Server Clusters : Storage Area Networks

Windows 2000 and Windows Server 2003

Microsoft Corporation

Published: March 2003

Abstract

This document describes what storage area networks (SAN) are, how server clusters can be deployed in a SAN, and how the Microsoft® Windows® platform, and Windows clustering in particular, take advantage of storage area network technology.

Contents

Introduction

A storage area network (SAN) is defined as a set of interconnected *devices* (for example, disks and tapes) and *servers* that are connected to a common communication and data transfer infrastructure such as Fibre Channel. The common communication and data transfer mechanism for a given deployment is commonly known as the *storage fabric*. The purpose of the SAN is to allow multiple servers access to a pool of storage in which any server can potentially access any storage unit. Clearly in this environment, management plays a large role in providing security guarantees (who is authorized to access which devices) and sequencing or serialization guarantees (who can access which devices at what point in time).

SANs evolved to address the increasingly difficult job of managing storage at a time when the storage usage is growing explosively. With devices locally attached to a given server or in the server enclosure itself, performing day-to-day management tasks becomes extremely complex; backing up the data in the datacenter requires complex procedures as the data is distributed amongst the nodes and is accessible only through the server it is attached to. As a given server outgrows its current storage pool, storage specific to that server has to be acquired and attached, even if there are other servers with plenty of storage space available. Other benefits can be gained such as multiple servers can share data (sequentially or in some cases in parallel), backing up devices can be done by transferring data directly from device to device without first transferring it to a backup server.

So why use yet another set of interconnect technologies? A storage area network is a network like any other (for example a LAN infrastructure). A SAN is used to connect many different devices and hosts to provide access to any device from anywhere. Existing storage technologies such as SCSI are tuned to the specific requirements of connecting mass storage devices to host computers. In particular, they are low latency, high bandwidth connections with extremely high data integrity semantics. Network technology, on the other hand, is tuned more to providing application-to-application connectivity in increasingly complex and large-scale environments. Typical network infrastructures have high connectivity, can route data across many independent network segments, potentially over very large distances (consider the internet), and have many network management and troubleshooting tools.

Storage area networks try to capitalize on the best of the storage technologies and network technologies to provide a low latency, high bandwidth interconnect which can span large distances, has high connectivity, and good management infrastructure from the start.

In summary, a SAN environment provides the following benefits:

Centralization of storage into a single pool. This allows storage resources and server resources to grow independently, and allows storage to be dynamically assigned from the pool as and when it is required. Storage on a given server can be increased or decreased as needed without complex reconfiguring or re-cabling of devices.

Common infrastructure for attaching storage allows a single common management model for configuration and deployment.

Storage devices are inherently shared by multiple systems. Ensuring data integrity guarantees and enforcing security policies for access rights to a given device is a core part of the infrastructure.

Data can be transferred directly from device to device without server intervention. For example, data can be moved from a disk to a tape without first being read into the memory of a backup server. This frees up compute cycles for business logic rather than management related tasks.

Because multiple servers have direct access to storage devices, SAN technology is particularly interesting as a way to build clusters where shared access to a data set is required. Consider a clustered SQL Server™ environment. At any point in time a SQL Server instance may be hosted on one machine in the cluster and it must have exclusive access to its associated database on a disk from the node on which it is hosted. In the event of a failure or an explicit management operation, the SQL Server instance may *failover* to another node in the cluster. Once failed over, the SQL Server instance must be able to have exclusive access to the database on disk from its new host node.

By deploying multiple clusters onto a single storage area network, all of the benefits of SAN technology described above can be brought to the cluster environment. The rest of this paper describes how clusters can be attached to storage area networks, what the requirements are and what is supported today in Windows 2000.

Storage Area Network Components

As previously discussed, the primary technology used in storage area networks today is Fibre Channel. This section provides a basic overview of the components in a fibre channel storage fabric as well as different topologies and configurations open to Windows deployments.

Fibre Channel Topologies

Fundamentally, fibre channel defines three configurations:

- Point-to-point
- Fibre Channel Arbitrated Loop (FC-AL)
- Switched Fibre Channel Fabrics (FC-SW).

Although the term “fibre channel” implies some form of fibre optic technology, the fibre channel specification allows for both fibre optic interconnects as well as copper coaxial cables.

Point-to-Point

Point-to-point fibre channel is a simple way to connect two (and only two) devices directly together, as shown in Figure 1 below. It is the fibre channel equivalent of direct attached storage (DAS).

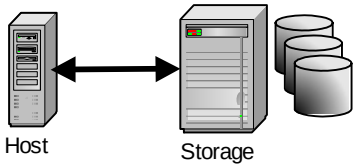

Figure 1: Point to point connection

From a cluster and storage infrastructure perspective, point-to-point is not a scalable enterprise configuration and we will not consider it again in this document.

Arbitrated Loops

A fibre channel arbitrated loop is exactly what it says; it is a set of hosts and devices that are connected into a single loop, as shown in Figure 2 below. It is a cost-effective way to connect up to 126 devices and hosts into a single network.

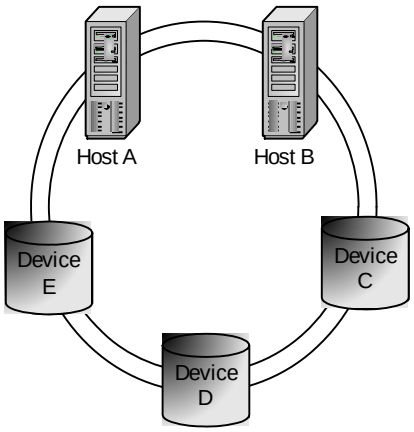

Figure 2: Fibre Channel arbitrated loop

Devices on the loop share the media; each device is connected in series to the next device in the loop and so on around the loop. Any packet traveling from one device to another must pass through all intermediate devices. In the example shown, for host A to communicate with device D, all traffic between the devices must flow through the adapters on host B and device C. The devices in the loop do not need to look at the packet; they will simply pass it through. This is all done at the physical layer by the fibre channel interface card itself; it does not require processing on the host or the device. This is very analogous to the way a token-ring topology operates.

When a host or device wishes to communicate with another host or device, it must first arbitrate for the loop. The initiating device does this by sending an arbitration packet around the loop that contains its own loop address (more on addressing later). The arbitration packet travels around the loop and when the initiating device receives its own arbitration packet back, the initiating device is considered to be the loop owner. The initiating device next sends an open request to the destination device which sets up a logical point-to-point connection between the initiating device and target. The initiating device can then send as much data as required before closing down the connection. All intermediate devices simply pass the data through. There is no limit on the length of time for any given connection and therefore other devices wishing to communicate must wait until the data transfer is completed and the connection is closed before they can arbitrate.

If multiple devices or hosts wish to communicate at the same time, each one sends out an arbitration packet that travels around the loop. If an arbitrating device receives an arbitration packet from a different device before it receives its own packet back, it knows there has been a collision. In this case, the device with the lowest loop address is declared the winner and is considered the loop owner. There is a fairness algorithm built into the standard that prohibits a device from re-arbitrating until all other devices have been given an opportunity, however, this is an optional part of the standard.

Note: Not all devices and host bus adapters support loop configurations since it is an optional part of the fibre channel standard. However, for a loop to operate correctly, all devices on the loop MUST

have arbitrated loop support[1]. Figure 3 below shows a schematic of the wiring for a simple arbitrated loop configuration.

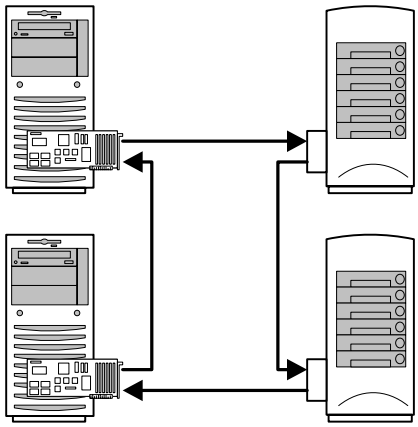

Figure 3: FC-AL wiring schematic

With larger configurations, wiring a loop directly can be very cumbersome. *Hubs* allow for simpler, centralized wiring of the loop (see section Hubs, Switches, Routers and Bridges). Communication in an arbitrated loop can occur in both directions on the loop depending on the technology used to build the loop, and in some cases communication can occur both ways simultaneously.

Loops can support up to 126 devices, however, as the number of devices on the arbitrated loop increases, so the length of the path and therefore the latency of individual operations increases.

Many loop devices, such as JBODs, have dip switches to set the device address on the loop (known as *hard addressing*). Most, if not all devices, implement hard addresses so it is possible to assign a loop ID to a device, however, just as in a SCSI configuration, different devices must have unique hard IDs. In cases where a device on the loop already has a conflicting address when a new device is added, the new device either picks a different ID or it does not get an ID at all (non-participating

Note: Most of the current FC-AL devices are configured automatically to avoid any address conflicts. However, if a conflict does happen then it can lead to I/O disruptions or failures.

Unlike many bus technologies, the devices on an arbitrated loop do not have to be given fixed addresses either by software configuration or via hardware switches. When the loop initializes, each device on the loop must obtain an Arbitrated Loop Physical Address which is dynamically assigned. This process is initiated when a host or device sends out a LIP; a master is dynamically selected for the loop and the master controls a well defined process where each device is assigned an address.

A LIP is generated by a device or host when the adapter is powered up or when a loop failure is detected (such as loss of carrier). Unfortunately, this means that when new devices are added to a loop or when devices on the loop are power-cycled, all the devices and hosts on the loop can (and

[1] Most devices today, except for some McData switches, support FC-AL.

probably will) change their physical addresses. This can lead to unstable configurations if the operating system is not fully aware of the changes.

For these reasons, arbitrated loops provide a solution for small numbers of hosts and devices in relatively static configurations.

Fibre Channel Switched Fabric

In a switched fibre channel fabric, devices are connected in a many-to-many topology using *fibre channel switches*, as shown in Figure 4 below. When a host or device communicates with another host or device, the source and target setup a point-to-point connection (just like a virtual circuit) between them and communicate directly with each other. The fabric itself routes data from the source to the target. In a fibre channel switched fabric, the media is not shared. Any device can communicate with any other device (assuming it is not busy) and communication occurs at full bus speed (1Gbit/Sec or 2Gbit/sec today depending on technology) irrespective of other devices and hosts communicating.

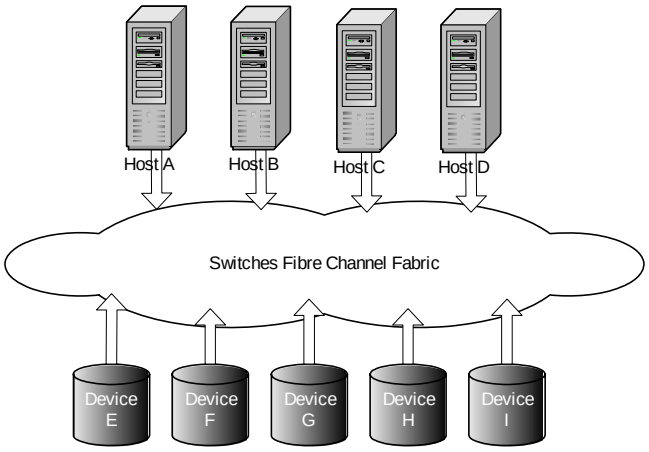

Figure 4: Switched Fibre Channel fabric

When a host or device is powered on, it must first login to the fabric. This enables the device to determine the type of fabric (there is a set of characteristics about what the fabric will support) and it causes a host or device to be given a fabric address. A given host or device continues to use the same fabric address while it is logged into the fabric and the fabric address is guaranteed to be unique for that fabric. When a host or device wishes to communicate with another device, it must establish a connection to that device before transmitting data in a way similar to the arbitrated loop. However, unlike the arbitrated loop, the connection open packets and the data packets are sent directly from the source to the target (the switches take care of routing the packets in the fabric).

Fibre channel fabrics can be extended in many different ways such as by federating switches or cascading switches, and therefore, fibre channel switched fabrics provide a much more scalable infrastructure for large configurations. Because device addresses do not change dynamically once a device has logged in to the fabric, switched fabrics provide a much more stable storage area network environment than is possible using an arbitrated loop configuration.

Fibre channel arbitrated loop configurations can be deployed in larger switched SANs. Many of the newer switches from vendors like Brocade incorporate functionality to allow arbitrated loop or point-to-point devices to be connected to any given port. The ports can typically sense whether the device is a loop device or not and adapt the protocols and port semantics accordingly. This allows platforms such as the Sun UE10000 or specific host adapters or devices which only support arbitrated loop configurations today, to be attached to switched SAN fabrics.

Note that not all switches are created equal. Brocade switches are easy to deploy; Vixel and Gadzoox switches behave more like hubs with respect to addressing.

Loops Versus Fabrics

Both fibre channel arbitrated loops and switched fabrics have pros and cons. Before deploying either, you need to understand the restrictions and issues as well as the benefits of each technology. The vendor's documentation provides specific features and restrictions; however, the following helps to position the different technologies.

FC-AL

Pros

Low cost

Loops are easily expanded and combined with up to 126 hosts and devices

Easy for vendors to develop

Cons

Difficult to deploy

Maximum 126 devices

Devices share media thus lower overall bandwidth

Switched Fabric

Pros

Easy to deploy

Supports 16 million hosts and devices

Communicate at full wire-speed, no shared media

Switches provide fault isolation and re-routing

Cons

Difficult for vendors to develop

Interoperability issues between components from different vendors

Switches can be expensive

Host Bus Adapters

A host bus adapter (HBA) is an interface card that resides inside a server or a host computer. It is the functional equivalent of the NIC in a traditional Ethernet network. All traffic to the storage fabric or loop is done via the HBA.

HBAs, with the exception of older Compaq cards and early Tachyon based cards, support both FC-AL and Fabric (since 1999). However, configuration is not as simple or as automatic as could be supposed. It is difficult to figure out if an HBA configures itself to the appropriate setting. On a Brocade fabric, it is possible to get everything connected, however, some of it might be operating as loop and still appear to work. It is important to verify from the switch side that the hosts are operating in the appropriate mode.

Note Be sure to select the correct HBA for the topology that you are using. Although some switches can auto-detect the type of HBA in use, using the wrong HBA in a topology can lead to data loss and can cause many issues to the storage fabric.

Hubs, Switches, Routers and Bridges

Thus far, we have discussed “the storage fabric” as a generic infrastructure that allows hosts and devices to communication with each other. As you have seen, there are fundamentally different fibre channel topologies and these different topologies use different components to provide the infrastructure.

Hubs

Hubs are the simplest form of fibre channel devices and are used to connect devices and hosts into arbitrated loop configurations. Hubs typically have 4, 8, 12 or 16 ports allowing up to 16 devices and hosts to be attached, however, the bandwidth on a hub is shared by all devices on the hub. In addition, hubs are typically half-duplex (newer full duplex hubs are becoming available). In other words, communication between devices or hosts on a hub can only occur in one direction at a time. Because of these performance constraints, hubs are typically used in small and/or low bandwidth configurations.

Figure 5 below shows two hosts and two storage devices connected to the hub with the dark arrows showing the physical loop provided by the hub.

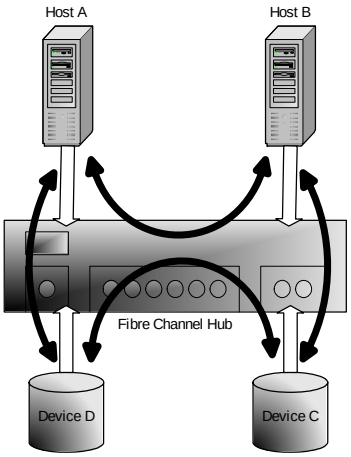

Figure 5: FC-AL hub configuration

A typical hub detects empty ports on the hub and does not configure them into the loop. Some hubs provide higher levels of control over how the ports are configured and when devices are inserted into the loop.

Switches

A switch is a more complex storage fabric device that provides the full fibre channel bandwidth to each port independently, as shown in Figure 6 below. Typical switches allow ports to be configured in either an arbitrated loop or a switched mode fabric.

When a switch is used in an arbitrated loop configuration, the ports are typically full bandwidth, bi-directional allowing devices and hosts to communicate at full fibre channel speed in both directions. In this mode, ports are configured into a loop, providing performance, arbitrated loop configuration.

Switches are the basic infrastructure used for large, point-to-point, switched fabrics. In this mode, a switch allows any device to communicate directly with any other device at full fibre channel speed (1Gbit/Sec or 2Gbit/sec today).

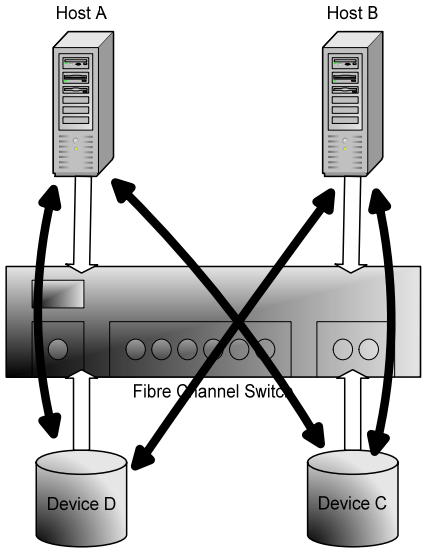

Figure 6: Switched fibre configuration

Switches typically support 16, 32, 64 or even 128 ports today. This allows for complex fabric configurations. In addition, switches can be connected together in a variety of ways to provide larger configurations that consist of multiple switches. Several manufacturers such as Brocade and McData provide a range of switches for different deployment configurations, from very high performance switches that can be connected together to provide a core fabric to edge switches that connect servers and devices with less intensive requirements.

Figure 7 below shows how switches can be interconnected to provide a scalable storage fabric supporting many hundreds of devices and hosts (these configurations are almost certainly deployed in highly available topologies, section Highly Available Solutions deals with high availability).

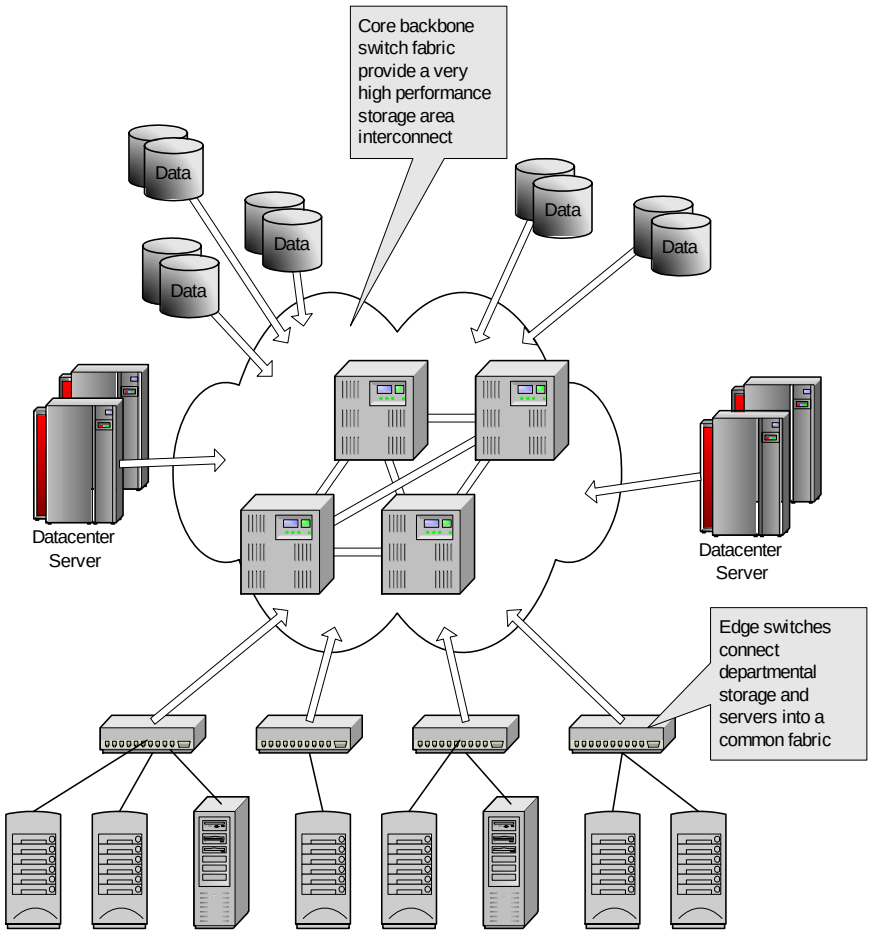

Figure 7: Core and edge switches in a SAN fabric

The core backbone of the SAN fabric is provided by high performance (and typically high port density) switches. The inter-switch bandwidth in the core is typically 8Gbit/sec and above. Large data center class machines and large storage pools can be connected directly to the backbone for maximum performance. Severs and storage with less performance requirements (such as departmental servers) may be connected via large arrays of edge switches (each of which may have 16 to 64 ports).

Bridges and Routers

In an ideal world, all devices and hosts would be SAN-aware and all would interoperate in a single, ubiquitous environment. Unfortunately, many hosts and storage components are already deployed using different interconnect technologies. To allow these types of devices to play in a storage fabric environment, a wide variety of *bridge* or *router* devices allow technologies to interoperate. For example, SCSI-to-fibre bridges or routers allow parallel SCSI (typically SCSI-2 and SCSI-3 devices) to be connected to a fibre network, as shown in Figure 8 below. In the future, bridges will allow iSCSI

(iSCSI is a device interconnect using IP as the communications mechanism and layering the SCSI protocol on top of IP) devices to connect into a switch SAN fabric.

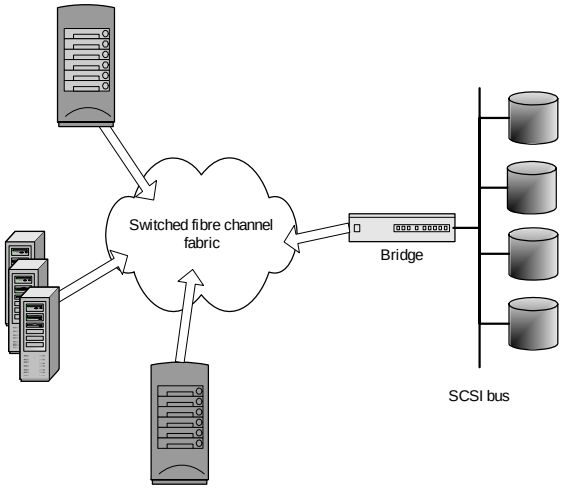

Figure 8: SCSI to Fibre Channel bridge

Storage Components

Thus far, we have discussed devices being attached to the storage bus as though individual disks are attached. While in some very small, arbitrated loop configurations, this is possible, it is highly unlikely that this configuration will persist. More likely, storage devices such as disk and tape are attached to the storage fabric using a *storage controller* such as an EMC Symmetrix or a Compaq StorageWorks RAID controller. IBM would refer to these types of components as Fibre RAID controllers.

In its most basic form, a storage controller is a box that houses a set of disks and provides a single (potentially redundant and highly available) connection to a SAN fabric. Typically, disks in this type of controller appear as individual devices that map directly to the individual spindles housed in the controller. This is known as a JBOD (just a bunch of disks) configuration. The controller provides no value-add, it is just a concentrator to easily connect multiple devices to a single (or small number for high availability) fabric switch port.

Modern controllers almost always provide some level of redundancy for data. For example, many controllers offer a wide variety of RAID levels such as RAID 1, RAID 5, RAID 0+1 and many other algorithms to ensure data availability in the event of the failure of an individual disk drive. In this case, the hosts do not see devices that correspond directly to the individual spindles; rather the controller presents a virtual view of highly available storage devices to the hosts called *logical devices*.

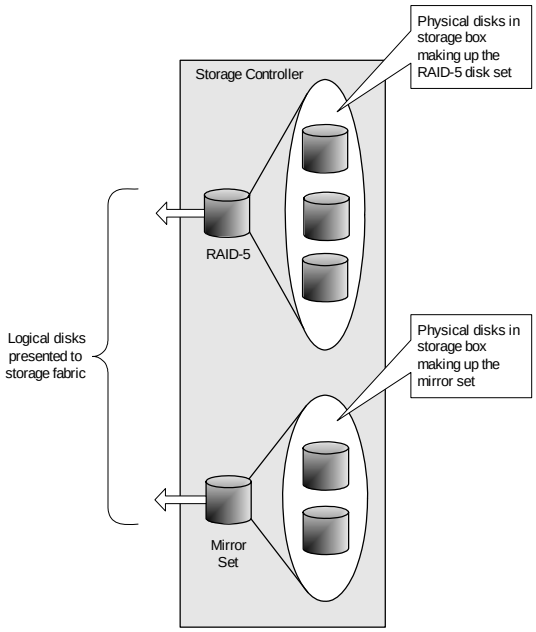

Figure 9: Logical devices

In the example in Figure 9, although there are five physical disk drives in the storage cabinet, only two logical devices are visible to the hosts and can be addressed through the storage fabric. The controller does not expose the physical disks themselves.

Many controllers today are capable of connecting directly to a switched fabric; however, the disk drives themselves are typically either SCSI, or more common now, are disks that have a built-in FC-AL interface. As you can see in Figure 10 below, the storage infrastructure that the disks connect to is totally independent from the infrastructure presented to the storage fabric.

A controller typically has a small number of ports for connection to the fibre channel fabric (at least two are required for highly available storage controllers). The logical devices themselves are exposed through the controller ports as logical units (LUNs).

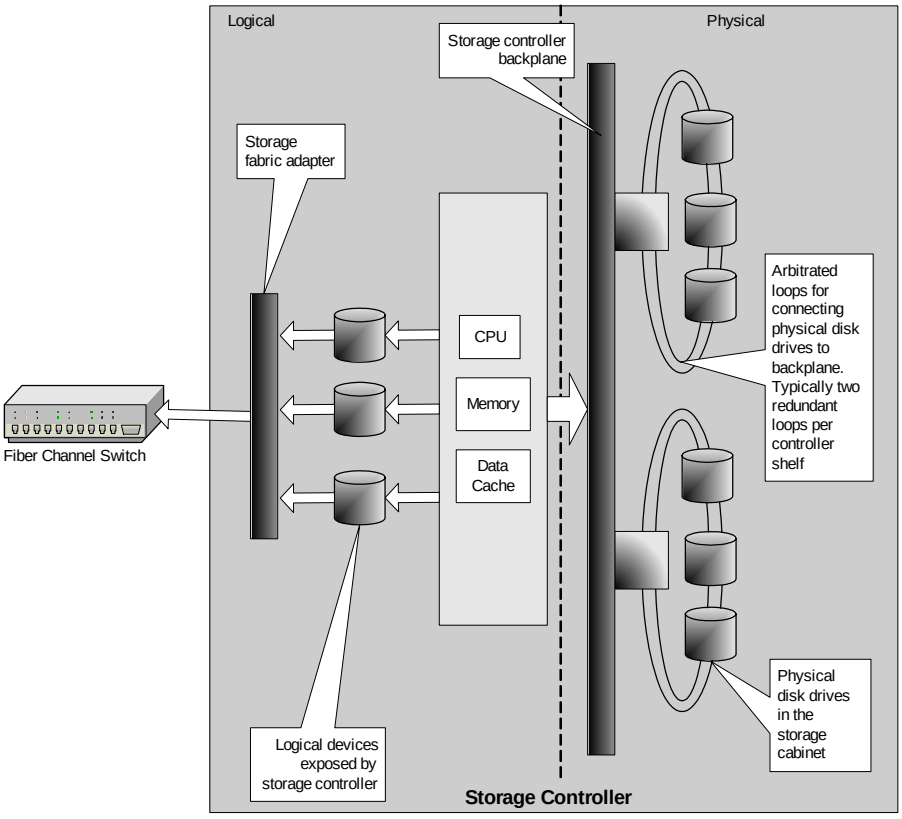

Figure 10: Internal components of a storage controller

Highly Available Solutions

One of the benefits of storage area networks is that the storage can be managed as a centralized pool of resources that can be allocated and re-allocated as required. This powerful paradigm is changing the way data centers and enterprises are built, however, one of the biggest issues to overcome is that of guaranteed availability of data. With all of the data detached from the servers, the infrastructure must be architected to provide highly available access so that the loss of one or more components in the storage fabric does not lead to the servers being unable to access the application data. All areas must be considered including:

No single point of failure of cables or components such as switches, HBAs or storage controllers. Typical highly available storage controller solutions from storage vendors have redundant components and can tolerate many different kinds of failures.

Transparent and dynamic path detection and failover at the host. This typically involves *multi-path* drivers running on the host to present a single storage view to the application across multiple, independent HBAs.

Built-in hot-swap and hot-plug for all components from HBAs to switches and controllers. Many high-end switches and most if not all enterprise class storage controllers allow interface cards, memory, CPU and disk drives to be hot-swapped.

There are many different storage area network designs that have different performance and availability characteristics. Different switch vendors provide different levels of support and different topologies, however, most of the topologies are derived from standard network topology design (after all a SAN is a network, just the interconnect technology is tuned to a given application). Topologies include:

- Multiple independent fabrics
- Federated fabrics
- Core Backbone

Multiple Independent Fabrics

In a multiple fabric configuration, each device or host is connected to multiple fabrics, as shown in Figure 11 below. In the event of the failure of one fabric, hosts and devices can communicate using the remaining fabric.

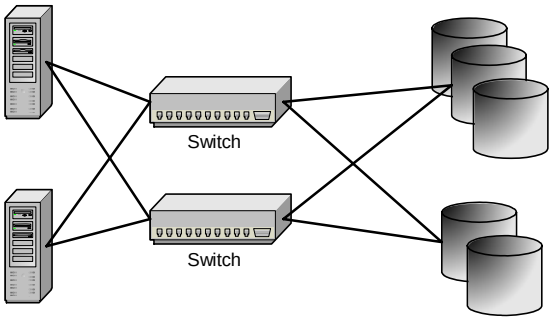

Figure 11: Multiple independent fabrics

Pros

Resilient to management or user errors. For example, if security is changed or zones are deleted, the configuration on the alternate fabric is untouched and can be re-applied to the broken fabric.

Cons

Managing multiple independent fabrics can be costly and error prone. Each fabric should have the same zoning and security information to ensure a consistent view of the fabric regardless of the communication port chosen

Hosts and devices must have multiple adapters. In the case of a host, multiple adapters are typically treated as different storage buses. Additional *multi-pathing* software such as Compaq SecurePath or EMC PowerPath are required to ensure that the host gets a single view of the devices across the two HBAs.

Federated Fabrics

In a federated fabric, multiple switches are connected together, as shown in Figure 12 below. Individual hosts and devices are connected to at least two switches.

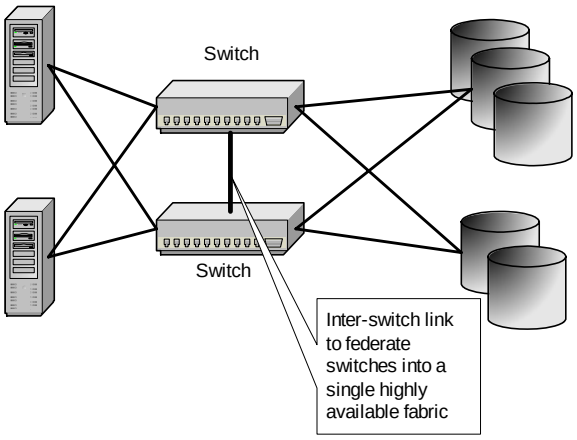

Figure 12: Federated switches for single fabric view

Pros

Management is simplified, the configuration is a highly available, single fabric, and therefore there is only one set of zoning information and one set of security information to manage.

The fabric itself can route around failures such as link failures and switch failures.

Cons

Hosts with multiple adapters must run additional *multi-pathing* software such as Compaq SecurePath or EMC PowerPath to ensure that the host gets a single view of the devices where there are multiple paths from the HBAs to the devices.

Management errors are propagated to the entire fabric.

Core Backbone

A core backbone configuration is really a way to scale-out a federated fabric environment. Figure 7 shows a backbone configuration. The core of the fabric is built using highly scalable, high performance switches where the inter-switch connections provide high performance communication (e.g. 8-10GBit/Sec using today's technology). Redundant edge switches can be cascaded from the core infrastructure to provide high numbers of ports for storage and hosts devices.

Pros

Highly scalable and available storage area network configuration.

Management is simplified, the configuration is a highly available, single fabric, and therefore there is only one set of zoning information and one set of security information to manage.

The fabric itself can route around failures such as link failures and switch failures.

Cons

Hosts with multiple adapters must run additional *multi-pathing* software such as Compaq SecurePath or EMC PowerPath to ensure that the host gets a single view of the devices where there are multiple paths from the HBAs to the devices.

Management errors are propagated to the entire fabric.

Management

As you can see from the previous section, storage area networks are increasingly complex and large configurations are becoming more and more common. While storage area networks certainly provide many benefits over direct attach storage, the big issue is how to manage this complexity.

Zoning

A storage fabric can have many devices and hosts attached to it. With all of the data stored in a single, ubiquitous cloud of storage, controlling which hosts have access to what data is extremely important. It is also important that the security mechanism be an end-to-end solution so that badly behaved devices or hosts cannot circumvent security and access unauthorized data.

Zoning is a mechanism, implemented at the switch level, which provides an isolation boundary. A port (either host adapters or storage controller ports) can be configured as part of a zone. Only ports in a given zone can communicate with other ports in that zone. The zoning is configured and access control is implemented by the switches in the fabric, so a host adapter cannot spoof the zones that it is in and gain access to data for which it has not been configured.

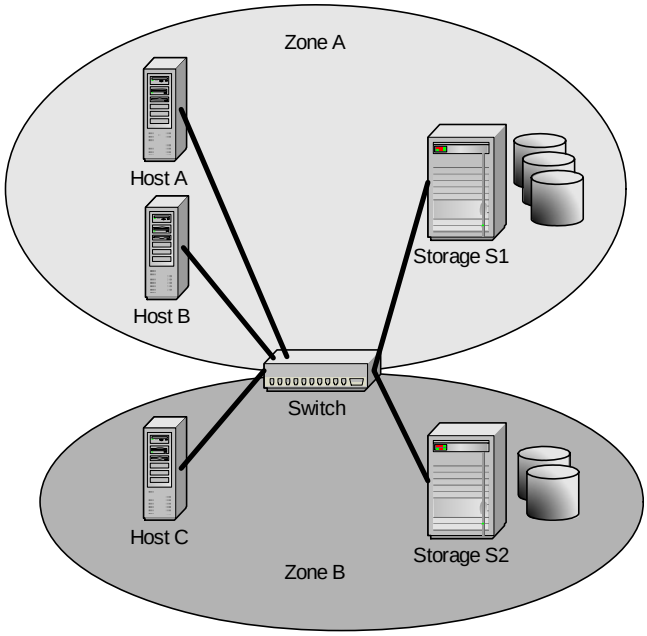

Figure 13: Zoning

In Figure 13 above, hosts A and B can access data from storage controller S1, however host C cannot as it is not in Zone A. Host C can access data from storage S2.

Many switches today allow overlapping zones. This enables a storage controller to reside in more than one zone, thus enabling the devices in that controller to be shared amongst different servers in different zones, as shown in Figure 14 below. Finer granularity access controls are required to protect individual disks against access from unauthorized servers in this environment.

Zoning can be implemented in either hardware or software. Hardware zoning is done by the ASIC in the switch ports themselves. Every packet is checked at line speed to ensure that it is authorized. Software zoning is done by the name server or other fabric access software. When a host tries to open a connection to a device, access controls can be checked at that time.

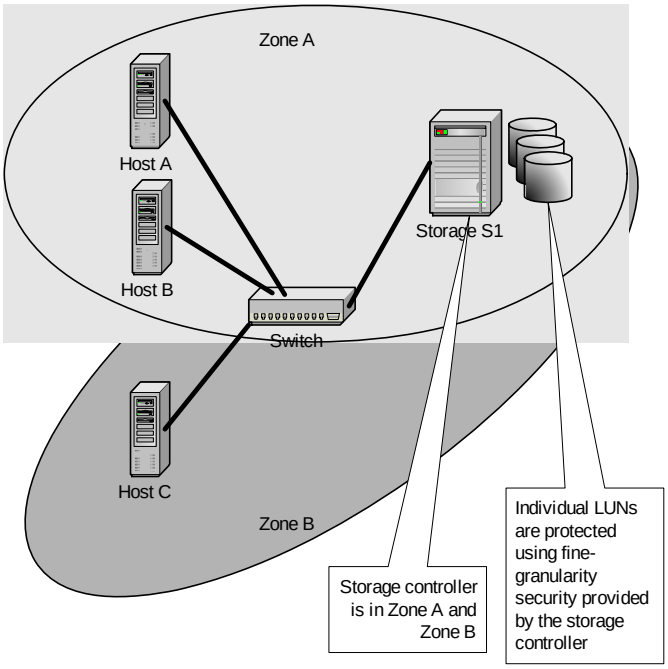

Figure 14: Storage controller in multiple zones

Zoning is an extremely important concept. Not only is it a security feature, but it also limits the traffic flow within a given SAN environment. Traffic (I/O requests and other storage requests) between ports is only routed to those pieces of the fabric that are in the same zone. Typically with modern switches, as new switches are added to an existing fabric, the new switches are automatically updated with the current zoning information.

I/Os (either read/write or such things as device reset or LIP) from hosts or devices in a fabric cannot "leak" out and affect other zones in the fabric causing "noise" or "cross-talk" between zones. As we shall see, this is fundamental to deploying Server clusters on a SAN.

Fine-Grain Security and Access Control

While zoning provides a high-level security infrastructure in the storage fabric, it does not provide the fine-grain level of access control needed for large storage devices. In a typical environment, a storage controller may have many gigabytes or terabytes of storage to be shared amongst a set of servers.

Storage controllers typically provide LUN-level access controls that enable an administrator to restrict access to a given LUN to one or more hosts. By providing this access control at the storage controller, the controller itself can enforce access policies to the data.

LUN masking is a host-based mechanism that “hides” specific LUNs from applications. Although the host bus adapter and the lower layers of the operating system have access to and could communicate with a set of devices, LUN masking prevents the higher layers from knowing that the device exists and therefore applications cannot use those devices. LUN masking is a policy-driven software security and access control mechanism enforced at the host. For this policy to be successful, the administrator has to trust the drivers and the operating systems to adhere to the policies.

SAN Management

SAN management is a huge topic on its own and is outside the scope of this document. Different vendors (both vendors that provide SAN fabric components as well as software vendors that provide storage management tools) provide a wide range of tools for setting up, configuring, monitoring and managing the SAN fabric, as well as the state of devices and hosts on the fabric.

Virtualized View of Storage

The previous section touched on virtualization of storage when describing various RAID levels. The logical devices presented by the controller to the storage fabric are some composite of the real physical devices in the storage cabinet. Moving forward, the panacea for storage management is that the devices presented to the storage infrastructure are not tied to any physical storage. In other words, the set of spindles in the cabinet is treated as a pool of storage blocks. Logical devices can be materialized from that storage pool with specific attributes such as “must survive a single failure, have xyz performance characteristics” etc. The storage controller is then free to store the data associated with the logical devices anywhere (and indeed change the placement at will) as long as the desired characteristics are maintained.

At this point, there are no real physical characteristics associated with a logical disk, any physical notions, such as a disk serial number or identity, are purely software-generated virtualized views. See Figure 15 below.

By taking this route, storage vendors can drive many value-added storage management features down into the storage infrastructure itself without having to have host involvement. We are seeing the first few steps down this path today with the notion of *snapshots* provided by some storage controllers today.

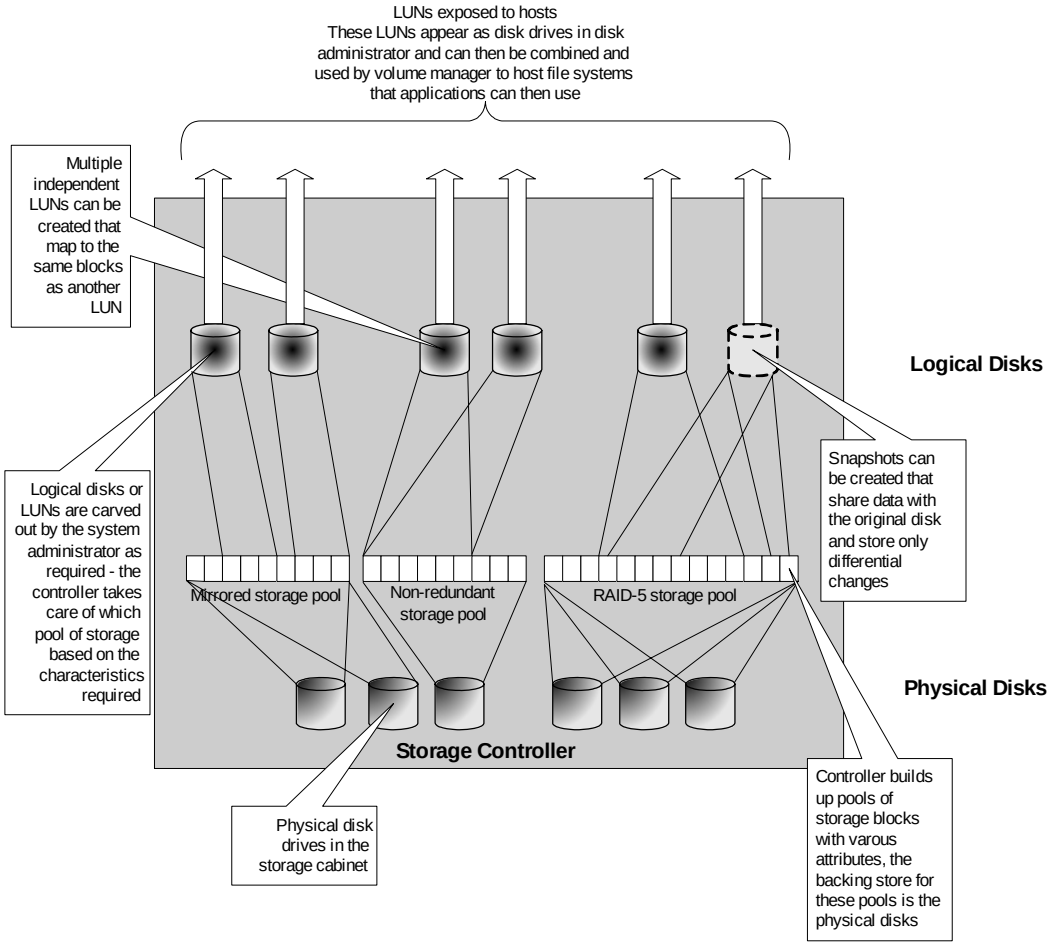

Figure 15: Storage virtualization by the controller

Deploying Server Clusters in a SAN Environment

This section covers best practices for MSCS in a SAN environment. Clusters are supported by Microsoft in a SAN environment; however, there are some specific requirements and restrictions placed on the configurations.

Note In a SAN environment, the storage fabric provides access to data for a wide range of applications. If the storage fabric stability is compromised, the availability of the entire data center could be at risk, no amount of clustering can protect against an unstable or unavailable storage fabric.

Qualified Configurations

As with all existing cluster configurations, only ***complete cluster solutions*** that appear on the Microsoft Hardware Compatibility List (HCL) will be supported by Microsoft. Clusters cannot be arbitrarily built up from device-level components (even those components such as RAID controllers, multi-cluster devices etc. that are qualified as cluster components) and put together into a supported configuration.

A single cluster can be qualified and placed on the HCL using fibre channel storage interconnects and switch technology and there are many examples of complete configurations on the HCL today. This, however, does not really constitute a storage area network (SAN) configuration.

Microsoft fully supports multiple clusters and/or servers deployed on a single fibre channel switched fabric and sharing the same storage controllers as long as the configuration adheres to the following rules:

- The storage controller must be on the Cluster/Multi-Cluster Device HCL list if it is shared between clusters.
- The complete configuration for any individual cluster must be on the Cluster HCL list.

Take, for example the following HCL lists:

Cluster/Multi-cluster device HCL list:

Storage Controller St1

Storage Controller St2

Cluster HCL list

2-node advanced server cluster AS1

Server 1: Server Box S1, 256Mb, 700Mhz PIII, HBA H1

Server 2: Server Box S2, 256Mb, 700Mhz PIII, HBA H1

Storage: Storage Controller St1

4-node advanced server cluster AS2

Server 1: Server Box S5, 512Mb, 1.2Ghz PIV, HBA H1

Server 2: Server Box S6, 512Mb, 1.2Ghz PIV, HBA H1

Server 3: Server Box S7, 512Mb, 1.2Ghz PIV, HBA H1

Server 4: Server Box S8, 512Mb, 1.2Ghz PIV, HBA H1

Storage: Storage Controller St1

2-node advanced server cluster AS3

Server 1: Server Box S10, 256Mb, 700Mhz PIII, HBA H2

Server 2: Server Box S11, 256Mb, 700Mhz PIII, HBA H2

Storage: Storage Controller St2

In this case, the 2-node AS1 and the 4-node AS2 configurations can both be placed on the same storage area network and can in fact share the same storage controller St1. It is also possible to have AS3 on the same storage area network as long as it uses storage controller St2 and not St1.

With Windows 2000, the storage area network fabric itself is not on the HCL and is not qualified directly by Microsoft. When building these configurations, you must ensure that the switches and other fabric components are compatible with the HBAs and the storage controllers.

Arbitrated Loops and Switched Fabrics

Fibre channel arbitrated loops can be configured to support multiple hosts and multiple storage devices, however, arbitrated loop configurations typically have restrictions due to the nature of the technology. For example, in some cases, a complete storage controller must be assigned to a given server or cluster. Individual devices in the controller cannot be assigned to different servers or clusters. While manufacturers and vendors allow multiple clusters to be hosted on a single arbitrated loop, due to the configuration restrictions and the mechanisms that the cluster service uses to protect disks in a cluster, Microsoft recommends that only one cluster is attached to any single arbitrated loop configuration and that arbitrated loop configurations are limited to small, relatively static cluster configurations.

Fabrics are fully supported by server clusters for both a single cluster and for multiple clusters and independent servers on the same storage fabric. Fabrics provide a much more stable environment where multiple server clusters are deployed using the same storage infrastructure. Nodes (and indeed storage devices) can leave or enter the SAN independently without affecting other parts of the fabric. Highly available fabrics can be built up, and in conjunction with multi-path drivers, can provide a highly available and scalable storage infrastructure.

Hints, Tips, and Don’ts…

This section describes the dos and don’ts of deploying one or more clusters in a SAN.

Must Do

Each cluster on a SAN MUST be deployed in its own zone. The cluster uses mechanisms to protect access to the disks that can have an adverse effect on other clusters that are in the same zone. By using zoning to separate the cluster traffic from other cluster or non-cluster traffic, there is no chance of interference. Figure 16 shows two clusters sharing a single storage controller. Each cluster is in its own zone. The LUNs presented by the storage controller must be allocated to individual clusters using fine-grained security provided by the storage controller itself. LUNs must be setup as visible to all nodes in the cluster and a given LUN should only be visible to a single cluster.

The multi-cluster device test used to qualify storage configurations for the multi-cluster HCL list tests the isolation guarantees when multiple clusters are connected to a single storage controller in this way.

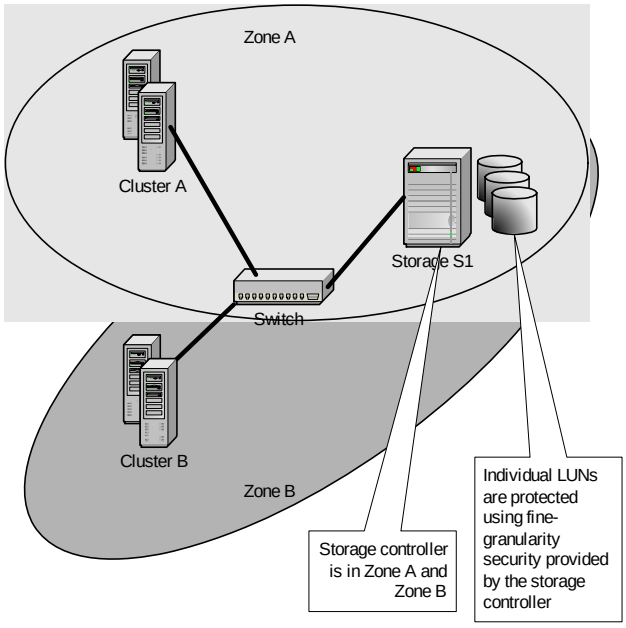

Figure 16: Clusters assigned to individual zones

All HBAs in a single cluster must be the same type and at the same firmware revision level. Many storage and switch vendors require that ALL HBAs on the same zone, and in some cases the same fabric, are the same type and have the same firmware revision number.

All storage device drivers and HBA device drivers in a cluster must be at the same software version.

SCSI bus resets are not used on a fibre channel arbitrated loop; they are interpreted by the HBA and driver software and cause a LIP to be sent. As previously described, this resets all devices on the loop.

When adding a new server to a SAN, ensure that the HBA is appropriate for the topology. In some configurations, adding an arbitrated loop HBA to a switched fibre fabric can result in widespread failures of the storage fabric. There have been real-world examples of this causing serious downtime.

The base Windows 2000 platform will mount any device that it can see when the system boots. The cluster software ensures that access to devices that can be accessed by multiple hosts in the same cluster is controlled and only one host actually mounts the disk at any one time. When first creating a cluster, make sure that only one node can access the disks that are to be managed by the cluster. This can be done either by leaving the other (to be) cluster members powered off, or by using access controls or zoning to stop the other hosts from accessing the disks. Once a single node cluster has been created, the disks marked as cluster-managed will be protected and other hosts can be either booted or the disks made visible to other hosts to be added to the cluster.
This is no different to any cluster configuration that has disks that are accessible from multiple hosts.

Note In Windows Server 2003 by using the new command mountvol/n you can disable dynamic scanning. It is recommended that dynamic scanning be disabled before the servers are connected to the SAN in a San environment. New Cluster Setup or adding and removing of nodes in a server cluster should be done while dynamic scanning is turned off. It is recommended that dynamic scanning remains turned off as long as the servers are connected to the storage infrastructure.

Must Not Do

NEVER allow multiple hosts access to the same storage devices unless they are in the SAME cluster. If multiple hosts that are not in the same cluster can access a given disk, this will lead to data corruption.

NEVER put any non-disk device into the same zone as cluster disk storage devices.

Other Hints

Highly available systems such as clustered servers should typically be deployed with multiple HBAs with a highly available storage fabric. In these cases be sure to ALWAYS load the multi-path driver software. If the I/O subsystem in the Windows 2000 platform sees two HBAs, it will assume they are different buses and enumerate all the devices assuming that they are different devices on each bus; where in fact, the host is seeing multiple paths to the same disks. Failure to load the multi-path driver will lead to data corruption. A simple manifestation of this is that the disk signature is re-written. If the Windows platform sees what it thinks are two independent disks with the same signature, it will re-write one of them to ensure that all disks have unique signatures. This is covered in KB article Q293778 *Multiple-Path Software May cause Disk Signature to Change*.

Note Windows Server 2003 will detect the fact that the same volume is being exposed twice. If such a situation arise Windows Server 2003 will not mount the volumes exposed by controller 2 that have were already exposed by the controller 1.

Many controllers today provide snapshots at the controller level that can be exposed to the cluster as a completely separate LUN. The cluster does not react well to multiple devices having the same signature. If the snapshot is exposed back to the host with the original disk online, the base I/O subsystem will re-write the signature as in the previous example, however, if the snapshot is exposed to another node in the cluster, the cluster software will not recognize it as a different disk. DO NOT expose a hardware snapshot of a clustered disk back to a node in the same cluster. While this is not specifically a SAN issue, the controllers that provide this functionality are typically deployed in a SAN environment.

Adding and Removing Disks from a Cluster

In Windows 2000 (SP3 onwards) and Windows Server 2003, adding a disk to the cluster is simple. Simply add the storage (in a SAN this probably means adding the physical drives to a storage controller and then creating a logical unit that is available in the correct zone and with the correct security attributes).

Once the disk is visible to the operating system, you can make the disk a cluster-managed disk by adding a physical disk resource in cluster administrator. The new disk will appear as being capable of being clustered. Note: Some controllers use a different cluster resource than physical disk, for those environments; create a resource of the appropriate type.

Only Basic, MBR format disks that contain at least one NTFS partition can be managed by the cluster. Before adding a disk, it must be formatted.

Remember that the same rules apply when adding disks as in creating a cluster. If multiple nodes can see the disk BEFORE any node in the cluster is managing it, this will lead to data corruption. When adding a new disk, first make the disk visible to only one cluster node and then once it is added as a cluster resource, make the disk visible to the other cluster nodes.

To remove a disk from a cluster, first remove the cluster resource corresponding to that disk. Once it has been removed from the cluster, the disk can be removed (either the drive can be physically removed or the LUN can be deleted or re-purposed)

There are several KB articles on replacing a cluster-managed disk. While disks in a cluster should typically be RAID sets or mirror sets, there are sometimes issues that cause catastrophic failures leading to a disk having to be rebuilt from the ground up. There are also other cases where cluster disks are not redundant and failure of those disks also leads to a disk having to be replaced. The steps outlined in those articles should be used if you need to rebuild a LUN due to failures.

Q243195 - Replacing a cluster managed disk in Windows NT® 4.0

Q280425 – Recovering from an Event ID 1034 on a Server Cluster

Q280425 – Using ASR to replace a disk in Windows Server 2003

Expanding Disks

Now you can expand volumes dynamically without requiring a reboot. Microsoft provided Diskpart tool can be used to expand volumes dynamically. Diskpart tool is available for each Windows 2000 and Windows Server 2003. You can download Windows 2000 version of Diskpart from the www.microsoft.com web site.

SAN Backup

Storage area networks provide many opportunities to offload work from the application hosts. Many of the devices in the SAN (either hosts or storage controllers) have CPUs and memory and are capable of executing complex code paths. In addition, any device can communicate with any other device, the SAN provides a peer-to-peer communication mechanism. This leads to such things as SAN-based backups. A storage controller can easily initiate the backup of a disk device to a tape device on the SAN without host intervention. In some cases, hybrid backup solutions are implemented where file system related information is provide by the host, but bulk copying of the data blocks is done directly from storage controller to tape device.

The cluster software uses disk reservations to protect devices that could be accessed by multiple computers simultaneously. The host that currently owns a disk protects it so that no other host can write to it. This is necessary to avoid writes that are in the pipeline when failover occurs from corrupting the disk. When failover occurs, the new owner protects the disk. This means that a cluster disk is always reserved and therefore can only be accessed by the owning host. No other host or device (including the controller that is hosting the disk) can access the disk simultaneously. This means that SAN-based backup solutions where data transfers from disk to tape are initiated by a 3rd party (i.e. initiated by a device other than the owning host) cannot be supported in a cluster environment.

Booting from a SAN

Microsoft supports booting from a SAN in limited environments. There are a set of configuration restrictions around how Windows boots from a storage area network, see KB article Q305547.

Windows 2000 Server clusters require that the boot disk, page file disk and system disk be on a different storage bus to the cluster server disks. To boot from a SAN, you must have a separate HBA for the boot, system and pagefile disks than the cluster disks. You MUST ensure that the cluster disks are isolated from the boot, system and pagefile disks by zoning the cluster disks into their own zone.

Note: Windows Server 2003 will allow for boot disk and the cluster server disks hosted on the same bus. However, you would need to use Storport miniport HBA drivers for this functionality to work. This is NOT supported configuration with in any other combination (for example., SCSI port miniport or Full port drivers)

Cluster Service Features in Windows Server 2003

The Windows Server 2003 release has a number of enhancements. The following enhancements are specific to supporting Server clusters in a SAN environment:

Targeted Reset

Historically, server clusters use the SCSI reservation mechanism to protect disks against access, guaranteeing that only the host that has the disk online can actually access it. To ensure that devices can be failed over in the event of failures, server clusters implements a challenge/response mechanism that can detect dead or hung server nodes even though they may not have crashed and therefore the storage fabric is unaware that the server is not responding. To do this, the reservations are periodically broken by other nodes in the cluster using SCSI bus reset commands. In a SAN fabric, SCSI reset commands can be very detrimental to the fabric since they are typically not implemented the same way by different vendors and they typically result in a LIP command that takes the fabric sometime to re-settle.

In Windows Server 2003, the server cluster code uses a new mechanism to first try targeted device reset, then LUN reset and if all else fails it will fall-back to a full bus reset. This feature requires the storage mini-port drivers to interpret the new control codes. At this time, several HBA vendors are modifying their mini-port drivers to provide this feature, thus enabling much more stable cluster configurations in a switched fabric environment.

This feature requires no administration to enable it. If the device driver supports the targeted reset functions they will be automatically used.

Single Storage Bus Configurations

As described previously, in Windows 2000, only storage devices on a different bus to the system disk will be considered eligible as cluster-managed devices. In a SAN environment, the goal is to centralize all storage into a single fabric accessible through a single port (actually in most cases the host will have multiple HBAs and multi-path drivers to provide a single port view to the cluster software).

In Windows Server 2003 Cluster server has a switch that when enabled, allows any disk on the system, regardless of the bus it is on, to be eligible as a cluster-managed disk. Using this, the system disk, boot disk, pagefile disks and any cluster managed disks can be attached to the same HBA. This feature is enabled by setting the following registry key:

```
HKLM\SYSTEM\CurrentControlSet\Services\ClusSvc\Parameters\ManageDisksOnSystemBuses
      0x01
```

This feature is enabled by a registry key to ensure that it is not accidentally enabled by customers that do not understand the implications of this configuration. It is intended for OEMs to ship qualified and tested configurations and not for a typical end-user or administrator to setup in an ad hoc manner.

A single storage bus configuration MUST have device drivers that support the targeted reset functionality previously defined.

Related Issues

Shared Disk Verses Shared-Nothing

You may see various documents that use terms like shared disk clusters and non-shared disk or shared-nothing clusters. These terms are very misleading and can cause confusion since they depend on the context of the discussion.

When talking about the physical connectivity of devices, shared disk clusters means that multiple computers have direct physical access to any given storage unit (for example, multiple hosts are directly connected to a disk drive on a SCSI bus that the computers are both connected to). Non-shared disk or shared-nothing clusters in this context means that any given disk is only physically connected to one computer. See Figure 17.

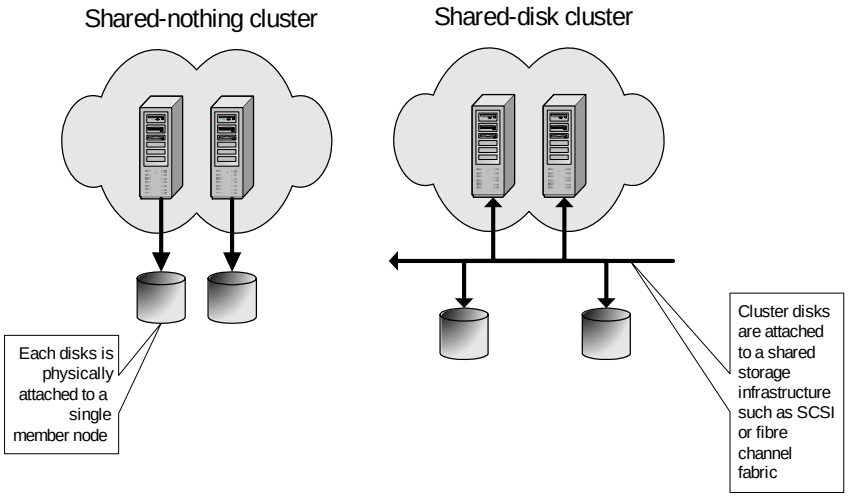

Figure 17: Physical view of cluster topologies

In the context of file systems or data access from applications, shared disk means that applications running on multiple computers in a cluster can access the same disk directly at the same time. To support this application, the file system must coordinate concurrent access to a single disk from multiple hosts (e.g. a cluster file system). Clearly, shared physical access is required for this configuration. When talking about application or data access, non-shared disk means that only applications running on one computer can access data on any given disk directly. In this case, the physical disk may or may not be connected to multiple computers, but if it is, then only the connection from one computer is in use at any one time. See Figure 18 below.

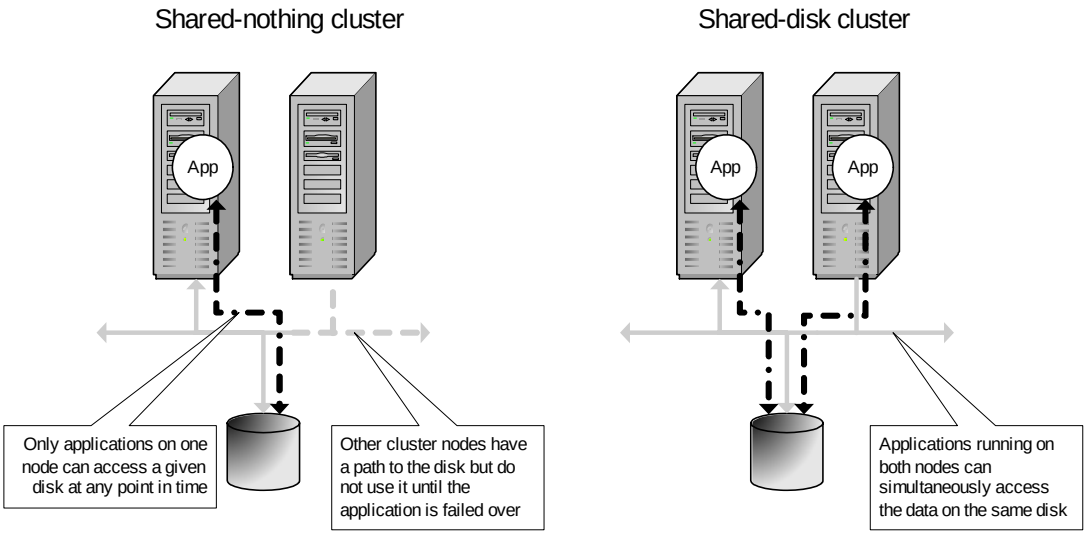

Figure 18: Application view of cluster topologies

SAN Versus NAS

There are two industry-wide terms that refer to externally attached storage:

- Storage Area Networks (SAN)
- Network Attached Storage (NAS)

Having two, similar sounding terms leads to some confusion and therefore it is worth discussing the differences between the two different technologies before delving into storage area network details.

Storage area networks (SANs), see Figure 19 below, are typically built-up using storage-specific network technologies. Fibre channel is the current technology leader in this space. Servers connect to storage and access data at the block level. In other words, to the server, a disk drive out on the storage area network is accessed using the same read and write disk block primitives as though it were a locally attached disk. Typically, data and requests are transmitted using a storage-specific protocol (usually based on the SCSI family of protocols). These protocols are tuned for low latency, high bandwidth data transfers required by storage infrastructure.

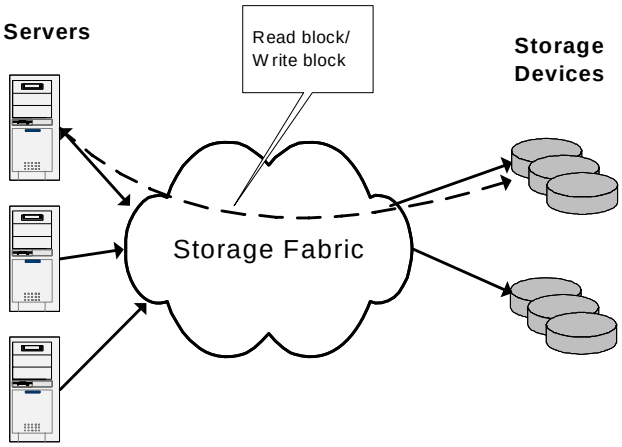

Figure 19: Storage Area Network

While fibre channel is by far the leading technology today, other SAN technologies have been proposed, for example SCSI over Infiniband, iSCSI (which is SCSI protocol running over a standard IP network). All these technologies allow a pool of devices to be accessed from a set of servers, decoupling the compute needs from the storage needs.

In contrast, network attached storage (NAS), see Figure 20 below, is built using standard network components such as Ethernet or other LAN technologies. The application servers access storage using file system functions such as open file, read file, write file, close file, etc.. These higher-level functions are encapsulated in protocols such as CIFS, NFS or AppleShare and run across standard IP-based connections.

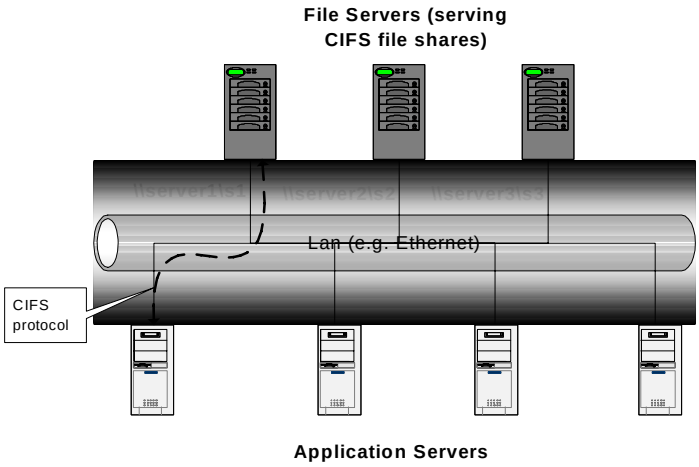

Figure 20: Network attached storage

In a NAS solution, the file servers hide the details of how data is stored on disks and present a high level file system view to application servers. In a NAS environment, the file servers provide file system management functions such as the ability to back up a file server.

As SAN technology prices decrease and the need for highly scalable and highly available storage solutions increases, vendors are turning to hybrid solutions that combine the centralized file server simplicity of NAS with the scalability and availability offered by SAN as shown in Figure 21 below.

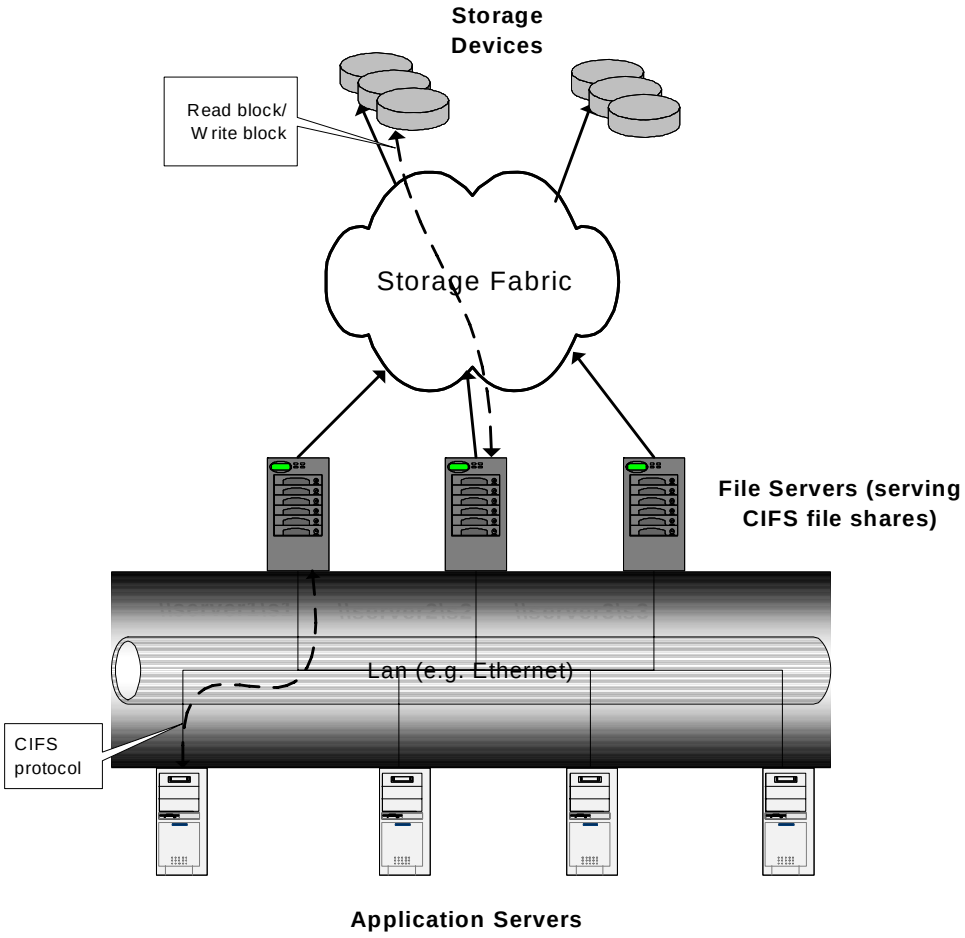

Figure 21: Hybrid NAS and SAN solution

The following table contrasts the SAN and NAS technologies

SAN Versus NAS Technologies

	Storage Area Network	Network Attached Storage
Application Server Access methods	Block-level access	File-level access
Communication protoc	SCSI over Fibre Channel iSCS (SCSI over IP)	CIFS, NFS; AppleShare

	Storage Area Network	Network Attached Storage
Network physical technology	Typically storage-specific (e.g. Fibre-channel) but may b high-speed Ethernet	General purpose LAN e.g. Gigabit Ethernet
Example Storage Vendors	Compaq StorageWorks SAN family; EMC Symmetrix	Network Appliance NetApp Filers; Maxtor NASxxxx; Compaq TaskSmart N-series

There are many camps that believe that in the future, various different technologies will win in the storage space and there are those that believe that in the end there will be a single network interconnect that will cover SAN and NAS needs, as well as basic inter-computer networking needs. Over time, the Windows platform and Windows Clustering technologies will support different interconnect technologies as they become important to end-customer deployments.

Summary

Storage area networks provide a broad range of advantages over locally connected devices. They allow computer units to be detached from storage units, thereby providing flexible deployment and re-purposing of servers and storage to suit current business needs. You do not have to be concerned about buying the right devices for a given server, or with re-cabling a datacenter to attach storage to a specific server.

Microsoft fully supports storage area networks both as part of the base Windows platform, and as part of a complete Windows Clustering, high availability solution. One or more server clusters can be deployed in a single SAN environment, along with standalone Windows servers and/or non-Windows-based platforms.

Related Links

See the following resources for further information:

- Technical Overview of Clustering Services at http://www.microsoft.com/windows.netserver/techinfo/overview/clustering.mspx
- What's New in Clustering Technologies at http://www.microsoft.com/windows.netserver/evaluation/overview/technologies/clustering.mspx
- Clustering Technologies at http://www.microsoft.com/windows2000/technologies/clustering/default.asp
- Microsoft Hardware Compatibility List at http://www.microsoft.com/hcl

For the latest information about Windows Server 2003, see the Windows Server 2003 Web site at http://www.microsoft.com/windows.netserver.

Vendors

- EMC Corporation at http://www.emc.com/
- Compaq at http://www.compaq.com/storage/bridge.html
- Brocade at http://www.brocade.com/
- McData at http://www.mcdata.com/
- Gadzoox at http://www.gadzoox.com/solutions/
- Emulex at http://www.emulex.com/
- QLogic at http://www.qlogic.com/
- Veritas at http://www.veritas.com