IP Storage
Internet Draft
Document: <draft-ietf-ips-framework-00.txt> November 17, 2000
Category: Informational

Mark A. Carlson
Sun Microsystems, Inc.

Satish Mali
StoneFly Networks

Milan Merhar
Pirus Networks

Charles Monia
Nishan Systems

Murali Rajagopal
LightSand Communications

A Framework for IP Based Storage

Status of this Memo

Table of Contents

1. Abstract

This document serves as a framework for the creation of standards in the IP based storage working group of the IETF. The environment surrounding IP based storage is explained and an overview of the applicable standards and protocols is provided. References to current and expected IP based storage standards in this area are provided. This document provides a background for participants who are experienced in either the storage industry or the networking industry but needs to understand the applicable standards and

conventions used in the other respective industry.

2. Conventions used in this document

The acronym _SCSI_ is typically used to describe both a parallel-wire bus interconnection used for storage attachment and the command/response protocol first developed for use over that interconnect.

Within this document, _SCSI_ will be used generically to refer to the protocol, independent of transport, while specific reference to its parallel bus transport will use the term _parallel SCSI._

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [RFC2119].

3. Overview

As computer systems grow in size and complexity, their non-volatile storage capacity is often expanded through the addition of external media, such as magnetic and optical disks and magnetic tape. Various interconnection methods have been developed to support this storage expansion, with the Small Computer Systems Interface, or parallel SCSI, being one of the most commonly used.

As originally conceived, SCSI defined both the parallel-wire bus interconnection between a computer system and its storage devices, and the command/response protocol used to transfer information over that interconnection. Over time, both aspects of the specification have been refined and extended, supporting higher speed transport technologies, and a more complex command syntax [SCSI-3]. The SCSI protocol has also been used over transport other than parallel SCSI, such as high speed serial Fibre Channel. This combination, commonly called a Fibre Channel Storage Area Network or Fibre Channel SAN [FCP], permits more flexible connectivity than allowed by parallel SCSI, while maintaining the common SCSI controller API [CAM] used by the computer system to access its storage devices.

These specialized transports provide substantial benefit in the local storage environment, but require dedicated fiber connections to extend to larger geographic distances. This has driven efforts

to reconcile SAN technology with ubiquitous IP transport, to provide a _best of both worlds_ environment.

4. Scope

The IP Storage Working Group is developing several different but complimentary solutions for SAN extension. They all share the goal of maintaining existing computer system and storage device APIs, although they each define different demarcation points where existing practice interfaces with new technology. Some of these proposed solutions are individual submissions and are not official working group work items. This document serves as a survey of these proposed techniques.

Briefly, all of these solutions allow SCSI operations to be performed transparently between an operating system driver and a target device controller, while making different trade-offs as to how much of the Fibre Channel SAN environment will be preserved unchanged, and how much will be translated into IP networking equivalents.

One set of proposed solutions maintains the switched Fibre Channel SAN environment, but introduces IP as a supported fabric or link technology. Due to the mixed nature of the resulting network, components of two different network architectures (Fibre Channel SAN and Internet) may need to be supported in the operational environment.

It should be noted that SCSI is only one of a number of higher-layer protocols having mappings to Fibre Channel, so solutions of this type have the potential of supporting applications beyond SCSI storage access.

FCIP provides a standard way of encapsulating FC frames within TCP/IP, allowing islands of FC SANs to be interconnected over an IP-based network. TCP/IP is used as the underlying transport to provide congestion control and in-order delivery of error-free data. All classes of FC frames are treated the same -- as datagrams. End-station addressing, address resolution, message routing, and other fundamental elements of the network architecture remain unchanged from the Fibre Channel model, with IP introduced exclusively as a transport protocol for an inter-network bridging function. [FCoverIP].

iFCP is a gateway-to-gateway protocol for the implementation of a

fibre channel fabric over a TCP/IP transport. Using iFCP, all traffic between fibre channel devices is routed and switched by TCP/IP network components instead of fibre channel components. Its goal is to bring IP technology to the considerable installed base of fibre channel storage products [iFCP]. This is an individual submission and is not currently a work item of the working group.

The mFCP protocol is a variant of iFCP that provides Fibre Channel fabric services to FCP-based Fibre Channel devices using a high-performance, reliable IP network. mFCP uses the UDP transport protocol to facilitate high performance, and assumes that reliability and flow control will be handled by the physical infrastructure. mFCP's primary objective is to allow interconnection and networking of existing Fibre Channel devices over an IP network [mFCP]. This is an individual submission and is not currently a work item of the working group.

A somewhat different solution maintains the SCSI protocol interface unchanged, with IP introduced as another transport beneath it, much as Fibre Channel was introduced as an alternative transport to parallel SCSI. In this model, end-station addressing, address

resolution, message routing etc. all follow the Internet model. This solution allows the existing IP support infrastructure to be leveraged in maintaining the SAN environment, supporting all operations that may be performed within the SCSI command syntax [iSCSI].

5. Applicable Protocols

It is recommended that implementers familiarize themselves with the applicable protocols and standards that exist from the IETF and Fibre Channel standards organizations. Appendix A is an overview of some of these standards and the procedures that are used in the IETF.

IP based storage standards will leverage these existing standards for transport, management, discovery and naming.

6. Environments

6.1 Naming and discovery

Every storage-related entity could find information about other

storage resources in several ways. The Fibre Channel based storage resources have a state transition where every node goes through the process of login for the fabric. The process of login in Fibre channel assigns a World Wide name to the device. This Worldwide Name is unique in that domain and is used by other devices to uniquely address this device.

It is required that the IP based storage shall support the naming architecture of SAM-2. In order to fulfill the SAM-2 architecture requirements, it is necessary to understand the naming and discovery process used in Fibre Channel based networks and IP based networks.

The naming and discovery architectures for IP based storage resources can be seen from two different views. One view is to provide naming and discovery process conventions using current IP based architecture. This may involve using various standards such as DNS, LDAP, and URL. The other view is to extend the naming and discovery process of the current Fibre Channel based architecture to fit within the boundaries of IP. Here is a summary about the various proposals and their issues.

Fibre Channel based Naming and Discovery:

The Fibre Channel based storage devices are connected in a meshed topology to provide redundant paths from initiator to target device. The initiator can identify these redundant paths by traversing different paths to the target and discovering that the multiple

paths lead to the same target. In case one of the paths becomes unavailable, the initiator conducts communication over the redundant path. It may be also necessary to take into account changes proposed by T10 committee that allows LUN renumbering. Refer to SPC-2 provisions for LU Identifiers (Vital product data page 83h [SPC-2, p. 203]).

IP based naming and discovery:

In IP based network, connected devices are also uniquely identified by their IP address and the DNS name. There maybe multiple paths from one device to the other. However, the main difference is that the routers along the path take care of managing a unique path from one end to another end. In case one of the paths has a problem, the alternate path is opened by the router, without end nodes being involved and aware of the transition-taking place for the in-between paths.

Here is one way of representing IP based storage resources.

URL based naming:

One of the iSCSI-naming schemes involves use of URL. The proposal here is to use the popular World Wide Web based URL encoding scheme to denote the storage entities within a target.

A URL for the target may have the following form:

scsi://hostname/path/with/

A URL referring to a specific LU has the following form:

scsi://hostname/path/with/?LUN=lunnumber?WWN=wwnnumber

If no LUN= term appears in the URL, then LUN 0 is assumed.

The WWN= term is optional. If present, the party should verify that the WWN in the LU's Device Identification Inquiry Page corresponds to the WWN.

An example of the above scheme will be:

scsi://ips.ietf.org/tape/?WWN=0a050a4bcdefa

Refers to LUN 0 of target scsi://ips.ietf.org/tape/. After connecting, the initiator verifies that the WWN of the LU is 0a050a4bcdefa.

Discovery:

Traditional SCSI discovery is based on a "bus walker" paradigm. Here every logical combination is searched for availability of a valid SCSI device. For a small number of devices connected to the parallel SCSI bus, this approach was not too time consuming. But when SCSI was implemented over Fibre Channel, the Fibre Channel introduced notion of World Wide Name (WWN) as device identifier. The WWN, which is made of 64 bits, could not iterate through every 64 bit WWN and still boot quickly. The Solution to that problem involved using a multi-round distributed address assignment scheme for Fibre Channel Arbitrated Loop environment and to query the name server for all known ports in Fibre Channel Switched environment. The Name Server in the Fibre Channel Switch kept track of each connected WWN.

IP based storage systems goal is not to alter any discovery mechanism in IP, as current mechanism based on DNS is quite evolved and the current management solutions use the existing structure to get IP address information, however, addition IP based mechanisms for discovery and naming may be specified by the working group. It is also required that there is a way to identify that the discovered entity does indeed have an IP based storage device. The problem is compounded by the existence of NAT and firewalls installed in the working environments. NAT and firewalls can hide a private network behind a public IP address, effectively shielding access to all the IP based storage devices from outside. The external device then cannot make an inquiry to the individual IP based storage device.

Here is a brief overview of the issues involved and likely reasons behind each option.

If the SCSI over IP discovery mechanism is maintained to be same as current DNS method, then there is no additional burden on understanding and administrating IP based systems. It can then be used in conjunction with current management solutions. The discovered storage entity can be verified by querying on iSCSI well-known port. This port is assigned to iSCSI by IANA and is used to service iSCSI based requests. Absence of any response from this well-known port will indicate absence of IP based SCSI device. NAT/Firewall is still a problem and need a different type of solution to reach to storage devices on the private side of the NAT. This can be handled by using IPsec based tunneling protocols that can set up a private tunnel to the private network, from where all the devices are now accessible. However, this means that IPsec based session may be required to be initiated during boot process to access boot drive inside the NAT/firewall area.

One issue that may have to be addressed is the reliability of DNS access. In mission critical environments it may not be enough to solely depend on the availability of primary or secondary DNS servers, and IP based access to volumes may be required.

Global context for third party names is an open issue in T10. In general, the Initiator of a 3rd party command must use names that resolve to the desired LUNs from the third party command Target's

naming perspective. It is required that both Initiator and Target may have to share the same naming context. This needs to be handled in the IP based storage devices similar to Fibre Channel based storage.

The two networks meet:

The issues that need to be resolved for IP based storage are:

* Naming scheme that takes into account current target LU and LUN based convention and maps it to IP based storage.
* Discovery process that can work with or without firewalls and NAT units.
* Authentication scheme that allows connections and management of the storage devices.
* Provide SCSI third party operations that allow hand over of naming schemes.
* Provide a security scheme in SCSI, which defines proxy-naming scheme that allows a given LU within a target to be addressed by different LUNs.
* Scalability of the storage domain such as storage Service Providers requiring large number of storage devices.

6.2 The Internet Storage Environment

Much of this section is based upon [iSCSI-REQ]. The material has been expanded, where appropriate, to reflect all the storage-related solutions being addressed by the IPS working group.

The use of IP technology for storage may be divided into three broad categories:

* The direct attachment of volume/block-oriented storage devices to an IP network,
* The implementation of IP Fabrics, for the interconnection of fibre channel storage devices using IP routing and switching elements.
* Backbones for the interconnection of storage clusters of fibre channel storage area networks across data centers.

For these applications, the IP/Ethernet infrastructure offers the following compelling advantages:

* Increasing performance and reduced cost driven by Internet economics and "IP convergence"
* Seamless conversion from local to wide area using IP routers
* Emerging availability of "IP datatone" service from carriers, in preference to ATM or SONET or T-1, T-3 services
* Protocols and middleware for management, security and QoS
* Economies arising from the need to install and operate only a single type of network

IP technology will be applied to the following storage applications:

* Storage area networks, providing local storage access, consolidation, clustering and pooling (as in the data center),
* The integration of storage area networks across data centers,
* Remote disk access (as for a storage utility),
* Local and remote synchronous and asynchronous mirroring between storage controllers,
* Local and remote backup and restore,
* Evolution with SCSI to support of emerging object-oriented storage model.

And the following connection topologies are contemplated:

* Point-to-point direct connections,
* Dedicated storage LAN, consisting of one or more LAN segments,
* Shared LAN, carrying a mix of traditional LAN traffic plus storage traffic,
* LAN-to-WAN extension using IP routers or carrier-provided "IP Datatone",
* Private networks and the public Internet.
* Backbones, interconnecting clusters of fibre channel or IP-based storage area networks.

Local-area storage networks will be built using Ethernet LAN switches. These networks may be dedicated to storage, or shared with traditional Ethernet uses, as determined by cost, performance, administration, and security considerations. In the local area, TCP's adaptive retransmission timers will provide for automatic and rapid error detection and recovery, compared to alternative technologies.

IP LAN-WAN routers will be used to extend the IP storage network to the wide area, permitting remote disk access (as for a storage utility), synchronous and asynchronous remote mirroring, and remote backup and restore (as for tape vaulting). In the WAN, TCP end-to-end will avoid the need for specialized equipment for protocol conversion, ensure data reliability, cope with network congestion, and automatically adapt retransmission strategies to WAN delays.

6.2.1 Internet Storage Migration Issues

This section discusses issues that arise as storage subsystems, migrate to the Internet and are released from the implicit constraints imposed by closed storage interconnects. These constraints have related to:

Connectivity _- the achievable device population,

Openness _ The potential for sharing network resources with other devices and applications,

Device address volatility _ The stability of the physical device address over time.

Reach _ The physical span of the network and its effect on I/O performance and behavior,

The storage subsystem model _ the process of defining and assembling a collection of storage devices dispersed throughout the network into a pool of resources available to an application.

The device's view of the storage network _ The assumptions a device may make regarding how the storage network appears to other devices.

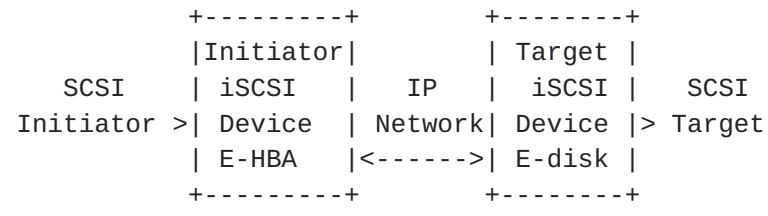
The paradigm for storage transactions _ The basic command-response model for I/O operations,

Security _ The effects of extending the storage subsystem beyond the physical confines of the chassis or computer room. These considerations are discussed in section [???].

6.3 iSCSI Environment

6.3.1 System models and addressing

The basic system model for iSCSI is that of an extended virtual cable, connecting a SCSI initiator device to a SCSI target device.

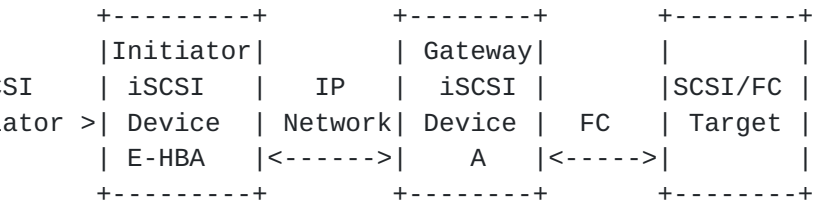


Both iSCSI initiator and iSCSI target are identified completely by their IP addresses, although some implementations may also assign a dummy SCSI or FC address to an internal interface for purposes of compatibility (e.g. _ to support an existing SCSI I/O driver.) _

Unfortunately, reality has not yet caught up with the simplicity of this model. In the near term, it appears there will be several implementations of initiator iSCSI adapters (so-called _Ethernet Host Bus Adapters_) but few disk drives with native Ethernet/IP interfaces.

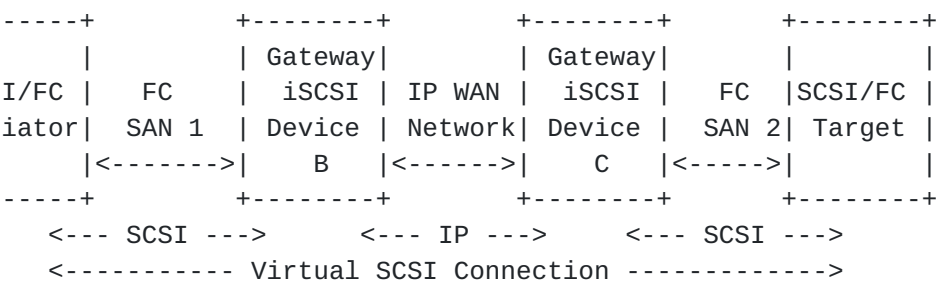
Instead, considerable effort has gone into the development of iSCSI

proxy devices, which provide a gateway from the IP Storage protocol to an existing storage interconnection such as Fibre Channel or parallel SCSI, which in turn attaches to conventional storage devices. For example, a server using a native iSCSI HBA may be attached to a conventional storage array via Fibre Channel in this manner:



As with the previous model, the initiator iSCSI device and the IP interface of gateway device A are identified by an IP address. In this example, gateway device A also has a Fibre Channel interface (i.e. _ either a N_Port or a NL port,) connected to a Fibre Channel fabric or arbitrated loop as an initiator, allowing iSCSI commands received from the IP network to be sent to the actual target device. As the gateway device only presents a single address to the Fibre Channel network, all iSCSI initiators using this gateway will appear on the Fibre Channel as originating from a single address.

Similar iSCSI gateway devices have also been proposed as WAN interconnections between existing Fibre Channel SANs.



In this configuration, gateway device B appears as a target device on FC SAN 1, accepting attachments as a proxy for the actual SCSI target residing on FC SAN 2. Similarly, gateway device C appears as an initiator device on FC SAN 2, initiating attachments as a proxy for the actual SCSI initiator on FC SAN 1.

SCSI operations occurring between the initiator and target are actually performed in three stages; by a SCSI operation occurring between initiator and Gateway B, by transfer of the request via iSCSI between Gateway B and Gateway C, and by a separate SCSI

operation performed by Gateway C on the target device. In other words, the initiator's SCSI requests are accepted locally, but the responses are delayed by the response time of the remote proxy operation, plus the WAN round-trip delay.

Note that independent local contexts (address space, Name server contents, Zoning configuration, etc.) are maintained for both SAN 1 and SAN 2, unless they are explicitly synchronized through configuration or other means outside the scope of this document.

This type of _remote access_ to SCSI devices across the WAN should be contrasted with the SAN extension solutions provided by FC/IP or

iFCP/mFCP (as described in sections 6.4 and 6.5 below.) Those solutions extend the context of a Fibre Channel SAN by transparently bridging protocol messages from one SAN to the other, without terminating and remotely recreating SCSI sessions.

6.3.2 Operation

iSCSI [iSCSI] is a connection-oriented command/response protocol. An iSCSI session begins with an iSCSI initiator connecting to an iSCSI target (typically, using TCP) and performing an iSCSI login. This login creates a persistent state between initiator and target, which may include initiator and target authentication, session security certificates, and session option parameters.

Once this login has been successfully completed, the iSCSI session continues in _full feature_ phase. The iSCSI initiator may issue SCSI commands encapsulated by the iSCSI protocol over its TCP connection, which are executed by the iSCSI target. The iSCSI target must return a status response for each command over the same TCP connection, consisting of both the completion status of the actual SCSI target device and its own iSCSI session status.

An iSCSI session is terminated when its TCP session is closed. This shutdown is graceful if all outstanding SCSI operations have been permitted to complete; closing a session while SCSI operations are outstanding may require implementation and target-specific cleanup actions to be performed.

6.3.2.1 iSCSI data flow

The same TCP session used for command/status is also used to transfer data and/or optional command parameters.

For SCSI commands that require data and/or parameter transfer, the (optional) data and the status for a command must be sent over the same TCP connection that was used to deliver the SCSI command.

Data transferred from the iSCSI initiator to iSCSI target can be either unsolicited, or solicited. Unsolicited data may be sent either as part of an iSCSI command message, or as separate data messages (up to an agreed-upon limit negotiated between initiator and target at login.) Solicited data is sent only in response to a target-initiated Ready to Transfer message.

Each iSCSI command, Data, and Ready to Transfer message carries a tag, which is used to associate a SCSI operation with its associated data transfer messages.

6.3.3 Auxiliary services

6.3.3.1 Discovery services

Close examination of the _inter-SAN gateway_ model described in section 6.3.1 above reveals several operations that rely on some form of discovery service.

If gateway devices B and C are to communicate across the IP network, one must know the IP address of the other. The definition of services to detect and identify these resources is the subject of ongoing work.

A different form of discovery occurs within the Fibre Channel regions of a mixed IP storage network. Consider the right port of gateway C, which for purposes of discussion is assumed to be attached to a Fibre Channel switch. To access that switch, the gateway must log into the fabric, identifying itself as an initiator device. It may then query the Name server on SAN 2 for the addresses of available target devices, and probe them to determine what Logical Units they contain.

Similarly, the left port of gateway B logs into its attached Fibre Channel switch, identifying itself as a target device, so that its address can be added to the Name server on SAN 1. If probed by a Fibre Channel initiator device, gateway B initiates an iSCSI session with gateway C, allowing the SCSI requests to be satisfied by the actual SCSI target device.

6.3.3.2 Relationship to NAT devices

As shown above, a single storage network may contain regions of IP, Fibre Channel, and parallel SCSI connectivity, each using different addressing schemes. Moreover, many networks of today's Internet are reachable only through NAT, for reasons both technical and administrative.

This fragmentation of addressing space imposes additional constraints on any resource identification or location service; the address of a resource may depend on who is asking, and what path a connection to that resource would take. For example, a target device might be identified by its Fibre Channel address to another device on the same fabric, by the local (e.g. _ Net.10) IP address of an iSCSI gateway to a local IP Storage initiator, and the public IP address of a NAT device to an remote iSCSI initiator.

This same concern applies to information passed within the iSCSI protocol itself; resource location strings and other meta-addressing information must be interpretable correctly within the context of the destination device.

Finally, address information presented through a network management interface, such as SNMP, must include sufficient contextual information to be unambiguous. In particular, it should be noted

that network management applications might aggregate data from multiple devices into a consolidated report.

6.3.4 Configuration issues

6.3.4.1 Address management

From the IP perspective, addresses can be assigned to iSCSI devices using conventional dynamic address assignment tools.

However, system-level concerns, such as allowing a server to boot via its iSCSI interface without the support of external services, may require static IP address assignment or other self-contained solution to be used in some environments.

iSCSI devices acting as gateways may also participate in Fibre Channel Arbitrated loops and switched fabric SANs as end-stations, acting as an initiator and/or target device. As such, they will also interact with existing SAN address management systems (e.g. _ the Fibre Channel SAN name server.)

6.4 FC over IP (FCIP) Environment

6.4.1 FCIP Environment

FCIP is a protocol specification that allows a FCIP device to transparently tunnel FC frames across an IP-based Network. This capability is applicable in edge devices that interface with a FC Switch [T11] or FC-BBW device [FCBB]. The term FCIP device generally refers to an edge device that encapsulates FC frames into TCP segments and re-assembles TCP segments to regenerate FC frames.

The motivation behind connecting remote sites using the FCIP device is to enable transparent disk or tape backup and live mirroring, or simply distance extension between two FC devices or FC Switch clusters (SAN islands) across an IP-based network. The transparency implies that the FCIP edge devices need not examine the FC frame content or even preserve FC state information and effectively provide a fast frame forwarding function just like FC switches. That is the FCIP device is a transparent translation point. This also means that the FC datagrams are expected to comply with existing FC specifications as far as delay latency is concerned. The FC traffic may span LANs, MANs and WANs, so long as this fundamental assumption is adhered to.

The IP network is not aware of the FC payload that it is carrying. Likewise, the FC fabric and the FC end nodes are unaware of the (TCP) IP-based transport.

6.4.2 Fibre Channel Backbone Switches Background

Fibre Channel (FC) Standards [T11] describe the operation of and interaction between FC Switches. Two distinct levels of switch interconnections are specified. Autonomous Regions (AR) are defined to allow clusters of FC Switches to be connected across a backbone network called an FSPF-backbone. An Autonomous Region is administratively defined with each AR encompassing one or more FC Domains. The FSPF-backbone network is formed from one or more Backbone Switches (BSW) that run the FSPF-backbone routing protocol. The FSPF-backbone routing protocol is based on OSPF and the FSPF backbone may consist of an arbitrary mesh network. A BSW may communicate with multiple neighbors. As specified in [3], native FC frames traverse the backbone between BSW neighbors. The FSPF-backbone Routing Protocol messages are exchanged between BSWs on the FSPF backbone.

An example network consisting of 4 ARs and an FSPF backbone consisting of 3 links is given in Fig. 6.4.1. There is no restriction

in adding other links to this network as needed. The connection between BSWs below may in fact form a fully connected mesh.

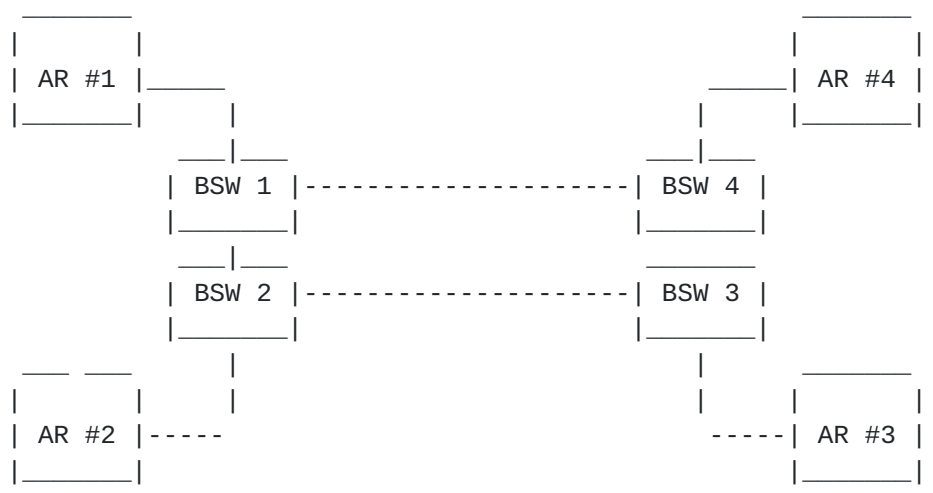


Fig. 6.4.1 Example Network showing FSPF-Backbone Switching Architecture

Note:
BSW 1 knows it is connected to BSWs 2 and 4;
BSW 2 knows it is connected to BSWs 1 and 3;
BSW 4 knows it is connected to BSWs 1.

An FCIP device provides a single, logical interface to the FSPF-backbone protocol connecting multiple BSW neighbors on the IP-network. From the FSPF-backbone routing's point of view, the connection to each neighbor on the IP-network is treated as a separate logical FC link.

In FCIP, the native FC frames are first encapsulated in TCP segments, which then traverse the IP-based network. The IP network provides a new transport path for each emulated FSPF-backbone link.

The IP network itself may consist of any number of hops between two FCIP devices. Also, the route taken by the IP packet between any two FCIP devices is dictated by normal IP routing.

A functional and logical diagram of an IP-based FSPF-backbone for the example network given in Fig. 6.4.3 is shown in Fig. 6.4.2. In this figure, each BSW is logically connected to other BSWs.

The IP-based network has transformed the FSPF-backbone into a fully connected network. From the perspective of each BSW all remote BSWs

therefore appear to be neighbors. The FSPF-backbone routing protocol computations would make the IP-based network topology appear as a fully connected mesh.

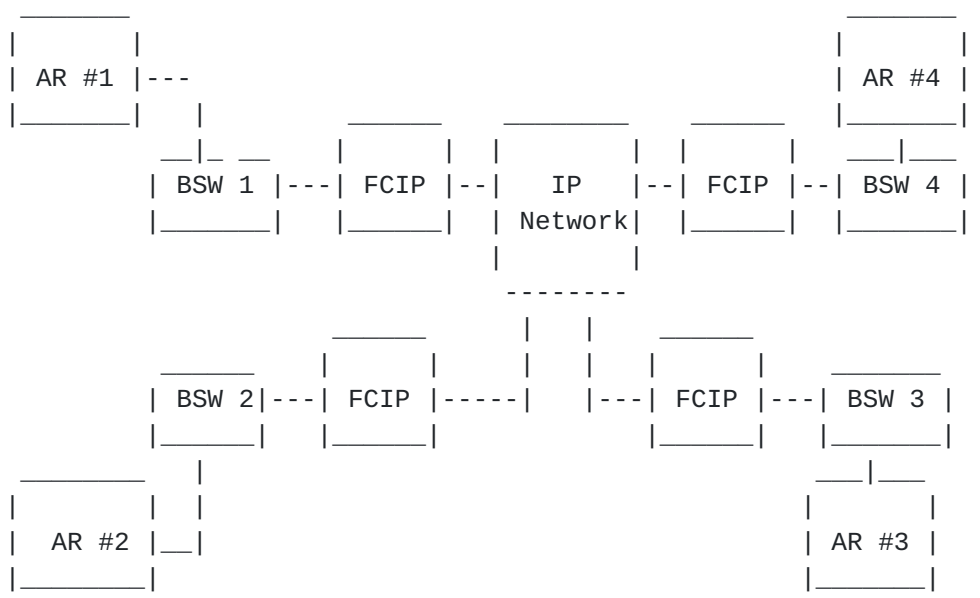


Fig. 6.4.2 Example Network showing an IP-based FC Backbone Switching Architecture

The FSPF-backbone routing protocol exchanges specified in [T11] between BSWs occur transparently to the FCIP devices. Encapsulated FC frames are routed on the IP network according to the normal IP routing procedures. In this mode, the FSPF backbone routing protocol lies over the IP network and has no knowledge of the underlying IP protocol and IP routing or the underlying technology that carries the IP datagram. This concept is shown in Fig.6.4.3

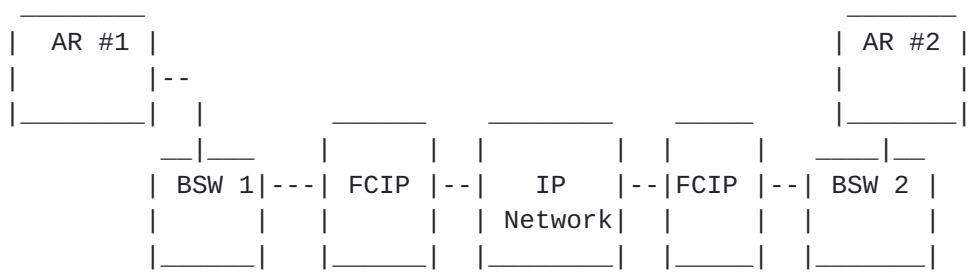


Fig. 6.4.3 FC packet routing over IP based backbone network

Note: IP Network routing may consist of multiple paths

6.4.3 Fibre Channel FC-BB Background

ANSI T11 FC-BB Standards [FCBB] specifies how ARs may be connected across a wide area. FC-BB is specifies a FC-BBW device that allows FC Switches to be connected to a FC-BBW device B_Port. The FC-BBW device has an interface to the wide area. More than one FC-BBW device may be connected to the wide area. Fig. 6.4.4 shows an example of the FC-BB Architecture showing an ATM Network. Currently, SONET is also specified in [FCBB]. In future, Gigabit Ethernet will be specified. FC-BB2 charter clearly states that any IETF protocol specified for carrying FC over IP-based network will be leveraged. It is therefore the intent of the FCIP specification to consider the FC-BB and FC-BB2 architectures while drawing this specification.

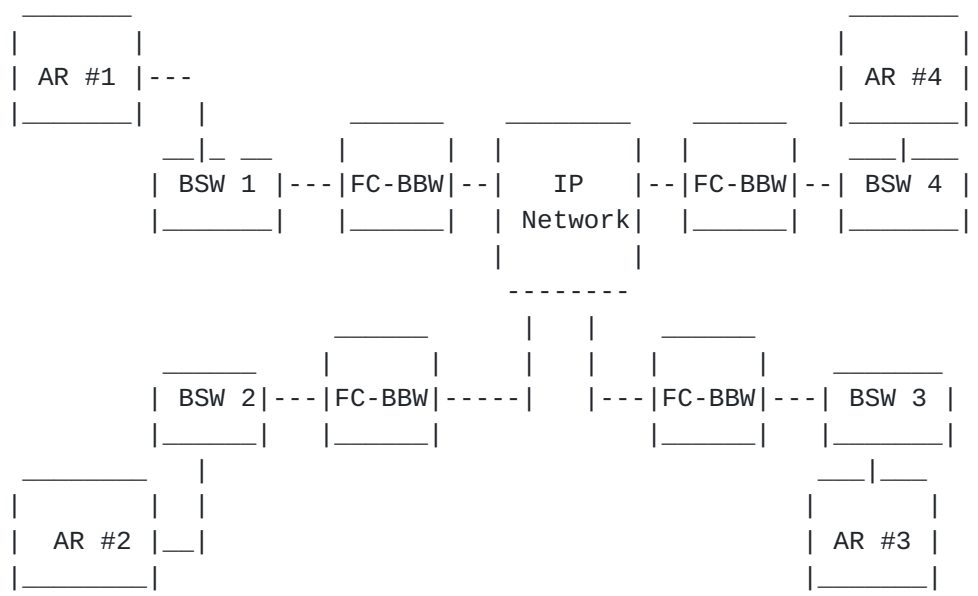


Fig. 6.4.4 Example Network showing an FC-BB Architecture

6.4.4 Introduction to FCIP Protocol

The purpose of the FCIP specification is to specify a standard way of encapsulating FC frames over TCP/IP and to describe mechanisms that allow islands of FC SANs to be interconnected over IP-based networks. FC over TCP/IP relies on IP-based network services to provide the connectivity between the SAN islands over LANs, MANs, or WANs. The FC over TCP/IP specification relies upon TCP for congestion control and management and upon both TCP and FC for data error and data loss

recovery. FC over TCP/IP treats all classes of FC frames the same -- as datagrams. Any FC concerns arising from tunneling FC traffic over an IP network, including security, data integrity (loss), congestion, and performance. This will be accomplished, where appropriate, by utilizing the existing IETF-specified suite of protocols.

A fundamental assumption made in this specification is that the FC traffic is carried over the IP network in such a manner that the FC fabric and all FC devices on the fabric are unaware of the fact. This means that the FC datagrams must be delivered in such time as to comply with existing FC specifications. The FC traffic may span LANs, MANs and WANs, so long as this fundamental assumption is adhered to.

FC operates at Gigabit speeds. This specification will be written such that FC traffic may be transported over an IP backbone that has been engineered to have equivalent or better bit-error-rate (BER) and line speed as the Fibre Channel environments being bridged. While the tunneling of Fibre Channel traffic over other IP networks not so engineered is not precluded, the above environment is an important one, and this specification has been written so that to optimize for such traffic, while not over-burdening other configured IP networks.

All FCIP protocol devices are peers and communicate using TCP/IP. Each FCIP device behaves like a TCP end-node from the perspective of the IP-based network. That is, these devices do not perform IP routing or IP switching but simply forward FC frames.

There is no requirement for an FCIP device to establish a login with a peer before communication begins. However, FCIP devices may authenticate the IP packet before accepting it using the IPSec protocols. Each IP datagram is treated independently and an FCIP device receiver simply listens to the appropriate socket value contained in the TCP header.

Each FCIP device may be statically or dynamically configured with a list of IP addresses corresponding to all the participating FCIP devices. Dynamic discovery of participating FCIP devices may be performed using Internet protocols such as LDAP, DHCP or other discovery protocols. It is outside the scope of this specification to describe any static or dynamic scheme for participating FCIP device IP address discovery.

Discovery of FC addresses (accessible via the FCIP device) is provided by techniques and protocols within the FC architecture. These techniques and protocols are described in Fibre Channel ANSI standards [T11]. The FCIP device does not participate in the discovery of FC addresses. Routing in the IP plane and the FC plane are largely independent.

The exact path (route) taken by an FC over TCP/IP encapsulated packet follows the normal procedures of routing any IP packet. From the

perspective of the FCIP devices this communication is between only two FCIP devices for any given packet.

An FCIP device may send FC encapsulated TCP/IP packets to more than one FCIP device. However, these encapsulated packets are treated as separate instances and are not correlated in any way by the FCIP protocol devices. The source FCIP device routes its packets based on the 3-byte FC destination Address Identifier (D_ID) contained in each FC frame.

An IP packet may make use of the IPSec protocol to provide secure communications across the IP-based network.

Any re-ordering of data link frames due to MTU fragmentation will be recovered in accordance with a normal TCP reliable delivery behavior.

Any re-ordering of FC frames due to IP packet re-ordering will be resolved via the standard TCP reliable delivery behavior.

FCIP relies on both TCP error recovery mechanism and normal FC recovery mechanisms to detect and recover from data loss due to any loss of IP packets.

FC over TCP/IP encapsulated IP packets shall indicate the use of the Premium Service in the DSCP bits in the IP header.

The TCP layer in the sending FCIP device shall package each FC frame handed down by the FC layer into a TCP segment and set the PSH control flag in the TCP header to ensure that the entire FC frame is sent in one TCP segment. If the FC frame cannot be packaged in one TCP segment (e.g. the FC frame size is greater than TCP MSS), the last part of the FC frame must occupy one TCP segment and the PSH of that segment must be set.

6.5 iFCP, mFCP and iSNS

The iFCP, mFCP and iSNS protocols are elements of a framework for the implementation of Fibre Channel fabric capabilities on an IP network. These protocols provide the technology for fabric implementation using TCP and IP routing and switching elements in place of fibre channel components.

The goals of the framework are to:

a) Support the subset of fabric services required by fibre channel storage devices,
b) Produce implementations that run at the speed and latency of gigabit IP transports.
c) Define the new interfaces to be standardized.
d) Identify the interfaces to existing storage standards.

The framework permits the transparent attachment of Fibre Channel storage devices to an IP-based fabric by means of lightweight gateways or edge switches.

This transparency is achieved through:

a) A process for efficiently re-mapping N_PORT addresses embedded in FC frames between the fibre channel and IP network address spaces.
b) Provisions for intercepting and emulating the fabric services required by an FCP device.

iSNS, the companion name service protocol, has been specially tailored to support both the fibre channel and iSCSI naming models.

The following section contains a brief summary of the architecture. The description assumes that the reader is familiar with basic Fibre Channel concepts. In that regard, the material in section 8 may be helpful.

6.5.1 Overview of the IP Storage Fabric Architecture

In an IP Storage fabric, a fibre channel device is attached to the network through an F_PORT interface that is part of an edge switch or gateway. To the attached device, the network appears as a fibre channel fabric.

N_PORT to N_PORT communications that traverse a TCP/IP or UDP network require the intervention of the mFCP or iFCP protocol layer in the gateway. This is done through the following operations on Fibre Channel frames:

a) Map addresses embedded in the fibre channel frame header between the fibre channel and IP address spaces.
b) Service requests for fabric-supplied link services addressed to one of the well-known fibre channel N_PORT addresses. These are handled entirely within the IP Storage Fabric.
c) Generate special control frames in response to certain link

service requests normally processed by a peer N_PORT. These require intervention by the sending and receiving iFCP layers in order to modify and process the frame payloads.

d) Encapsulate frames for injection into the IP network and de-encapsulate frames received from the IP network.
e) Direct de-encapsulated frames to the appropriate N_PORT.

6.5.2 iSNS _ the Storage Naming Service

The iSNS protocol is designed to provide the name services required by fibre channel devices. In addition, since many storage objects are independent of the transport protocol, iSNS has been extended to support iSCSI as well.

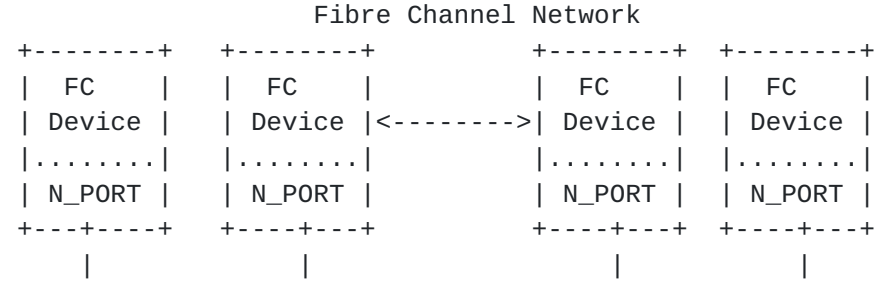
7. Fibre Channel Network Overview

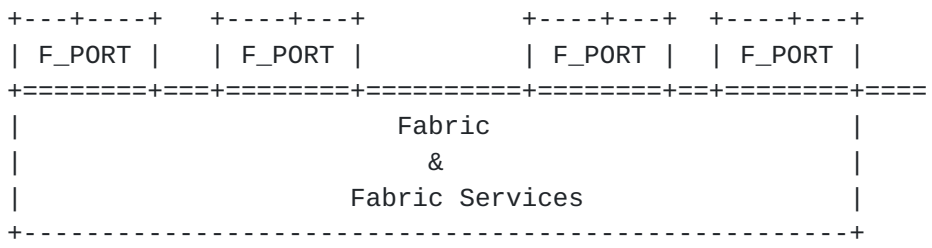
This section contains a brief discussion of the fibre channel concepts needed to understand the architectures described in this document. The reader is advised to consult the documents in section 10 for a thorough treatment of the technology.

7.1 The Fibre Channel Network

The fundamental entity in fibre channel is the fibre channel network. As shown in the diagram below, a fibre channel network is comprised of the following elements:

a) N_PORTs -- The end points for fibre channel traffic,
b) FC Devices _ The fibre channel devices to which the N_PORTs provide access.
c) F_PORTs -_ The ports within a fabric that provide fibre channel attachment for an N_PORT,
d) The fabric infrastructure for carrying frame traffic between N_PORTs,
e) Within the fabric, a set of auxiliary services and a name service for device discovery and network address resolution.

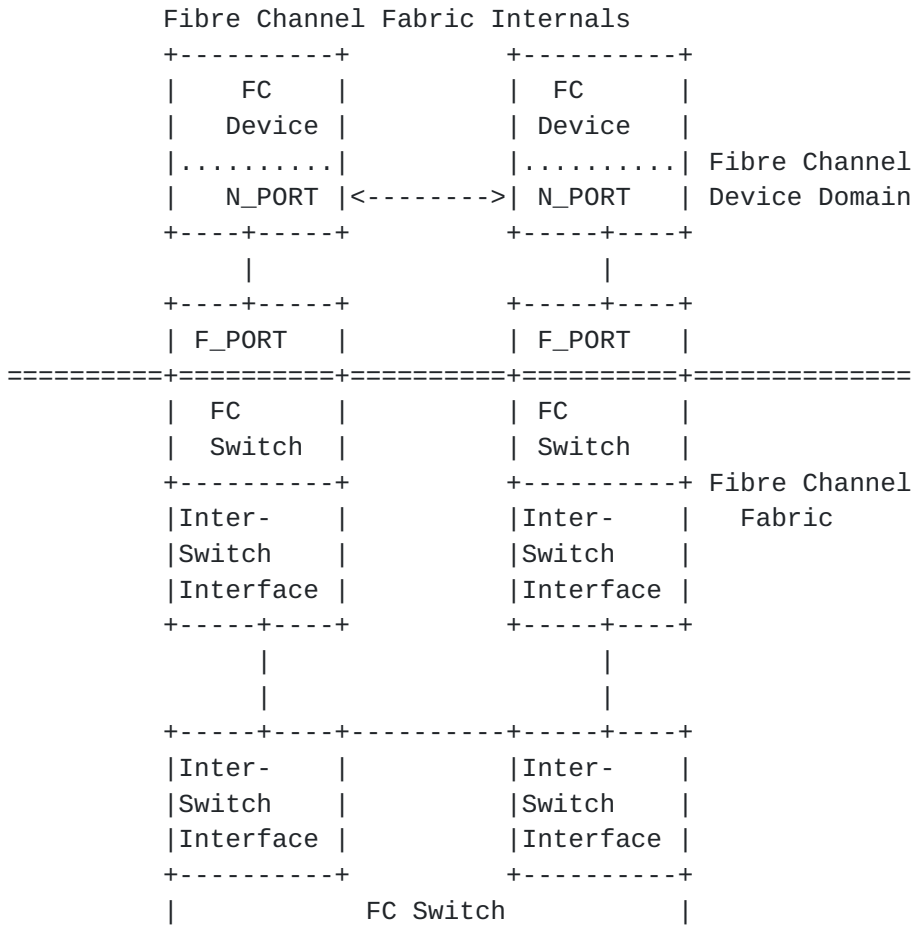




The following sections describe the internals of a fibre channel fabric and give an overview of the fibre channel communications model.

7.1.1 Multi-Switch Fibre Channel Fabric

The internals of a multi-switch fibre channel fabric are shown below.



The interface between switch elements is either proprietary or a standards-compliant E_PORT interface described by the FC-SW2 specification.

7.2 Fibre Channel Layers and Link Services

Fibre channel consists of the following layers:

FC0 -- The interface to the physical media,
FC1 _- The encoding and decoding of data and out-of-band physical link control information for transmission over the physical media,
FC2 _- The transfer of frames, sequences and exchanges comprising protocol information units.
FC3 _- Common Services,
FC4 _- Application protocols, such as FCP, the fibre channel SCSI protocol.

In addition to the layers defined above, fibre channel defines a set of auxiliary operations, some of which are implemented within the transport layer fabric, called link services. These are required to manage the fibre channel environment, establish communications with other devices, retrieve error information, perform error recovery and other similar services. Some link services are executed by the

N_PORT. Others are implemented internally within the fabric. These internal services are described in the next section.

7.2.1 Fabric-Supplied Link Services

Servers internal to the fabric handle certain classes of Link Service requests. The servers appear as N_PORTS located at well-known N_PORT fabric addresses. Service requests use the standard fibre channel mechanisms for N_PORT-to-N_PORT communications.

All fabrics must provide the following services:

Fabric F_PORT server _ Services an N_PORT request to access the fabric for communications.

Fabric Controller -- Provides state change information to other N_PORTS. Used to inform other FC devices when an N_PORT exits or enters the fabric.

Directory/Name Server _ Allows N_PORTs to register information

in a database or retrieve information about other N_PORTs.

The following optional services are defined:

Broadcast Address/Server _- Transmits single-frame, class 3 sequences to all N_PORTs.

Time Server _- Intended for the management of fabric-wide expiration timers or elapsed time values and is not intended for precise time synchronization.

Management Server _ Collects and reports management information, such as link usage, error statistics, link quality and similar items.

Quality of Service Facilitator _ For fabric-wide bandwidth and latency management.

7.3 Fibre Channel Devices

A fibre channel device has one or more fabric-attached N_PORTs. The device and its N_PORTs have the following associated identifiers:

a) A world-wide unique identifier for the device,
b) A world-wide unique identifier for each N_PORT attached to the device,
c) For each N_PORT, a fabric-assigned N_PORT fabric address provided when the device is granted fabric access. This address is unique within the scope of the fabric.

Information about a fibre channel device, such as the fibre channel addresses and world wide names of its N_PORTs, can be discovered through the appropriate name service queries.

7.4 Fibre Channel Information Elements

The fundamental element of information in fibre channel is the frame. A frame consists of a fixed header and up to 2112 bytes of payload having the structure described in section 74.4.1. The maximum frame size that may be transmitted between a pair of fibre channel devices is negotiable up to the payload limit, based on the size of the frame buffers in each fibre channel device and the MTU supported by the fabric.

Operations involving the transfer of information between N_PORT pairs are performed through exchanges. In an exchange, information is transferred in one or more ordered series of frames referred to

as sequences.

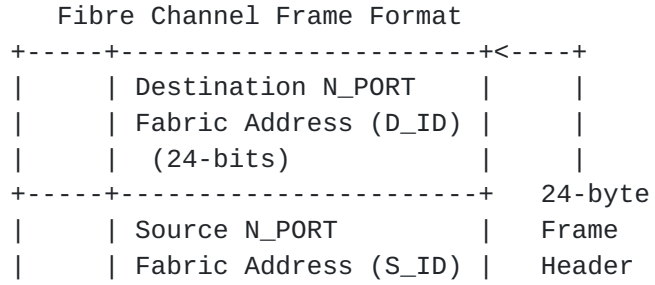
Within a sequence, frames flow from the sequence originator to the sequence recipient. Control information within the frame:

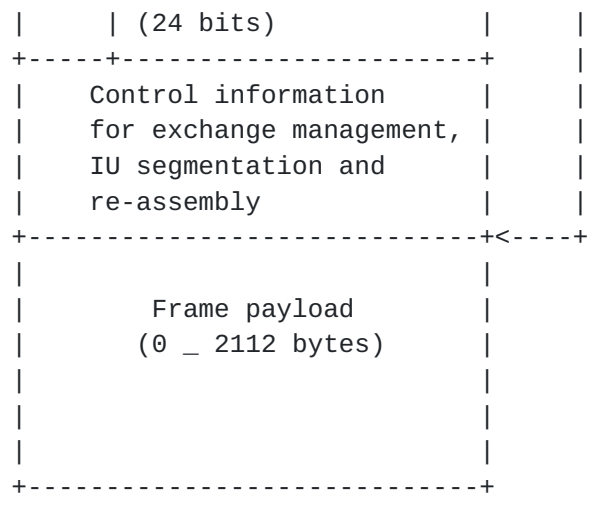
a) Delimits sequence boundaries,
b) Identifies the position of a frame within a sequence,
c) Provides for the initiation of a new sequence reversing the direction of data flow when required by the upper layer protocol.

Within this framework, an upper layer protocol is defined in terms of transactions carried by exchanges. Each transaction, in turn, consists of protocol information units, each of which is carried by an individual sequence within an exchange.

7.4.1 Fibre Channel Frame Format

A fibre channel frame consists of the payload and a header containing the control information necessary to route frames between N_PORTs and manage exchanges and sequences. The following diagram gives a highly simplified view of the frame.





The source and destination N_PORT fabric addresses are embedded in the S_ID and D_ID fields respectively.

7.5 Fibre Channel Transport Services

The fibre channel standard defines the following classes of service provided by a fabric implementation:

Class 1 _ A dedicated physical circuit connecting two N_PORTs.

Class 2 _ A frame-multiplexed connection with end-to-end flow control and delivery confirmation.

Class 3 _ A frame-multiplexed connection with no provisions for end-to-end flow control or delivery confirmation.

For class 2 or class 3 service, the fabric is not required to preserve frame ordering.

Class 3 service is equivalent to UDP or IP datagram service. Fibre channel storage devices using this class of service rely on the ULP implementation to detect and recover from transient device and transport errors.

In addition to the above services, fabrics may implement additional quality of service policies within the framework of class 2 or class 3 delivery mechanisms.

7.6 N_PORT to N_PORT Communication

An N_PORT joins the fabric and establishes a session with another

N_PORT by invoking the following series of services:

a) F_PORT login _- The device invokes the fabric login service to register its presence on the fabric and obtain an N_PORT fabric address.
b) Name Service Lookup - The device obtains the N_PORT fabric address of another device through a name service query.
c) N_PORT login - The device issues a port login request to establish an N_PORT-to-N_PORT session.
d) Process Login _ The device performs a process login to establish a ULP session with a peer process on the remote N_PORT.

An N_PORT issues a corresponding set of logout requests to gracefully terminate the ULP and N_PORT sessions and fabric login.

8. Definitions

Fabric _ A network interconnecting devices that implement the Fibre Channel communications model defined in the FC-FS standard. A fabric may be implemented in the IP framework by means of the protocols discussed in this document.

FC-2 _ The Fibre Channel transport services layer described in the FC-FS specification.

FCP Portal - An IP-addressable entity representing the point at which an iFCP or mFCP node is attached to the IP network.

F_PORT - The interface through which an N_PORT is attached to a fibre Channel fabric.

N_PORT _ An iFCP or Fibre Channel entity representing the interface to Fibre Channel device functionality. This interface implements the Fibre Channel N_PORT semantics specified in the FC-FS standard [FC-FS].

9. Security Considerations

Security considerations for IP Storage protocols are driven not only by the same general concerns expressed regarding other Internet application protocols, but also by the historical expectations of storage users. IP Storage emulates SCSI, and there is a risk that users will treat it as such, ignoring the potential security issues introduced by this new technology.

Parallel SCSI interconnections between computer systems and storage devices inherently rely on the physical security of the computer equipment room. It is easy to perform a security audit for such a network; you determine connectivity by following where the wire goes, verify addressing by noting the value dialed into each device's selector switch, and ascertain privacy by confirming there are no unauthorized connections.

The introduction of SCSI over Fibre Channel removed the distance limitations that kept parallel SCSI within the equipment room, and active SAN devices such as repeater hubs and switches removed SCSI's restriction to a linear bus topology. Even so, security considerations for a Fibre Channel SAN often still relies on physical isolation of the network, supplemented by configuration features such as zoning, the Fibre Channel equivalent of Virtual LANs. It has been said that the relatively limited deployment of Fibre Channel SAN connections (relative to the ubiquitous presence of the LAN,) gives the SAN a false patina of _security through obscurity._

Such assumptions are incompatible with an underlying network architecture such as that of the Internet, which inherently provides ubiquitous connectivity across both private and public network segments. Perhaps the _worst case_ scenario would be perfect emulation of SCSI-attached storage, where some or all of the underlying connectivity traversed a public IP network. In that situation, the end-station behavior would be that of a parallel SCSI system which assumes underlying physical security, while the network behavior would be that of an open transport environment that defers any layered security needs to the end-stations.

A partial solution may be obtained by providing enhanced security services at the interfaces from the existing storage network to the IP storage network. There, services such as encryption and authentication may be applied to the non-physically secure portion of the path, without requiring end-station changes.

Overall security for such a solution may be no better than that of a SAN presumed to be physically secure, but at least it will not be perceived as worse.

Link-level security need not be the only solution. In many applications, privacy of data transmission across the network may be less significant an issue than controlling who accesses the storage resources. In such situations, access-control solutions such as the resource login provided by iSCSI can add value.

Finally, it should be noted that new security concerns may be raised

merely by allowing SANs to be scaled to larger size and complexity. Introducing many milliseconds of WAN delay into storage protocols accustomed to microseconds of latency may expose previously undetected behavioral problems. Discovery algorithms which work

perfectly for fifteen target devices, may fail spectacularly with fifteen hundred. Practices that are acceptable in small systems, such as static assignment of passwords to resources, may become unacceptable in large ones, where devices are frequently added and removed. And malicious behavior, such as _spoofing_ of Fibre Channel source addresses, take on new significance when the malicious device is located in some remote SAN, who's operational capabilities and management policies may be different than your own.

10. References

RFC2026 Bradner, S., "The Internet Standards Process -- Revision 3", BCP 9, RFC 2026, October 1996.

RFC2119 Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

SCSI-3 SCSI-3 Standards Architecture (Diagram) http://www.t11.org/t10/scsi-3.htm

FCP American National Standard of Accredited Standards Committee NCITS, "Fibre Channel Protocol for SCSI, Second Version (FCP-2) " ftp://ftp.t11.org/t10/drafts/fcp2/fcp2r04b.pdf

CAM American National Standard of Accredited Standards Committee X3, "SCSI-2 Common access method transport and SCSI interface module". ftp://ftp.t11.org/t10/drafts/cam/cam-r12b.pdf

T11 NCITS 321-200x (ANSI) T11/Project 1305-D/Rev 4.8 "Fibre Channel Switch-Fabric-2", (FC-SW-2) October 29, 2000 (www.t11.org)

FCBB NCITS T11/Project 1238-D/Rev4.7 "Fibre Channel Backbone", (FC-BB) June 8, 2000 (www.t11.org)

SPC-2 NCITS T10/Project 1236-D/Rev 18, "SCSI Primary Commands - 2 (SPC-2)", 21 May 2000.

FCoverIP Rajagopal, M., Bhagwat, R., Rodriguez, E., Chau, V., Berman, S., Wilson, S., O'Donnell, M., Carlson, C., "Fibre Channel Over TCP/IP (FCIP)", draft-ietf-ips-fcovertcpip-00.txt, October 2000.

iFCP Mullendore, R., Monia, C., Tseng, J., "iFCP - A Protocol for Internet Fibre Channel Storage Networking", draft-monia-ips-iFCP-00.txt, November 2000.

mFCP Mullendore, R., Monia, C., Tseng, J., "mFCP - Metro FCP protocol for IP Networking", November 2000.

iSCSI Satran, J., Sapuntzakis, C., Wakeley, M., Von Stamwitz, P., Haagens, R., Zeidner, E., Dalle Ore, L., Klein, Y., "iSCSI", draft-ietf-ips-iscsi-00.txt, November, 2000.

iSCSI-REQ Haagens, R., "iSCSI (Internet SCSI) Requirements", draft-haagens-ips-iscsireqs-00.txt, July 2000.

RFC1718 IETF Secretariat, Malkin, G., "The Tao of IETF", RFC 1718, CNRI, Xylogics, Inc., November 1994.

RFC2418 Bradner, S., "IETF Working Group Guidelines and Procedures", RFC 2418, Harvard University, September 1998.

RFC1157 Case, J., M. Fedor, M. Schoffstall and J. Davin, "The Simple Network Management Protocol", STD 15, RFC 1157, May 1990.

RFC2578 McCloghrie, K., Perkins, D. and J. Schoenwaelder, "Structure of Management Information Version 2 (SMIv2)", STD 58, RFC 2578, April 1999.

RFC2837 Teow, K. S., "Definitions of Managed Objects for the Fabric Element in Fibre Channel Standard", RFC 2837, May 2000.

RFC2625 Rajagopal, M., Bhagwat, R., and Rickard, W., "IP and ARP over Fibre Channel", RFC 2625, June 1999.

FCMIB Carlson, M., Bowlby, G., and Hu, L., "A Framework for Fibre Channel MIBs", draft-ietf-ipfc-mib-framework-03.txt, July 2000.

FIMIB Blumenau, S., "Fibre Channel Management Framework Integration MIB", draft-ietf-ipfc-fcmgmt-int-mib-04.txt, May 2000.

RFC2143 Elliston, B., "Encapsulating IP with the Small Computer

System Interface", RFC 2143, May 1997.

RFC760 Information Sciences Institute, "Internet Protocol", RFC 760, January 1980.

RFC761 Information Sciences Institute, "Transmission Control Protocol", RFC 761, January 1980.

RFC2960 Stewart, R., Xie, Q., Morneault, K., Sharp, C., Schwarzbauer, H., Taylor, T., Rytina, I., Kalla. M., Zhang, L., Paxson, V., "Stream Control Transmission Protocol", RFC 2960, October 2000.

RFC1180 Socolofsky, T., and Kale, C., "A TCP/IP Tutorial", RFC 1180, January 1991.

RFC813 Clark, D., "Window and Acknowledgement Strategy in TCP", RFC 813, July 1982.

RFC879 Postel, J., "The TCP Maximum Segment Size and Related Topics", RFC 879, November 1983.

RFC955 Braden, R., "Towards a Transport Service for Transaction Processing Applications", RFC 955, September 1985.

RFC962 Padlipsky, M., "TCP-4 Prime", RFC 962, November 1985.

RFC2001 Stevens, W., "TCP Slow Start, Congestion Avoidance, Fast Retransmit, and Fast Recovery Algorithms", RFC 2001, January 1997.

RFC2101 Carpenter, B., Crowcroft, J., and Rekhter, Y., "IPv4 Address Behaviour Today", RFC 2101, February 1997.

RFC2330 Paxson, V., Almes, G., Mahdavi, J., and Mathis, M., "Framework for IP Performance Metrics", RFC 2330, May 1998.

RFC2415 Poduri, K., and Nichols, K., "Simulation Studies of Increased Initial TCP Window Size", RFC 2415, September 1998.

RFC2414 Allman, M., Floyd, S., and Partridge, C., "Increasing TCP's Initial Window", RFC 2414, September 1998.

RFC2581 Allman, M., Paxson, V., and Stevens, W., "TCP Congestion Control", RFC 2581, April 1999.

RFC2151 Kessler, G. and Shepard, S., "A Primer On Internet and TCP/IP Tools and Utilities", RFC 2151, June 1997.

RFC2398 Parker, S. and Schmechel, C., "Some Testing Tools for TCP Implementors", RFC 2398, August 1998.

RFC2140 Touch, J., "TCP Control Block Interdependence", RFC 2140, April 1997.

X500 CCITT Recommendation X.500, "The Directory: Overview of Concepts, Models and Service", 1988.

RFC2251 Wahl, M., Howes, T. and Kille, S., "Lightweight Directory Access Protocol (v3)", RFC 2251, December 1997.

RFC2830 Hodges, J., Morgan, R. and Wahl, M., "Lightweight Directory Access Protocol (v3): Extension for Transport Layer Security", RFC 2830, May 2000.

RFC2256 Wahl, M., "A Summary of the X.500(96) User Schema for use with LDAPv3", RFC 2256, December 1997.

RFC2255 Howes, T. and Smith, M., "The LDAP URL Format", RFC 2255, December 1997.

RFC2254 Howes, T., "The String Representation of LDAP Search Filters", RFC 2254, December 1997.

RFC2252 Wahl, M., Coulbeck, A., Howes, T. and Kille, S., "Lightweight Directory Access Protocol (v3): Attribute Syntax Definitions", RFC 2252, December 1997.

RFC2247 Kille, S., Wahl, M., Grimstad, A., Huber, R. and Sataluri, S., "Using Domains in LDAP/X.500 Distinguished Names", RFC 2247, January 1998.

RFC1591 Postel, J., "Domain Name System Structure and Delegation", RFC 1591, March 1994.

SNS Gibbons, K., Tseng, J. and Monia, C., "iSNS Internet Storage Name Service", draft-tseng-ips-isns-00.txt, October 2000.

RFC2914 Floyd, S. "Congestion Control Principles", RFC 2914, BCP 41, September 2000.

RFC2979 Freed, N., "Behavior of and Requirements for Internet Firewalls", RFC 2979, October 2000.

RFC1631 Egevang, K. and Francis, P., "The IP Network Address Translator (NAT)", RFC 1631, May 1994.

RFC2663 Srisuresh, P. and Holdrege, M., "IP Network Address Translator (NAT) Terminology and Considerations", RFC 2663, August 1999.

RFC1918 Rekhter, Y., Moskowitz, B., Karrenberg, D., de Groot, G. J. and Lear, E., "Address Allocation for Private Internets", RFC 1918, BCP 5, February 1996.

RFC2050 Hubbard, K., Kosters, M., Conrad, D., Karrenberg, D. and Postel, J., "INTERNET REGISTRY IP ALLOCATION GUIDELINES", RFC 2050, BCP 12, November 1996.

RFC2766 Tsirtsis, G. and Srisuresh, P., "Network Address Translation - Protocol Translation (NAT-PT)", RFC 2766, February 2000.

RFC2504 Guttman, E., Leong, L. and Malkin, G., "Users' Security Handbook", RFC 2504, February 1999.

RFC2411 Thayer, R., Doraswamy, N. and Glenn, R., "IP Security Document Roadmap", RFC 2411, November 1998.

RFC2828 Shirey, R., "Internet Security Glossary", RFC 2828, May 2000.

RFC2709 Srisuresh, P., "Security Model with Tunnel-mode IPsec for NAT Domains", RFC 2709, October 1999.

RFC2401 Kent, S. and Atkinson, R., "Security Architecture for the Internet Protocol", RFC 2401, November 1998.

RFC1511 Linn, J., "Common Authentication Technology Overview", RFC 1511, September 1993.

RFC2402 Kent, S. and Atkinson, R., "IP Authentication Header", RFC 2402, November 1998.

RFC2406 Kent, S. and Atkinson, R., "IP Encapsulating Security Payload (ESP)", RFC 2406, November 1998.

RFC1630 Berners-Lee, T., "Universal Resource Identifiers in WWW", RFC 1630, June 1994.

RFC1738 Berners-Lee, T., Masinter, L. and McCahill, M., "Uniform

Resource Locators (URL)", RFC 1738, December 1994.

RFC2624 Shepler, S., "NFS Version 4 Design Considerations", RFC 2624, June 1999.

RFC2224 Callaghan, B., "NFS URL Scheme", RFC 2224, October 1997.

NFS4 Shepler, S., Beame, C., Callaghan, B., Eisler, M., Noveck, D., Robinson, D. and Thurlow, R., "NFS version 4 Protocol", draft-ietf-nfsv4-07.txt, June 2000.

11. Acknowledgments

Thanks to Randy Haagens for text taken from the iSCSI requirements document.

12. Author's Addresses

Mark A. Carlson
Sun Microsystems, Inc.
Email: Mark.Carlson@Sun.COM

Satish Mali
StoneFly Networks
Email: satish@stoneflynetworks.com

Milan Merhar
Pirus Networks
Email: milan@pirus.com

Charles Monia
Nishan Systems
Email: cmonia@nishansystems.com

Murali Rajagopal
LightSand Communications
Email: muralir@lightsand.com

Appendix A: Existing Internet Standards and Procedures

A.1 IETF and RFC overview

Various working groups formed out of interested parties define Internet Protocols. The Internet Engineering Task Force (IETF) forms these working groups. The IETF is a large open international community of network designers, operators, vendors, and researchers concerned with the evolution of the Internet architecture and the smooth operation of the Internet. The spirit of IETF is explained in _The Tao Of IETF_.(To borrow from _The Tao of IETF_, the mission of

IETF is

* Identifying, and proposing solutions to, pressing operational and technical problems in the Internet,
* Specifying the development or usage of protocols and the near-term architecture to solve such technical problems for the Internet
* Making recommendations to the Internet Engineering Steering Group (IESG) regarding the standardization of protocols and protocol usage in the Internet
* Facilitating technology transfer from the Internet Research Task Force (IRTF) to the wider Internet community; and
* Providing a forum for the exchange of information within the Internet community between vendors, users, researchers, agency contractors and network managers.

The IETF working group members interact with each other and produce documents called _Request For Comments_. These RFCs are now subdivided into FYIs (For Your Information also know as _Informational_) and STDs (Standard). The _Informational_ RFC sub-series provides overviews and topics that are introductory. This document, for example is an Informational RFC. STD RFCs identify those RFCs that specify Internet standards.

Every RFC, including FYIs and STDs, have an RFC number by which they are indexed and by which they can be retrieved. FYIs and STDs have FYI numbers and STD numbers, respectively, in addition to RFC numbers

A.2 RFC summary:

The Internet Engineering Task Force (IETF) is responsible for developing and reviewing specifications intended as Internet

Standards. IETF activities are organized into working groups (WGs). RFC 2418 describes the guidelines and procedures for formation and operation of IETF working groups [RFC2418].

Because there are so many RFCs, a summary is provided for RFCs by an RFC itself. You will find that generally every n99 RFC provides a summary of RFCs from n00 to n99. Here n99 stands for 1 to 27. Therefore, RFC 2799 will provide summary for all RFCs from 2700 to 2798. Also most of the RFCs that end in 00 provide a one-line

summary up to that RFC. Thus, 2700 will be one line summary for all RFCs from 2600 to 2699.

A.3 Management of TCP/IP based devices:

Management of TCP/IP based networked element in enterprises is routinely done using Simple Network Management Protocol (SNMP) [RFC1157]. This SNMP protocol allows any networked entity that supports SNMP, to be managed using a standardized method. The parameters that can be managed and monitored for this network entity is defined by another standard called the Structure of Management Information Version 2 (SMIv2)[RFC2578] that defines a collection of managed objects, residing in a virtual information store, termed the Managed Information Base (MIB). Every vendor normally supports a subset of the standard MIB and also provides a vendor dependant private MIB. This private MIB being written in standard format can be used by SNMP management software. MIBs are proposed from time to time, and added to the standard MIBs to have a generic way of managing generic parameters and interfaces.

A.4 Fibre Channel related standards:

[RFC2837] defines the objects for managing the operations of the Fabric Element portion of the Fibre Channel Standards. While [RFC2625] specifies a way of encapsulating IP and Address Resolution Protocol (ARP) over Fibre Channel and to describes a mechanism(s)

for IP address resolution. A Framework for Fibre Channel MIBs [FCMIB] discusses technical issues and requirements for the management information base (MIB) for Fibre Channel and storage network applications. The Fibre Channel Management Framework Integration MIB [FIMIB] provides an integrated management environment to enable interoperability among the various vendors involved in the Fibre Channel marketplace. [RFC2143] outlines a protocol for connecting hosts running the TCP/IP protocol suite over

a Small Computer System Interface (SCSI) bus. This RFC defines an experimental protocol and is not a proposed standard.

For a general discussion of Fibre Channel standards, see section 7 of this document.

A.5 Standards related to TCP and IP:

There are many TCP/IP related RFCs dating back to 1980. The original RFC that formulated Internet Protocol (IP) is [RFC760]. [RFC761] is for Transmission Control Protocol (TCP). To get a good TCP/IP bare-bones overview you might want to start with [RFC1180]. [RFC813] describes implementation strategies using sliding window protocol. It presents a field-tested flow control algorithm. [RFC879]

discusses the TCP Maximum Segment Size Option and related topics. The purpose here is to clarify some aspects of TCP and its interaction with IP. [RFC869] provides ways of congestion control in IP/TCP networks. It is suggestive in nature and not a standard. Two RFCs ([RFC955] and [RFC962]) provide early thoughts on transaction processing on TCP/IP based networks.

Information about TCP/IP workings are provided in [RFC2001] (TCP Slow Start, Congestion Avoidance, Fast Retransmit, and Fast Recovery Algorithms), [RFC2101] (IPv4 Address Behaviour Today), [RFC2330] (Framework for IP Performance Metrics), [RFC2415] (Simulation Studies of Increased Initial TCP Window Size), [RFC2414] (Increasing TCP's Initial Window), [RFC2581] (TCP Congestion Control). A primer on Internet and TCP/IP Tools and Utilities is provided by [RFC2151] along with [RFC2398](Some Testing Tools for TCP Implementors). [RFC2140] provides information about TCP Control Block Interdependence.

A.6 Standards related to Naming and Discovery topics:

LDAP is used as a directory service for querying and accessing resources in a networked domain. LDAP can be used as one of the means to identify iSCSI resources. Another possible method would be to use a DNS server. A third alternative is to use a Name server specifically designed for iSCSI.

A.6.1 LDAP:

The Open System Interconnect (OSI) defined a _Directory_ protocol that provides a powerful infrastructure for the retrieval of information objects. This infrastructure can be used to do lookups in white pages, or applications or other informational objects. It was standardized by CCITT as X.500 [X500].[RFC2251] describes access to the X.500 Directory while not incurring the resource requirements of the Directory Access Protocol (DAP). This protocol is specifically targeted at simple management applications and browser applications that provide simple read/write interactive access to the X.500 Directory, and is intended to be a complement to the DAP itself.

One approach to discovery of iSCSI devices would be to query LDAP services.

[RFC2830] provides extension for transport layer Security to LDAP version 3. RFCs that cover other areas of LDAP are [RFC2256] (User Schema for use with LDAPv3), [RFC2255] (The LDAP URL Format), [RFC2254] (The String Representation of LDAP Search Filters), [RFC2252] (Attribute Syntax Definitions), [RFC2247] (Using Domains in LDAP/X.500 Distinguished Names).

A.6.2 DNS:

DNS (Domain Name Service) is used to find networked nodes using a name. The DNS server maps this name to an IP address.

The Internet Assigned Numbers Authority (IANA) is the overall authority for the IP Addresses, the Domain Names, and many other parameters, used in the Internet. [RFC1591] provides information on the structure of the names in the Domain Name System (DNS), specifically the top-level domain names and on the administration of domains.

It is possible to use DNS as one way to identify iSCSI devices. The DNS named devices then can be addressed in URLs or other addressing schemes.

A.6.3 iSCSI Name Server:

There are currently no IETF standards. There is a proposed IETF standard draft [iSNS] for an iSCSI Name Server that describes one proposal.

A.7 Flow Control related Standards:

Any IP based storage protocol needs to account for the existing congestion control mechanisms of the Internet. Internet Protocol based networks provide congestion control using TCP/IP and are documented in [RFC2581] and [RFC2001]. The Congestion Control Principles are explained in [RFC2914].

A.8 Standards related to Firewall, NAT:

A firewall is a device that is inserting on the path between the Internet and the internal network and it screens network traffic in some way, blocking traffic it believes to be inappropriate, dangerous, or both.

[RFC2979] defines behavioral characteristics of and interoperability requirements for Internet firewalls.

A NAT (Network Address Translation) is a special purpose router that generally has a private IP addresses on one interface and a public

IP address on the second interface. The private IP address is mapped to a public IP address by the NAT, and is managed by using port numbers. This allows a company to use private IP addresses within the company and only use a few public IP addresses. An overview about NAT is documented in [RFC1631]. The terms used in NAT are provided in [RFC2663]. The private side network address allocation is provided by [RFC1918]. The public side address allocation guide

is provided by [RFC2050]. [RFC2766] provides information about translation involved from Ipv4 to Ipv6.

Please note that the firewall functions are disjoint from NAT functions. Neither implies the other, although, many times, both are provided by the same device.

iSCSI devices will be accessed through either firewalls or NAT devices or both. The addressing, naming and discovery schemes should be designed to work with existing firewalls and NAT devices.

A.9 Security and Authentication related Standards:

IP based storage devices may be available over Internet which may mean that they are publicly accessible. It is also envisioned that the IP based connectivity is provided between isolated private SANs. Security is then a major concern and needs to be addressed at every level of access.

There are many RFCs related to security. iSCSI will mainly involve use of IPsec based security. The RFCs provided here may be of interest to many users as security in an open environment like IP is very important.

If you want to understand security, you may want to start with [RFC2504] (Users' Security Handbook). IPsec protocol suite is used to provide privacy and authentication services at the IP layer. Several documents are used to describe this protocol suite. The interrelationship and organization of the various documents covering the IPsec protocol are discussed in [RFC2411]. The security glossary is provided by [RFC2828].

A secure path is established between two Internetworked ends by encrypting packets between the two ends. This allows complete privacy of any data that is sent from one end to another. The establishment of this private channel is called tunneling and is defined by IPsec standard. [RFC2709] provides the security model for Tunnel-mode IPsec for NAT domains.

The _Security Architecture for the Internet Protocol_ is provided in [RFC2401] and specifies the base architecture for IPsec compliant systems.
Authentication Mechanisms are covered by [RFC1511] (Common Authentication Technology Overview).

Sending and receiving special secret codes called keys provide authentication. [RFC2402] provides details of _IP Authentication Header_.

Encrypting data provides privacy. It is detailed in [RFC2406] (IP Encapsulating Security Payload (ESP)).

A popular method of providing security is by use of a proxy server. Devices inside a company, wanting to communicate to outside world will communicate with the proxy server and the proxy server translates their requests to outside world.

A.10 Addressing and other Miscellaneous Standards:

One proposed to access ISCSI devices is by using URLs.

Uniform Resource locator (URL) is commonly used as a compact string representation for a resource available on Internet. The concept of URL was introduced by the World Wide Web global information initiative in 1990 and is described in "Universal Resource Identifiers in WWW", [RFC1630]. The Hypertext Transfer Protocol (HTTP) is an application-level protocol that is used for communications between these URL specified resources on Internet. HTTP is also used as a generic protocol for communication between user agents and proxies/gateways to other Internet protocols, such as SMTP, NNTP, FTP, Gopher, and WAIS, allowing basic hypermedia access to resources available from diverse applications and simplifying the implementation of user agents.

[RFC1738] specifies the Uniform Resource Locator (URL), the syntax and semantics of formalized information for location and access of resources via the Internet.

A.11 Network File related protocols:

Though these protocols are not related to iSCSI or other block data protocols, they are enumerated here for interested parties.

Network File System (NFS) protocol provides access to shared file-systems across networks. It is designed to be machine, operating system, network architecture, and transport protocol independent.

The latest NFS version will be version 4 and the issues in the design of NFS 4 are provided in [RFC2624]. The NFS Version 4 protocol is described in [NFS4]. [RFC2224] defines 'nfs' as a new URL scheme. The `nfs' URL will refer to files and directories on NFS servers using the general URL syntax as per [RFC1738].

Full Copyright Statement

HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE."