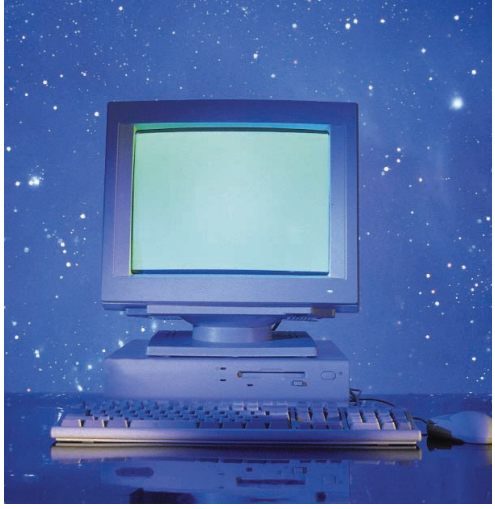


Technology allows any network server to transparently consolidate and access data stored in multiple physical locations.

Thomas C. Jepsen

The Basics of Reliable Distributed Storage Networks

Storage networks increase storage efficiency and data availability by providing shared storage access to computers and servers in multiple locations. Companies can use storage networks to logically pool different storage devices—which might use different access protocols. With low system overhead, storage networks permit users to quickly and efficiently perform information management functions such as backup and recovery, data mirroring, disaster recovery, and data migration.

Efficient management of stored data has become imperative as total disk storage exploded. IDC estimated disk storage to be 500,000 Tbytes worldwide in 2002 and expects this figure to climb to 1.4 million Tbytes by 2005 (R.C. Gray, J. McArthur, and V. Turner, *Storage Consolidation: A Business Value Analysis*, report no. 02-072STORAG3437, IDC, Aug. 2002). The information management divisions at many companies now manage hundreds of terabytes of data. However, the traditional "islands of storage" management approaches are vastly inefficient, wasting or underusing as much as 50 percent of storage capacity.

Besides efficiency, enterprises need the increased reliability of distributed storage systems to curtail expensive downtime. Thus, using storage networks to manage access to data increases performance and survivability, and also controls costs. IDC estimates the worldwide networked storage market to grow from $2 billion in 1999 to more than $25 billion this year. Maturing business-to-business and business-to-consumer e-commerce will create even greater demand for stored data management.

WHY USE A DISTRIBUTED STORAGE NETWORK?

Companies are increasingly distributing storage networks over wide geographical areas to ensure data survivability and to provide data synchronization over large distances. This distribution also helps businesses comply with recently introduced legislation mandating reliable backup and recovery of critical data. In the US, the Sarbanes-Oxley Act of 2002 requires businesses to maintain secure backups of financial data over extended periods; the Health Insurance Portability and Accountability Act (HIPAA) similarly requires the backup of healthcare-related data. Also, the US Securities and Exchange Commission requires financial institutions to remotely mirror transaction data. To automate these functions in conformance to these new regulatory requirements, IT managers are increasingly turning to distributed storage networks.

Serial optical-fiber-based storage protocols such as Escon (Enterprise Systems Connection) and Fibre Channel greatly increase the distances among processors and storage devices across which systems can transfer data. The introduction of these protocols allowed the development of dis-

tributed storage applications. Although bus-based protocols such as the mainframe bus/tag interface and parallel Small Computer System Interface (SCSI) limited this distance to a few meters, native-mode Escon and Fibre Channel enable data transmission over distances of 10 kilometers or more. Repeaters or link extenders allow transport as far as 100 kilometers. Storage data can also travel over metropolitan area networks (MANs) or wide area networks (WANs) for a virtually unlimited distance using suitable MAN or WAN transport protocols, such as asynchronous transfer mode (ATM), Synchronous Optical Network (Sonet), wavelength-division multiplexing (WDM), or Internet Protocol (IP). Proper networking techniques allowed companies to develop storage applications that were not limited to a specific geographical area, and could span the distance of a MAN or WAN.

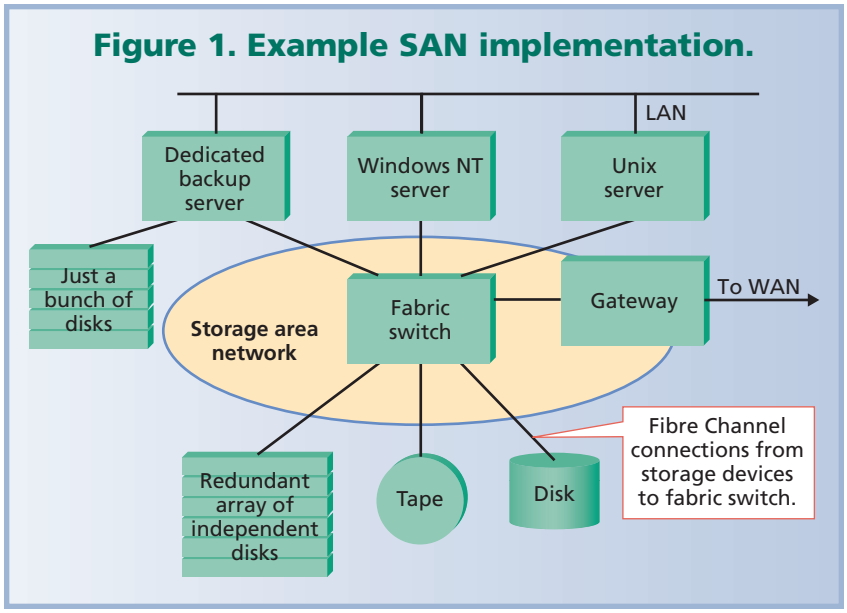


Figure 1. Example SAN implementation.

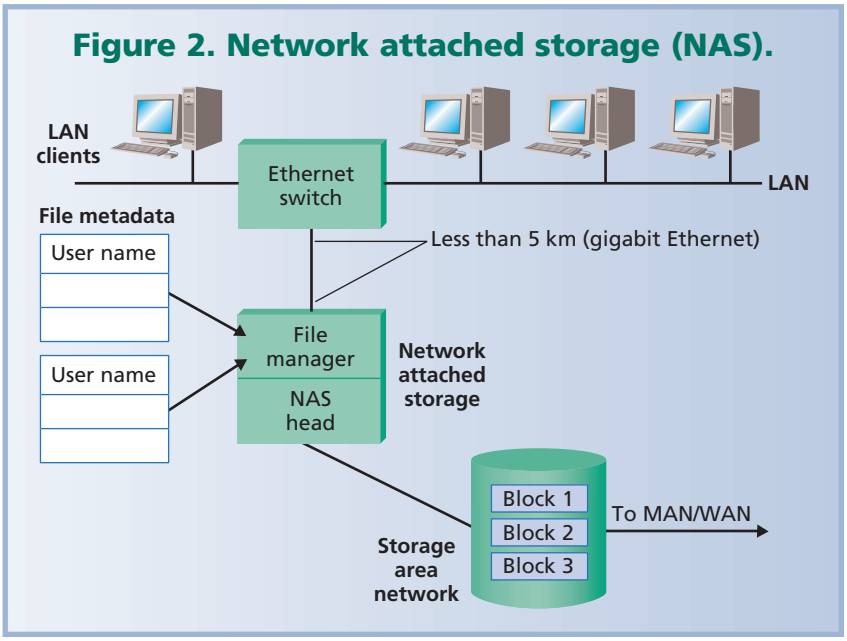


Figure 2. Network attached storage (NAS).

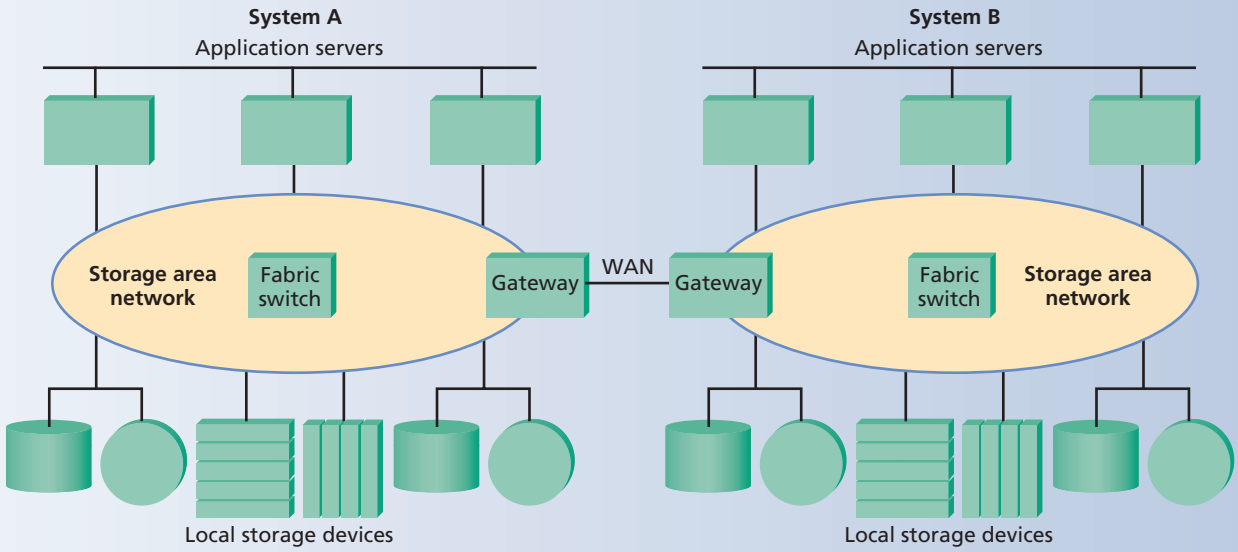
What is a storage area network?

One common storage network architecture is the storage area network. A SAN using a switched-fabric topology consists of computing and storage nodes interconnected via a fabric of network switches. With a *switched fabric*, any storage device can connect to any computing device throughout a data transfer operation. A switched fabric can support multiple simultaneous full-bandwidth connections among storage and computing nodes. A *storage director* is a specialized type of fabric switch with enhanced management and reliability features, such as duplicated **fabric switches** and power supplies. With a switched fabric, a common backup server can connect to any storage device for scheduled or manual backup. The network could employ a *gateway* to provide MAN or WAN interworking and protocol translation in distributed SAN applications. Figure 1 shows an example SAN implementation.

What is network-attached storage?

Network-attached storage (NAS) is another managed-storage technique. NAS consists of a file manager (or filer) attached to a local area network (LAN) that manages and provides access to stored data. The primary difference between NAS and SAN architectures is that NAS serves file-structured data to clients while a SAN serves block-structured data to application servers. This neat distinction blurs somewhat because the NAS file manager must manage data at the block level in its own attached storage. The NAS file manager contains metadata—in the form of directories and data structures—that maps file requests to blocks of data in disk storage. Some NAS implementations use a *NAS head*, a file manager front end that serves files to its clients and attaches to a back-end SAN to manage block-structured data. Figure 2 shows a NAS implementation that uses a NAS head with a SAN to manage block-level storage.

Figure 3. Distributed SAN for storage sharing and integration.



NAS uses one of several standard file-sharing protocols, so multiple applications can share access to files and support functions such as a locking mechanism that activates when more than one application accesses a single file. For file sharing, Unix typically employs the Network File System protocol; Windows NT or 2000 environments can use NFS or Common Internet File Sharing.

This article is not a comparison of SAN and NAS. SAN and NAS address different problems, and are actually complementary technologies. This article is, however, about techniques for distributing storage networks. A storage network can be a storage area network (SAN), network attached storage (NAS), or a mainframe extended channel. This article addresses two issues related to extending the geographic range of NAS:

- the distance limitations inherent in LAN protocols, and
- the use of a distributed SAN to provide backend storage for a NAS.

When implementing a storage network, IT planners must often decide whether to implement a SAN or NAS architecture. However, think of SAN and NAS as complementary technologies that solve different problems for the enterprise planner, rather than as mutually exclusive choices.

DISTRIBUTED SAN APPLICATIONS

Distributed SAN applications provide the following functionality:

- storage integration,
- remote backup and restore,
- disk mirroring,
- data migration,
- business continuity and disaster recovery,
- remote operation of peripheral devices, and
- mainframe and open-systems connectivity.

Storage integration

Storage integration or storage sharing refers to using distributed storage networks to share disks and integrate storage across a wide geographic area. This approach is particularly useful in applications where the amount of stored data is so large that duplicating it would be infeasible or uneconomical. Applications that use distributed storage networks for storage sharing or integration include databases of genetic information for genome research; multimedia or video servers; and e-commerce applications where multiple servers update a common database.

In the shared-storage configuration, as Figure 3 shows, each system can access its own storage and the other system's storage. System A and System B could be SANs, similar to the configuration in Figure 1, connected across the WAN by means of gateway devices. Many switched fabrics have a remote-device mapping capability that makes a remote storage device appear to be part of the local SAN configuration. Thus, system A can access data blocks on system B's storage devices, as well as its own storage. Likewise, system B can access data blocks on System A, as well as its own storage. Users of either system would be unaware of their data's actual physical location.

Shared access is not problematic if access is for read-only data. If a user writes or modifies data, however, the systems must prevent corruption by limiting data access to one user modifying the data at any time. File systems and database

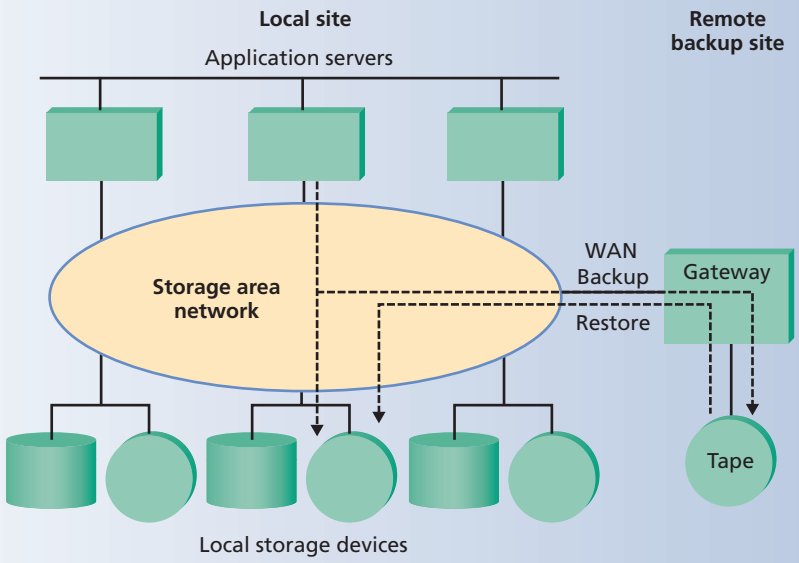
managers typically employ a locking mechanism to ensure that only one user updates data at any time.

Remote backup and restore

Remote backup and restore refers to using distributed storage networks to provide non-real-time backup and restore of user data from a remote location. *Electronic tape vaulting* is another term that people sometimes use for this task. Traditionally, companies have performed remote backup by writing data to tape at the primary site, and then using a truck or other vehicle to transport the tapes to the remote site.

In the event of lost or corrupted data at the primary site, the company retrieves tapes from the remote storage site and brings them to the primary site for recovery. (Data processing people sometimes refer to this method as the Chevy Truck Access Method.) Most ordinary business and financial applications use some backup and restore method to increase the reliability and availability of business data.

Figure 4. Remote backup and restore using distributed storage network.



Manual system data backup is a time-consuming and sometimes disruptive process, because it requires that access to certain data sets must stop during the backup period. It also requires dedicated personnel and equipment. The trend toward 24/7 uptime in datacenters has greatly reduced the size and frequency of backup windows, segments of time set aside for backup operations. Application processing stops during this time. Using distributed storage networks to automatically back up data to a remote site eliminates much of the overhead associated with manual backup. Data simply backs up automatically across the MAN or WAN to tape or disk devices at the remote site, using a backup scheduling algorithm. Lost data at the primary site is restorable from the backup copy across the MAN/WAN network. Figure 4 shows an example architecture for remote backup. The local site could be a SAN similar to Figure 1.

Disk mirroring

Using distributed storage networks, disk mirroring copies user data to multiple disks at near-real-time. Unlike backup, disk mirroring creates a remote copy at the time a transaction commits and writes to the local disk, rather than at a predefined backup time. Thus, mirroring provides greater data availability than backup, but at the cost of duplicating storage capacity. Many finance or banking applications store mission-critical data redundantly in a location that is physically separate from the primary site. Often the backup, or secondary, site obtains power from a different power grid than the primary site, to ensure that power loss at one site does not affect operations at the other. Remote disk mirroring can be processor- or storage-centric. The company must ensure that the remote mirroring technique does not adversely affect system performance so the user experience is one of near-real-time data access.

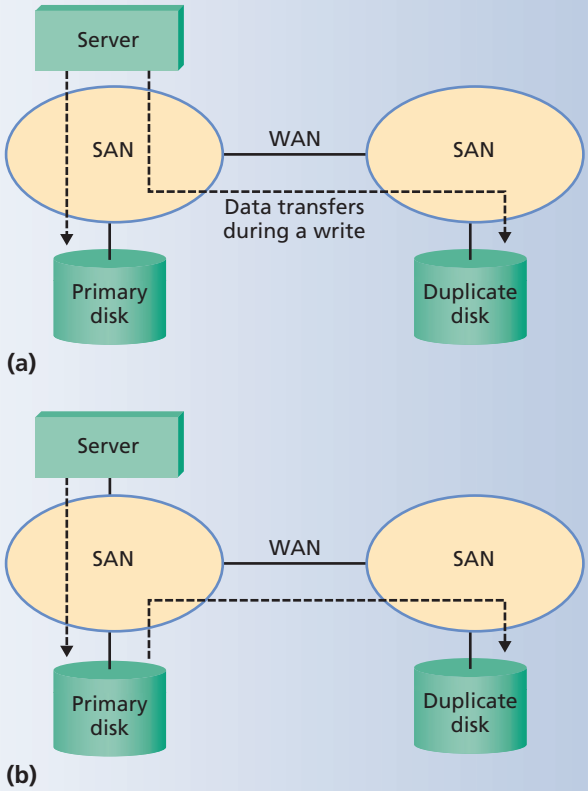
Mirroring can be synchronous or asynchronous. A synchronous system mirrors every transaction as it commits. This requires a data link that can handle the full transfer bandwidth or system performance will degrade noticeably. In contrast, asynchronous mirroring depends less on the data link's speed. The remote system's data update might lag behind that of the local system. The local system logs changes to data as they occur, and then periodically applies the logged changes to the remote system. Applying logs typically takes between 5 minutes and 1 hour.

Various forms of disk mirroring back up data differently. In processor-centric remote disk mirroring—as Figure 5a illustrates—the primary processor or server is aware of both the primary and remote secondary disks' existence. The server sends separate write commands to each disk to write data to each individually, and receives separate responses from each.

In storage-centric remote disk mirroring—as Figure 5b shows—the primary processor or server is only aware of the primary disk. The primary disk's controller or the network appliance copies data to the remote mirror.

Split-mirror copy is a specialized form of disk mirroring that increases reliability and has a minimal impact on

Figure 5. Processor- (a) and storage-centric (b) remote disk mirroring.



In processor-centric disk mirroring, both disks connect to a processor or server that issues a separate write command to each disk. In storage-centric disk mirroring, the data automatically mirrors to the duplicate disk.

application performance. Split-mirror copy uses a third mirror disk, often remotely located, to increase reliability and minimize the possibility of data loss. (Some implementations might use more than three mirrors, but the basic concept remains the same.) During normal operation, all three mirrors contain identical data. When copying data is necessary, one of the mirrors splits, that is, removes itself from the configuration. Using split-mirror copy, data backup is possible without needing a backup window, and a copy of a database can be created for emergency recovery. Split-mirror copy's journaling and mirroring techniques provide optimum performance and minimal disruption to running applications. Figure 6 shows an example split-mirror implementation.

For applications that process a high volume of transactions, it is difficult to find a convenient time to perform a copy operation. Ideally, copying would occur precisely when all current transactions have completed and written to disk, and no new transactions have begun; in other words, when the database contents are consistent. One way to achieve this consistency is to simply stop the application from performing further processing during backup; however, doing so greatly reduces performance. A better solution is to continue processing, suspend writes to the mirrors, and temporarily cache all changes. Read-only operations can continue normally.

When it is necessary to make a backup copy or a duplicate database, the administrator suspends the applications that normally write data to the mirrors, so that copying is possible while the database is in a consistent state. One of the mirrors then splits, and an administrator or automated function can make a backup copy from its content. (Or, if a duplicate database is being created, the administrator can activate the new database, using the copied data.) This copy, often called a frozen image, represents a snapshot of the database contents at a specific point in time. The split mirror then returns to the mirrored configuration. Although suspending writes or application processing guarantees consistency at the time of the split, it is possible that an application has written additional changes to the other two mirrors during the backup of the third mirror. There are two basic ways to resynchronize the third mirror with the other two disks when it merges back into the configuration. One is to simply overwrite all the data on the third disk with data from the other two; however, this is an inefficient use of I/O processing. A more efficient way is to use a transaction log to selectively update changed blocks on the third mirror. This way, each mirror image contains a transaction log as well as stored data. The transaction log shows all transactions that have written to the mirror since the last resynchronization. It is possible to resynchronize the split mirror by simply playing back all the changes recorded in one of the other mirrors' transaction logs.

Data migration

A company can use storage networks to efficiently move large volumes of data from one physical location to another when, for example, the company moves to a larger building. This was previously an awkward process in which the company had to copy volumes of data to tape and physically transport it to the new location, often disrupting day-to-day processing.

Storage networks provide an elegant solution to this problem. Mirroring or backup techniques can create a new copy of the data at the new location while processing continues as usual at the old location. At a specified time, the application archives or deletes the copy at the old location, and the copy at the new location becomes the primary copy. Telecommunications companies, for example, have used this technique to move billing data from one processing center to another while continuing operations.

Business continuity/ disaster recovery

Distributed storage networks are important to business continuity planning and in recovering from disaster-caused data losses. Business continuity typically requires companies to maintain redundant storage and complete processing capabilities at a remote location, which provides a multilevel solution. If a company loses data at the primary site, it can recover the data from the secondary site using standard backup and restore techniques, or by mirroring. If a company loses processing capability at the primary site, processing can continue at the secondary site. If either site becomes completely inoperative, the other site can continue full processing.

Figure 6. Split-mirror copy using three mirrors.

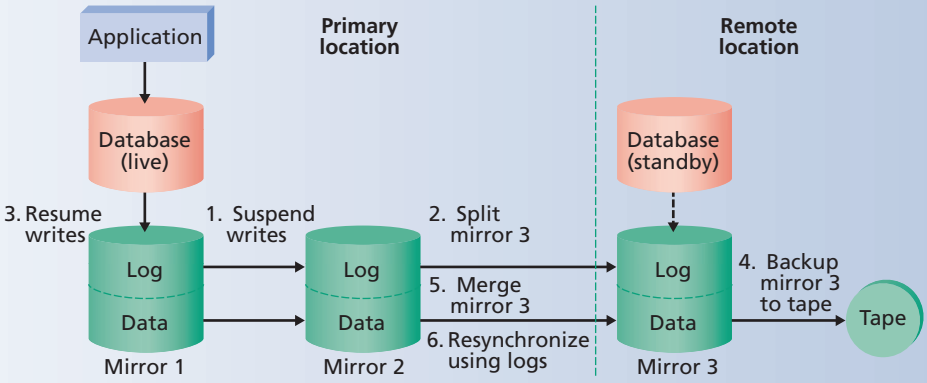
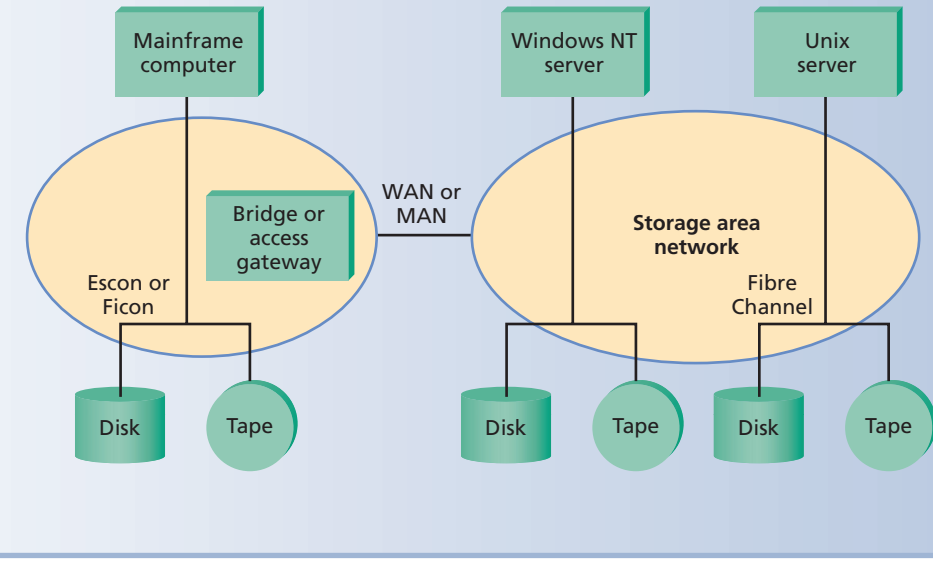


Figure 7. Using a storage network for mainframe and open-systems connectivity.



Remote operation of peripheral devices

Companies can use storage protocols to remotely operate peripheral devices, such as printers or check sorters, which are not at the same location as the server or processor. This can allow the location of peripherals near the department that uses them or can keep paper chaff and ink from entering a filtered air system. Companies can perform remote-peripheral operation using SAN or mainframe storage architectures; in mainframe terminology, this function is a *channel extension*. For example, a financial services company might use this technique to remotely operate check readers and sorters in a remote processing center.

Mainframe/open systems connectivity

Storage networks can provide connectivity among mainframe systems and open systems running Unix or Windows NT or 2000 operating systems. Applications using mainframe and open-systems connectivity include

- the integration of e-commerce and mainframe-based applications,
- data warehousing, and
- backup and recovery.

Moving data among mainframe applications and open systems is possible by using LAN bandwidth or by creating a flat file and moving the data via FTP, but these approaches are slow and complex. It is faster to use the bandwidth available in the mainframe channel and the open-systems storage network to transfer the data directly between storage subsystems. Normally, a gateway or bridge translates between the Escon or Ficon (Fiber Connection) storage protocol used in the mainframe channel subsystem and the SCSI or Fibre Channel used in the open sys-

tems storage network, as Figure 7 illustrates. The storage protocols extend across the WAN or MAN to provide remote connectivity.

DISTRIBUTED NAS APPLICATIONS

Companies can distribute NAS applications over a wide geographic area in two ways. Using switched gigabit Ethernet enables NAS file managers to be 5 kilometers from the production LAN without needing repeaters (see Figure 2). This would enable, for example, the colocation of a NAS file manager or head with remotely located back-end storage.

For the NAS head configuration, an implementation could distribute the back-end SAN across multiple sites, using any of the distributed SAN applications described earlier.

SAN and NAS are beginning to converge.

SAN OR NAS?

NAS is generally appropriate when a company needs a short-term tactical solution to a storage problem and can use an existing Ethernet LAN infrastructure to support the storage application. NAS performance is generally adequate for smaller installations of less than 5 Tbytes. SAN is usually the architecture of choice for large installations (greater than 5 Tbytes of stored data) that will likely accommodate continued data growth and where management and reliability are key factors. The use of SANs for multisite data sharing and replication is also an important factor. Think of SAN as the long-term strategic solution enabling the enterprise to manage continued growth.

However, the two architectures are beginning to converge as NAS file managers become more specialized and use managed SANs for back-end storage. A NAS head with a SAN back end is functionally identical to a SAN using a metadata controller to provide file-based access. In the future, the distinction might disappear completely, and you might think of NAS and SAN as simply two different views of the same stored data.

FUTURE DIRECTIONS: IP STORAGE

Transporting block storage data using the Internet and the TCP/IP (Transmission Control Protocol/Internet Protocol) suite is desirable, because of the ubiquity and availability of these technologies. Storage solutions using TCP/IP would provide the cost and performance advantages of Internet technology, and extend how far the protocol would operate. Using TCP/IP for storage solutions requires mapping storage protocols, such as SCSI or Fibre Channel, to the standard TCP/IP protocol stack.

Applications using storage over IP technology would enable consolidation, pooling, and clustering of local storage, and would provide network client access to remote storage. Applications could support data mirroring as well as local and remote backup and recovery.

The Internet Engineering Task Force (IETF) IP Storage Working Group has produced a request for comment—RFC 3347, "Small Computer Systems Interface Protocol over the Internet (iSCSI) Requirements and Design Considerations"—defining the requirements for SCSI over the Internet. Drafts are also in progress for Fibre Channel Over TCP/IP (FCIP) and Internet Fibre Channel Protocol (iFCP).

Many believe that Fibre Channel's primary competition will come from these IP storage protocols as they mature. However, it is unlikely that iSCSI, FCIP (Fibre Channel Over TCP/IP), or iFCP (Internet Fibre Channel Protocol) will ever completely replace Fibre Channel as the predominant storage protocol, because they cannot offer the same constant delay and data loss guarantees as a Fibre Channel network. Rather, it is likely that each protocol will occupy a different market niche on the cost-versus-performance continuum and will coexist in many enterprise networks. In fact, think of IP storage as a compatible technology for cost-effectively extending Fibre Channel fabrics over the Internet.

Recent changes in regulatory requirements have made creating a strategic data backup and recovery plan a must for most IT organizations. Because of storage protocols that operate over extended distances, various distributed storage applications that improve the efficiency and reliability of data storage are now possible. Distributed storage applications improve efficiency by allowing any network server to transparently consolidate and access data stored in multiple physical locations. Remote backup and mirroring improve the system's reliability by copying critical data. These processes improve efficiency by eliminating backup downtime and manual backup operations. Business continuity and disaster recovery capabilities enable enterprises to recover quickly and transparently from system failure or data loss. Storage protocols and gateway devices enable rapid and transparent data transfer between mainframe applications and open-systems applications. NAS applications provide shared file access for clients using standard LAN-based technology, and can integrate with SAN architectures to provide truly distributed network capabilities. All these distributed storage network applications enable IT managers to improve data availability and reliability while minimizing management overhead and costs. ■

Thomas C. Jepsen *is an IT consultant based in Chapel Hill, N.C., and IT Professional's editor for programming languages. Contact him at tjepsen@mindspring.com.*