IP STORAGE: A TECHNOLOGY OVERVIEW

Dan McConnell, Storage Architecture and Technology Evaluation
Enterprise Systems Group

The growth of storage area networks (SANs) and the pervasiveness of the Internet Protocol (IP) are driving interest in using IP-based networks to transport block storage traffic. Referred to as IP storage, early implementations are beginning to appear in the market, and the Internet Engineering Task Force (IETF) standards body is developing standards in this area. This white paper reviews IP storage and compares file- and block-level access. The paper also presents some of the factors driving development of IP storage, surveys the technologies involved, identifies factors that will enable or obstruct its widespread adoption, and takes a look at possible adoption strategies.

What is IP Storage?

IP storage refers to a group of technologies that allows block-level storage data to be transmitted over an IP-based network. There are two key concepts in this definition: "the use of IP" and "block-level storage."

Transferring block-level storage data over a networked topology is not a new concept. Today's SANs use the Fibre Channel (FC) technology to do just that. The promise of the new IP storage protocols is the interconnection, as well as the complete construction, of these SANs with prevalent IP-enabled technologies such as Ethernet. The use of IP to transfer data is also not a new concept. Familiar protocols such as Common Internet File System (CIFS) and Network File System (NFS) have been used to access file-level storage data over IP networks for years. The difference between these protocols and the IP storage protocols lies in how the data is accessed—at the "file level" or at the "block level."

CIFS and NFS issue file-level requests to a server that "owns" the file system. Thus, these requests rely on the existence of a network entity (either a typical file server or a network attached storage [NAS] device) that owns the file system and serves its files to other network hosts. As shown in Figure 1, when a file-level request (such as a request to open **myfile.txt**) is received, the

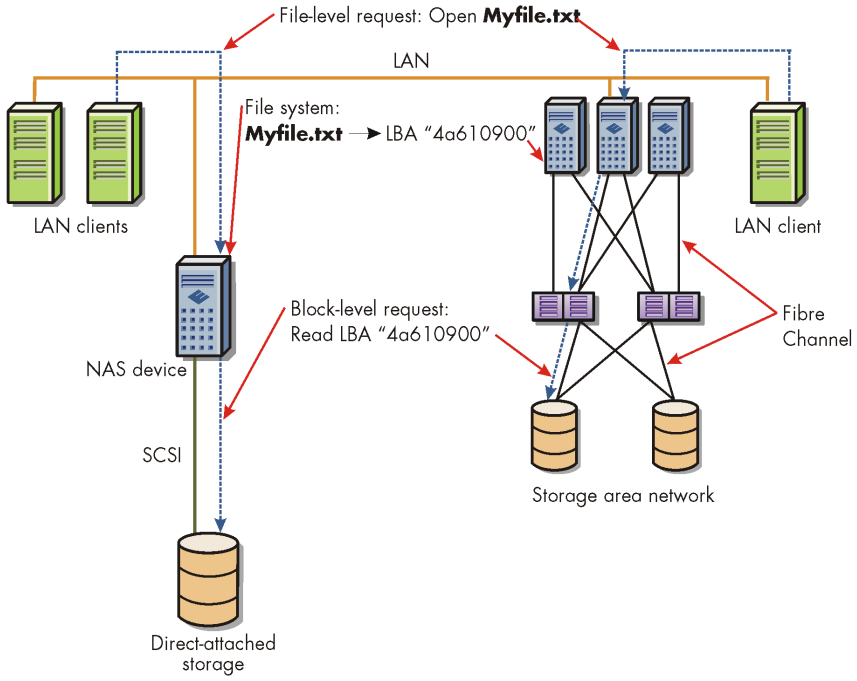


Figure 1. File- and Block-Level Data Access

file server or NAS device refers to its file tables and translates the logical file name to a list of the physical block addresses corresponding to the location of the data on the physical medium, and performs the physical block access. This process requires CPU cycles on the host system, and can add considerable latency to the file operation because of the overhead involved. Applications such as transactional databases that rely on low-latency data access cannot tolerate this additional overhead. Instead, they require direct block-level access to the physical hardware. SANs are used to issue these direct block-level requests to the storage device.

The IP storage protocols provide the means to encapsulate these block-level requests for transmission over the IP network using the standard Transmission Control Protocol (TCP). This allows the direct block-level requests used by SANs to take place over an IP-based network.

Why IP Storage?

The concept of IP storage has emerged as networked storage requirements have grown and as IP has become firmly established as the predominant general-purpose networking protocol.

Growth of Networked Storage

Storage requirements are growing at explosive rates. International Data Corporation (IDC) estimates that storage capacity will increase at nearly 75 percent a year in the 2001-2003 time frame (IDC Worldwide Disk Storage Systems Forecast and Analysis, 1999-2004). This growth highlights two very important issues: the increasing importance of data and the difficulties of managing burgeoning storage resources. To address these issues, networked storage in the form of SANs is increasingly being deployed to store, access, protect, and manage mission-critical data. IDC forecasts that by 2004, 67 percent of all storage will be networked. (IDC Worldwide Disk Storage Systems Forecast and Analysis, 1999-2004).

The typical SAN protects data by allowing redundant paths between host and storage devices, enabling remote mirroring solutions for disaster recovery, and allowing backups to be performed over the SAN with minimal impact on application servers or the host network. These three advantages allow networked storage to provide protection across device or site failures, and to ease backup difficulties.

Networked storage also allows storage to be consolidated, which reduces management complexity. Centralized management of a consolidated storage pool can be more efficient than managing separate direct-attached storage subsystems. The ability to easily allocate storage from the consolidated pool where and when it is needed simplifies administration and helps to eliminate underutilization of storage.

IP: Established Networking Protocol

IP is the prevailing general-purpose networking protocol. Because of its worldwide acceptance and ability to run on virtually any subnetworking technology, IP has gained a critical mass that gives it many advantages over other networking protocols.

IP has essentially become a requirement for corporate networking. There are IP-enabled backbones that span the globe and a large pool of technical workers with IP experience. With this omnipresence comes the large and ever-growing development base behind IP. The existing quality of service, link prioritization, and security protocols that are available for IP networks prove that this large development base continues to drive the technology forward at a rapid pace. Finally, IP is relatively inexpensive, because it runs over commodity subnetworking technologies such as Ethernet.

IP Storage

In the early days of IP development, there was a vision of "IP over everything"—Ethernet, Token Ring, Asynchronous Transfer Mode (ATM), and so forth. With video, voice, and now even block-level storage being transported over IP, it seems the vision is now "everything over IP." The existing LAN/WAN infrastructure, support for, and knowledge base surrounding IP make it a very attractive storage networking protocol. The vision of a single networking technology for the LAN and SAN is compelling. No longer would IT departments have to maintain equipment, technical staff, and expertise in both the IP and FC technologies. In addition, many smaller companies that want to take advantage of the advanced features of networked storage may be in-

clined to implement IP-based, rather than FC-based, SANs because IP is a familiar technology. Enabling block storage over prevalent IP-based networks would also allow easy access to storage over long distances.

There are current implementations that encapsulate storage traffic over IP for applications such as remote mirroring. Many are proprietary, but implementations are beginning to appear that are based on draft IETF standards (such as Internet SCSI [iSCSI]) designed to provide a standardized way to transport block-level storage over existing IP networks.

IP Storage Standards

The IETF is currently working on three IP storage encapsulation protocols:

- iSCSI
- FC Over TCP/IP (FCIP)
- Internet FC Protocol (iFCP)

iSCSI

iSCSI will provide the necessary mapping to make IP a transport for SCSI commands, just as FC today is a transport for SCSI commands. iSCSI is designed to be a host-to-storage end-to-end solution. Similar to the FC SAN architecture today, iSCSI technologies will include iSCSI-enabled hosts that will communicate through IP switches to iSCSI-enabled storage arrays. (The drives will probably still be native SCSI drives, because iSCSI is not currently a disk-attach technology.) Figure 2 is a simplified view of the protocol layers involved.

The server on the left would contain an iSCSI-enabled device. This could be a special-purpose iSCSI Host Bus Adapter (HBA) or a software layer running on the host, which is equipped with a standard Ethernet NIC. (See "IP Storage Issues" later in this paper for a discussion of issues with standard NICs.) The SCSI command is encapsulated into an iSCSI Protocol Data Unit (PDU). As defined by the IETF, the iSCSI protocol will use TCP as its underlying transport layer to provide a reliable transport with guaranteed in-order delivery. Once the TCP/IP headers are added, the encapsulated SCSI command is treated the same as any other IP packet. It can be routed to its end destination (based on its IP address) over standard IP infrastructure. Once the destination device receives the packet, it strips off each layer until it eventually returns the SCSI command to the SCSI layer just as if the source and destination were attached locally.

iSCSI takes advantage of the global addressing scheme that IP enables. iSCSI devices will have two types of identifiers: an iSCSI name and an iSCSI address. Similar to the FC worldwide name (WWN), all iSCSI initiators and targets will be given a permanent iSCSI name by an existing naming authority. This name will identify the device, regardless of its location or IP address. The iSCSI address specifies the location of an iSCSI initiator

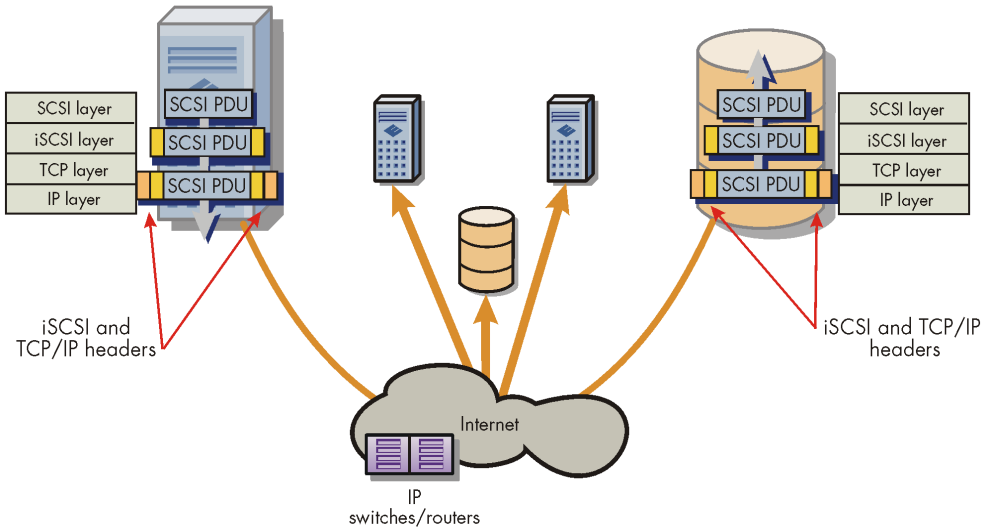


Figure 2. iSCSI Protocol Layers

or target and is composed of an IP address, port number, and the iSCSI name of the device. For example:

iSCSI address format:
iSCSI://<insert domain name>:<insert port>/<insert iSCSI name>

iSCSI Name: fqn.com.disk-vendor.diskarray.45678

iSCSI Address:
iSCSI://diskfarm1.acme.com:80/fqn.com.disk-vendor.diskarray.45678

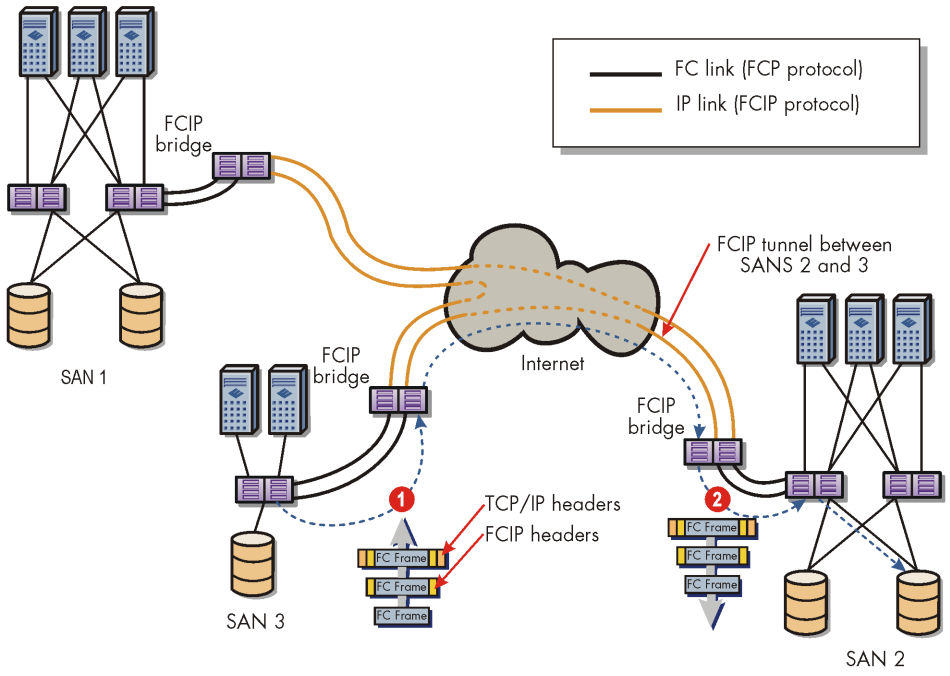
The iSCSI specification, currently at version 0.6, has been through multiple iterations and is currently scheduled to be submitted for approval in March 2002. Though the specification is not final, the underlying basic protocol is stable, and reference products have already been announced and released. There are two other specifications related to the iSCSI protocol that are scheduled to be submitted for approval in April 2002:

- iSCSI Management Information Base (MIB) for Simple Network Management Protocol (SNMP)-based management of iSCSI devices
- Internet Storage Name Service (iSNS), the naming service for IP storage environment. (iSNS will also be the naming service for iFCP.)

FCIP

As its name implies, the aim of the FCIP protocol is to transport FC frames over an IP infrastructure. FCIP provides the mechanisms to allow islands of FC SANs to be interconnected over IP-based networks to form a single, unified FC SAN fabric. The extended FC SAN fabric continues to use standard FC addressing. Essentially, IP tunnels are set up between FCIP end points. Once these tunnels are in place, FC devices view these extended links as standard FC links and use FC addressing. Typical implementations will use the FCIP end points to connect two (or more) FC switches in an interswitch link (ISL) fashion over a standard IP infrastructure. The result will be to combine two separate SAN fabrics into one.

Figure 3 shows three SAN islands connected with FCIP end points (or gateways) to form a single FC SAN fabric. The FCIP gateways encapsulate the FC frames, then use TCP as the underlying transport. (Although there are User Datagram Protocol [UDP] implementations, the IETF specification calls for the use of TCP.) Once the FC frame has been mapped onto IP, it can be routed through the IP infrastructure, just as any IP packet, to the destination device. After the FCIP tunnels have been



1. FC frame is encapsulated into FCIP packet, then into TCP/IP packet, and sent over FCIP tunnel to destination device in SAN 2.
2. FC frame is unencapsulated and forwarded to correct local device in SAN 2.

Figure 3. FCIP Scenario

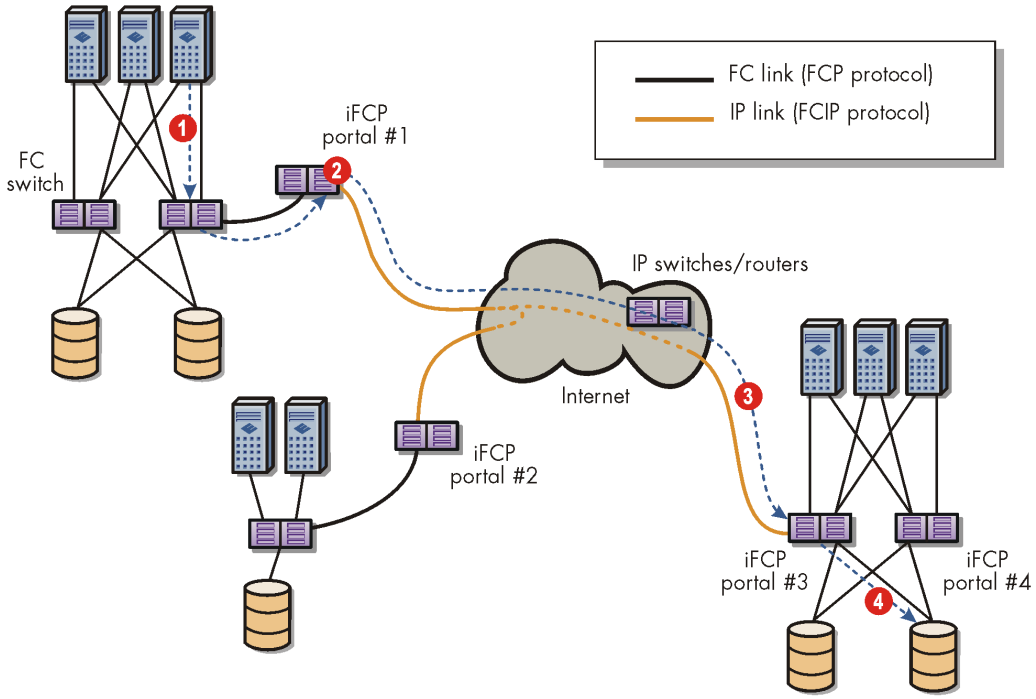
established, the links are transparent to the FC devices. FC switches see the links as standard ISLs and, therefore, communicate their name server information and establish a single FC fabric namespace.

The FCIP draft is currently scheduled to be submitted for approval in March 2002. Because FCIP uses many preexisting features of FC, it is likely that products using this protocol will precede products based on the iSCSI protocol.

iFCP

The last IP storage protocol, iFCP, lies somewhere in between the two previously discussed protocols. Like FCIP, iFCP encapsulates FC frames to be sent over the IP infrastructure. Because of this, the IETF chose to specify a common FC encapsulation format. The main difference between the two protocols lies in their addressing schemes. The FCIP protocol establishes point-to-point tunnels that can be used to connect two FC SANs together, with Ethernet, to create a single, larger SAN. In contrast, iFCP is a gateway-to-gateway protocol that combines FC and IP addressing to allow the FC frames to be routed to the appropriate destination address. Unlike the addressing scheme of the FCIP protocol, the current iFCP addressing scheme allows each interconnected SAN to retain its own independent namespace.

Figure 4 shows three remote SANs connected across an IP infrastructure via iFCP end points (or portals). Each iFCP portal presents all the devices on its local FC fabric to other iFCP portals attached to the IP network. Each portal maintains a table of the remote devices and presents them as local devices on its local FC fabric. This process is transparent to actual devices on the local SAN; the remote devices appear to be local. When a local device needs to access a remote device, it issues an FC frame to the remote device's local FC address. This frame goes to the local iFCP portal, which encapsulates it into an IP packet and routes the packet to the appropriate remote iFCP portal. The remote portal unpackages the FC frame and delivers it to the designated device. Because these portals must understand both FC and IP addressing, most implementations will



1. Request addressed to "presented" device on the local FC fabric.
2. iFCP portal encapsulates FC frame into IP packet addressed to the appropriate remote iFCP portal IP address. The packet also contains the remote FC address of the requested device.
3. Packet is routed to remote iFCP portal.
4. iFCP portal unpackages the FC frame and forwards to correct local device. Request addressed to actual FC address of remote FC device.

Figure 4. iFCP Scenario

also function as standard FC or IP switches, as shown in the SAN on the right of Figure 4.

The iFCP draft is currently scheduled to be submitted for approval in March 2002. Implementations using a variant of this protocol called the Metropolitan FCP (mFCP), which uses UDP instead of TCP for the underlying transport, are currently available.

IP Storage Issues

IP storage is still a very young technology. Although the foundations of the standards are almost complete and early implementations are emerging, there are issues that must be resolved before widespread adoption is feasible.

TCP Offload

Because IP does not guarantee delivery, all three IP storage protocols rely on TCP as the underlying transport to guarantee reliable, in-order delivery in the potentially congested long-haul IP space. This means that even though IP packets may arrive out of order, the TCP layer must deliver the data to the upper-layer protocol (in this case SCSI) in the correct order. To do this, the TCP layer typically uses a reorder buffer. This buffer stores out-of-order sequences until the full in-order sequence is obtained. Once the sequence is in the correct order, the TCP layer sends the data to the next layer.

This can be a complicated process that consumes host CPU cycles and adds latency to the transaction. The result is much more I/O overhead than a typical FC or SCSI block transfer. A mechanism is required to offload this work from the host processor. This mechanism has been termed a TCP Offload Engine (TOE). TOEs are also young in their implementation, but are becoming available and will help to solve this issue.

Price/Performance

Even though these protocols will run on IP, it will not be feasible to use standard off-the-shelf Ethernet NICs. While technically possible, their use is not plausible when performance is an issue. As mentioned earlier, TOEs will be required to offload the IP storage I/O workload from servers. These TOE devices, early in their implementation, will add hardware cost and complexity to today's standard NICs. Widespread adoption of IP storage technologies will rely on the price-to-performance ratio of these enhanced "iHBAs" being comparable to established technologies such as FC.

Security

In a world where storage devices can be physically located anywhere and connected through standard IP infrastructure, security becomes a bigger issue than when SANs resided in data centers. The industry is struggling with the age-old question: What is an acceptable level of security and how much overhead is acceptable to ensure it? This question, currently being studied by the IETF, must be addressed before these technologies are readily accepted. (Among the proposals under consideration by the IETF is the use of a subset of the IPSec protocol for iSCSI.)

Interoperability

Just because these technologies are based on IP does not mean that they will live up to the Internet's promise of interoperability. Nor does the fact that standards for the protocols are published by the IETF ensure that products from vendor X will interoperate with those from vendor Y. In order for IP storage to succeed, the various vendors that implement these protocols must work together to ensure that their implementations are interoperable.

Possible Adoption Strategies

It is uncertain when IP storage solutions will gain widespread adoption, but they are likely to appear in three phases:

- Phase 1: SAN extender
- Phase 2: Limited-scope IP storage
- Phase 3: IP SAN

Phase 1: SAN Extender

Now that SANs are deployed worldwide, there is a need to connect these geographically separated SANs over long distances. Whether it is to share data across sites of a large corporation, to remotely mirror vital data to another facility, or any number of other reasons, this is an immediate requirement that these IP storage technologies are positioned to address. The concept behind these solutions is to add FC-to-IP bridges/routers to ex-

tend the SAN links across an IP infrastructure, thus joining two remote SANs. FCIP and iFCP are best suited to this application because of their ties to FCP, but iSCSI devices may be used as well.

Figure 5 demonstrates what a typical FC extender configuration may resemble. In the figure, two geographically disparate FC SANs are connected through the use of FC-to-IP gateways. To ensure no single point of failure, two gateways are shown on each site.

Phase 2: Limited-Scope IP Storage

Phase 2 will bring IP storage to small, cost-sensitive environments. In this phase, the vision of a true global SAN will not be realized, but the emerging technologies will enable limited-scope, IP-based SAN connectivity. Solutions may appear in which iSCSI cards are integrated into NAS devices, because the technologies and required TOE devices complement NAS solutions. This would provide a single multifunction device that offers either block-level or file-level data access. This combination block- and file-level NAS device would be an easy way for previous direct-attached environments to begin to make the switch to networked storage.

Phase 2 will also introduce workgroup-oriented or small business IP-based SANs in environments where networked storage is required, but the tasks involved with introducing a new networking technology may be too difficult. iSCSI is best suited to this type of environment. These iSCSI-based SANs will not displace FC SANs, but instead may emerge as a low-cost way to achieve the benefits of networked storage.

Figure 6 shows what these limited-scope IP storage products might resemble. The figure on the left shows a workgroup-oriented, IP-based SAN. The figure on the right shows how the combination file- and block-level NAS device might be implemented. Notice that the Exchange server on the left is able to access its storage on the block level while the application server on the right has file-level access to its storage.

Phase 3: IP SAN

Over time, as many of the obstacles associated with IP storage are overcome, we may begin to see complete end-to-end, IP-based global SANs. Of the three IP storage technologies, the iSCSI protocol is best suited to this type of implementation. These iSCSI-based IP SANs will consist of iSCSI HBAs that offload much of

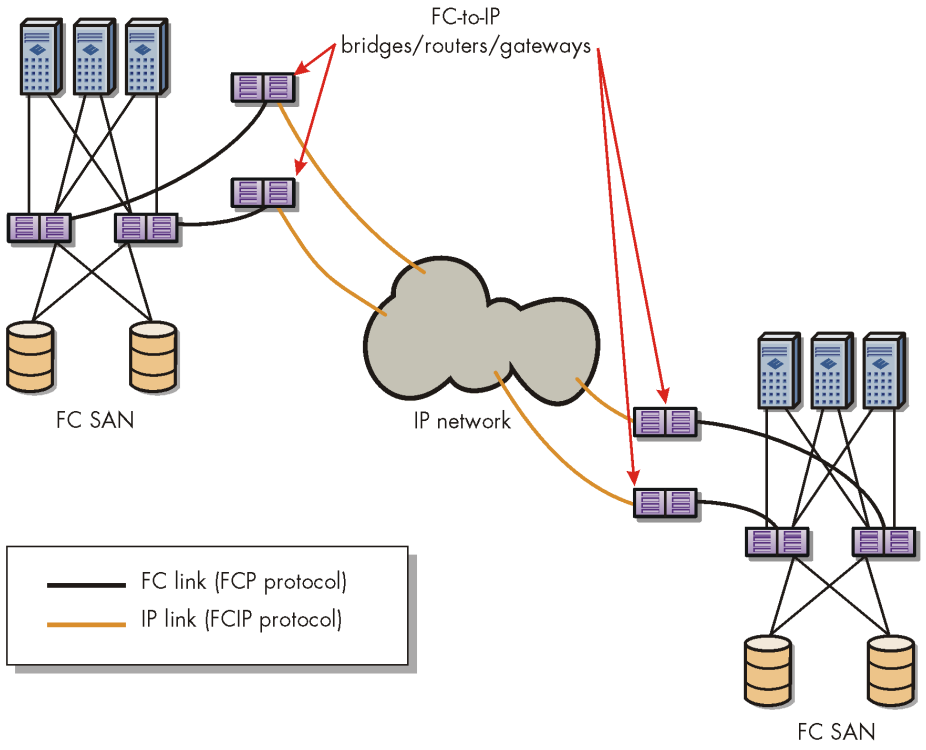


Figure 5. SAN Extender in Redundant Configuration

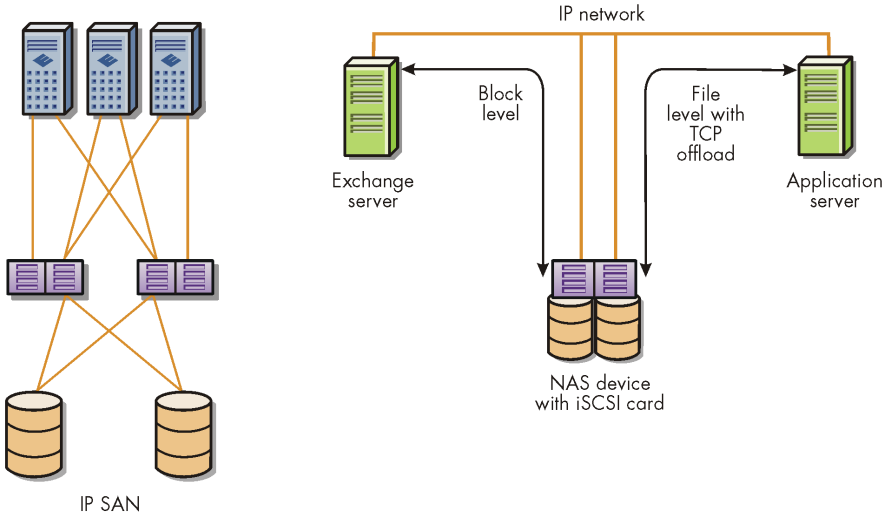


Figure 6. Limited-Scope IP Storage

the TCP overhead, and native iSCSI storage devices all communicating through a standard IP infrastructure. Once this is achieved, many of the advanced IP functions previously confined to the LAN space, such as bandwidth aggregation, quality of service guarantees, and so forth will begin to enter the SAN space.

Figure 7 shows what a true global IP SAN may resemble. With IP being the underlying SAN transport, many distributed configurations will be possible. For example, SANs may be easily interconnected for disaster recovery and resource sharing, or remote servers will be able to access consolidated data pools in remote SAN installations.

Proof of Concept

On September 24, 2001, eight of the leading storage and networking vendors successfully demonstrated the viability of these IP storage protocols. Dell, Adaptec, Hitachi Data Systems, IBM, Intel, Nishan Systems, Qlogic, and Quest Communications took part in the Promontory Project, which connected servers and storage on the east and west coasts of the United States. The project was named after Promontory Summit, the location where America's first transcontinental railroad was joined.

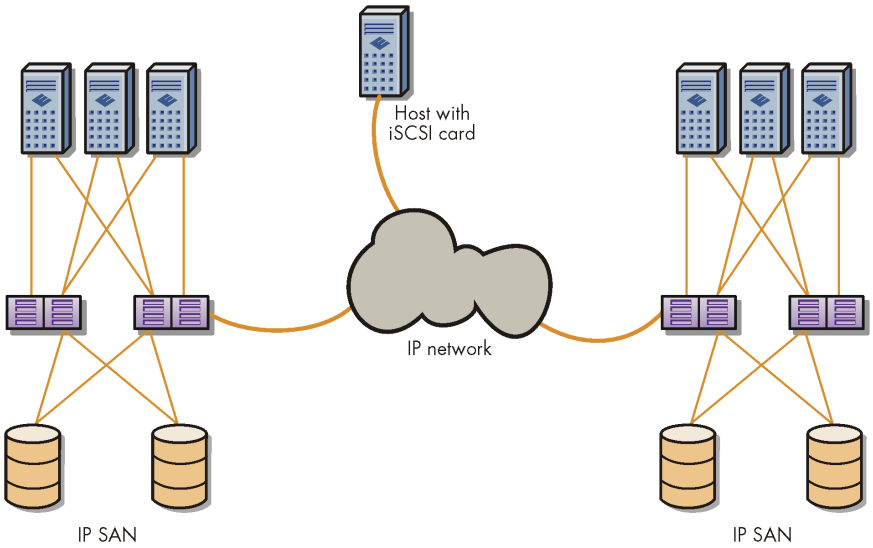


Figure 7. IP SAN

The project consisted of connecting remote SANs from sites in Sunnyvale, California and Newark, New Jersey over an IP network at gigabit speeds. Each local site consisted of both Fibre Channel and IP-based SANs. These sites were connected across the nation over a pair of OC-48 (2.488 Gbps) links provided by Qwest Communications. The IP-based SANs used the iSCSI protocol, while the Fibre Channel SANs were connected into the IP network via FC-to-IP storage gateway devices. Both the iSCSI and iFCP protocols were used for the long-haul transmission of the storage data over the IP network. With the connection in place, servers in Sunnyvale could access storage in Newark (and vice versa) as if those devices were attached to the local SAN.

The success of the Promontory Project proves the viability of these IP storage protocols, shows that interoperability issues can be overcome, and paves the way for the entrance of IP into the SAN market.

Conclusion

The incorporation of IP networking into storage networks, in whatever form it takes, will extend SANs from their FC technology base today to a broader range of globally deployed technologies that have been enabled with IP (for example, Ethernet, ATM). Whether used for interconnecting remote FC SANs, creating complete end-to-end IP SANs, or somewhere in between, iSCSI-, IFCP-, and FCIP-based IP storage are expected to play a key role in the future of the storage industry.