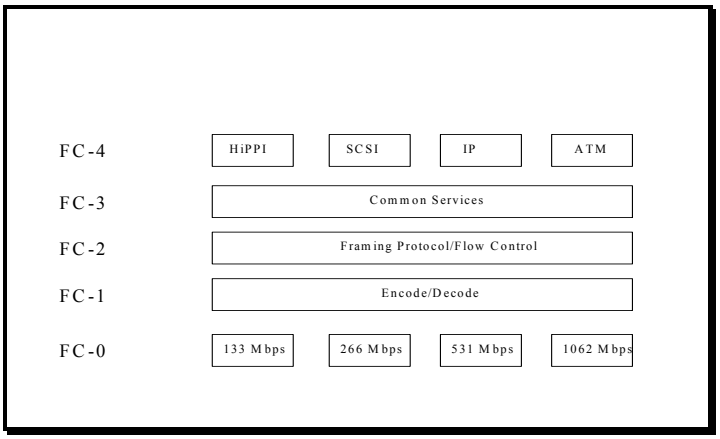

Figure 5: Fibre Channel Protocol Architecture.

FC-0 Layer

The FC-0 level describes the physical interface. The purpose of the physical interface is to take a stream of bits in at the transmitter, send them over the media, receive them and convert them back to bits at the output. Essentially, the physical interface represents a point to point link between two ports. It describes the requirements for the transmitter, the transport media and the receiver hardware. Layer 0 also describes the data rates that are supported by the different media types.

The FC-0 layer has an analog interface to the transmission medium and a digital interface to the FC-1 layer. The receiver must always be in an operational state but the transmitter may be on or off for several reasons. One of the safety features of the optical implementation requires the transmitter to stop transmitting if a fiber optic cable is disconnected or broken. This is to prevent any chance of the laser from causing eye damage. The transmitter is responsible for generating the transmission clock. Since the clock is encoded in the data stream, the receiver must be able to decode it. Also, the receiver includes a mechanism to detect the Special Character code known as the "comma" which is used for bye and word alignment.

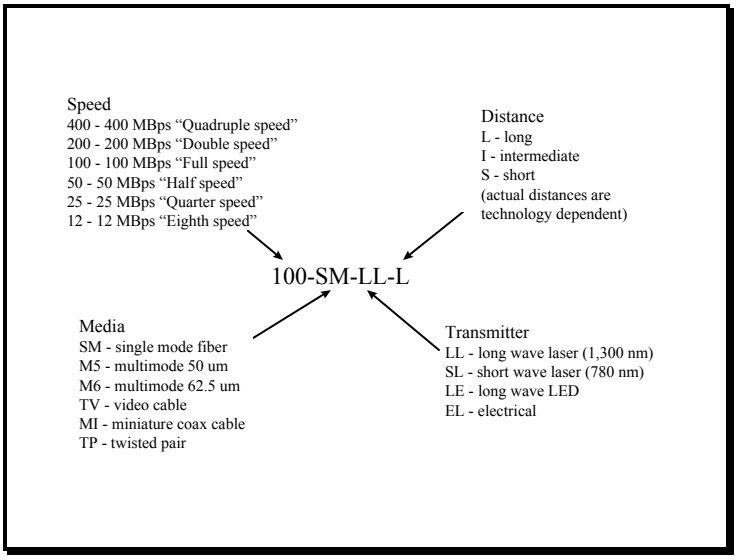

Figure 6: Nomenclature for describing FC-0 plant options.

FC-1 Layer

The FC-1 level describes the means by which user data is encoded for transmission and decoded at the other end. FC uses an 8/10 bit encode/decode scheme originally developed by IBM. This code accomplishes several things. It attempts to minimize errors by equalizing the number of 1's and 0's transmitted and not allowing more than 5 consecutive bits of the same type in a row. The code also allows for distinguishing "Special Characters" and also provides for simplifying byte and word alignment. Additionally, the evening out of 1's and 0's also has the effect of minimizing low frequency components in the transmitted signal. This allows for the design of relatively inexpensive transmitter/receiver circuitry which can perform at the required bit error rate (10^{-12}).

All data transmitted through a Fibre Channel network are sent as 10 bit chunks known as Transmission Characters. The 8/10 bit code used in encoding data for transmission over an FC link is a very powerful technique. 8 bits of user data at a time are translated to a 10 bit word. The encoding scheme ensures that the resulting word has no more than 6 total of one type (either 1 or 0) and that no more than 4 of any one type appear in a row. There is a special case where there are 2 characters of one type followed by 5 characters of the opposite type known as the "comma". This feature is important when looking at the characteristic of the transmitted signal. If there are many 1's, for example, this would appear as a DC voltage. This would make it extremely difficult for the receiver to know where bit boundaries are. If the receiver can't synchronize with the bit stream, it cannot pull the clock from the data stream. If it can't detect the clock, then it will never be able to use the data stream. This feature is the basis for keeping track of the Running Disparity.

A TC is composed of a 6 bit subgroup and a 4 bit subgroup. The running disparity is computed on a per-subgroup basis and can be positive, negative or even. Positive is when there are more 1s than 0s in a subgroup. Negative is more 0s than ones than 0s. Even is when there are equal amounts of each. For each TC, there is a corresponding value it may be encoded into for both positive and negative running disparities. The transmitter keeps track of the previous running disparity and will choose the appropriate encoding for the TC based upon keeping the running disparity as close to equal as possible. For example, if the previous disparity was negative, the encoding logic would encode the TC using its positive value.

Transmission Characters (TC) come in two flavors, Data Characters (DC) and so called Special Characters (SC). Data characters consist of user data encoded into TCs via the 8/10 bit encoding scheme. SCs, used for control, are distinguished from DCs via a special indicator known as the control variable. The notation

format of a TC is Zxx.y where Z is the control variable, xx is the decimal value of the binary number composed of the bits E, D, C, B and A. The y is composed of the remaining 3 bits, H, G and F. Note: the order is important when doing the converting. For all 256 data values, the encoding scheme produces valid TCs, that is, they meet the qualification of having the proper ratio of 1s to 0s. Of the SCs, only 12 meet that criteria and of the 12, only one is currently used. That is the K28.5 also known as the "comma". Once the TC has been transformed to the Zxx.y format, the Z, xx and y components are used to "look up" their corresponding encoded bit patterns. The following example show the conversion process for the SC K28.5. The process is exactly the same for a DC except that a different set of tables are used for the encoding (this is the purpose of the Z variable).

```
FC-2 byte notation:     0xBC  -- Special Character

                        -------------   -------------
FC-2 bit notation:      7   6   5   4   3   2   1   0       variable
                        1   0   1   1   1   1   0   0             K

                        ---------   -----------------
FC-1 un-encoded:        H   G   F   E   D   C   B   A             Z
                        1   0   1   1   1   1   0   0             K

                        Z   xx                  .y
                        -   -----------------   ---------
FC-1 Reordered for      Z   E   D   C   B   A   H   G   F
      Zxx.y notation:   K   1   1   1   0   0   1   0   1

Yields the TC:          K28.5

To get the encoded version, look up K28 in the 5B/6B encoding table for
SCs and look up .5 in the 3B/4B encoding table for SCs. At this point,
the hardware would use the running disparity to determine whether to
select the positive or negative version. For this example, we will say
that the previous disparity was positive.

                    5B/6B (negative)        3B/4B (positive)
                    --------------------    -------------
FC-1 encoded:       a   b   c   d   e   i   f   g   h   j
                    0   0   1   1   1   1   1   0   1   0
```

Figure 7: Example: Represent a byte with TC nomenclature then convert to its 10 bit encoded value.

Detection of valid characters actually begins in FC-0 where the hardware detects an incoming K28.5 (comma) SC. This character only appears at the beginning of an Ordered Set which is aligned on Transmission Word (TW) boundaries. A TW is a set of 4 TCs. After this character is received, the receiver can now derive byte and word order and is able to decode TCs.

Ordered Sets allow the FC-1 level to distinguish data from control information. The type of control information is designated by the 3 TCs directly following a K28.5. Currently, there are 3 types of Ordered Sets: Frame Delimiters, Primitive Signals and Primitive Sequences. Frame delimiters mark the beginning and end of frames and are used to specify the class of connection as well. Primitive Signals are used to indicate the Idle condition (maintain link when no data is flowing) and Receiver Ready (low level flow control). There are also special events used with the Arbitrated Loop topology only. Primitive Sequences relay the state of FC ports such as Not Operational, Offline, Link Reset and Link Response. As with Primitive Signals, there is also a set of sequences reserved for use with the Arbitrated Loop topology.

Error detection at the FC-1 layer occurs during decoding. There are two types of errors that can occur. The first is when 10 bits cannot be found in the encoding tables. This results in an invalid TC and a "code violation" is logged. The second type of error is when a valid TC is received, yet it was not the TC that was transmitted. This type of error occurs when due to an error in the transmission hardware, a bit or bits gets flipped resulting in a valid TC, but when the running disparity is computed, it is not the expected value (for example negative following negative instead of negative following positive).

FC-2 Layer

The FC-2 level is, by far, the most complex layer of the protocol. The major elements of FC-2 are the encapsulation of data using frames, flow control and classes of service. Additionally, FC-2 provides error control. In order to serve as a transport mechanism, FC must be able to encapsulate user data and deliver it to the intended recipient. Frames are the basic package used to encapsulate and transport the data. Generally speaking, there are two basic types of frames, the Data Frame and the Link Control Frame. Both types have the same basic format. A frame is composed of 6 basic sections.

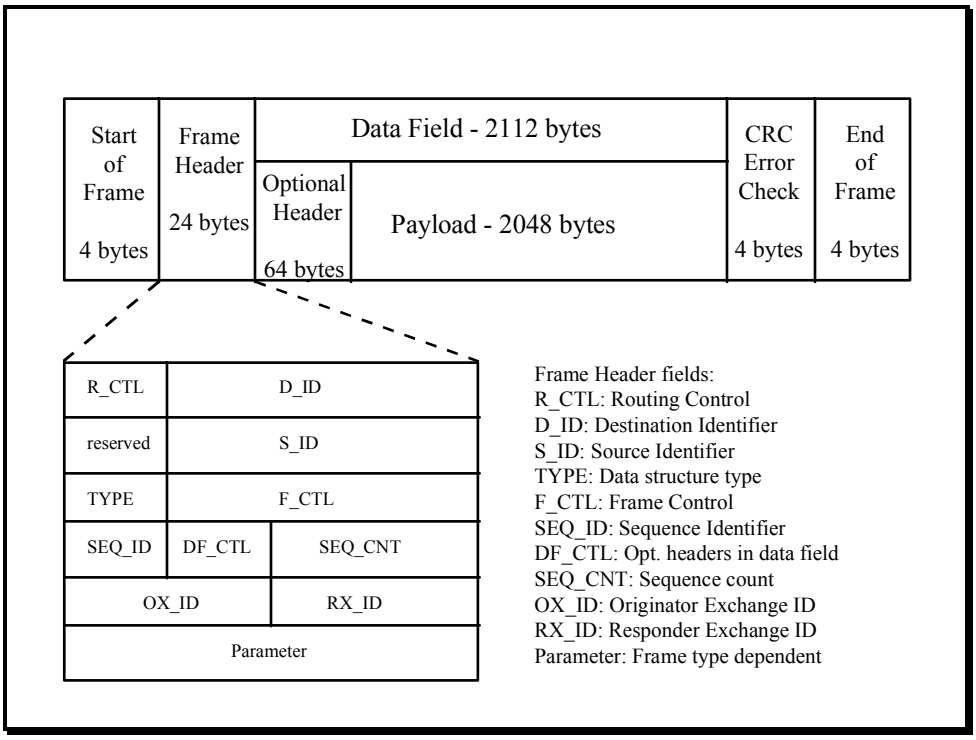

Figure 8: Frame and Frame Header formats.

The first section is the Start of Frame (SOF). The SOF is comprised of 4 bytes, the "comma" and 3 bytes indicating the type of connection service. Next is the Frame Header (FH). The FH contains information to control link operations, perform routing, do protocol processing and detect missing or out of order Frames. One of the major functions of the FH is to uniquely identify active Frames. A Frame is uniquely identified by a combination of the following fields: S_ID, D_ID, OX_ID RX_ID, SEQ_ID and SEQ_CNT. A Sequence Qualifier identifies active and open Sequences. The Sequence Qualifier is composed of all the fields listed above except the SEQ_CNT. The SEQ_CNT is used to identify the Frame's location within the Sequence. The Optional Header section can contain up to 4 types of headers. These are 1) Expiration Security Header (how long until Frame should be discarded), 2) Network Header (contains info pertaining to network related ULPs e.g. IP), 3) Association Header (associates Frames with different Exchange Id's) and 4) Device Header (contains other ULP info e.g. SCSI). Next comes the Payload. For a Data Frame, the payload contains the user Data. For a Link Control Frame, the payload is not used. The CRC field is used to verify the data integrity of the FH and Payload. Finally, the End of Frame (EOF) delimiter is an ordered

set that designates the end of the Frame content. Additionally, the EOF contains information about the validity of the Frame's content, whether to keep the connection open or normally close and if necessary, abort the connection.

R_CTL:	Routing Control - used for categorizing the Frame's function
D_ID:	Destination Identifier - address identifier of the Frame's destination port
S_ID:	Source Identifier - address identifier of the Frame's source port
TYPE:	Data structure type - categorization of the Frame's data
F_CTL:	Frame Control - control information on Frame handling
SEQ_ID:	Sequence Identifier - unique identifier for the Frame's Sequence
DF_CTL:	Opt. headers in data field - indication of optional header inclusion
SEQ_CNT:	Sequence count - number of the Frame within its Sequence or Exchange
OX_ID:	Originator Exchange ID - identification of Frame's Exchange at Originator
RX_ID:	Responder Exchange ID - identification of Frame's Exchange at Responder
Parameter:	Relative offset (applicable to Data Frame) -OR- Frame information (applicable to Link Control Frame)

Figure 9: Frame Header field descriptions.

One logical level up from the Frame is the Sequence. Sequences provide the means for ensuring data integrity of blocks of data that are transmitted and received. A Sequence is composed of a group of related Frames transmitted in one direction. As noted in previous sections, there are several other functions performed by the Start of Frame and End of Frame Ordered Sets. One of these additional functions is to initiate and terminate Sequences. Other information pertaining to the control of Sequences is found in dedicated fields in the Frame Header. Exchanges are groups of related Sequences. Exchanges are used to provide a means for reliably sending Sequences. They do this by detecting and recovering from Sequence errors.

Before data is sent across an FC Link, the N_port performs Login procedures. Before a port connects to a Fabric or another port, it is initialized with a default set of operating parameters. These are not typically optimal. The login procedures allow the port to determine what type of environment it resides in and to adjust its operating parameters based on what it finds out. There are login procedures for Fabrics and for other ports. Logging into the Fabric lets the requesting port know 1) what type of topology it is connected to 2) if attached to a Fabric, validates the requesting port's port identifier and 3) if attached to a Fabric, provides the Fabric's current operating parameters and buffer credit value (for flow control) to the requesting port. Logging into another N_port lets the requesting port know 1) the port's current operating parameters and 2) initializes the end-to-end credit value and if Point to Point buffer-to-buffer credit value (for flow control). When a port logs out, its settings are set back to the default values.

FC-2 controls the flow of Frames between ports so that receiver buffers are not overrun. A count of the number of buffers available is maintained by the Sequence Initiator (transmitter) and is used to throttle the transmission of Frames. There are two basic types of flow control. These are End to End and Buffer to Buffer. End to End flow control is used in cases of N_port to N_port communications. In this type, the receiver responds to all valid Frames it receives with an ACK Frame. In the case of no buffers available on the receiver, it responds with a Busy Frame. In the case of an invalid Frame, it sends a reject frame. The Sequence Initiator is responsible for maintaining the current amount of credit (EE_Credit_CNT). It decrements the credit count when it sends a Frame and increments it only when an ACK response is received. Buffer to Buffer flow control is used in the case of an N_port talking to a Fabric or in the case of an N_port to N_port connection in a Point to Point topology. In this case each side is responsible for maintaining its own BB_Credit_Count. As with End to End, the BB_Credit_CNT is decremented for every Frame transmitted, but is incremented only when it receives a R_RDY Ordered Sequence (Primitive Signal). This indicates that the receiver has resources available to accept another Frame.

FC-2 provides up to 5 Classes of Service (CoS). The different CoS represent different levels of delivery guarantee, bandwidth and connectivity. Different applications require different levels of these services. For

example, a tape backup might desire a long term dedicated connection (on the order of hours) with high bandwidth whereas a video editing application might require only a temporary connection (on the order of seconds - file transfer, or minutes - scene playback) and less bandwidth. As stated, there are 5 CoS but only 3 have been implemented.

Class 1 provides a dedicated connection between two ports including Fabric ports known as the Sequence Initiator and the Sequence Recipient. The connection has the full bandwidth available. Additionally, in most cases, a Class 1 connection does not allow other ports to access either of the participants. This ensures that the ports will not be busy. It also ensures that the Frames all arrive in the order they were sent which speeds up the re-assembly at the end. A Class 1 connection will remain active until either the Sequence Initiator or the Recipient closes it. This Class is good for Fabric type topologies, where the dedicated connection facilitates routing on a per-connection basis instead of a per-Frame. This simplifies Fabric operations by reducing its involvement because the N_ports can manage the link control. It also may result in utilizing less of the Fabric's resources. Note: There is a major option to Class 1 known as Intermix. It allows Class 1 and Class 2 Frames to be sent over a Class 1 connection during periods of inactivity.

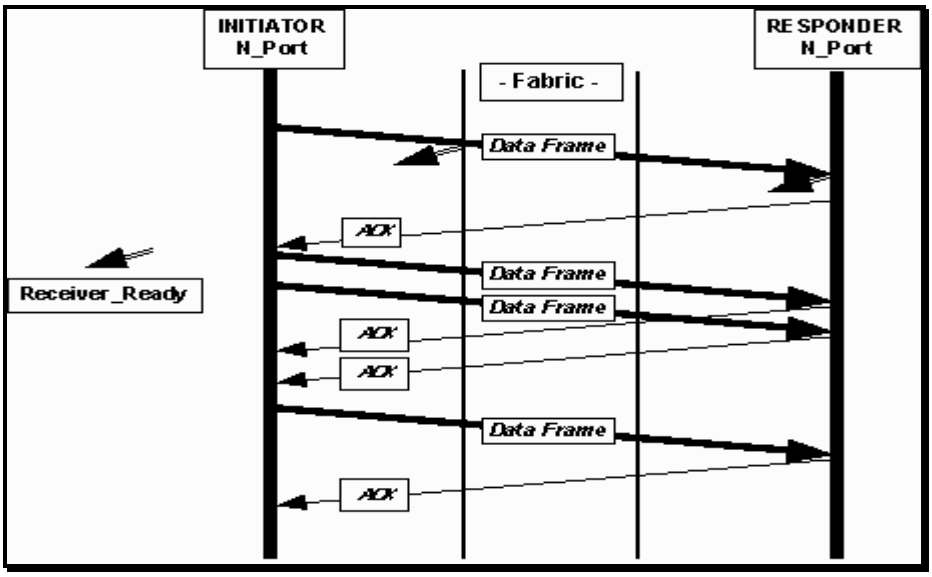

Figure 10: Class 1 data flow. Note R_RDY on Connect Request only.

Class 2 provides a multiplexed service which allows for routing and flow control on a Frame by Frame basis. Additionally, it provides acknowledgment of Frame delivery. This allows interleaving of Sequences over the single connection from multiple N_ports. However, it does not guarantee any resources. Frames are forwarded as resources allow. Frames may arrive in a random order depending on traffic conditions such as congestion or in the case of multiple routes. A Class 2 connection may specify in-order-delivery to ensure that Frames arrive in the proper order. Of course there is a tradeoff - the basic service requires more processing at higher levels whereas the in-order-delivery service will demand more of the lower levels. This Class is good for applications that want the transport to ensure data integrity. It is also good for applications that have smaller data transactions with bursty traffic.

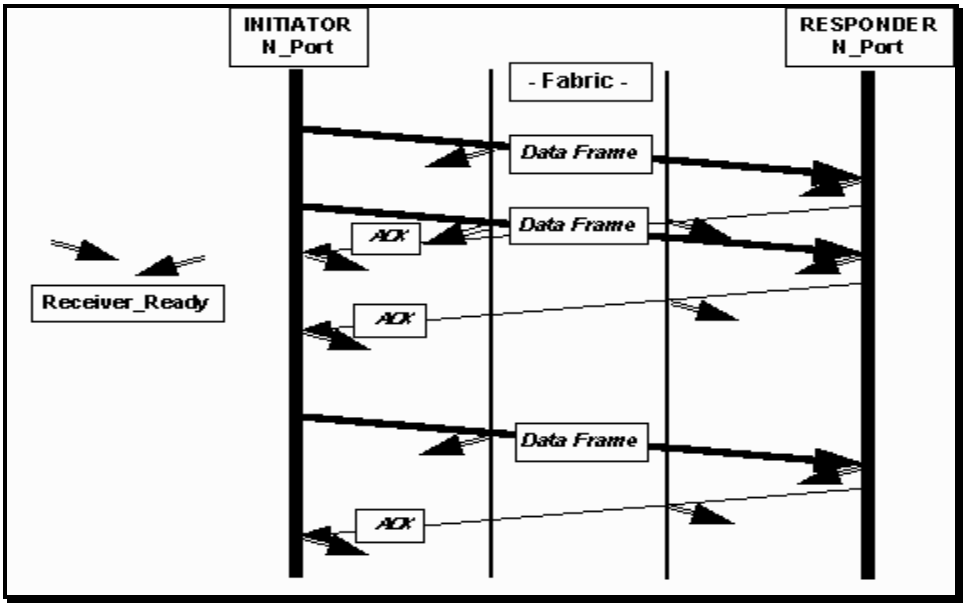

Figure 11: Class 2 data flow. Note the ACK for every Frame. Also R_RDY.

Class 3 provides a connectionless service with no acknowledgment. This means that there is no confirmation that a Frame was received by the destination N_port. This also means that there are no indications of Busy or Reject forms of acknowledgment either. This class breaks a major design objective of Fibre Channel which is to not deliver corrupt data (for example a Frame Header error) without some form of indication that it is corrupt. However, despite this and other potential pitfalls, support for a connectionless service is valuable enough that it is included. It essentially sacrifices reliability for performance. It is used in cases where the ULP it is supporting needs to perform its own error detection and correction. It is used widely in Arbitrated Loop (FC-AL) environments.

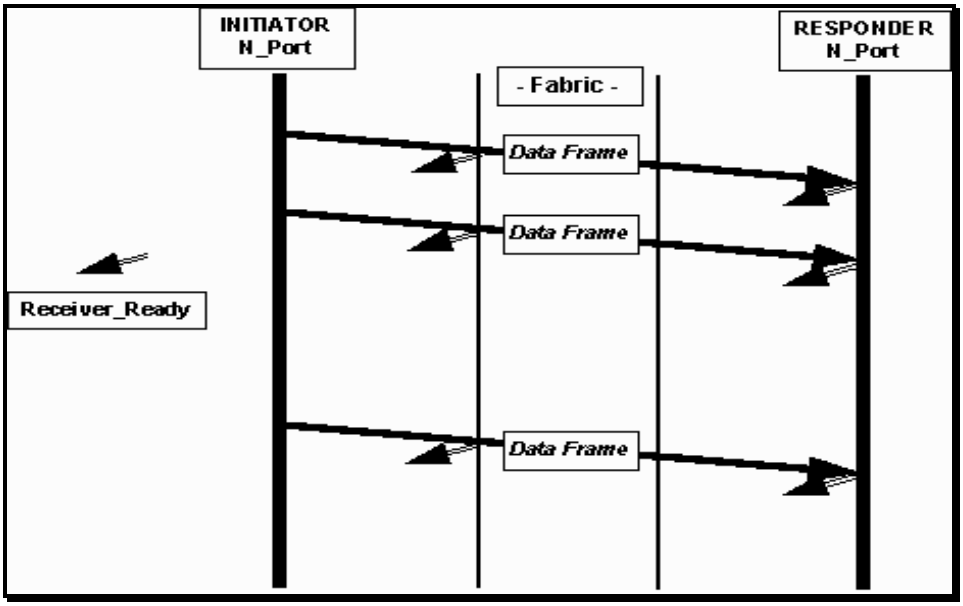

Figure 12: Class 3 data flow. Note the lack of ACK. Only R_RDY for link maintenance.

The FC-2 layer provides for error detection and recovery at the Sequence level. Error detection at this level is concerned with data integrity within Frames and Sequences and Sequences within Exchanges. The basic types of errors that can occur at this level are corrupt data in the control information in Frames, Frames out of order in a Sequence, dropped Frames, corrupt Sequence control information and Sequences out of order in Exchanges.

The basic method used by this layer to detect errors is the time-out mechanism. Initiators and Recipients expect specific events to occur within time-out periods in the delivery of Frames and Sequences. When an error is detected, the Exchange Error Policy (EEP) is used to determine what to do. EEPs specify whether to "discard" Sequences or to "process" them with the errors. In some cases, Sequences with certain errors are still valid to the application. Selection of the EEP to be used is negotiated on a connect operation.

Once an error has been detected and action has been taken, the last step is to perform recovery. Error recovery happens at the Sequence and Exchange level only, individual Frames are not recovered. This means that if 1 Frame is dropped from a Sequence, the recovery procedure may be (at minimum) to retransmit the entire Sequence. In addition to having FC-2 handle retransmission, the FC-4 or the ULP may control retransmission as well.

FC-3 Layer

The FC-3 level is not currently fully defined. Its intended purpose is to support "common services". The term "common services" means a service that would utilize multiple N_ports working together on a single node. An example would be "striping" where data is sent out of a node on several N_ports such as data acquired by a high speed data acquisition process being "striped" to several RAIDs. This would increase the effective bandwidth by sending data out in parallel.

FC-4 Layer

The FC-4 level supports the mapping of Upper Level Protocols (ULP) onto Fibre Channel data structures. This allows FC to be a lower level transport mechanism for ULPs. Because of its high speed, low overhead and reliability, FC can be used to carry channel, peripheral interface and networking protocols. There are mappings for several major protocols:

- SCSI (Small Computer Systems Interface)
- IPI-3 (Intelligent Peripheral Interface-3)
- HiPPI (High Performance Parallel Interface)
- IP (Internet Protocol) - IEEE 802.2 (TCP/IP) data
- ATM/AAL5 (ATM adaptation layer for computer data)
- SBCCS (Single Byte Command Code Set)

Generally speaking, the way that FC serves as a transport for ULPs is by mapping the ULP messages (known as Information Units) into FC Sequences and/or Exchanges. The flexibility of Sequences and Exchanges are what make mapping many existing ULPs to FC feasible. An important requirement necessary for transport of ULPs is the ability for the ULP to be able to excersize some level of control at the FC level as in the case of a connectionless service. Following is a brief description of how IP and SCSI are mapped and transported over FC.

Transporting IP over Fibre Channel is accomplished by using 2 Information Units. These are the IP datagram and the ARP datagram. The IP datagram actually does the work, moving between nodes on networks using the IP protocol stack while the ARP datagram is used during network configuration to map IP addresses to Media Access Control addresses (used for routing). These datagrams are mapped into Sequences. At this point they are in a format that the FC layers can work with. IP over FC uses a connectionless service. This service is provided by both service Class 2 and 3. Class 2 is typically chosen because it has better performance and reliability than Class 3 even though Class 3 is closer to a true UDP

than Class 2. IP packets are combined into Sequences for transmitting. Error handling is set up to discard Sequences with no Retransmit upon detecting a Frame error. This allows the IP and TCP levels to control the retransmission. They will also be responsible for controlling the order for the Sequences. Finally, in order to perform address resolution (which in an Ethernet network is normally done by the nameserver), a dedicated ARP server must be set up at a "well known" address. The ARP server performs the translation of IP addresses to FC addresses so that routing can be achieved. The ARP server must be at the well known location due to the limitations of FC's broadcast ability (it can't locate the nameserver just by broadcasting as is done in an IP network, it must know apriori where it is).

A primary use of FC technology is as a transport for SCSI. It is so significant that an acronym has been created for it. FCP stands for Fibre Channel Protocol for SCSI. The transport is accomplished by wrapping SCSI command, response, status and data blocks. SCSI operations have bi-directional communications scheme and thus are suited for operations using FC Exchanges. When transporting SCSI, N_ports are termed FCP_ports. FCP_ports can be a part of any topology although FC-AL is the most prevalent. Once the SCSI operation (ex: read operation) is mapped, the FCP_port originates an Exchange and initiates an FCP_CMND Sequence that transports a SCSI command. This Sequence is then transmitted to the target FCP_port (ex: SCSI disk drive) where it is received and handled. When the operation is finished, the SCSI status is returned to the initiator via an FCP_RSP Sequence. Note that the FCP_CMND and FCP_RSP plus any data transfer sequences for the read make up an Exchange. At this point the FCP_RSP Sequence will contain the EOF Ordered Set which will terminate the Exchange. The SCSI status is passed up to the application that initiated the I/O to indicate that the I/O is complete and is ready for the next request. Error handling for SCSI operations is handled by the normal mechanisms in the SCSI protocol. If a Frame error occurs at the FC level, the entire Exchange is discarded. If this occurs, SCSI is notified and is responsible for requesting a retry if one is required.

Fibre Channel Equipment

As the market for Fibre Channel networks grows, so too does the industry that manufactures the equipment. Consequently, there is a strong field of manufacturers that build FC equipment. They are continually improving on the products using new technology to increase the speed and reliability of their products. The main components that are utilized in a FC network are Gigabit Interface Converters (GBICs), Host Bus Adapters (HBAs), FC Redundant Arrays of Independent Disks (RAIDs), FC JBODs (Just a Bunch Of Disks), FC-AL Hubs, FC Switches and FC to SCSI Bridges.

Gigabit Interface Converters are small transceivers used in interconnecting devices. GBICs can handle a variety of media. This means that a GBIC can be used with optical or electrical cabling. GBICs take an optical cable or electrical cable (such as a DB9 connector) in on one side and have an interface on the other side to connect with other equipment. Most FC equipment such as switches and hubs can accept GBICs. They are simply "plugged into" ports in the equipment. Advances in laser technology have lead to the replacement of the original lasers (which were of the same type as Compact Disc lasers) with a newer technology Vertical Cavity Surface Emitting Lasers (VCSEL). VCSELs are less difficult to manufacture, consume less power, radiate less heat and are more stable than the older CD type lasers. Another advance is in the overall network management area. GBICs have a specified feature, Serial ID which permits the GBIC to answer queries about identifier information such as serial #, manufacture date, manufacture id etc. GBICs can also provide limited status information such as power consumption. This all depends on the vendor implementation. It also implies that devices such as HBAs, switches or hubs can perform the query.

Host Bus Adapters are similar to LAN Network Interface Cards (NICs). It is a card that is connected to the host computer's (either server or workstation) bus. In addition to providing a physical means of connecting a node to the FC network, it also implements the FC-0 through FC-4 layers. At the FC-0 layer, the HBA has connectors that will support various media types. Typically, a given HBA has connector for a single type of media, manufacturers produce different variants of the HBA to support different media. The HBA has transmitter/receiver circuitry, clock and data recovery circuitry, serializing/deserializing circuitry

and retiming circuitry. FC-1 is realized by the 8/10bit encoder/decoder logic, flow control logic and error detect logic (Ordered Set logic). Also, if the HBA supports FC-AL, it has a Loop Port State Machine. At the FC-2 layer is support for Frame segmentation/reassembly, class of service and credit algorithms (for flow control) and port and fabric login/logout link services. The FC-4 layer is implemented in software drivers provided by the HBA manufacturer. Advances in HBAs are being made in the area of simplifying the physical construction of the board by putting more and more functionality on Application Specific Integrated Circuits as well as offloading some of the functionality onto other components such as FC RAIDs.

Redundant Arrays of Independent Disk storage subsystems are a mature technology that provides large amounts of storage and safeguarding of data from loss by specifying methods of storing data to multiple disks. The RAID specification provides for different levels of security versus performance. A RAID enclosure consists of a RAID controller, an internal bus and several disk drives. A FC attached RAID is a RAID which has a FC interface in front of the RAID controller. SCSI-3 commands are sent to the RAID controller via the FC network utilizing the FCP for SCSI ULP. While this is the typical configuration for FC RAIDs, it is also possible to incorporate FC technology behind the RAID controller. This means that there is a small FC-AL network within the RAID enclosure. The RAID controller and the FC enabled disk drives are all nodes on this small internal FC network. RAIDs can be added to the network with minimal impact on resources. For example a RAID only appears as 1 NL_port on a FC-AL network. Since a FC-AL network can have only 126 nodes, adding several FC RAIDs has little impact on available ports while adding a huge amount of storage. Additionally, FC RAIDs are not tied to a specific server as are SCSI RAIDs or JBODs. This makes them very good for clustering applications.

FC JBODs (Just a Bunch Of Disks) is essentially a “bunch of” FC enabled disk drives that are daisy chained to form a FC-AL network. However, the loop is not a complete loop. There are input an output ports which must be connected to a server. Note: JBODs typically have 2 loops for use either as 2 separate loops or for redundancy. Due to their “incomplete loop” nature, JBODs have a large impact on the FC-AL loop’s resources. This is because each disk in the JBOD is counted as an NL_port. JBODs may be equipped with a controller that can configure the disks and the loops. If configured for 2 loops it is possible to divide the drives into separate partitions, one on each loop. Additionally, JBODs can be used with RAID software which can increase the performance and also to provide the redundancy for safeguarding data. However, in this case, the JBOD can only be accessed by the server it is connected to. In general, JBODs are much cheaper than RAIDs but are more limited in their capabilities as well.

FC-AL Hubs in principal take the basic loop topology and move it inside a box. One consequence of this is that cabling becomes much simpler. Instead of having cables running between devices, cabling simply runs from the device to the central location of the hub. Hubs have auto bypass circuitry which enable the loop to maintain its integrity if a port is empty or a device fails. This feature also allows the loop topology to be self-configuring. A limitation of the hub/loop implementation is that there can only be 126 NL_ports on the loop. When other hubs are added to the loop, each device on the new hub takes up an available NL_port. With the self-configure ability, FC-AL hubs take some of the burden of configuration of the network administrators. Hubs are required to only implement the capabilities and features found in the FC-0 layer and FC-AL series of specifications. Vendors must meet these basic specifications but then may build in other value added features such as more or less number of ports, fixed media ports (optical or copper) vs GBIC based ports, management features etc. so they may target specific markets from entry level through enterprise class level configurations.

FC Switches are high speed routing engines that provide full bandwidth to every port on the switch simultaneously. Switches, unlike hubs, perform services such as Fabric login/logout and flow control at each port. Switches use two methods to route frames, “cut-through” and “store-and-forward”. In cut-through, only the header must be processed in order to determine where the Frame will be sent. The D_ID or destination address is read from the Frame Header and is used to route the Frame. The store-and-forward method buffers the entire Frame before routing it. Switches have latencies in the low 10s of microseconds per Frame range with newer products pushing into the nanosecond range. Switches provide buffering capability to support anywhere from 2 to 16 Frames per port. FC switches support different port types for

attaching N_ports (nodes), E_ports for fabric expansion and NL_ports for loop attachment. Fabric expansion may be achieved by several means. In one method, switches may be simply cascaded by using an E_port to connect to another switch. This method is a relatively inexpensive means of adding more switched ports but can fall prey to bottlenecking at the E_port. Another method has a switch with all ports configured as E_ports which acts as a dedicated interconnect. This method is more expensive but provides much greater expansion capability with minimal problems due to bottlenecking. As with the other FC products, manufacturers produce equipment with differing levels of capability to target different markets.

FC to SCSI Bridges provide a link between a FC network and legacy SCSI devices. A bridge provides a FC interface on one side and several SCSI ports on the other. The bridge must convert the physical signals from whichever FC media they came across to the SCSI bus format. Also they must convert the SCSI-3 serial protocol to the SCSI protocol of the SCSI devices (SCSI-2, SCSI wide, SCSI ultrawide etc.). This equipment is an intermediate solution that will decline as newer native FC devices gain market share.

Fibre Channel As Applied to SAN

Introduction to SAN

A Storage Area Network (SAN) is a high performance network with the primary goal of transferring data between computer systems and storage devices and between multiple storage devices. A SAN consists of a communications infrastructure which provides physical connections and transport capabilities and a management layer which organizes the connections storage devices and computer systems so that data transfer is secure and robust.

The SAN provides a common link between multiple servers and storage devices. This is beneficial from a the standpoint of access to data and redundancy for data safeguarding. Having the storage on a network instead of directly attached to a server allows much greater flexibility in scaling server and storage. Additionally, having the storage devices on a dedicated network removes storage specific traffic, such as in the case of tape backup, entirely off the local area network (LAN). This reduces the impact on users and other traffic on the LAN.

Scalability allows SANs support small users up to large enterprises. A small FC-AL loop can economically support a small business leaving room to grow. An enterprise covering a large campus may use long distance links, FC-AL loops, multiple switches, large banks of storage accessible by clusters of servers and serverless/network-less backup capabilities.

SAN enjoys a large support base with industry standards associations and many manufacturers of networking and storage products. The Storage Network Industry Association and individual vendors in addition to numerous World Wide Web based resources provide a host of information on how to build and manage SANs, what equipment is available and new technology developments applicable to SAN.

Fibre Channel for SAN

While SANs are not required to use Fibre Channel technology, FC is currently the predominant technology used to implement SANs. There is a large industry base of FC equipment manufacturers. The reason for this is that FC technology provides numerous advantages which enable SANs.

The FC-AL topology is the most commonly used configuration for FC SANs. The primary reason for this is because Switched Fabric SANs are significantly more expensive. FC-AL is scalable and therefore can

suit small to medium size configurations. One of the tradeoffs using FC-AL is price to performance. FC-AL is a shared medium which means all traffic on the network must share the bandwidth. However, large storage configurations are possible by keeping the number of servers down. The servers directly contribute to the level of activity on the loop and therefore the bandwidth degradation can be controlled by controlling the number of them. FC-AL loops may be cascaded to increase the number of ports available for devices.

While a "true FC-AL loop" is very simple to construct, it is not the most desirable loop configuration for implementing SANs. More desirable is the use of a FC Hub "star topology" with the loop in the center. This configuration can simplify both the physical installation as well as troubleshooting problems. Additionally Hubs have the capability of bypassing empty or bad ports. This feature makes the loop self configuring and allows hot-swapping of devices. This type of configuration is very popular due to its simplicity, flexibility and low cost.

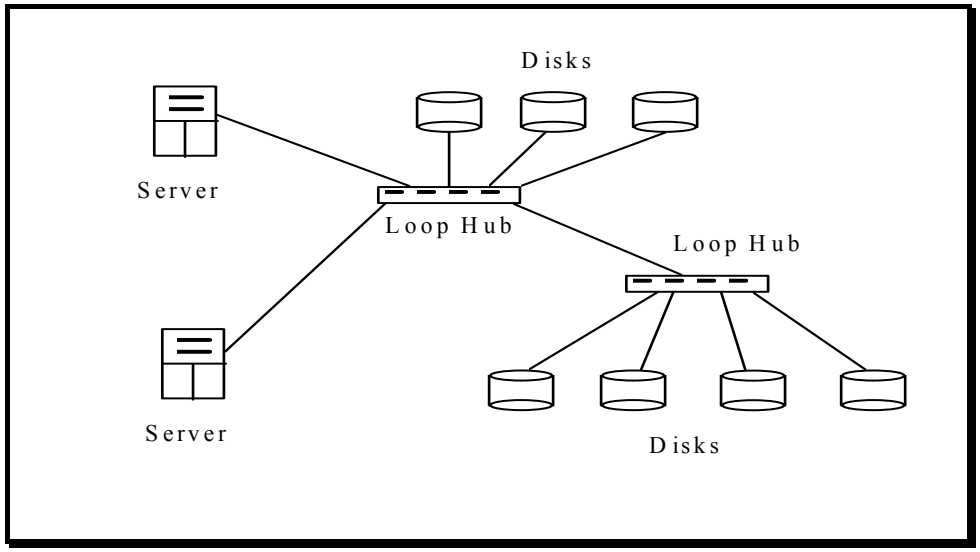

Figure 13: Cascaded Arbitrated Loop hubs.

The Fabric topology has had a slow start. This is due to several factors. Switches are very complex devices and thus have required more development time to produce a viable product. This also means that on a per-port basis, switches are much more expensive than FC-AL hubs. However, switches are essential for large scale, high speed configurations. They provide a full bandwidth connection for each port. As is the case with hubs, switches have the capability for self-configuration allowing for minimal disruption to SAN operations.

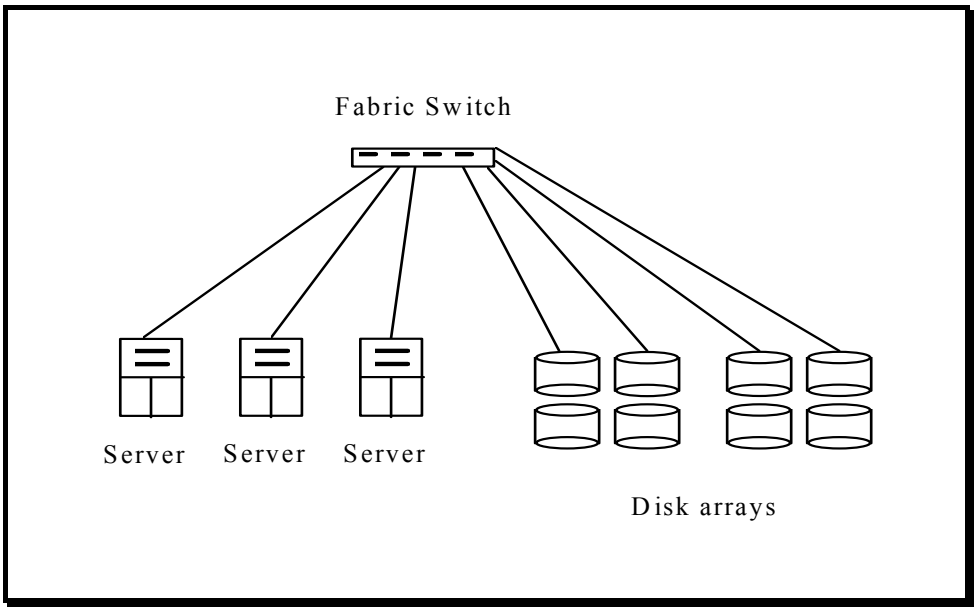

Figure 14: A simple Switched Fabric topology.

Switches also provide simple nameserver services. Since in a fabric topology, there may be up to 16 million address spaces, performing target location by logging on to all ports on every connection would be impossible. The nameserver is a small database stored in the switch that contains information about nodes in the fabric. This information can be used to more quickly locate devices in the fabric. Being able to quickly locate and connect to a device in a large fabric configuration is essential for implementing SANs.

Switches incorporate functionality to be cascaded. This is important for switch networks to be able to be expanded. Switches may be interconnected in a "meshed topology". This topology supports multiple paths to all the devices in the mesh. The switch in a fabric topology is capable of supporting all other types of topologies. This adds to the complexity and cost of the fabric topology but in turn is what gives it the flexibility to support many configurations desirable in a SAN. And with introduction of ASICs into switch design, performance and cost are being reduced. Considering that they tie together hundreds of thousands of dollars worth of servers and storage devices, FC interconnect devices, even expensive ones such as switches, are among the least expensive components in a SAN.

Conclusion

Storage Area Networks answer the information infrastructure and application needs of today's corporations both large and small, government or civilian. SANs provide scalability for cost effective configurations as well as expandability. They provide flexibility for ease of implementation and maintenance. SANs are also reliable and redundant, providing integrity and security for data. SANs reduce traffic on standard LANs by moving storage specific traffic off the LAN and onto the SAN. Finally, SANs allow clusters of storage devices to reside in different geographic locations for support of offsite data storage.

Fibre Channel is a mature networking technology that is ideally suited for SANs. FC is a gigabit technology supporting speeds up to 1gigabit per second with faster rates being realized in the future. FC supports different transport media such as copper for lower cost lower capability configurations or fiber optics for greater speed and distance at a higher cost. FC products support a SANs need for reliability by incorporating self-configuring capabilities that allow reconfiguring of networks, faulty equipment isolation and maintenance of the network all with minimal to no impact on SAN operations.

Both SAN and FC technologies enjoy wide industry support with industry associations, well tested standards and multiple vendors. Storage Area Networks using Fibre Channel technology will be with us for some time to come.

Bibliography

Books, White papers, Tech journals:

Benner, Alan F. “Fibre Channel: Gigabit Communications and I/O for Computer Networks”. McGraw-Hill, New York, 1996.

Clark, Tom. “Designing Storage Area Networks: A Practical Reference for Implementing Fibre Channel SANs”. Addison-Wesley, Pearson Education, Upper Saddle River, NJ, 2001.

Emulex Corp. “Fibre Channel: The Future of High-Speed Connectivity”. Emulex Corporation, Costa Mesa, CA.

Nowak, Stephen G. “Fibre Channel Basics”. StorageTek, Louisville, CO. 1999.

Shultz, Greg. “Introduction to Fibre Channel SANs”. INFOSTOR June 2001. Penwell Corp, Tulsa OK 2001.

Stallings, William. “Data and Computer Communications”. Prentice-Hall Inc., Upper Saddle River, NJ. 2000.

Web Resources:

The Fibre Channel Industry Associaction:
www.fibrechannel.com

The Storage Network Industry Association:
www.snia.org

National Committee for Information Technology Standards, Technical Committee T11
http://www.t11.org/index.htm

Internet Engineering Task Force - IP over Fibre Channel Charter:
http://www.ietf.org/html.charters/ipfc-charter.html

allSAN.com Online SAN resource:
http://allsan.com/sanoverview.php3

University of New Hampshire Interoperability Lab Fibre Channel Tutorial:
http://www.iol.unh.edu/training/fc/fc_tutorial.html

The High Speed Interconnect page at CERN Fibre Channel Overview:
http://www1.cern.ch/HSI/fcs/spec/overview.htm#b1