



IP Storage Inside White Paper

- IP Storage Basics: iSCSI, FCIP, iFCP
- NAS and how it differs
- IP Storage in your IT Environment
- iSCSI Deep Dive

Preface

The Storage Networking Industry Association (SNIA) promotes the use of storage networking systems by providing vendor-neutral architectures, educational avenues and other services to encourage the effective use of IP-based storage technologies. This whitepaper about IP storage is just one example of the SNIA's commitment.

This paper is aimed at a rather wide audience ranging from executives to technical specialists. For this reason, it is presented in two sections.

- The first section is a high-level overview of what is IP Storage, why consider using IP Storage and briefly discuss how IP Storage is different from other storage networking technologies. A significant portion of the material presented comes from earlier Storage Networking World Conferences, particularly the “*IP Storage Technologies*” presentation by David Dale and Peter Hunter and the presentation of “*NAS and iSCSI Technology Overview*” by Wolfgang Singer.
- The second section contains a more technical discussion of the three main technologies deployed in IP Storage; namely iSCSI, iFCP and FCIP. All three of these protocols are now fully ratified and vendors have a wide range of products in the market that are utilizing them.

About the SNIA IP Storage Forum (IPSF)

The SNIA IP Storage Forum offers a vendor-neutral environment for end users to become informed on the current and future directions of IP-based storage technology. The IPSF is composed of leading IP storage vendors driving the broad adoption of IP-based SAN storage solutions around the world. The IPSF goals are to drive the adoption of IP-based Storage Area Network (SAN) solutions through global marketing initiatives and end-user education. For more information, visit www.snia.org/ipstorage.

About the SNIA Europe IP Storage Initiative

The SNIA-Europe IP Storage Initiative is the sister organization of the IP Storage Forum and focused on driving the broad adoption of IP-based SAN storage solutions in Europe through evangelism, education and promotion. For more information, visit <http://www.snia-europe.org/ipstorage>

Acknowledgements

We would like to take this opportunity to acknowledge the many contributions that contributed many hours, dedication and persistence to make this document possible.

- The SNIA's IP Storage Forum
- Jim Blue, IBM
- David Dale, NetApp
- Janice Kall, Stonefly Networks
- Doug Rainbolt, Alacritech
- Wolfgang Singer, IBM

Introduction

Today more than ever, business is dependent on the constant availability of business applications and the rapidly growing data they create. This is driving increasingly stringent requirements of IT organizations, as they struggle to maintain and improve quality of service while operating with tight IT budgets, an already lean IT staff, and limited time available to perform data management functions.

IP SANs featuring contemporary iSCSI-based storage systems can provide a new network storage architecture that meets the availability, recoverability, and manageability requirements your business applications, with the cost structure and ease of use appropriate for Windows.

Thousands of IT organizations around the globe have been upgrading the storage infrastructure supporting this class of applications to a contemporary IP SAN to better support their business goals. This has accelerated their adoption of IP Storage solutions to the point that the technology entered the IT mainstream. Today, you can find iSCSI-based IP SAN deployments in IT production environments of large, medium and small enterprises around the globe, and the existence of an ever-increasing number of customer references and broad platform vendor support is further accelerating the rate of deployment.

Consequently, over the past few years we saw enormous growth in the market for IP-based SAN storage technology, as leading server and storage vendors released products, and IT organizations around the world installed IP Storage solutions to consolidate, simplify, bullet proof, and reduce data management costs in their distributed business critical applications environments.

The market for IP Storage technologies continues to be the fastest-growing segment of the storage market, growing from an estimated 2500 IP SANs deployed at the end of 2004 to more than 10,000 IP SANs deployed by the end of 2005 – and more than 25,000 IP SANs by the end of 2006. This growth is expected to continue to accelerate through 2007 and beyond, as iSCSI-based SANs continue to replace direct-attached storage in the Windows market, and as a broad range of robust affordable IP SAN solutions become available for the other host OS environments.

The key factors driving this growth have been the continuing need for IT organizations to do more with less – less capital cost, fewer admins per Terabyte, less complexity – coupled with continuing data growth, emerging best practices requiring strict and comprehensive data retention, and the recognition that even tier 2 applications are mission critical in today's 7x24 world. Interestingly, these factors affect enterprises of every size, and most IT managers now recognize that yesterday's direct-attached storage architecture simply cannot meet the needs of today's business environment.

Initial IP SAN deployments were most often at the departmental level of larger enterprises, particularly in Windows environments comprised of smaller servers where limited admin support, host attach costs and infrastructure complexity have traditionally inhibited Fibre Channel SAN deployment.

In large enterprises, IP SANs are also popular where multiple separate data centers need to be affordably interconnected and where the management of corporate data needs to be integrated. Many companies have a large robust Fibre Channel SAN environment, but they may still have large numbers of smaller "stranded" servers each with their own internal or direct-attached storage. There are clear advantages in being able to leverage the existing SAN investment if an affordable way to connect these servers into the SAN can be found.

More recently, we have seen the adoption of IP Storage solutions in medium and even in small organizations – once again replacing direct-attached storage and supporting Windows-based business applications. In many cases these applications are mission critical, and data growth concerns, data availability concerns, the need to more efficiently manage the environment with existing admin resources, and limited IT budgets are the major customer concerns.

Although most IP SAN deployments to date have been to replace legacy direct-attached storage, we have over the past year seen a trend towards IP SANs replacing first-generation (1 Gb) Fibre Channel SANs. Typically here, the legacy SAN is approaching the end of its lifecycle, the customer needs a more cost-effective solution requiring less admin support, and the applications don't need the additional bandwidth of an upgraded Fibre Channel SAN solution.

Chapter I – Storage Basics

Storage and its associated storage networks are growing phenomenally as businesses generally trend towards pooled storage for flexibility. Regardless of the type of business, the growth of data is exploding. Data must be stored and protected while being properly managed to meet the demands of accessibility, capacity and performance. To make use of this growing data, an organization may increase the number of servers it uses. However, the increasing number of servers leads to a different set of data access problems such as the case of multiple servers attempting to use the same data concurrently. This issue has been addressed by a number of methods, but the most prevalent is storage networks.

There are three basic elements of data storage that need to be understood, including connectivity, media and I/O protocol. Connectivity describes how the servers and their storage resources are physically connected. Media is the type of cabling and the associated low-level protocols needed to manage traffic over the cabling that provide the connection; while I/O protocol describes how the I/O requests are transmitted over the media.

Of the three storage elements, media and the associated low-level traffic management protocols offers the widest range of choice for customers.. Storage media technologies have evolved, providing far more choice and potential complexity for customers, so the decision about which media is right is not as clear as it once was. One must realize that no one solution is right for all companies since each have different organizational objectives, current infrastructure, and future goals for their storage needs.

The principal I/O protocols in use today are SCSI (Small Computer Systems Interface), FCP (SCSI over Fibre Channel), iSCSI (SCSI over TCP/IP) NFS (Network File System), and CIFS (Common Internet File System). SCSI is often called a “block level” protocol while NFS and CIFS are file-level protocols. But with the use of the logical volume manager in various operating systems, SCSI is still the common point of storage operations, it is just a matter of how far into the I/O stream and/or where the I/O operations occur.

The three most common storage topologies are shown in Figure I-1.

Direct-attached storage has historically been the most prevalent method. In single server or small cluster environments (such as two servers), direct attached storage is viable. However, from the departmental or enterprise viewpoint, direct attached storage is not an attractive option for a number of reasons such as storage utilization, total costs, and lack of a centralized management point. Consequently, networked storage has gradually been displacing direct-attached storage environments.

Storage area networks (SAN) while successfully addressing the need for better storage utilization when compared to direct-attached storage have evolved, providing more intelligence in an attempt to further simplify and automate storage operations. The evolution of SANs has resulted in more methods, protocols and product configurations available to customers. Fibre channel protocol is one such method available for the storage network and was the protocol of choice in the early evolution of SANs finding its way into mission critical data-centers; but it has come with a level of complexity, requiring special skills and additional costs for storage administrators to effectively administer.

This has largely prevented its adoption across the distributed enterprise and limited the number of servers that can take full advantage of a shared storage model. The customer’s cry for easy-to-deploy storage networking that leverages existing IT skills while delivering on the promise of improved storage utilization and simplified management- across the enterprise- has been answered by a significant

development in the evolution of SANs: the availability of iSCSI-based SANs which use Ethernet as the transport medium.

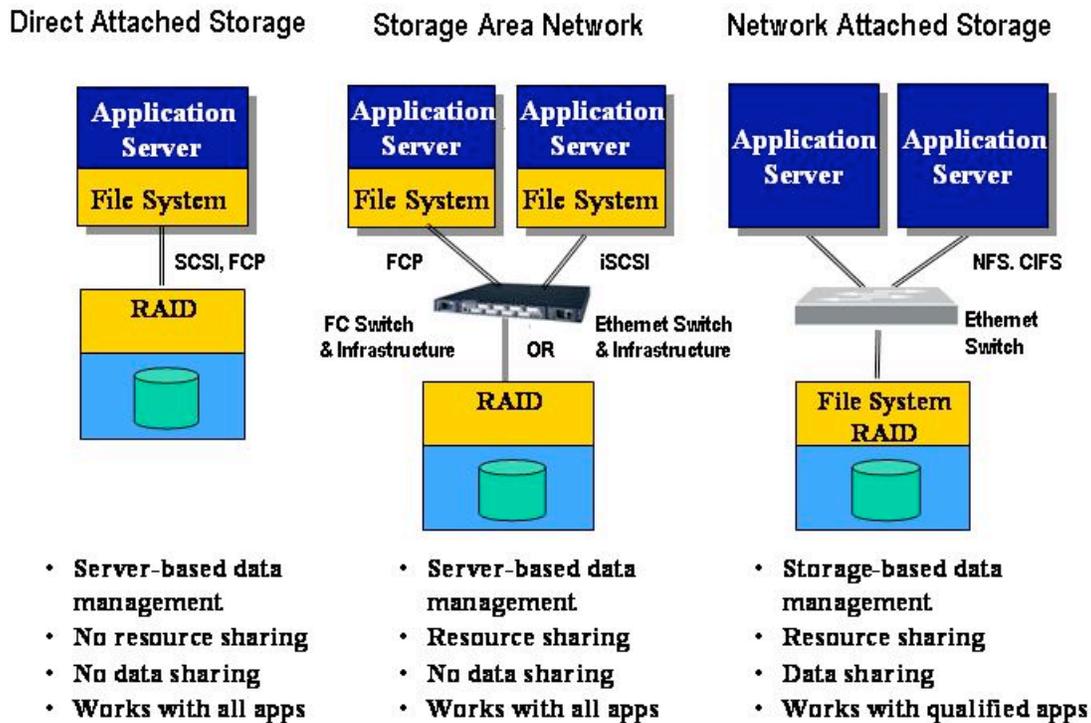


Figure I-1. Storage Connectivity Methods

Simply stated, host servers can access external storage using the iSCSI protocol via TCP/IP networks, and this method of storage access is called IP Storage. IP Storage eliminates the need for additional costly equipment such as special adapter cards or switching devices, while it leverages existing administrative LAN skills and tools. The use of IP storage also significantly addresses distance limitations; almost to the point of claiming unlimited distances between storage and servers.

Network Attached Storage is a storage appliance that directly attaches to an IP network and it uses a heterogeneous environment which features file sharing, not storage sharing. Typically, a NAS box shares files with clients of the same operating system, but some are capable of working in a mixed environment. There are also NAS gateways which differ from the appliance in that the NAS gateway has minimal, if any, internal storage but does have a fibre channel connection for external storage as shown in Figure I-2. The IP network side of the gateway is talking file-I/O while the fibre channel side deals with block-I/O.

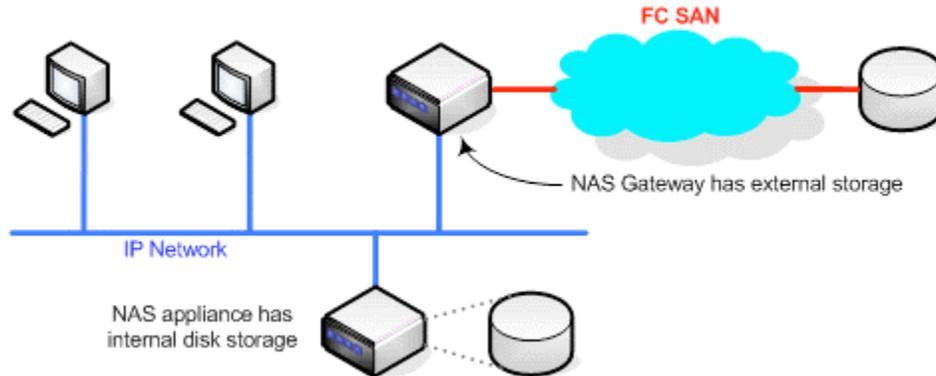


Figure I-2. NAS Appliance and Gateway Comparison

In building an IP SAN model, servers and storage are interconnected across an IP network. Servers initiate requests using iSCSI block protocol through the IP network through standard Ethernet switches to controllers optimized for processing IP traffic. The model is very similar to SANs built using fibre channel, the difference is that IP is the transport and the tools used to manage the network are standard and familiar to most network administrators. Therefore IP SANs are generally very easy to install and administer. There are several models of iSCSI implementation. Some integrate the controller with the disks within the same enclosure. Another model is to implement the iSCSI “target” in a gateway, as shown in Figure I-3.

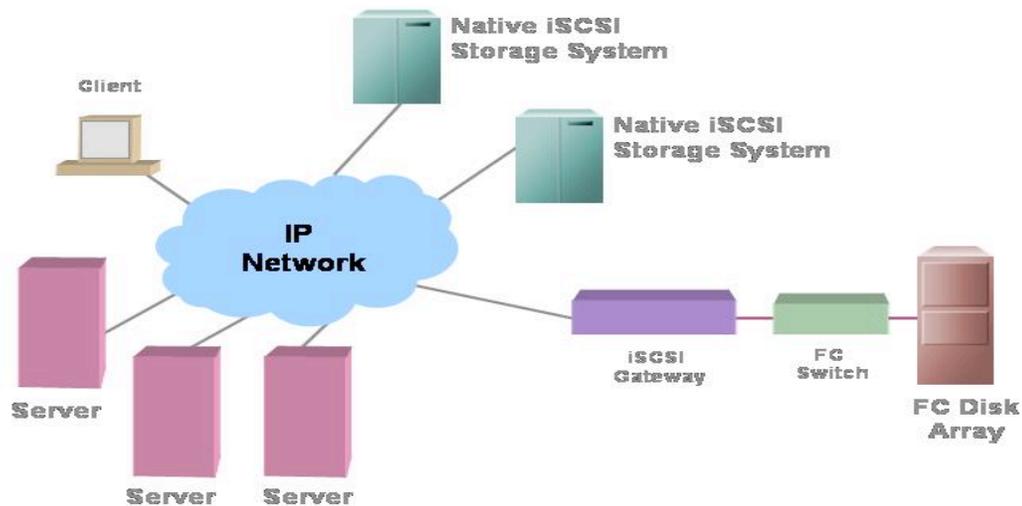
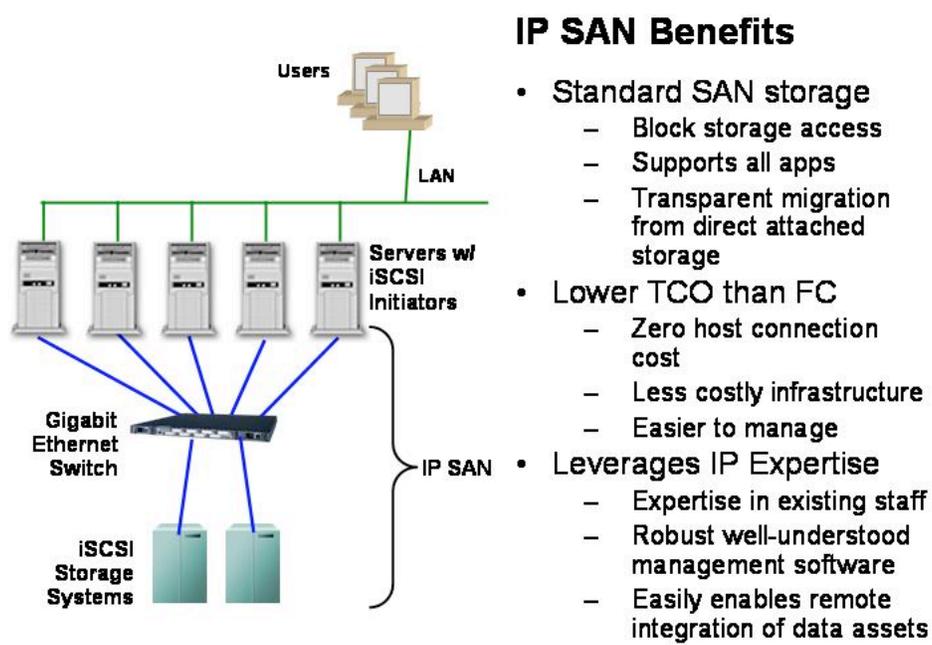


Figure I-3. iSCSI Implementation Options

Despite these options the vast majority of iSCSI-based SANs are built in much the same way as Fibre Channel SANs – see Figure I-4.



IP SAN Benefits

- **Standard SAN storage**
 - Block storage access
 - Supports all apps
 - Transparent migration from direct attached storage
- **Lower TCO than FC**
 - Zero host connection cost
 - Less costly infrastructure
 - Easier to manage
- **Leverages IP Expertise**
 - Expertise in existing staff
 - Robust well-understood management software
 - Easily enables remote integration of data assets

Figure I-4. Typical iSCSI-based SAN Topology

Chapter 2 – Benefits of IP Storage

So what are the benefits of TCP/IP as a storage interconnect? First of all, IP networking is a ubiquitous technology with the use of Ethernet networks. Ethernet is already widely deployed in companies of all sizes, and even many residential homes. Ethernet is well understood by most system administrators, whether they are server or storage. There are a wide range of vendors involved in the Ethernet ecosystem, so costs are economical because of the market size. As a result, IP Storage costs are amortized over more ports by several orders of magnitude when compared to the number of fibre channel SAN ports and thus much more affordable to the point of being considered as a commodity.

IP networking is mature. It is based on the Open Systems Interconnection (OSI) standards of the International Standards Organization (ISO). These standards include definitions of protocols for addressing, formatting and sequencing of data transmissions across networks. Thus, IP Storage does not suffer from many interoperability issues. Another advantage of IP networking maturity is that manageability of the IP network infrastructure is readily addressed with many years of well-proven tools. Since these same tools can be used for IP Storage network management, administrators can use the same well-developed skills they have already been using.

IP networking incorporates additional levels of security with storage operations with IP security protocols and encryption for data protection. IP networking includes mechanisms that deal with many conditions not normally addressed by fibre channel network. Some of these features resolve issues with congestion, dynamic rerouting with minimal network impact, datagram retransmission and session integrity. The net effect of this functionality is that it is possible to recover from interruptions within the IP network.

Even though IP networking is mature, it is still evolving as vendors introduce newer and better products. Because it is so wide spread, vendors have large research and development efforts in this technology to increase their share of this large market. Therefore, IP Storage has a strong roadmap into the future - 10 GbE is available today, and there are now with plans for 100 GbE. Network bandwidth has grown significantly over the past few years and indications are that this growth will continue for many more years.

IP networks provide a cost effective means for a distance solution that are not readily implemented in fibre channel fabrics. With the widely accepted OSI standards from the International Standards Organization, IP storage is not limited by the factor of distance or location. Whether local area, Metropolitan Area (MAN) or Wide Area (WAN) networking; TCP/IP technology and the underlying protocols work consistently around the globe. Applications such as remote data replication and disaster recovery, expect things to happen with IP Storage regardless of the length of the communications path.

With storage able to be consolidated more readily, the effectiveness, management and total costs of stored data improves. By consolidating the data, IP Storage leads to better utilization of resources. The utilization is further increased by the ability of IP Storage devices to support multiple operating systems within the same device. There are additional benefits with this data consolidation besides reduced costs of storage, and one of the biggest is the overall integrity of the data. Imagine having only one copy of a data file for multiple file systems instead of multiple copies of that data file where each copy is for a single file system. Not only is the amount of necessary storage reduced, but all of the various servers are using a consistent image of the data.

Using IP networking protocols to interconnect fibre channel networks is attractive for many reasons. By using existing IP networks, companies are able to control TCO for expansions of their storage solutions regardless of the distance. For larger storage network solutions, IP networks can grow beyond the current limitations of fibre channel networks with the total number of switches in a fabric.

IP Storage leads to enhanced data backup processes in a number of ways. First, the volume of data to be copied is reduced which translates into smaller time frame for the data backup application. IP Storage provides flexibility to data backup with a broad range of copy tools and services built into the device that does not require interaction by other servers. Another advantage of IP Storage is that the storage administrator can manage more capacity when compared to many other storage topologies.

Chapter 3 – Other Considerations of IP Storage

There are some factors to consider on the converse side of IP storage decisions, but these factors reduce to basically four items. They are:

- IP Network bandwidth
- Server performance impact
- Data security
- Application conformance

IP networks, specifically Ethernet, favor short bursty traffic patterns rather than large continuous data transmissions; and the movement of a large block of data can consume significant overhead. Network congestion worries arise with the heavy utilization of a network's bandwidth by IP storage. However, there are a number of mechanisms that already exist in the IP networking world that addresses this item. First, network speeds have evolved during the 1990's and they are fast enough for IP storage with the wide spread use of gigabit Ethernet. And now 10GbE enables the aggregation of Gigabit Ethernet ports. Further, network administrators can easily and quickly allocate IP storage traffic to dedicated network segments using a variety of tools and functionality in the existing networks.

Another consideration is the additional load that might be placed on the server's CPU to process the extra protocols, particularly iSCSI. When iSCSI started to be implemented several years ago, this was a reasonably valid concern. But the introduction of significantly higher processor speeds has helped address the CPU load factor from software-only iSCSI processing. This concern is also addressed by the introduction of TCP/IP Off-load Engine (TOE) network interface cards as well as iSCSI HBAs which handles some or all of the protocol processing in hardware, as shown in Figure 3-1. With the current CPU capacity, software-only protocol handling should be sufficient for all but the most demanding performance requirements implementations.

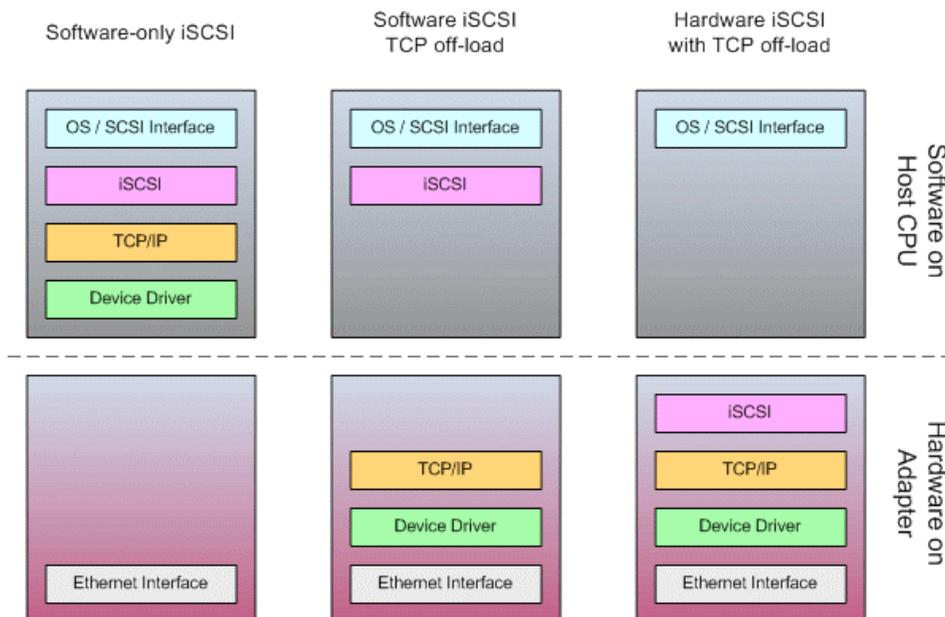


Figure 3-1. Comparison of TOE Adapters

Data security seems to be a concern to many when talking about IP Storage because the perception is that the internet is not secure. In reality, the opposite is true considering the large number of tools and best practices that have been developed while IP networking matured. From the hardware perspective, IP networks can be implemented as private networks in a similar fashion as fibre channel SANs are now. In addition, however, the iSCSI Specification developed and ratified by the IETF (Internet Engineering Task Force) has always included protocol-level authentication built in – a feature only now emerging for Fibre Channel. To cover long distances, IP based firewalls and VPN (virtual private networks) are considered best practices. Finally, the data stream can be fully protected with the inclusion of IPSec for data encryption, privacy and to prevent eavesdropping and/or IP address spoofing. So the bottom line is there are a number of options to choose from depending on the IP network and storage infrastructures.

Application conformance was not a major concern, since iSCSI is based on the same SCSI protocol used by direct-attached and fibre channel SAN arrays – the default way that host operating systems talk to storage. NAS systems, in comparison, needed to be qualified by operating system and application vendors to ensure support.

Chapter 4 – Types of IP Storage

As described earlier, IP storage involves the transmission of SCSI block storage traffic across an IP network and there are two basic transportation mechanisms. One method is to encapsulate a fibre channel frame within an IP datagram using either the iFCP or FCIP protocols. The other method is the encapsulation of SCSI frame within an IP datagram, which uses the iSCSI protocol. All of these protocols have overcome the distance limitation of SCSI and fibre channel with the use of IP networks. Figure 4-1 shows a comparison of the various protocol models. It should be noted that all three of these protocols are ratified IETF standards.

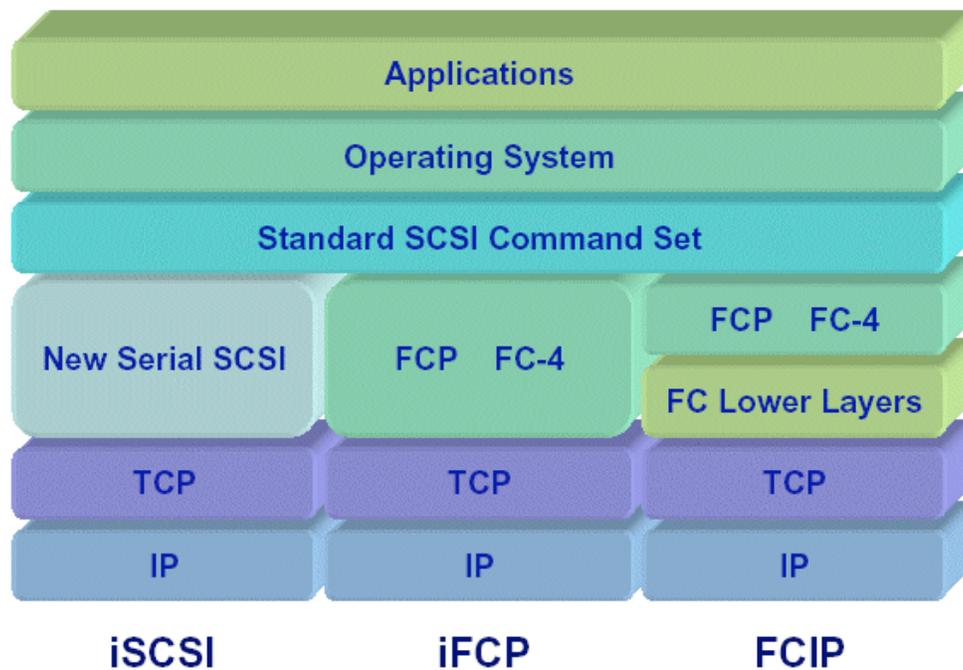


Figure 4-1. Comparison of iSCSI, iFCP and FCIP

Making a comparison of the three protocols based on closeness to FCP or TCP/IP, FCIP is significantly based on the fibre channel protocol. Meanwhile, Internet Small Computer Systems Interface (iSCSI) is at the opposite end with mostly IP-based content and very little from fibre channel. iFCP resides somewhat in the middle but leans more to the fibre channel side. FCIP's encapsulation of FC frames should be considered more as an extension of fibre channel. Conversely, iSCSI could easily represent a complete shift from fibre channel based storage networks with an IP network path all the way from a server to its storage resource.

There are a number of situations where IP Storage is a good solution. One of the first places is SAN bridging to solve distance problems. An increasingly popular example is connecting two widely separated fibre channel SAN environments, such as a remote site with the core data center, for a disaster recovery solution making use of a TCP/IP connection for remote copy or mirroring. The core data center likely has a fibre channel SAN core which will stay there for the foreseeable future and continue to grow; while the remote site is a mixture of FCP and IP storage. Thus, IP Storage is not the death knell of fibre channel SANs, but rather an enhancement and new opportunity to efficiently use shared storage.

The extra benefit to use IP storage for distance solutions is the TCO is typically less than that of fibre channel solutions in terms of equipment, administration skills and management tools.

However, regional data centers or departmental data centers which do not have a SAN can also be considered as candidates for an IP Storage environment. This is due to the fact that the cost and complexity associated with the core data center prevents the implementation of these solutions in some of the other secondary and/or smaller data centers. One option for these locations that need storage networks is iSCSI. Simply stated, iSCSI is the protocol that unites storage and IP networking.

iSCSI is a transport protocol that transports SCSI command and data traffic over TCP/IP networks. iSCSI protocol has a number of advantages in that it is relatively simple, high speed, low cost and it is a solution that works over long distances. From a network point of view, iSCSI is well behaved since it has TCP/IP header so iSCSI frames can be routed, along with guaranteed delivery. Thus, existing network components such as the LAN/MAN/WAN infrastructure as well as the management tools can become the storage network. From a storage perspective, iSCSI makes use of the SCSI-3 protocol standards that have been in existence for a number of years and are considered to be a mature technology.

IP Storage is also appropriate for more dispersed fibre channel environments with SAN islands. IP Storage technologies can couple the SAN islands together, rather than using dark fiber or other expensive proprietary implementations. For these solutions, there are two IP Storage technologies to select from: FCIP and iFCP. The main advantages of iFCP and FCIP are that they overcome distance limitation and can connect distributed SANs using existing IP infrastructure.

Internet fibre channel Protocol (iFCP) is a protocol for transmitting data between fibre channel devices in a SAN over the Internet using TCP/IP. iFCP allows for the incorporation of already existing SCSI and fibre channel SANs. iFCP replaces the lower-level transport mechanism of FCP with TCP/IP whose frames can then be routed. iFCP is a gateway-to-gateway protocol, where each gateway appears as if it were an end device in the SAN. The gateways maps the fibre channel addresses at one end to one or more IP addresses at the other end. Once the IP links are established, iFCP only needs to keep track individual fibre channel sessions.

Not only does iFCP have the advantage of using existing IP networks (LAN/MAN/WAN) as the storage infrastructure, but the iFCP interconnects multiple SAN environments without merging them into one big fabric, as shown in Figure 4-2. iFCP provides fibre channel device-to-device communications via the IP network instead of network to network communications. As a result of the level of granularity, individual communication sessions can have their own, distinct Quality of Service (QoS) over the same physical link. As a side note, network congestion or any interruptions of the TCP/IP link will only affect individual communication sessions instead of all traffic on the link. Another feature of iFCP is that storage networks can continue to grow past various limitations imposed within fibre channel solutions.

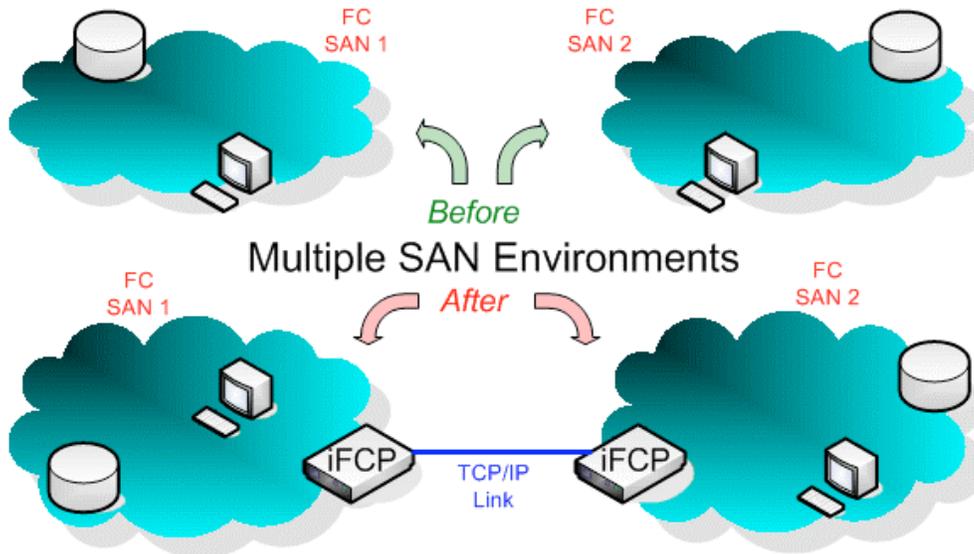


Figure 4-2. iFCP Implementation

Fibre channel over IP (FCIP) is simply a tunneling mechanism that allows fibre channel information, such as fabric building protocols, name server data and state change notifications, to be tunneled across an IP link. From a fibre channel perspective, it does not know that an IP link exists. When connecting two SAN islands using FCIP, the result becomes a single SAN environment with only one name server scheme from the merged fibre channel SAN, as seen in Figure 4-3. FCIP fully supports fibre channel equipment over long distances through IP networks. FCIP relies on the best of both worlds, fabric services from fibre channel while TCP/IP provides established, robust mechanisms for congestion control, network management and recovery from data error and data loss.

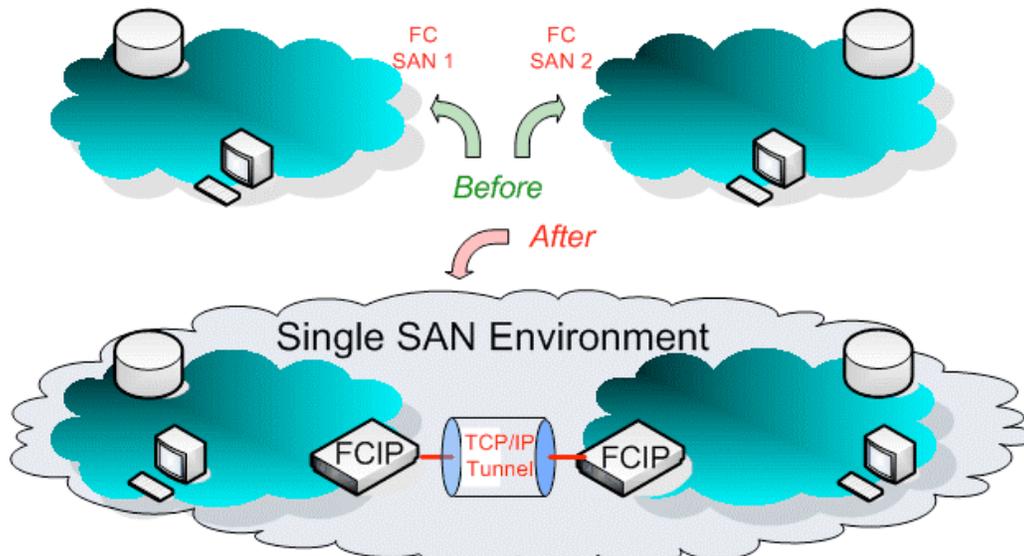


Figure 4-3. FCIP Implementation

Chapter 5 – iSCSI

The Internet Small Computer Systems Interface (iSCSI) protocol defines a mechanism for the transmission of storage traffic across an IP network as described in IETF RFC 3720. iSCSI is the encapsulation of SCSI frame within an IP datagram, which uses the iSCSI protocol. iSCSI takes full advantage of the maturity of both SCSI and TCP/IP technologies as well as addresses some minor functionality gaps of the SCSI protocol. The iSCSI protocol overcomes the distance limitations of SCSI and fibre channel with the use of an IP network. At the same time, iSCSI takes advantage of the existing Internet infrastructure and management facilities.

iSCSI architecture is fundamentally derived from SCSI in that it is based on a client/server model. iSCSI is a mapping of the SCSI remote procedure invocation model (SAM2) over the TCP protocol. SCSI commands and responses are carried within iSCSI protocol data units (iSCSI PDU). Figure 5-1 shows the iSCSI protocol model.

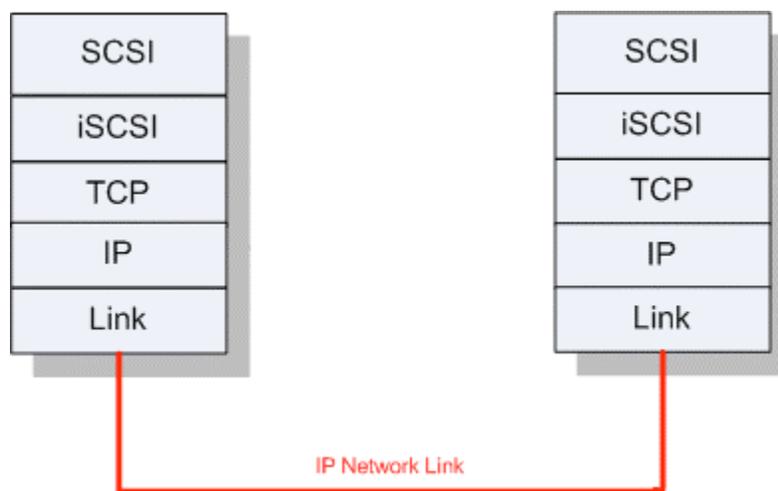


Figure 5-1. iSCSI Protocol Model

TERMS AND CONCEPTS

iSCSI is the transport of SCSI frames over TCP/IP. There are basically two parts of iSCSI, the initiator and the target, as shown in Figure 5-2. The iSCSI initiator sends a SCSI request over the IP network to iSCSI target. The target can be a device with storage, or presents SAN-based storage resources to the initiator. iSCSI protocol bases all traffic flow directions on the initiator, where outbound traffic is from the initiator to the target and inbound traffic comes into the initiator. A layered concept, or model, of the protocol is where the upper SCSI layer builds/receives the command blocks and passes them to the lower iSCSI layer. The communications that form between an initiator-target pair are called a session. In turn, the iSCSI layer builds the iSCSI PDU and is responsible for its transportation across one or more TCP connections that comprise the session. Across all connections within a session, the initiator will only see one target. To keep track of potential multiple initiator-target sessions, an initiator task tag is used in conjunction with the initiator-target-LUN nexus (ITL).

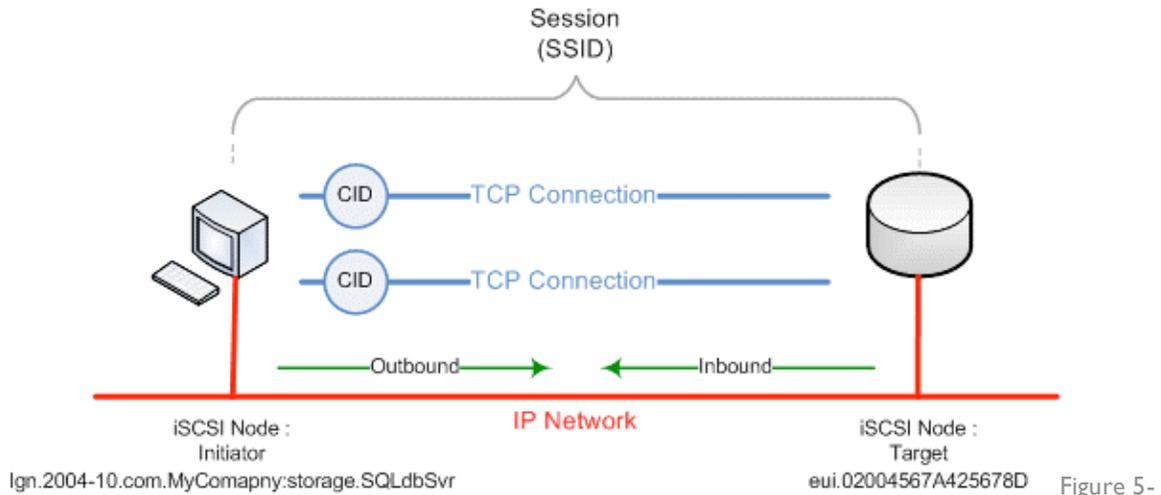


Figure 5-

2. iSCSI Concepts and Terminology

ISCSI FRAME FORMAT

iSCSI PDUs, as shown below in Figure 5-3, are padded to an integer number of 4 byte words. All iSCSI PDUs begin with one header segment, called the Basic Header Segment (BHS), which is a fixed-length 48-byte header segment. The BHS can be followed by any number, if any at all, Additional Header Segment (AHS). An optional header digest may come after the entire header segment group. In a similar manner, actual data and a data digest can then follow. The minimal frame size is 48 bytes (WN and BHS).

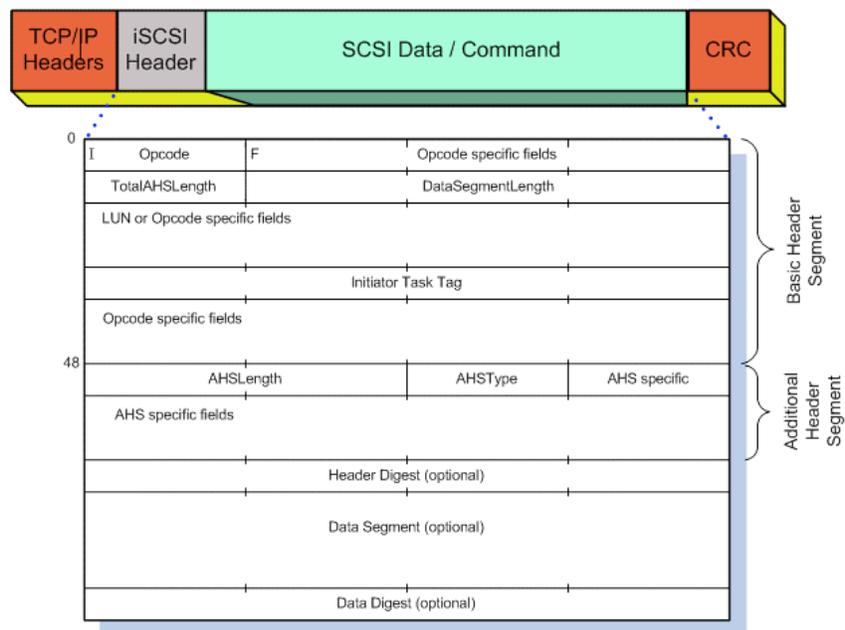


Figure 5-3. Basic Structure of iSCSI Frame

The optional header and data digest fields are used for integrity and authenticity checks of header and data segments. The digest is basically a CRC check of the preceding segment. The digests, if present, appear as trailers. The Basic Header Segment is 44 bytes long and is described in better detail further in this document.

The opcode indicates the type of iSCSI PDU. There are two categories of opcodes, which are specific to the initiator and target. Initiators and targets can not use opcodes from the other category. Tables 5-1 and 5-2 summarize the opcodes used by initiators and targets respectively.

The very first bit of the iSCSI header is a flag, called the I bit, which is used to mark the iSCSI PDU for immediate delivery. Another bit flag within the first word of the header, called the Final (F) bit indicates if the PDU is the last, or only, PDU of a sequence. Both the I and F bits are enabled when set to a value of 1.

Initiator Opcode	iSCSI PDU Contents
0x00	NOP Out
0x01	SCSI Command (encapsulated SCSI CDB)
0x02	SCSI Task Management function request
0x03	Login Request
0x04	Text request
0x05	SCSI Data-Out (for write operations)
0x06	Logout Request
0x10	SNACK Request
0x1C to 0x1E	Vendor specific codes

Table 5-1. iSCSI Initiator Opcodes

Target Opcode	iSCSI PDU Contents
0x20	NOP Out
0x21	SCSI Response - contains SCSI status, may include sense data or other response information
0x22	SCSI Task Management function response
0x23	Login Response
0x24	Text response
0x25	SCSI Data-In (for read operations)
0x26	Logout Response
0x31	Ready To Transfer (R2T) – target sends when ready for data
0x32	Asynchronous Message – target sends to indicate special conditions
0x3C to 0x3E	Vendor specific codes
0x3F	Reject

Table 5-2. iSCSI Target Opcodes

When there is at least one Additional Header Segment in the PDU, the TotalAHSLength field provides the length of all AHS segments based on units of four byte words. Otherwise, this field has the value of 0x00. The DataSegmentLength field indicates the amount of data within the payload based on the number of bytes. This field does not include the length of any padding that may have been added to the data segment to keep it on a four byte word boundary.

A subset of the opcodes performs an operation on a specific Logical Unit (LU). In these cases, the LUN field contains the 64-bit Logical Unit Number that identifies the specific LU. If the opcode does not apply to an LU, this field may either be ignored or used to contain additional information, which is specific to the opcode. The Initiator Task Tag field contains a unique identifier for every iSCSI task that is issued by the initiator. The scope of the Initiator Task Tag is the session between the initiator and target.

If an Additional Header Segment is present in the iSCSI PDU, the AHSLength field indicates the length of the individual AHS header minus the AHSType and AHSLength fields. The AHSType field denotes the content and purpose of the additional header.

The iSCSI PDU contains an optional header and data digests to protect the integrity of the information. The use and type of digests are negotiated during the Login Phase.

ISCSI NAMING AND ADDRESSING

The World-Wide Unique Identifier (WWUI) is the recognized means for identification of individual iSCSI entities. Initially, the IP address and TCP port number were considered a major component of the WWUI. However, it was recognized that this approach was not the most reliable method with the use of DHCP services within a network. When DHCP is utilized, a specific network entity did not necessarily retain the same IP address all the time. However, an alias can be associated with a target to simplify operations and to help organizations associate their own semantic meanings with the aliases within their network.

The WWUI corresponds to the logical operating system of the iSCSI host. A WWUI really names a logical software entity, and not generally tied to a port or other hardware that can be changed. Thus multiple interfaces on the same system will present the WWUI to all others since the iSCSI agent behind them is one and the same entity. One very noticeable feature about WWUI is they are text-based. This was done for several reasons:

- Easier for a user to differentiate and/or transcribe
- WWUI are only used during login and discovery phases, so overhead has minimal impact
- iSCSI communicates via text strings already

Besides names, iSCSI devices also have addresses. An iSCSI address specifies a single path to an iSCSI entity, and includes the system's WWUI as a portion of the address. The iSCSI address is not generally used within normal connections; but it is used primarily during discovery. As shown above, the IP address portion of the iSCSI address can be either the IPv4 or IPv6 numerical address or the fully qualified domain name (i.e. example.raleigh.ibm.com).

One example of an iSCSI address is shown below:

Computingcenter.acme.com:4002/com.gateways.yourtarget.24

If the TCP port is not designated, then the well-known port (3939) will be used. This port number has not been fully ratified by the IETF, but no major opposition has been expressed to using this port number.

To address targets and logical units within a target, SCSI uses a uniform addressing scheme of 8 byte fixed length addresses. iSCSI refers to this address as the SCSI reference address (SRA). To provide the target with the protocol specific addresses, iSCSI relies on the SCSI aliasing mechanism. This aliasing support enables an initiator to associate protocol specific addresses with SRAs, which can be used in subsequent commands.

SESSION MANAGEMENT

There are three mandatory phase for an iSCSI connection between an initiator and target. The first phase is the iSCSI login, which builds a TCP connection, authenticates both devices, negotiates the session's parameters, and marks the connection as part of an iSCSI session. One additional parameter that is optional is the agreement between the initiator and target on a security protocol for the session. The session transitions to the Full Feature Phase once the Login phase is completed for the first TCP connection between an initiator and target. There is an optional stage that can precede the Login phase, called Discovery, where an iSCSI initiator can query an iSNS server for information about targets.

During the Full Feature phase, the initiator may send SCSI commands and data to the various LUs on the target by wrapping them in iSCSI messages that are then transferred over the established iSCSI session. For SCSI commands that require some form of data transfer, the data and the status for the commands must be sent over the same TCP connection that was used to deliver the command. This situation is called connection allegiance. For example, an initiator issues a read command and the target must send the requested data, as well as any status responses, over the same connection. However, consecutive commands that linked can use different connections, if they exist. Command allegiance is per-command, not per-task.

The final phase of an iSCSI connection is the Connection Termination phase. All TCP connection resets are assumed to be exceptional events. When an endpoint receives a TCP frame to close the connection, it waits for all outstanding tasks to complete and then shuts down its side of the connection. As a prelude to certain error condition recoveries, connection termination is required. By terminating the connection, the initiator and target will discard all in-transit PDUs as part of recovery operations.

ISCSI TRAFFIC FLOWS

Before an initiator can begin communications with a particular target, the initiator must know the target's address. At the iSCSI protocol level, this address is the world-wide unique identifier (WWUI). Unfortunately, the initiator may not necessarily know the WWUI of the target and thus there is a need by the initiator to be able to find iSCSI target entities. Since the initiator and target can have multiple physical connections just as easily as multiple client-server sessions, it can be critical for smooth communications to mark a connection as belonging to a given iSCSI session. At a logical level, the initiator – target nexus could be the same for several sessions yet the commands, data, and messages must be kept separated to their own traffic stream. To help start a communication session, the targets listen on a well-known TCP port for incoming connection requests. The iSCSI target will conclude the Login process by sending a Login Response with “login accept” or “login reject” to the initiator.

iSCSI uses numbering schemes for command and status messages, and a sequencing scheme for data flows. Command numbering is session wide. By following a consistent pattern, the flow of commands

from the initiator to the target can be tracked very easily. Using an example of multiple connection paths between the initiator and target, and the initiator sends an ordered set of commands to the target. With command numbering, the target can easily reconstruct the sequence of commands back into its original order with a little difficulty. Status numbering is per connection and is used to enable recovery in case of connection failure. Thus, the initiator and the target are able to correlate various responses, or status messages. This multilevel numbering mechanism also has the side benefit of being able to determine when a connection failure has occurred.

Data PDUs that are transferred as part of a command execution must be sequenced. For read or write operations, the sequence starts at 0 for the first data PDU and advances the count by 1 for each subsequent PDU. This method enables the receiving endpoint to determine if a data packet has arrived out of order, or is missing. Unlike command and status PDUs, data PDUs are not acknowledged except as covered by a status message.

The initiator and target have three counters that define the numbering mechanism for commands. These counters are:

- CmdSN – current command sequence number advanced by 1 on each command
- ExpCmdSN – next expected command by the target – acknowledges all commands up that value
- MaxCmdSN – maximum command sequence number that the target will accept

For response and status numbering, the target uses the StatSN counter for responses in transit from the target to the initiator. At the other endpoint of the connection, the initiator uses the ExpStatSN counter to acknowledge status. Status numbering starts after Login. During login, there is always only one outstanding command per connection and status numbering is not needed until the connection is established. The login response will contain an initial value for status numbering.

For SCSI commands that require data and/or parameter information, the data and status for a command must be sent over the same TCP connections. For example, an initiator issues a WRITE command and has data included after the command. The target must return R2T and status over the same connection for the remainder of the data, if any. By utilizing multiple connections, numerous commands with their associated data (optional) can be interleaved over the session.

Targets operate in either solicited (R2T) or unsolicited (non R2T) data mode. If operating in the unsolicited data mode, a target is allowed to issue R2T PDUs. An error condition will result if an initiator sends unsolicited data within a command's data stream to a target operating in the solicited mode. The principle difference between these two data modes occurs mainly when an initiator starts a new write command and includes data to be written in the same PDU as the command. Figure 5-4 shows an example of solicited write operation.

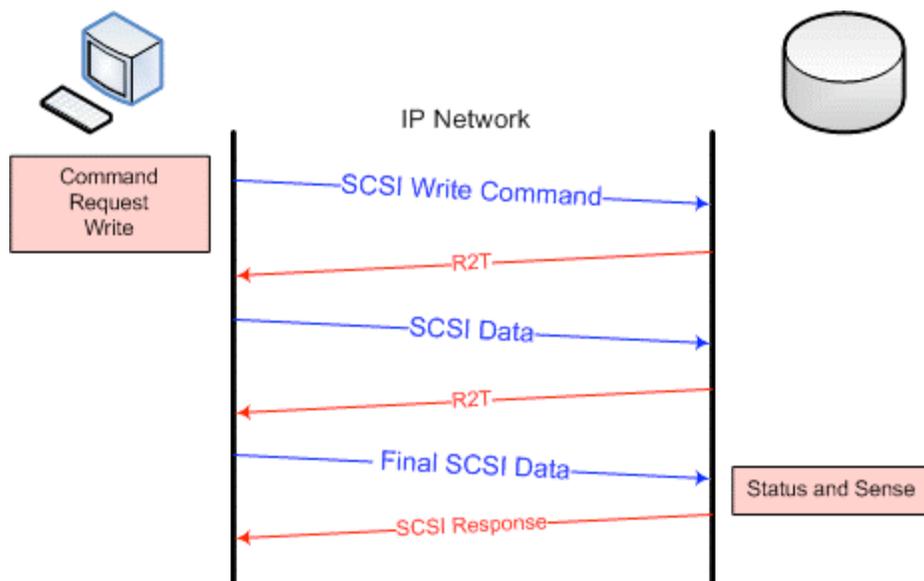


Figure 5-4. Solicited Write Operation Flow

The status of the read or write SCSI data streams is verified by using the status and data sequence numbers. During write operations, the target uses the transfer length field to indicate the amount of data that it is capable of receiving from the initiator and thus control the data flow.

TCP does not have a built-in mechanism for signaling message boundaries at the TCP layer. This obstacle is overcome by iSCSI placing the message length in the iSCSI message header. As long as IP packets are delivered in order over the network, it is not an issue. However, with sophisticated networks and multiple paths, it is possible for a packet to be delivered out of order or even dropped. With the inclusion of the message length in the iSCSI message header, the iSCSI layer can direct the SCSI contents of the PDU into its proper place in the SCSI buffer and not have to temporarily buffer the contents. This action minimizes the amount of system resources required as well as saves on later use of processor cycles of having to move contents from a buffer once the missing PDU(s) are received.

There is an optional mechanism to assist with message synchronization, called Synch and Steering. This method has special markers at regular intervals in the data stream. Within these markers are pointers to the start of the next iSCSI message header. This mechanism repeats various pointers within the markers in case the marker is split in conjunction of segmentation and reassembly process of TCP. Thus, if a marker is split and one of the pieces is lost, the iSCSI agent should still be able to retrieve a valid pointer and continue its reassembly of the SCSI data stream with minimal impact.

The primary purpose of the optional synch and steering layer, with its relative position shown below in Figure 5-5, is to assist with the segmentation and reassembly of the SCSI traffic stream.

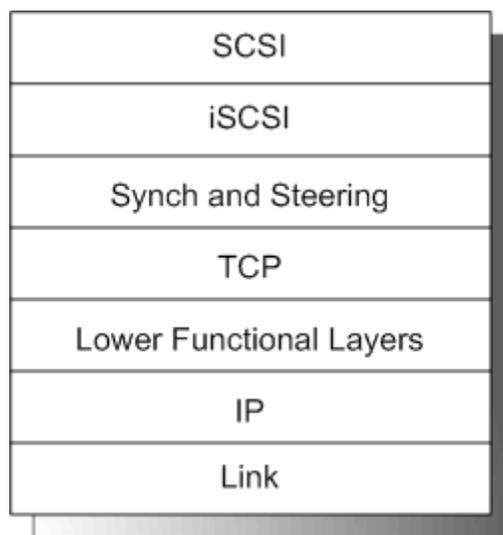


Figure 5-5. Synch and Steering Layer Relationship

ISCSI ERROR HANDLING

Since iSCSI can function over a variety of IP networks, it must be able to detect as well as recover from a wide variety of errors. For this reason, iSCSI enables recovery from any level of communications, such as within an individual command, a specific connection, or the entire session. Command errors can result from potential corruption of an iSCSI PDU that indicated by either an invalid header or data digest.

To facilitate iSCSI error handling within an individual command, both the initiator and target must have the ability to buffer commands and responses until they are acknowledged. Using a write operation example, the iSCSI initiator would keep data in its buffer that has already been transmitted until reception of the R2T from the target. Not only does the R2T iSCSI PDU indicate that the previous data has been received, it signifies that the target is ready to receive further data. Another benefit from buffering the data streams is that only missing and/or corrupted data PDUs need to be retransmitted.

From the connection recovery point of view, there are a number of ways in which a specific connection appears to have issues. If multiple paths exist between the initiator and target across the IP network, then it possible to suffer a failure on one of the paths which has at least one established connection. When a TCP connection fails, the initiator must close the connection while sending a message PDU with the reason code “remove the connection for recovery”. If the session still has multiple connections that are operational, the message can be sent asynchronously over a functional connection about the failed connection.

Session recovery is the most drastic and implies that all other recovery attempts have failed. One other possibility is that the iSCSI device is not very robust and session recovery is implemented as the default method to recovery from any form of error condition. When a session recovery process is invoked, all existing TCP connections will be closed while all tasks (queued or currently being implemented) are aborted. Once all connections are cleared and commands have been flushed, the initiator can attempt to rebuild its TCP connections to the target.

iSCSI defines two session timeout values to assist with recovery of a connection and/or task error condition. The first value, called Time2Wait, is the initial delay period (in seconds) before attempting an

explicit/implicit logout for the connection in question. The other value, called Time2Retain, is the maximum time (in seconds) that a target will wait for potential recovery efforts from the initiator. Thus, the initiator's recovery attempts delays for Time2Wait to be started and those efforts must be completed before Time2Retain.

ISCSI SECURITY

The iSCSI protocol allows for multiple security methods to be employed over a variety of IP networks. IPSec, or other, encryption solutions can be enabled outside of the iSCSI layer and thus do not require special negotiations during the Login phase. Further, this encryption is transparent to the upper layers.

For authentication methods, iSCSI allows are a discrete set of parameters to be negotiated during the Login phase of operations using text fields. This mechanism uses the format of:

parameter=<value>

The only allowed parameters at Login are:

- SessionType
- InitiatorName
- TargetName
- TargetAddress
- InitiatorAlias
- TargetAlias
- TargetPortalGroupTag
- AuthMethod

The vendor neutral methods and their associated keys are shown in Table 5-3. The iSCSI protocol also allows for the implementation of vendor unique methods.

Key	Description of Method
KRB5	Kerberos V5 (IETF RFC 1510)
SPKM1	Simple Public-Key GSS-API Mechanism (IETF RFC 2025)
SPKM2	Simple Public-Key GSS-API Mechanism (IETF RFC 2025)
SRP	Secure Remote Password (IETF RFC 2945)
CHAP	Challenge Handshake Authentication Protocol (IETF RFC 1994)
None	No authentication

Table 5-3. Authentication Keys and Methods

Chapter 6 – iFCP

Internet Fibre Channel Protocol (iFCP) is a gateway-to-gateway protocol which supports fibre channel layer 4 services over TCP/IP which is described in IETF RFC 4172. iFCP uses TCP services for congestion control, error detection and error recovery. To facilitate the interconnection of end devices, iFCP implements a method of frame address translation within the transparent iFCP gateways. The main function of the iFCP protocol is the transportation of fibre channel frames between FC end devices by encapsulating the FC frame, as shown below in Figure 6-1, and then sending the frame across the IP network via a predetermined TCP connection.

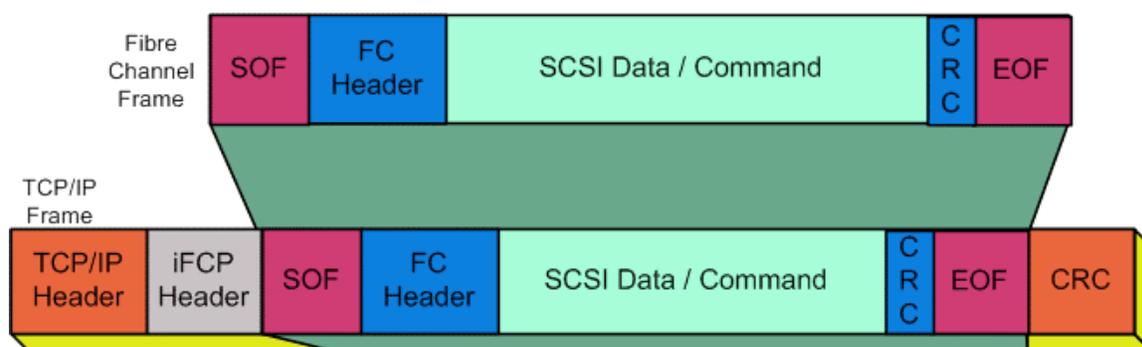


Figure 6-1. FC Frame Encapsulation by iFCP

An iFCP fabric supports FC Class 2 and Class 3 transport services. The iFCP protocol does not support FC Class 1, Class 4 or Class 6 services.

TERMS AND CONCEPTS

There are a few concepts and terminology that are introduced with the iFCP protocol, as shown in figure 6-2. iFCP is a gateway-to-gateway protocol, and the iFCP gateway exists within the fibre channel and IP network domains. The iFCP gateway appears as a fibre channel switch to FC end devices, yet it contains the IP components and functions for at least one IP network interface. The subset of fibre channel end devices which are locally attached to a given iFCP gateway make up a gateway region. Thus, two devices within the same gateway region will not use the iFCP portal and/or the IP network to communicate with each other.

If a global Domain ID and N_Port service is used by end devices regardless of gateway region, the iFCP fabric is called bounded. In this situation, the fibre channel storage name service functions are provided by the Internet Storage Name Service (iSNS), which is an IP protocol. The iSNS server is the repository for all iFCP storage objects such as zoning information, FC N_Port assignments, Domain ID assignments along with the maintenance of security and access control information. The iSNS server can be either an isolated server within the IP network or a distributed server entity with the individual iFCP gateways. When the iSNS functionality is within the iFCP gateways, the gateway is responsible for the N_Port addresses allocated to its local Gateway Region, which has the advantage of scalability due to a lack of central addressing authority. Thus, the iFCP fabric remains stable as the network grows in size. However, the local gateway must now respond to Extended Link Service (ELS) traffic while it maintains and updates the address tables.

Referring to Figure 6-2, fibre channel devices within FC SAN “A” are called locally attached devices that iFCP gateway and all fibre channel devices which are accessed via another iFCP gateway are called remotely attached devices.

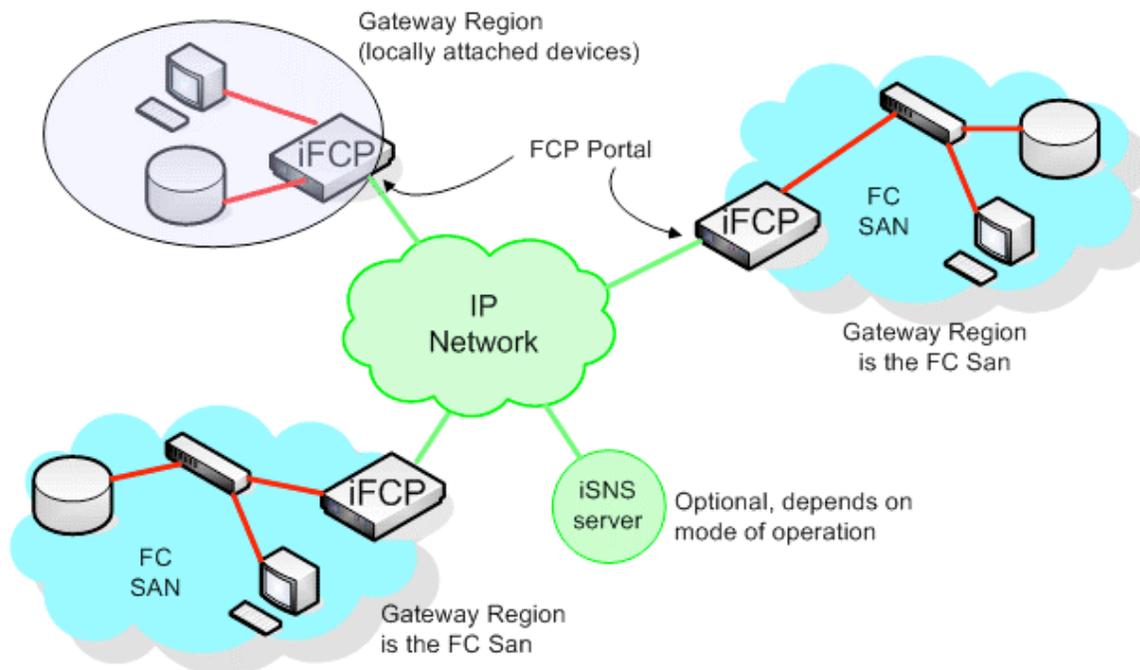


Figure 6-2. iFCP Concepts and Terminology

ADDRESSING AND USE OF INTERNET STORAGE NAME SERVICE (ISNS)

As previously discussed, the iSNS protocol is used in place of the native fibre channel name server functions. The iSNS server maintains a database of information for all fibre channel devices in a manner similar to the FC name server. Additional database information is required for devices that will be accessing other devices over the IP network. This extra information is required for the mapping of the FC device addresses to their related IP network addresses. If address translation mode of operations are used, then FC N_Port aliases that will be used by remotely attached devices.

The address translation database is based on the iFCP N_Port Address Model shown in Table 6-1, and it is built and updated by the iSNS. When a fibre channel device logs into the FC fabric, the database creates a new entry for the device. The mapping of the N_Port address to an IP addresses occurs at this time as well as any alias. This information is then distributed between all of the iFCP gateways. Thus, FC frames can now be mapped to the appropriate TCP connection and N-Port address on the fly where the process is described and shown below in Figure 6-4.

Key	FC Device WWPN
N_Port ID	24-bit N_Port address of the locally attached device
N_Port Alias	24-bit N_Port address assigned as of local device to be used by remotely attached devices
N_Port network address	Tuple of gateway IP address, TCP port number

Table 6-1. iFCP N_Port Address Model Information Elements

IFCP FRAME FORMAT

The format for an encapsulated frame is shown in Figure 6-3. This format is used for regular FC traffic as well as FC Extended Link Services (ELS) frames. Some of the fields in the iFCP Encapsulation header are ones complement duplicates of previous fields. This repetition of select fields was incorporated to assist with error detection.

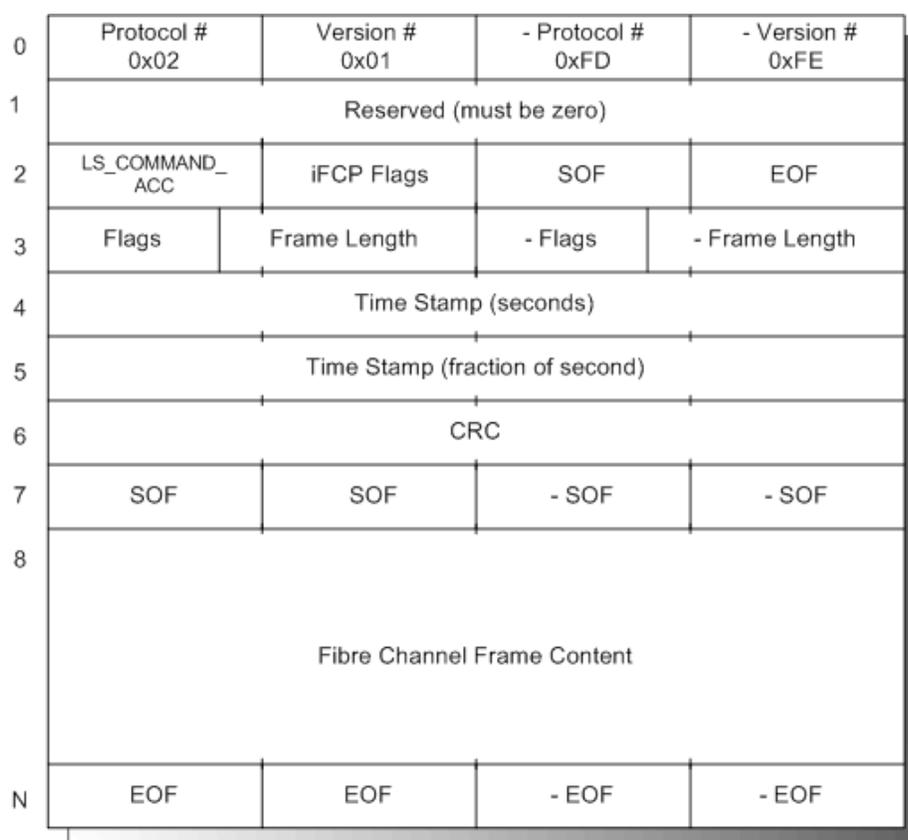


Figure 6-3. iFCP Frame Format

The Internet Assigned Numbers Authority (IANA) has assigned the protocol number value of 2 for iFCP. The version field contains the encapsulation version as specified by RFC 3643 – Fibre Channel (FC) Frame Encapsulation with a value of 1. The frame length field contains the length of the entire FC encapsulated frame, and the length is based on the 4-byte boundary.

The Time Stamp fields are used to compute the IP network transit times experienced by the encapsulated FC frames. When a frame is readied for transmission, it can be accompanied by a time stamp value. If both of the iFCP gateways support synchronized time services, then the time stamp fields are available for use. Otherwise, these fields are set to a value of zero.

The iFCP flags field uses only 3 of the bits. The Session Control frame flag, SES, indicates whether the frame is used for control messages. The address transparent flag, TRP, denotes the mode of operation for the originating iFCP gateway. If set to the value of one, then address transparent mode is enabled. Finally, the Special frame flag, SPC, flags whether the encapsulated FC frame is part of a link service message that requires special processing by the iFCP protocol.

Only one bit within the Flags field is used to indicate whether the CRC field contains a valid value. Since the iFCP protocol standards state that the CRC must be valid, the CRCV flag must be set to one.

The SOF and EOF fields contain an encoding of the SOF and EOF that should be used within the fibre channel frame's EOF and SOF. Tables 6-2 and 6-3 show valid values for the encoded SOF and EOF respectively.

Fibre Channel SOF	SOF Encoded Value
SOFi2	0x2D
SOFn2	0x35
SOFi3	0x2E
SOFn3	0x36

Table 6-2. Encoded Values for Fibre Channel SOF Signals

Fibre Channel EOF	EOF Encoded Value
EOFn	0x41
EOFt	0x42

Table 6-3. Encoded Values for Fibre Channel EOF Signals

The LS_COMMAND_ACC field has a non-zero value when a special link service ADD response is being transmitted. When used, the field contains the LS_COMMAND to which the ACC applies.

MODES OF OPERATION

There are two modes of operation in which a collection of iFCP gateways interoperate and all of the iFCP gateways must run in the same operational mode. The modes are either address transparent mode for bounded iFCP fabrics or address translation mode for unbounded iFCP fabrics. In address transparent mode, the fibre channel address for an end device is fabric-wide. Domain IDs are issued by an iSNS server which provides the name server functionality for all fibre channel devices. Thus, address transparent mode of operation has most FC services as if all of the iFCP gateway regions were part of one large SAN. While the address transparent mode has less impact on gateway configuration and provides a long distance solution, it does introduce a limitation in regard to the maximum number of switches within a FC fabric. Support for address transparent mode of operation is optional by the iFCP standards.

The second mode of operation is address translation. In this mode, the Port ID for an end device is local to its gateway region. If traffic between gateway regions is necessary, the local and remote gateways must perform N_Port address translation within each fibre channel frame that transverse the IP network.

When a FC frame must transit the iFCP network, the FC frame is sent from the locally attached device to the iFCP gateway in the normal manner of traffic within a fibre channel network. At the iFCP gateway, the address lookup table is consulted for the IP address and any associated alias (Figure 6-4, line 1). This information is used during the FC frame encapsulation within the IP datagram for the IP destination address (Figure 6-4, line 3). If the iFCP fabric is working in the address translation mode, the alias for the remotely attached device is inserted into the FC frame header (Figure 6-4, line 2) for the destination N_Port. The encapsulated frame is then transmitted by the FCP portal over the IP network to the destination iFCP gateway.

During the de-encapsulation process within the remote iFCP gateway in address translation mode, the address translation table is referenced once again (Figure 6-4, line 4). At this point, the source N_Port address is translated from its locally attached address to the appropriate alias (Figure 6-4, line 5) that will be used by the remotely attached device for any responses. Thus, the Source N_Port identifier in the FC frame is modified during the FC frame extraction, which is then transmitted via the iFCP gateway's FC fabric port.

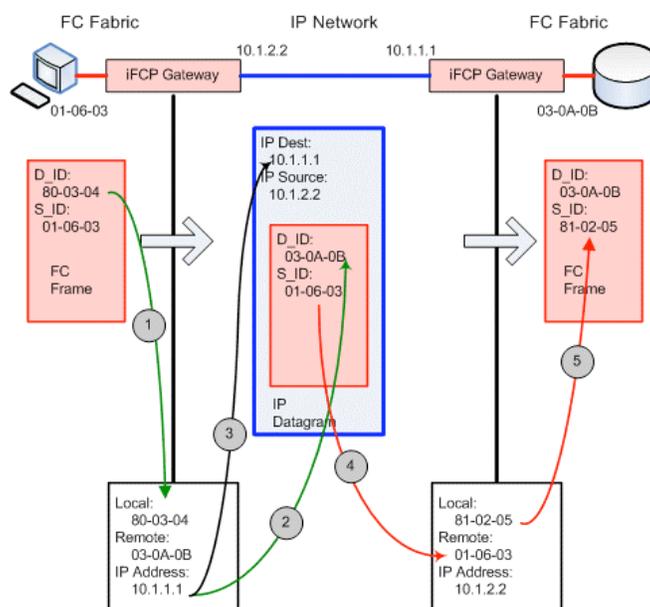


Figure 6-4. iFCP Mapping

Whenever the iFCP gateway modifies the S_ID and or D_ID fields of an encapsulated fibre channel frame, the iFCP gateway must calculate a new FC CRC on the reformatted frame. Otherwise, the frame contents are not modified and the gateway simply encapsulates and then transmits the frame.

TRAFFIC ACROSS IFCP LINKS

When transporting FC frames to a remote FC port, the iFCP encapsulates and routes the frame via a predetermined TCP connection for transport across the IP network. Only one fibre channel session

between a pair of FC devices is allowed per TCP connection so that the FC traffic has its own dedicated connection for the duration of the session. TCP connections that are used within an iFCP session are either “bound” or unbound”. An unbound connection is a TCP connection that is not currently active with at iFCP session. The iFCP protocol allows a gateway to establish a pool of unbound connections to reduce the overall connection creation time between a pair of N_Ports. When a N_Port pair needs to communicate, an unbound TCP connection is allocated to an iFCP session via the use of a CBIND message.

In order to keep track of the individual iFCP sessions, the iFCP gateway builds a descriptor about the remotely attached device. This descriptor contains the WWPN, iFCP Portal’s IP address, N_Port ID of the remote device that was assigned by its local iFCP gateway and the N_Port Alias of the remote device. When the descriptor is created, the local iFCP gateway will query the iSNS name server.

When a fibre channel frame is a PLOGI ELS, then an iFCP session will be created. To create an iFCP session between a pair of FC N_Ports, the local iFCP gateway queries the iSNS name server to build the remote N_Port descriptor if one does not already exist, as shown in Figure 6-5. Once the IP address of the remote iFCP gateway is known, local iFCP gateway will allocate a TCP connection to the remote gateway. After the successful creation of the TCP connection between the gateways, the local gateway issues a CBIND session control message and the session assumes the state of “Open Pending” until the remote gateway responds with a CBIND status message of “SUCCESS”. At this point, the iFCP session is now placed into the “Open” state and an iFCP session descriptor is created. This iFCP session descriptor contains the local N_Port ID, the remote N_Port ID and the remote N_Port alias. At this point, communications between the pair of N_Ports can now begin normal fibre channel processing, such as the PLOGI command from the initiator FC port.

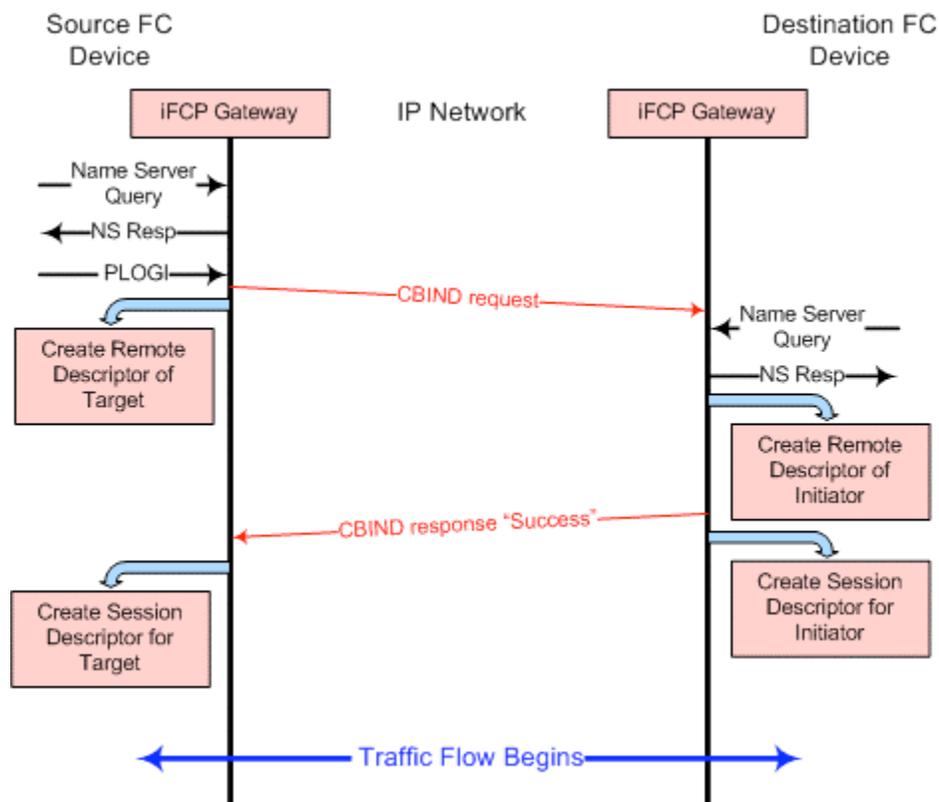


Figure 6-5. iFCP Session Creation

The remote iFCP gateway will not return a “SUCCESS” message for a number of reasons. If the TCP connection is terminated for any reason, the local gateway can attempt to establish the TCP connection again. If an iFCP session already exists between the N_Port pair, then TCP connection is returned to the unbound state since connection is already in the “OPEN” or “Open Pending” state.

iFCP sessions can be terminated for a number of reasons, such as error recovery, FC N_Port rejection or log-out or in response to issues with the TCP connection. Whatever the reason, the iFCP session being terminated stops sending fibre channel traffic over the TCP connection as well as discard any later incoming traffic except for the UNBIND message. If the UNBIND is successful, the gateways may allow the TCP connection to remain in the unbound condition; otherwise, the TCP connection will be terminated with a connection reset command.

Session Control

TCP session control messages are used to create and manage an iFCP session. These messages are passed between peer iFCP portals and the messages are handled by the iFCP layer only. The iFCP gateway transmits messages in the format of a fibre channel extended link service message, as shown in Figure 6-5.

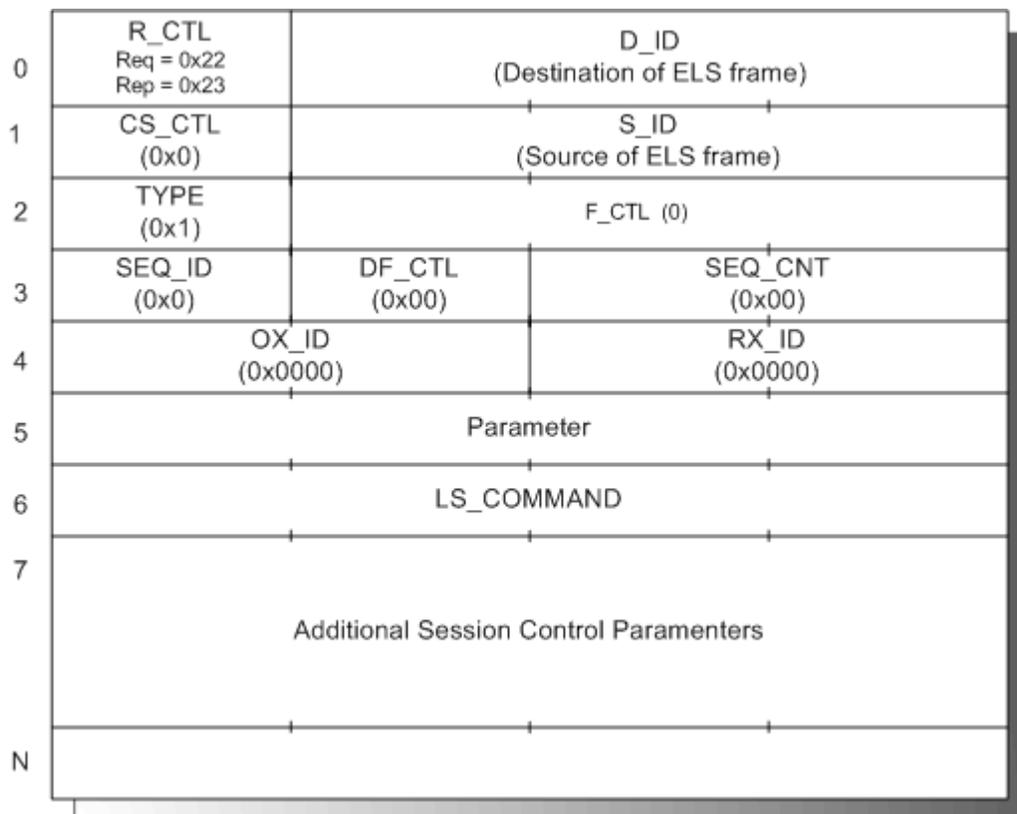


Figure 6-5. iFCP Session Control Message Format

The FC payload area contains the link services command along with any additional session control parameters. Once the FC frame is created, it is encapsulated and transmitted to the remote iFCP gateway for processing.

During extended periods of time, an iFCP session may be terminated due to a lack of traffic or to some failing condition. To check on the connectivity status between a pair of gateways, one of the iFCP gateways can generate a low level of traffic by using LTEST messages. This message requests the LTEST heartbeat by specifying a time interval within the message. Once accepted, the gateway providing the heartbeat will send messages at the specified frequency. If a LTEST message is not received within twice the time period, the receiving iFCP gateway will terminate the iFCP session. A side benefit of the LTEST message is that it can be used to compute the degree of any IP network propagation delay.

ERROR DETECTION AND RECOVERY

Since iFCP has portals in the FC and IP fabrics, it makes use of both fibre channel and TCP/IP error detection functions. From the FC perspective, the iFCP gateway will enforce the R_A_TOV timeout value for the fabric. If there is an error condition within the FC fabric, the gateway will discard all frames that exceed the specified time limits. One component of the propagation time between the FC end devices is the delay associated with the transit time through the IP network, and the limit for this delay is referred to as IP_TOV.

In contrast to the FCIP protocol, iFCP requires the use of a synchronized time base for all gateways such as the Simple Network Time Protocol (SNTP). The resolution of the time reference must be 125 milliseconds or better. If an iFCP gateway loses its synchronized state with the time base, it will abort all TCP sessions and not permit the creation of new sessions until reliable time synchronization is established again.

IFCP SECURITY

iFCP relies on the IPsec protocol suite for data confidentiality and authentication services along with the Internet Key Exchange (IKE) as the key management protocol. Since iFCP can be deployed in a wide variety of IP network configurations, the iFCP security architecture is capable of leveraging the protective services of existing security within the IP network infrastructure such as firewalls, NAT and NAPT services and IPsec VPN services.

With iFCP being a peer-to-peer protocol, iFCP sessions can be initiated by the iFCP gateway at either end of the connections. N_Port identities that are used during the port login process (PLOGI) are considered to be authenticated as long as the PLOGI request comes from the remote gateway over a secure connection with IPsec protection. With the possibility of iFCP traffic traveling over public IP networks, encryption is implemented on a per-packet basis.

Access control properties stem from the requirement that communicating gateways will be known to the iSNS server before they start a communication session. iSNS is an integral part of an iFCP implementation since iSNS is required for discovery services and optional security policies. The level of authentication can vary depending on which authorization scheme is implemented within the iFCP fabric. Examples of authentication schemes include discovery domains, Identity Payloads, and certificate-based authentication.

Chapter 7 - FCIP

Fibre Channel over TCP/IP (FCIP) protocol transports encapsulated fibre channel frames across an IP network and FCIP is described in IETF RFC 3821. FCIP provides a mechanism for a high degree of interconnection between fibre channel islands over IP-based networks, such that the networks coalesce into a unified SAN fabric. FCIP relies heavily on TCP/IP services to be able to work over local area networks (LAN), metropolitan area networks (MAN) as well as wide area networks (WAN).

A key concept of FCIP is the FCIP Entity which encapsulates fibre channel frames within a TCP/IP frame, as shown below in Figure 7-1. FCIP Entities are connected in peer pairings to forward fibre channel frames between regular FC devices, but the FCIP link does not fully mimic the mechanism of an Inter-Switch Link (ISL) to transport FC frames.

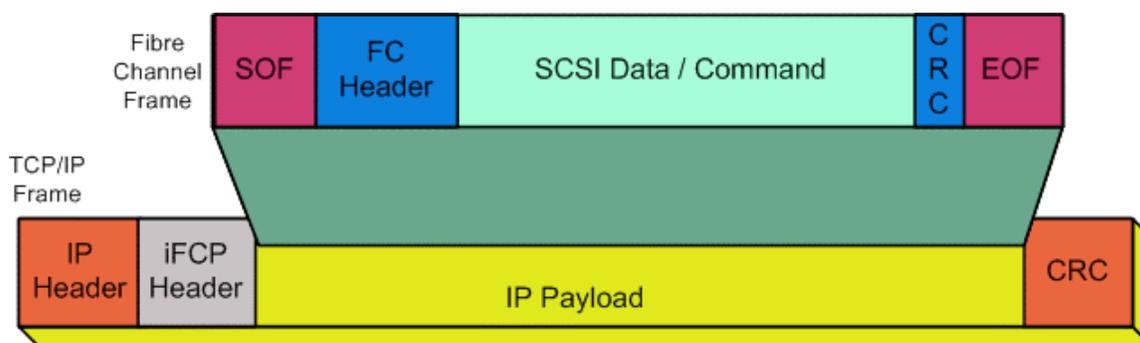


Figure 7-1. FC Frame Encapsulation by FCIP

TERMS AND CONCEPTS

There are a number of terms and concepts that are introduced with the FCIP protocol, as shown in Figure 7-2. A FC End Node is simply a fibre channel device, such as a server or storage, which uses the various services provided by a FC fabric. The FC End Nodes are not aware of the presence of any FCIP Links within the FC fabric. A FC Fabric Entity is the device which contains one or more FC ports as well as one or more FC Entity and FCIP Entity pairs. A physical example of a FC Fabric Entity could be a SAN router. The FC Entity can be considered to be the FC component which transmits and receives FC frames which communicates with an FCIP Entity.

The FCIP Entity is composed of at least one FCIP Link Endpoint (FCIP_LEP) with its associated FCIP Data Engine (FCIP_DE). The FCIP Entity is responsible for forwarding encapsulated FC frames on the IP network along with the FCIP control and services functions. The control and services includes FCIP link initialization, FCIP link dissolution and an interface to certain key IP network functions.

FCIP Entities are paired at both ends of the FCIP link and communicate using TCP/IP. However, the FCIP Entity does not actively participate in FC frame routing, nor do they participate in the discovery of FC source and destination identifiers; which are service features already provided by the fibre channel architecture. The FCIP Entity processes TCP connection requests from the FCIP Link Endpoints that it manages. The FCIP Entity is identified by at least one TCP port/IP address combination. The Internet

Assigned Numbers Authority (IANA) has assigned TCP port 3225 as the well-known port number for FCIP.

The FCIP Link Endpoint is the component which handles a single FCIP Link. The FCIP_LEP contains one FCIP Data Engine for each TCP connection in the FCIP Link. A given FCIP_LEP communicates with exactly one other FCIP_LEP. A FCIP_LEP uses normal TCP based flow control mechanisms for internal resource management as well as setting TCP window sizes based the available local buffer resources and desired throughput. Flow control is critical since the mechanisms of both TCP (window size) and fibre channel (credit based) work independently of each other. If flow constraints do exist along the FCIP link, then the FC Entity and FCIP Entities are responsible for appropriate actions with the fibre channel and IP networks respectively.

Each time a TCP connection is created, the FCIP Entity creates a new FCIP Link Endpoint with a new FCIP Data Engine. The FCIP_DE is the logical component that processes fibre channel frame encapsulation, de-encapsulation and transmission of FCIP frames through a single TCP connection. The FCIP_DE can also detect some data transmission errors and attempt minimal error recovery. There is one FCIP_DE for each TCP connection in the FCIP Link.

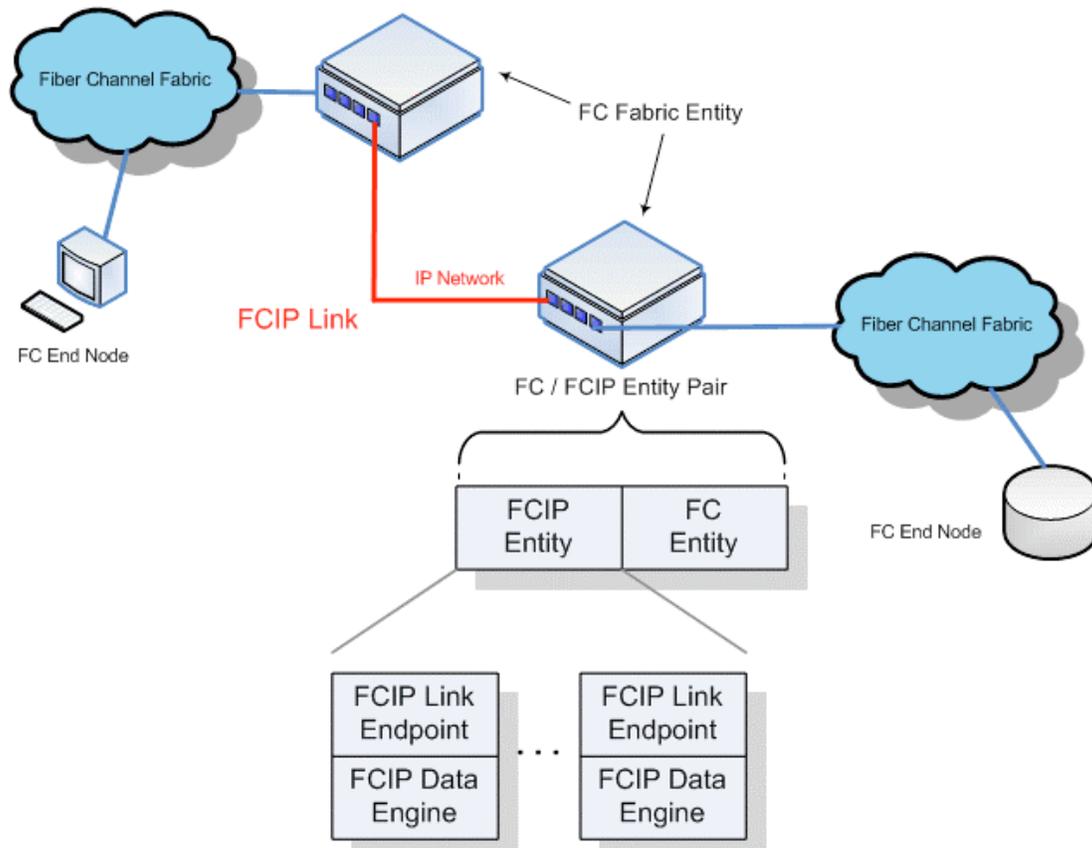


Figure 7-2. FCIP Concepts and Terminology

PROTOCOL MODEL

Although FCIP will merge separate FC Fabrics into a single SAN, not all FC traffic can be transmitted across a FCIP Link. Some examples are FC Primitive Signals, Primitive Sequences and Class 1 FC frames since they can not be encapsulated. Each FCIP Entity must be configured with the IP address and TCP port numbers of other FCIP Entities with which it communicates. This configuration can be accomplished either statically or dynamically. If dynamic configuration is employed, then the Service Location Protocol (SLPv2) must be used during the discovery process. FC Entities must support IP Network security using cryptographically protected authentication, cryptographic data integrity keyed to the authentication process and data confidentiality security features.

There are seven steps for a one-way data frame flow. First, the fibre channel frame is received from the FC Entity at FC Frame Receiver Portal and the frame is passed to the FCIP_DE's encapsulation engine. Then the FC frame encapsulated with the necessary FCIP header information. The new FCIP frame will be passed to the FCIP Entity's Encapsulated Frame Transmitter Portal where it is inserted into the TCP byte stream. Once the FCIP frame is in the IP network, TCP rules of operation are enforced for the in-order delivery of data to the partner FCIP Entity. At this point, the FCIP frame is received by the partner FCIP_DE's Encapsulated Frame Receiver Portal and passed to the De-Encapsulation Engine. The embedded FC frame is extracted from the FCIP frame and checked for validity. If the FC frame is valid, it will be passed to the FC Frame Transmitter Portal and delivered to the FC Entity to continue passing the FC frame to the designated FC End Node. Figure 7-4 shows the FCIP Protocol model.

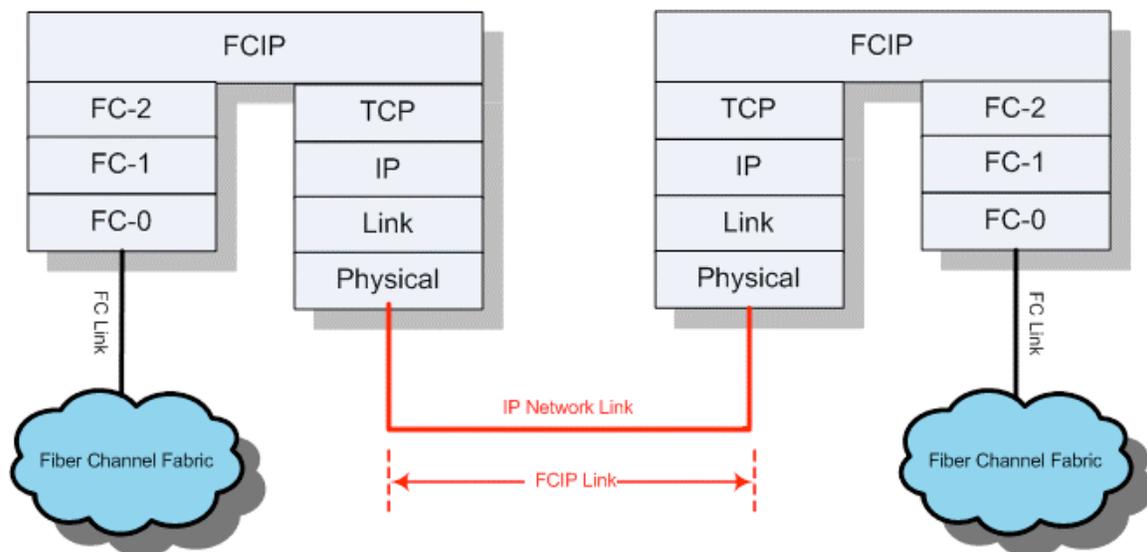


Figure 7-3. FCIP Protocol Model

FCIP FRAME FORMAT

The normal FCIP data frame has fixed format, as shown in Figure 7-4. This format does not account for FCIP special frames (FSF) that are used only once during the establishment of a new FCIP link. With the exception of the FC frame content, all fields are repeated within the same word boundary with the ones

complement of the field used for the second instance. This repetition of the fields was incorporated to assist with error detection.

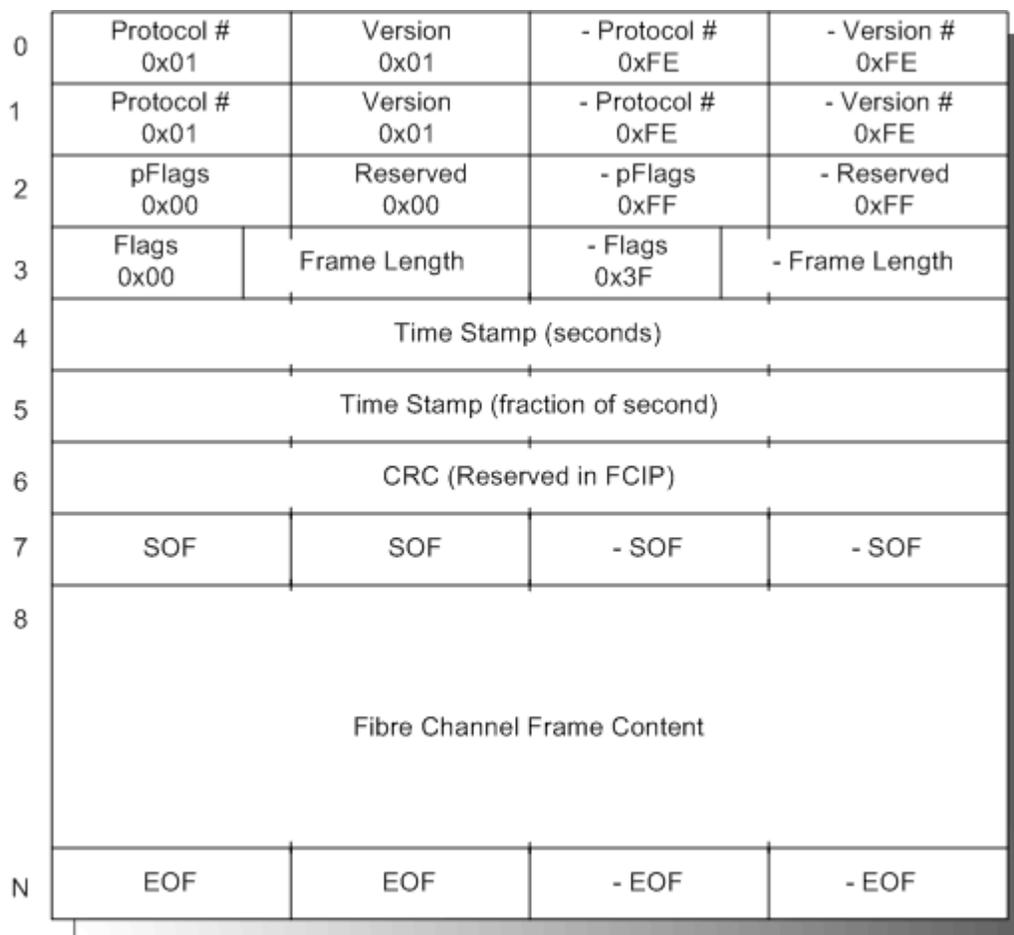


Figure 7-4. FCIP Frame Format

The pFlags field (protocol specific flags) uses one bit to provide information about the type of frame, either FCIP Special Frame (FSF) or regular fibre channel frame. There is a second bit that is used for FSF only to indicate whether the echoed FSF frame has been intentionally altered. The CRC field, word six, is not used for a checksum of the FC Frame Encapsulation header and thus set to the value of 0x00.

The Frame Length field indicates the number of words, based on the fibre channel 4-byte boundary, that are in the FC frame content area. The frame length field is used to determine where in the data stream the next FC Encapsulated Header is located which is used in conjunction with error recovery.

FCIP does not address certain types of error conditions since TCP already has well-established error detection functions. For example, TCP requires in order delivery along with the generation and verification of TCP checksums. The FCIP_LEP relies on these TCP functions working correctly at all times. Frames with errors in the FCIP Encapsulation header will not be passed to the FC Entity. In addition, there are sixteen additional, optional tests on the FCIP Encapsulation Header that may or may

not indicate a loss of synchronization of the byte stream. At least three of these optional tests are required by the FCIP standards, but which tests are used is determined by the vendor. If any of the tests do fail, then an encapsulation error is indicated and the entire FC frame will be discarded.

The Time Stamp Fields are used to compute the IP network transit time experienced by the FC frames. Whenever a FC frame is received by the FCIP_DE for transmission, it may be accompanied by a time stamp value. The time stamp value is only valid if both FC Entities support synchronized time clocks, fibre channel time services, or a SNTF Version 4 time server. Otherwise, the Time Stamp field is set to zero. If the originator and recipient FCIP Entities have synchronized clocking, then the transit time for the FC frame will be compared to the configured "Time to live" value. If the FC frame traversed the FCIP link in less than the maximum time, the FC frame is allowed onward to the FC End Node. Otherwise, the frame will be discarded.

The encoding of the SOF and EOF bit streams from the FC frame in an IP network byte stream requires special formatting and 8-bit code definitions. Thus, there are fields in the FCIP header for both portions of the FC frame. The redundancy of the SOF and EOF fields in the format is a result of concerns that the information must be protected from transmission errors over the IP network. The definitions for various encoded SOF and EOF values are shown in Tables 7-1 and 7-2 respectively.

Fibre Channel SOF	SOF Encoded Value	FC Class
SOFF	0x28	F
SOFi2	0x2D	2
SOFn2	0x35	2
SOFi3	0x2E	3
SOFn3	0x36	3
SOFi4	0x29	4
SOFn4	0x31	4
SOFc4	0x39	4

Table 7-1. Encoded Values for Fibre Channel Start of Frame Signal

Fibre Channel EOF	EOF Encoded Value	FC Class
EOFn	0x41	2, 3, 4, F
EOFt	0x42	2, 3, 4, F
EOFnI	0x49	2, 3, 4, F
EOFa	0x50	2, 3, 4, F
EOFdt	0x46	4
EOFdti	0x4E	4
EOFrt	0x44	4
EOFrti	0x4F	4

Table 7-2. Encoded Values for Fibre Channel End of Frame Signal

ERROR DETECTION AND RECOVERY

Whenever a failure has been detected, the FCIP Entity will notify the FC Entity of the failing condition along with a description of the reason. The burden to recover from discarded frames is then the

responsibility of the FC fabric and/or the FC End Node engaged in the communication session over the FCIP link. If a loss of synchronization is detected, the receiving end will attempt resynchronization by scanning the incoming byte stream for a valid FCIP Frame header by searching for the known pattern of the first twelve bytes of the FCIP header. The resynchronization process is based on the assumption that IP network security and authentication procedures are able to protect the data stream from being replaced by an intruding data stream.

Once resynchronization is started, the FCIP Data Engine checks for the first twelve bytes of the Frame Encapsulation Header, and potential hits from the searching are called a candidate header. After a candidate header is identified, the FCIP_DE uses the frame length field to skip incoming bytes to the expected location of the next FCIP header. The FCIP_DE will continue using this method to successfully predict the occurrence of additional header candidates for at least 4,352 bytes, or twice the maximum length of a FCIP frame. After successfully predicting the next series of encapsulated frames, the data stream is again allowed to continue. However, any failure to regain synchronization will cause the FCIP Entity to terminate the TCP/IP connection.

FCIP CONNECTION CREATION

When a FCIP Entity needs to create a new connection to another FCIP Entity, the first step (shown in Figure 7-5) in the process is to create a new TCP connection to the FCIP Entity. To accomplish this action, the originating FCIP Entity must know the IP address of the destination FCIP Entity; where the IP address has been statically configured or via dynamic discovery using the Service Location Protocol (SLPv2). The originating FCIP Entity sends a TCP connect request to the FCIP well-known port of 3225 at the specified IP address of the destination FCIP Entity. Once the TCP connect request is accepted, the originating FCIP Entity sends a FCIP Special Frame (FSF) as the first frame on the new connection. The format for the FSF is shown in Figure 7-6.

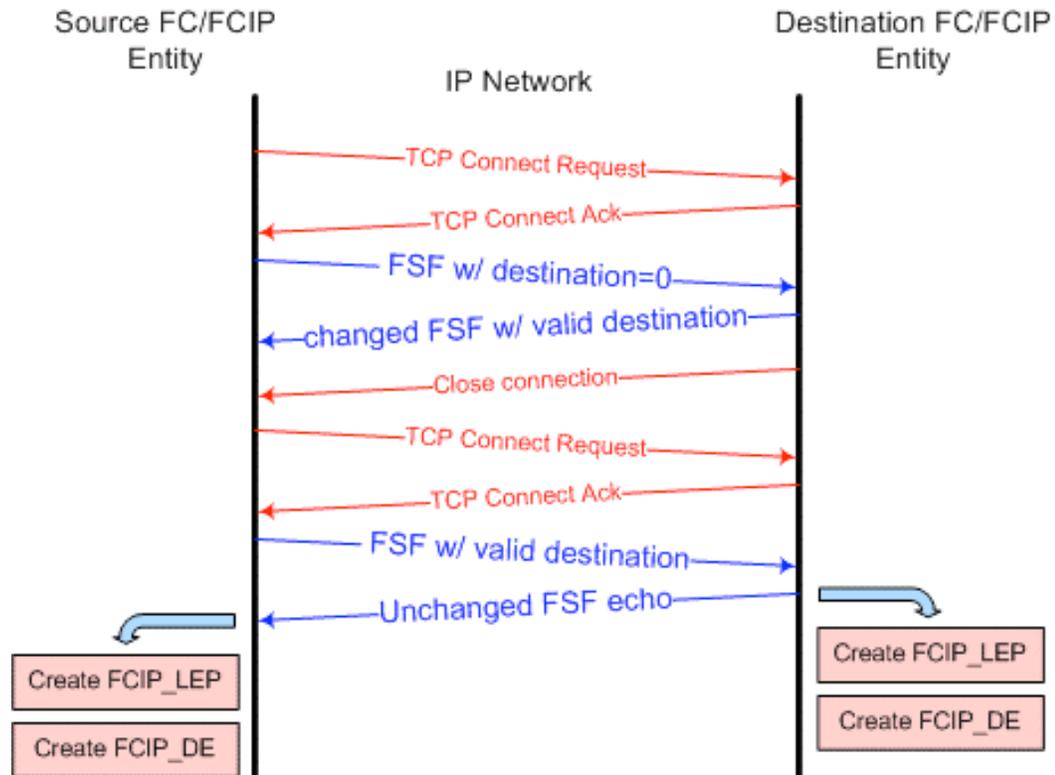


Figure 7-5. FCIP Connection Setup

The purpose of the FSF is to provide information about itself to the destination FCIP Entity, such as the source FC Fabric Entity World-Wide Name, a 64-bit random number Connection Nonce, and the expected destination's FC Fabric Entity World-Wide Name if it is known. If the destination FC Fabric Entity WWN is not known, that field in the FSF will contain 0. If the destination FC Fabric Entity WWN field is 0, then the destination FD Fabric Entity will change the field to its own WWN along with the "Changed" bit flag. Once the FSF is transmitted, the originator waits for the FSF to be echoed back as the first bytes received on the newly formed connection. If the FSF has been changed, the destination will close the TCP connection and the originator FCIP Entity must start the process open connection again. However, it now knows the destination FC Fabric Entity's WWN to be used in the FSF. The destination FCIP Entity may modify the destination FC Fabric Entity WWN field of the received FSF if that field contains 0.

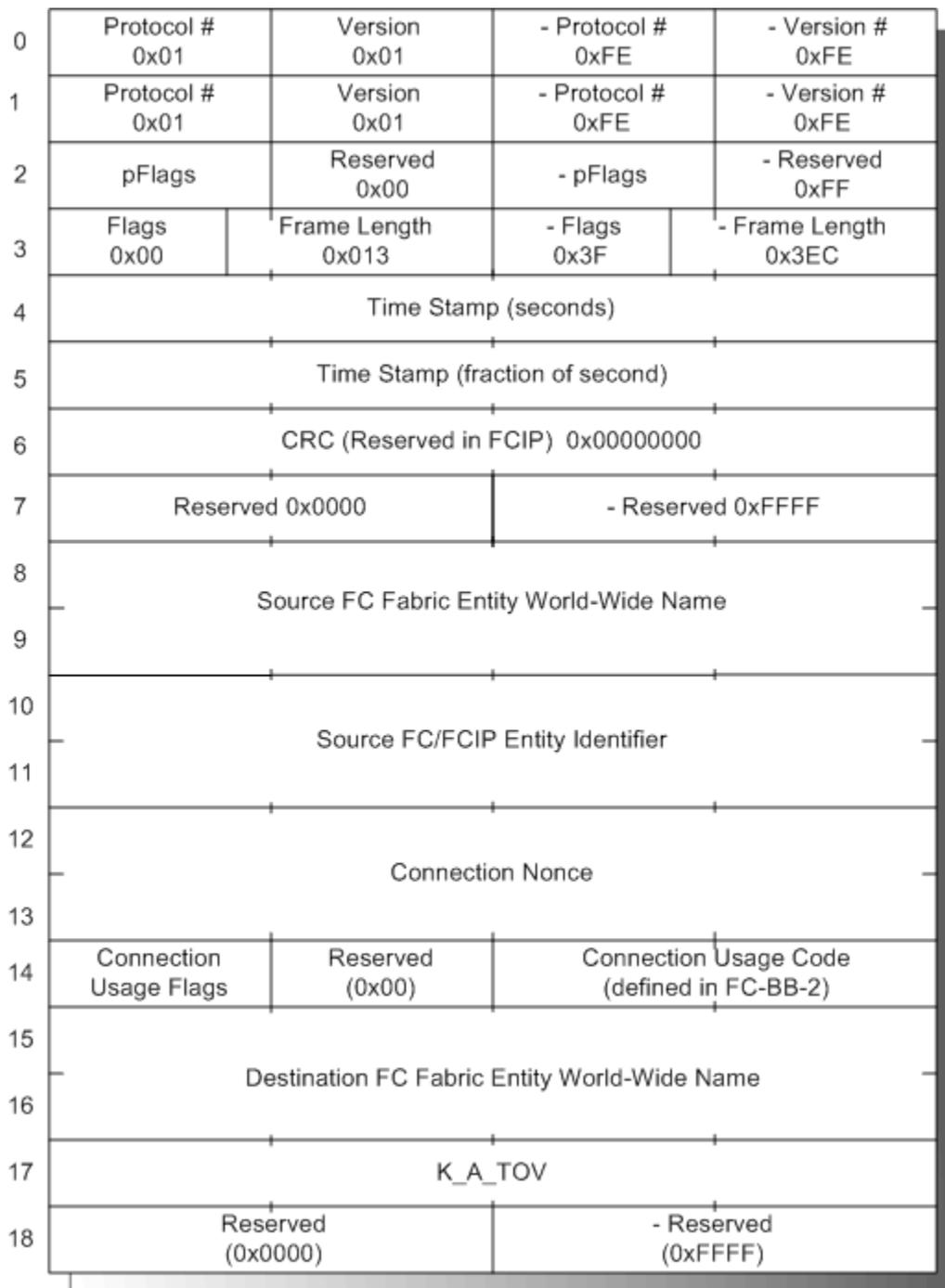


Figure 7-6. FCIP Special Frame Format

The Connection Usage Flags identifies the type of SOF values to be carried on the connection. However, these flags are not necessarily used as a test mechanism of the type of frames being transmitted once normal traffic begins. Instead, these flags are used mainly for setting the appropriate QoS of the newly

formed TCP connection. The K_A_TOV field contains the fibre channel Keep Alive Timeout Value, which is applied to the new TCP connection as specified in FC-BB-2.

If there are no other issues with the FSF, the destination FCIP Entity will echo the FSF back to the originator. If the destination FCIP Entity detects any errors within the FSF, it will discard the special frame and then close the TCP connection. When the originator receives its echoed FSF successfully, it then completes the new connection process by implementing the appropriate Quality of Service condition on the new TCP connection, create a new FCIP_LEP with a new FCIP_DE to service the new TCP connection, and finally inform its local FC Entity of the new connection. The destination FCIP Entity will also take the same steps for the completion of the new TCP connection after it has transmitted the echo FSF. If the destination FCIP Entity requests another new TCP connection from the originator FCIP Entity with an existing connection, it will request the FC Entity to authenticate the source of the TCP connection request.

FCIP SECURITY

FCIP uses the IPsec protocol suite for data confidentiality and authentication services along with Internet Key Exchange (IKE) as the key management protocol. The use of an IP network can introduce some security problems not normally encountered in fibre channel fabrics. The overall effect is that the fibre channel fabric security is only as good as the security of the entire IP network that is used by FCIP links. There is a broad range of potential issues associated with an IP network, such as unauthorized access by various methods, denial of service attacks as well hardware related issues with the TCP link.

The use of FCIP tunneling between FC fabrics and the existing IPsec Security Architecture and protocol suite offers protection from potential threats from an IP network perspective. Since FCIP is based on tunneling the peer-to-peer relationship, the degree of exposure from the IP network is kept to a minimum. In this regards, FCIP entities can implement IPsec Encapsulating Security Payload (ESP) for data integrity.

When establishing a peer-to-peer connection, IKE is used for peer authentication, negotiation of Security Associations (SA) and key management using the IPsec Domain of Interpretation of ISAKMP (DOI). Manual keys are not allowed, but pre-shared keys and digital certificates are permitted. For the purpose of establishing a secure FCIP link, the two entities consult a security Policy Database (SPD) for authentication of the endpoint IP addresses. When the TCP connection is established between two FCIP_DEs, a pair of SAs are created for the connection that are unidirectional. Once established, the FCIP entities will refresh the SAs routinely. Once routine traffic begins flowing, if a datagram fails an integrity check, the receiver drops the datagram which will trigger a retransmission. If too many datagrams are dropped, the FCIP entity may close the TCP connection and then restart the connection as a recovery mechanism.

Appendix A – Applicable Standards

PUBLISHED AND/OR COMPLETED STANDARDS

IETF RFC3643 Fibre Channel (FC) Frame Encapsulation. R. Weber, M. Rajagopal, F. Travostino, M. O'Donnell, C. Monia, M. Merhar. December 2003. (Format: TXT=39980 bytes) (Status: PROPOSED STANDARD)

IETF RFC3720 Internet Small Computer Systems Interface (iSCSI). J. Satran, K. Meth, C. Sapuntzakis, M. Chadalapaka, E. Zeidner. April 2004. (Format: TXT=578468 bytes) (Updated by RFC3980) (Status: PROPOSED STANDARD)

IETF RFC3821 Fibre Channel Over TCP/IP (FCIP). M. Rajagopal, E. Rodriguez, R. Weber. July 2004. (Format: TXT=165907 bytes) (Status: PROPOSED STANDARD)

IETF RFC4171 Internet Storage Name Service (iSNS). J. Tseng, K. Gibbons, F. Travostino, C. Du Laney, J. Souza. September 2005. (Format: TXT=310636 bytes) (Status: PROPOSED STANDARD)

IETF RFC4172 iFCP - A Protocol for Internet Fibre Channel Storage Networking. C. Monia, R. Mullendore, F. Travostino, W. Jeong, M. Edwards. September 2005. (Format: