

IBM Storage Tank™
A Distributed Storage System

IBM Corporation

January 24, 2002

Overview

IBM Storage Tank™ provides a complete storage management solution in a heterogeneous distributed environment. IBM Storage Tank is designed to provide performance that is comparable to that of file systems built on bus-attached, high-performance storage. In addition to high performance, the goal of IBM Storage Tank is to provide high availability, increased scalability, and centralized, automated storage and data management.

Storage Area Network Technology

Storage Area Network (SAN) technology allows an enterprise to connect large numbers of devices, including clients, servers, and mass storage subsystems, to a high-performance network.

On a SAN, clients can access large volumes of data directly from storage devices, using high-speed, low-latency connections. IBM Storage Tank is designed to be independent of the actual SAN fabric technology. IBM Storage Tank will work with Fibre Channel networks as well as new emerging storage networking technologies such as Gigabit Ethernet (iSCSI) and Infiniband.

By using SAN technology, IBM Storage Tank can meet the needs of general data sharing in a distributed environment, as well as the needs of special, data-intensive applications, such as imaging, animation, digital video, and large-scale distributed applications.

IBM Storage Tank vs. Traditional Distributed Systems

Traditional distributed file systems use a client/server data access model that requires servers to access data from storage devices, and then send the data to clients. They have the additional limitation of using conventional network bandwidth to transfer the data. While these systems allow users to share data, they do not provide the high performance required for data-intensive applications.

In contrast, IBM Storage Tank uses a data access model that requires clients to obtain only metadata from a server. Clients can then access data directly from storage devices using the high-bandwidth provided by a Fibre Channel or other high-speed network. Direct data access helps eliminate server bottlenecks and provides the performance necessary for data-intensive applications.

Storage Virtualization

Storage virtualization masks the physical characteristics of storage devices and presents users and applications with a unified, logical pool of shared storage. It gives storage administrators the flexibility to create virtual disks that better meet the needs of users and their applications.

IBM Storage Tank provides storage virtualization through the use of storage pools. A storage pool can consist of multiple disks that reside on any combination of heterogeneous storage devices. To an application, a storage pool appears as a single storage space in which it can store data without the need to know anything about the characteristics or boundaries of the physical disks.

A storage administrator sets up storage pools to meet specific needs of an enterprise. For example, an enterprise might want to have storage pools that consist of disks located on fast devices for transactional data and storage pools of disks on slower devices for backup data. An enterprise might also want to have multiple storage pools that provide different availability characteristics, or, perhaps, separate storage pools for each department within the enterprise.

After setting up storage pools, an administrator can increase or decrease the sizes of specific storage pools to meet changing needs, and can easily move data from one storage pool to another. All of these tasks are transparent and non-disruptive to users and applications.

System-Managed Storage

The IBM Storage Tank architecture makes it possible to bring the benefits of system-managed storage (SMS) to a distributed environment. Features such as policy-based allocation, volume management, and file management have long been available on IBM mainframe systems. However, the infrastructure for such centralized, automated management has been lacking in the open systems world. The centralized storage management architecture of IBM Storage Tank makes it possible to realize the advantages of system-managed storage for all of the data that the IBM Storage Tank system manages.

On conventional systems, storage management is platform dependent. While storage devices may be attached to a SAN, they are still allocated to specific server machines and cannot be centrally or consistently managed.

IBM Storage Tank provides a single, centralized point of control to better manage storage devices and data. Centralized storage and data management simplifies administration and can result in lower cost of ownership.

IBM Storage Tank Architecture

Figure 1 illustrates the basic IBM Storage Tank architecture.

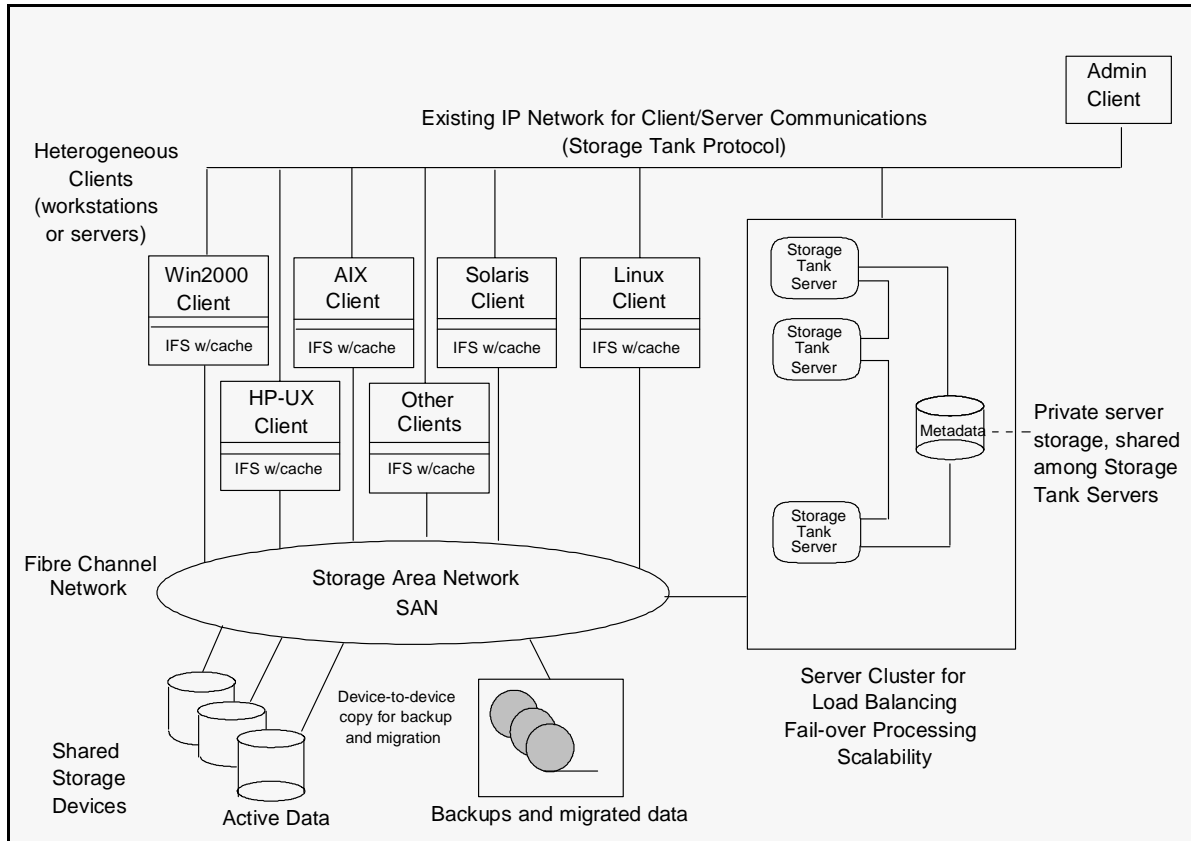


Figure 1. IBM Storage Tank Architecture

Figure 1 shows that IBM Storage Tank clients and the administrative client communicate with IBM Storage Tank servers over an enterprise's existing IP network using the IBM Storage Tank protocol. It also shows that IBM Storage Tank clients, servers, and storage devices are all connected to a high-speed Storage Area Network (SAN).

The IBM Storage Tank administrative client serves as the administrative control point. An administrator can perform almost all administrative tasks online with no service interruption to clients.

An installable file system (IFS) is installed on each IBM Storage Tank client. An IFS directs requests for metadata and locks to a IBM Storage Tank server and sends requests for data to storage devices on the SAN. Storage Tank clients can access data directly from any storage device attached to the SAN.

IBM Storage Tank clients aggressively cache file data, as well as metadata and locks that they obtain from a Storage Tank server, in memory. They do not cache files to disk.

An enterprise can use one IBM Storage Tank server, a cluster of IBM Storage Tank servers, or multiple clusters of IBM Storage Tank servers. Clustered servers provide load balancing, fail-over

processing, and increased scalability. The IBM Storage Tank servers in a cluster are interconnected on their own high-speed network or on the same IP network they use to communicate with IBM Storage Tank clients. The private server storage that contains the metadata managed by a cluster of IBM Storage Tank servers can be attached to a private network connected only to the cluster of servers, or it can be attached to the IBM Storage Tank SAN.

IBM Storage Tank Protocol

The IBM Storage Tank protocol is a locking and data consistency model that allows the IBM Storage Tank distributed storage system to look and behave like a local file system. The objective of the IBM Storage Tank protocol is to provide strong data consistency between clients and servers in a distributed environment.

The IBM Storage Tank protocol provides locks that enable file sharing among IBM Storage Tank clients, or, when necessary, provides locks that allow clients to have exclusive access to files. A IBM Storage Tank server grants the locks to clients. With the IBM Storage Tank protocol, when a client reads data from a particular file, it always reads the last data written to that file from anywhere in the IBM Storage Tank distributed storage system.

To open a file in the IBM Storage Tank distributed storage system, a client does the following:

1. Contacts a IBM Storage Tank server to obtain metadata and locks.

Metadata supplies the client with information about a file, such as its attributes and location on storage device(s).

Locks supply the client with the privileges it needs to open a file and read or write data. The IBM Storage Tank locking scheme is designed to ensure strong data consistency.

2. Accesses the data for the file directly from a shared storage device attached to a high-performance SAN.

IBM Storage Tank Clients

One of the goals of IBM Storage Tank is to enable full, transparent data sharing of files among heterogeneous clients, such as those running the Windows 2000, AIX, Solaris, Linux, and HP-UX operating systems.

All IBM Storage Tank clients can access the same data using IBM Storage Tank's uniform global namespace. A *uniform global namespace* provides the capabilities for all clients to have a consistent view of the IBM Storage Tank name tree.

File Server Support

A file server, such as an NFS, CIFS, or HTTP server, can also be an IBM Storage Tank client. For these clients, IBM Storage Tank provides the following:

- **Scalability** — A file server can access all of the files in the IBM Storage Tank distributed storage system. IBM Storage Tank is highly scalable and, therefore, can provide a file server access to a vast amount of data.

- Reliability and fail-over processing — Because many servers can be IBM Storage Tank clients and can export the same files, requests from clients of a failed server can be transferred to another server using any technique supported by the file server, such as high availability cluster multi-processing (HACMP) or IP address stealing.

Installable File Systems

IBM Storage Tank requires an installable file system (IFS) on each IBM Storage Tank client. The IFS software, which is easy to install, is available through a Web interface to an IBM Storage Tank server.

An IFS is a subsystem of an IBM Storage Tank client's file system (or in the case of supported UNIX clients, the client's virtual file system (VFS) layer). The IFS is designed to direct all metadata operations to an IBM Storage Tank server, and direct all data operations to storage devices attached to a high-speed network. It makes the metadata that is visible to the client's operating system, as well as any applications that the client runs, look identical to metadata read from a native, locally-attached file system.

Note that special purpose applications, such as digital libraries and databases, can also access data from the IBM Storage Tank distributed storage system by using an application programming interface (API) to the IBM Storage Tank protocol. Because these applications do not use the file system to access their data, the clients on which they run do not need to have the IBM Storage Tank IFS installed.

Storage Tank Client Cache

The IBM Storage Tank client cache is used to achieve low-latency access to metadata and data. A client can cache the following:

- Data — Caching data allows a client to perform reads and writes for smaller files locally, eliminating I/O operations to storage devices attached to the SAN.
- Metadata — Caching metadata allows a client to perform multiple metadata reads locally without contacting a Storage Tank server.

Note that all metadata writes are done by a Storage Tank server.

- Locks — Caching locks allows a client to grant multiple opens to a file locally without contacting an IBM Storage Tank server.

An IBM Storage Tank client performs all data caching in memory. If there is not enough space in the client's cache for all of the data in a file, the client simply reads the data from the shared storage device on which the file is stored. Data access is fast because the client has direct access to all storage devices attached to IBM Storage Tank's high-speed network. There is no need for a client to cache data to a private disk.

IBM Storage Tank Servers

IBM Storage Tank servers can run on different operating systems, such as AIX and Linux. The IBM Storage Tank server is a portable, user-level, C++ application. It could be ported to additional operating systems, for example, Windows.

Support for multiple operating systems allows an administrator to choose from a wide range of server machines on which to install the server programs. This allows the administrator to provide the appropriate level of performance for an enterprise. For example, an administrator can choose to install the server programs on computers built on Intel processors running Linux for cost-effective scalability or on IBM SP2 supercomputers running AIX for high-end scalability.

IBM Storage Tank servers provide the following services:

- Metadata services
- Administrative services
- Storage management services

Metadata Services

An IBM Storage Tank server is designed to perform these metadata services:

- Manage allocation and placement of data in storage pools on storage devices
- Perform metadata writes
- Serve file system metadata to clients
- Grant file and data locks to clients
- Detect client failures and perform client recovery

An enterprise can use a single server, a cluster of servers, or multiple clusters of servers. An administrator can move data between clusters using the IBM Storage Tank administrative interface.

Using IBM Storage Tank servers in a cluster configuration has the following benefits:

- Load balancing — The workload and data structures for the IBM Storage Tank distributed storage system are partitioned and allotted to the servers in the cluster. This is a continuous process that keeps the cluster workload balanced at all times.
- Fail-over processing — IBM Storage Tank servers are designed to redistribute the load evenly among servers in the cluster if one of the servers should fail.
- Scalability — An administrator can add more servers to a cluster or add more server clusters to the SAN to serve more data and more clients. Note that multiple server clusters cooperate to maintain the Storage Tank uniform global namespace.

The IBM Storage Tank servers in a specific cluster must all be of the same type. However, an installation can have multiple clusters of different types. For example, an enterprise might have one server cluster in which all the servers run AIX, and another server cluster in which all the servers run Linux.

IBM Storage Tank servers in a cluster can be interconnected on their own high-speed network or on the same IP network through which they communicate with IBM Storage Tank clients. The

metadata managed by a cluster of servers resides on private server storage that is shared among the servers in the cluster.

Administrative Services

As part of a server cluster, there can be one or more administrative clients that an administrator can use to control the IBM Storage Tank servers. An administrative client is connected to the IBM Storage Tank servers via an IP network. To perform administrative tasks, an administrator can choose to use either a graphical user interface or a command line interface.

From an administrative client, an administrator can do the following:

- Manage physical devices (for example, an administrator can add or remove disks and can commission or decommission server nodes).
- Create and maintain storage pools based on Quality of Service (QoS) requirements. For example, an administrator can create a storage pool that consists of RAID or striped storage devices to meet reliability requirements, and can create a storage pool that consists of random or sequential access or low-latency storage devices to meet high performance requirements.
- Create containers that are subtrees of the directory tree in a cluster's global namespace.
- Manage quotas that control the size of containers or the portion of any given storage pool that can be used by a particular container. Although IBM Storage Tank does not directly support user or group quotas, an administrator can map users and/or groups to containers to achieve the same results.
- Manage groups that are a collection of users to which an administrator can grant specific rights and permissions.
- Take snapshots of the IBM Storage Tank metadata.

A *snapshot* creates a point-in-time view of IBM Storage Tank's entire uniform global namespace or any portion of it. Storage Tank clients support metadata snapshots by implementing a copy-on-write facility. Any data a client writes to a file between snapshots is written to a different location in storage. The data that existed when the last snapshot was taken remains in the original location.

Data mining applications and backup utilities can access the point-in-time view of Storage Tank data without interrupting normal data operations on the IBM Storage Tank SAN. While a snapshot is being taken, all data remains online and can be read and written to by clients.

The IBM Storage Tank design allows for these administrative tasks to be performed online with no interruption in service to clients.

Storage Management Services

Based on storage management policies set up by an administrator, IBM Storage Tank is designed to automatically perform a variety of storage management services. It performs these services across the SAN with no client involvement.

IBM Storage Tank Shared Storage

An administrator can choose to use various types of storage for the IBM Storage Tank SAN. Data storage can be any SAN-enabled disk or subsystem (such as RAID, JBOD, or hierarchically managed devices), and will ultimately include tape and optical devices. An administrator can group disks (LUNs) into storage pools by attribute (for example, by grouping all RAID subsystems into a high availability storage pool or by grouping all cached subsystems in a high performance storage pool).

All storage devices attached to the IBM Storage Tank SAN can be accessed by all clients that are part of the IBM Storage Tank system. This enables data sharing among the heterogeneous clients supported by IBM Storage Tank.

IBM Storage Tank Security

The IBM Storage Tank security architecture offers techniques for both homogeneous and heterogeneous environments. An administrator must choose the technique to be used at install time. However, if an administrator chooses the homogeneous technique at install time, the administrator can choose to switch to the heterogeneous technique at a later time if desired.

IBM Storage Tank Features and Benefits

This section highlights the most significant features and benefits in the IBM Storage Tank design:

Exploitation of Storage Area Network technology on a Fibre Channel network

A SAN environment allows large numbers of devices, such as clients, servers, and storage devices, to connect to a Fibre Channel network using Fibre Channel adapters. A Fibre Channel network is a high-performance interconnect standard that can transfer large volumes of data between storage devices and clients. On a Fibre Channel network, Storage Tank clients can establish high-speed, low-latency connections with intelligent, mass storage subsystems that each manage thousands of terabytes of information. A Fibre Channel network is ideal for data-intensive applications.

While IBM Storage Tank will initially run on a Fibre Channel SAN, the IBM Storage Tank design is independent of the SAN fabric technology and can run on any high-speed storage network, such as Gigabit Ethernet (iSCSI) or Infiniband.

Support for heterogeneous clients

An enterprise can choose to use any combination of supported clients (for example, Windows 2000, AIX, Linux, Solaris, HP-UX, etc.).

Support for multiple server platforms

An administrator can choose from a wide range of servers, such as Intel servers for cost-effective scalability or IBM SP2 servers for high-end scalability. This flexibility allows an administrator to choose the appropriate server platform for an enterprise's needs.

Data sharing through a uniform global namespace

All IBM Storage Tank servers in a cluster and all server clusters in an IBM Storage Tank installation cooperate to provide a uniform global namespace for the data they manage. Maintaining a uniform global namespace for all data managed by Storage Tank servers allows all IBM Storage Tank clients to have a consistent view of IBM Storage Tank data at all times.

Highly-scalable storage pools and server clusters

A SAN is designed to implement large-scale systems. An administrator can:

- Add storage devices to the IBM Storage Tank high-performance network as necessary. Each new device is available to all clients on the SAN.
- Add servers to an IBM Storage Tank server cluster as necessary to serve more clients. When an administrator adds a new server to a cluster, that server is integrated into the cluster automatically and can absorb a portion of the overall cluster's workload.
- Add new server clusters to the SAN to serve more clients and data and increase performance.

High availability of data

The goal of an IBM Storage Tank server cluster is to perform load balancing and fail-over processing to ensure that data is available to users quickly and continuously if there is a server failure.

If a server fails, IBM Storage Tank is designed to move portions of the file system workload to other servers in the same server cluster, and redirects clients of the failed server to those servers. The IBM Storage Tank protocol allows clients to reassert the locks they held with the failed server with other servers in the cluster. By reasserting locks, clients preserve open files and retain cached data. They avoid the expense of obtaining new locks and reading data from storage devices again. Lock reassertion is a technology designed to allow data to be available at the client continuously with little or no performance impact if a server fails.

IBM Storage Tank also provides backup for recovery from media failures and data corruption. In addition, Storage Tank can use RAID devices and storage devices that perform mirroring to help enable recovery from device failures.

High system availability

IBM Storage Tank is designed so that an administrator can perform SAN administration tasks, such as adding or deleting a disk, allocating storage, moving data, or initiating a backup, without interrupting system availability. This design allows an administrator to perform SAN administration tasks with all IBM Storage Tank clients and servers remaining operational, and all data remaining online and available to clients and servers.

Uniform support for a wide range of file sizes

Aggressively caching data at the client provides low-latency access to smaller files, while the Fibre Channel or other high-speed network provides high-bandwidth for larger files.

Storage virtualization

Storage virtualization offers cross-platform storage control that masks the characteristics of physical devices and provides flexible, easily managed storage pools.

The storage virtualization provided by IBM Storage Tank enables administrators to manage disk space more efficiently than would be possible without it, thus increasing an enterprise's return on investment (ROI) in storage devices and decreasing the total cost of ownership (TCO).

System-managed storage

The IBM Storage Tank architecture allows centralized, automated storage management in a distributed environment. This centralized management makes it possible for IBM Storage Tank to provide similar capabilities to those found on IBM mainframes but in heterogeneous SAN environments.

Reduced administration costs

IBM Storage Tank simplifies storage administration and can reduce the total cost of ownership for storage.

Storage administration is easier because all storage is centralized and available to all clients. Storage resources are not fragmented, where some storage devices belong only to a specific workstation or only to a specific server. An administrator does not have to move storage devices or reallocate storage among servers to balance space or workload.

To perform administrative tasks, IBM Storage Tank provides an administrative client. An administrator can choose to use a graphical user interface or a command line interface. Storage Tank requires only a minimal number of administrative commands to manage the SAN. Fewer commands are required to manage the IBM Storage Tank distributed storage system than are required to manage a traditional file server.

Data integrity

The IBM Storage Tank protocol provides distributed file access that is strongly consistent and cache coherent. It uses distributed data locks for cache consistency and file access locks to synchronize multiple, concurrent open instances of the same file.

IBM Storage Tank differs from previous distributed storage systems in that it provides exact local file system semantics through the installable file system installed on an IBM Storage Tank client. IBM Storage Tank locks are unique because their semantics are rich enough to fully describe and enforce local file system semantics in a distributed environment. In addition, IBM Storage Tank clients cache file access locks in memory, which allows them to service open requests locally and avoid the overhead of additional messages to the server.

When clients or servers fail, IBM Storage Tank is designed to enforce cache coherency and file system semantics. The design uses a lease-based safety protocol that helps protect the consistency and structural integrity of the storage system. The lease-based safety protocol is tightly integrated with failover processing to help ensure high availability and data integrity if failures occur.

Summary

When compared to a conventional distributed environment, IBM Storage Tank's data access model improves data sharing performance. In a conventional distributed system, servers obtain data from storage devices, and then send the data to clients. Using the Storage Tank distributed storage system, clients obtain only metadata and locks from a server, and then read data directly from storage devices. The improved data access model and the SAN technology used by Storage Tank provide heterogeneous data sharing in a distributed environment with performance that is comparable to that of a local file system.

Centralized storage and server clustering increase the scalability of the IBM Storage Tank SAN. An administrator can add more storage devices as needed to serve all clients, and add more servers to a server cluster or add more server clusters to the SAN to manage more data and serve more clients. IBM Storage Tank server clusters are designed to increase the availability of data to clients by performing load balancing and fail-over processing.

Additionally, IBM Storage Tank is designed to be easy to manage. An administrator can perform SAN management tasks, such as adding or removing disks, creating storage pools based on Quality of Service requirements, and taking snapshots of the IBM Storage Tank data tree, without interrupting service to clients.

Finally, because IBM Storage Tank's architecture makes it possible to realize the advantages of open system-managed storage, Storage Tank provides a storage management solution that automates many aspects of storage management and can reduce the total cost of ownership for storage in a heterogeneous SAN environment.

References

The following are references to IBM Storage Tank technical papers written by members of the IBM Almaden Research Center Storage Tank team:

R. C. Burns, R. M. Rees, and D. D. E. Long
Efficiently Distributing Data in a Web Server Farm
To appear in: *IEEE Internet Computing*, 2001.

R. C. Burns, R. M. Rees, and D. D. E. Long
An Analytical Study of Opportunistic Lease Renewal
In *Proceedings of the 16th International Conference on Distributed Computing Systems (ICDCS)*, IEEE, 2001.

R. C. Burns and W. C. Hineman
A Bit-Parallel Search Algorithm for Allocating Free Space
To appear in: *Proceedings of the 9th International Symposium on Modeling, Analysis, and Simulation in Computer and Telecommunication Systems (MASCOTS)*, IEEE, 2001.

R. C. Burns, R. M. Rees, L. J. Stockmeyer, and D. D. E. Long
Scalable Session Locking for a Distributed File System
In *Cluster Computing Journal*, Volume 4, Number 4, Dec. 2001.

R. C. Burns, R. M. Rees, and D. D. E. Long
Safe Caching in a Distributed File System for Network Attached Storage
In *Proceedings of the International Parallel and Distributed Processing Symposium (IPDPS)*, IEEE, 2000.

R. C. Burns, R. M. Rees, and D. D. E. Long
Semi-Preemptible Locks for a Distributed File System
In *Proceedings of the 2000 International Performance Computing and Communication Conference (IPCCC)*, IEEE, 2000.

R. C. Burns, R. M. Rees, and D. D. E. Long
Consistency and Locking for Distributing Updates to Web Servers Using a File System
In *Performance Evaluation Review*, 28(2), ACM, 2000.

R. C. Burns
Data Management in a Distributed File System for Storage Area Networks
A dissertation in completion of the Doctor of Philosophy degree,
Department of Computer Science, University of California, Santa Cruz, March 2000.

Special Notices

© International Business Machines Corporation 2001

IBM Corporation
Storage Systems Group
5600 Cottle Road
San Jose, CA 95193

Produced in the United States of America
All Rights Reserved

The following terms are trademarks of the International Business Machines Corporation in the United States and/or other countries:

AIX
IBM ®
IBM Storage Tank
SP2 ®

The following terms are trademarks of other companies:

Microsoft, Windows, Windows 2000, and the Windows logo are trademarks of Microsoft Corporation in the United States and/or other countries.

Solaris is a trademark of Sun Microsystems, Inc. in the United States and/or other countries.

HP-UX is a trademark of the Hewlett Packard Corporation in the United States and/or other countries.

UNIX is a registered trademark in the United States and/or other countries licensed exclusively through X/Open Company Limited.

Other company, product, and service names may be trademarks or service marks of others.

IBM may have patents or pending patent applications covering subject matter in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to the IBM Director of Licensing, IBM Corporation, North Castle Drive, Armonk, NY 10504-1785.

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PAPER "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes may be made periodically to the information herein; these changes may be incorporated in subsequent versions of the paper. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this paper at any time without notice.