IBM @server xSeries Clustering Planning Guide

- Describes the xSeries and Netfinity clustering hardware
- Covers Windows, Linux and NetWare clustering
- Helps you plan a cluster implementation

David Watts
Roger Bullard
Marius Kalmantas

Redbooks

ibm.com/redbooks

SG24-5845-01

International Technical Support Organization

IBM @server xSeries Clustering Planning Guide

October 2000

Take Note!

Before using this information and the product it supports, be sure to read the general information in Appendix B, “Special notices” on page 269.

Second Edition (October 2000)

This edition applies to clustering products supported by the IBM @server xSeries and Netfinity servers.

Comments may be addressed to:
IBM Corporation, International Technical Support Organization
Dept. HZ8 Building 678
P.O. Box 12195
Research Triangle Park, NC 27709-2195

Contents

Preface

This redbook will help you to implement practical clustered systems using the IBM @server xSeries and Netfinity families of Intel-based servers. Clustering as a technology has been used for many years but has only recently started to become popular on Intel-based servers. This is due, at least in part, to the inclusion of clustering technology in Windows NT 4.0 Enterprise Edition and now Windows 2000 Advanced Server and Windows 2000 Datacenter Server operating systems. After discussing the reasons why clustering has become such an important topic, we move on to specific clustering technologies.

Clustering comes in many forms. It can appear natively within an operating system, as is the case with Microsoft Cluster Server, or it can be offered as an add-on product by either the operating system vendor itself or by third-party vendors. We examine several of these offerings that are compatible with the xSeries and Netfinity product families.

A third variation is clustering that is built into specific applications. Generally, these have been complex products, such as relational databases, and we look at two of them, Oracle and IBM's DB2. In yet another implementation, Lotus Domino Server offers a powerful form of application clustering and this is also reviewed.

By describing the characteristics and benefits of each approach, the book will help you to make informed decisions about the clustering solutions most appropriate for your applications.

In the second half of the book, we look at some of the tools and issues relating to the maintenance of your cluster to help keep it running smoothly, and at the Netfinity hardware that is ideally suited for implementing real-life clusters. The disk subsystem and interconnecting network are important elements of any cluster, so we examine the different options that are available for constructing xSeries and Netfinity-based clusters.

The final chapter provides guidance in putting all of the necessary pieces together to form a successful cluster. It includes a checklist that will help you decide whether a clustered solution is appropriate for you, and hints and tips for planning a successful implementation.

The team that wrote this redbook

This redbook was produced by a team of specialists from around the world working at the International Technical Support Organization, Raleigh Center.

David Watts is an Senior IT Specialist at the IBM ITSO Center in Raleigh. He manages residencies and produces redbooks on hardware and software topics related to xSeries and Netfinity servers. He has authored over 20 redbooks; his most recent books include *Tuning Netfinity Servers for Performance* and *Migrating from Netfinity Manager to Netfinity Director*. He has a Bachelor of Engineering degree from the University of Queensland (Australia) and has worked for IBM for over 11 years. He is an IBM Professional Server Specialist and is an IBM Certified IT Specialist.

Roger Bullard is a Lead Customer Support Specialist at the IBM Netfinity HelpCenter in Raleigh. He has three years of experience in xSeries and Netfinity servers. He holds a degree in Mechanical Engineering from the University of Maryland and has worked for IBM for more than three years. His areas of expertise include Microsoft clustering, Fibre Channel technologies and IBM Advanced Systems Management products. He is a Microsoft Certified Systems Engineer, a Certified Novell Administrator, and an IBM Professional Server Expert.

Marius Kalmantas is an IBM Personal Systems Group (PSG) Products Manager for xSeries and Netfinity in IBM Lietuva (Lithuania). He has five years of experience in xSeries and Netfinity systems. He has worked at IBM for five years. His areas of expertise include the full xSeries and Netfinity product ranges and solutions for Windows 2000, Novell, and Linux. He is an IBM Professional Server Specialist.

Figure 1. The team (l-r): David, Roger, Marius

Thanks to the following people from IBM for their invaluable contributions to the project:

Shawn Andrews, World Wide Level 2 Support, Raleigh
Paul Branch, Microsoft Technical Alliance Manager, Raleigh
Mike Brooks, Lead Customer Support Specialist, Raleigh
Dave Coffman, WW Program Manager, Microsoft Relationship, Raleigh
Raymond Cook, World Wide Level 2 Support, Raleigh
Rufus Credle, Senior IT Specialist, ITSO, Raleigh
Craig Elliot, Personal Solutions Center, Dallas
Chris McCann, World Wide Level 2 Support, Raleigh
Darryl Miles, Systems Management Network Services, Melbourne
Alex Pope, World Wide Level 2 Support, Raleigh
Torsten Rothenwaldt, Netfinity Technical Support, Frankfurt
Don Roy, Netfinity Marketing, Raleigh
Steve Russell, Senior IT Specialist, ITSO, Raleigh
Ann Smith, World Wide Level 2 Support, Raleigh
John Walczyk Jr, PartnerWorld Performance & Technology Center Manager
Jim Wilkerson, World Wide Level 2 Support, Raleigh

Comments welcome

Your comments are important to us!

We want our Redbooks to be as helpful as possible. Please send us your comments about this or other Redbooks in one of the following ways:

- Fax the evaluation form found in “IBM Redbooks review” on page 295 to the fax number shown on the form.
- Use the online evaluation form found at **ibm.com**/redbooks
- Send your comments in an Internet note to redbook@us.ibm.com

Chapter 1. Introduction

You have been there: for the third time this month, the accounting server has stopped responding and, as the person responsible for keeping it up and running, you are getting calls from management asking why the problem has not been fixed.

Sooner or later, servers used for business become critical enough that any downtime, scheduled or unscheduled, is expensive. Important maintenance work and upgrades to hardware or software may be delayed to avoid loss of service.

One solution to this type of systems management problem is clustering.

In UNIX or mainframe environments, clustering is not a new technology, but it comes at a price. For Intel-based servers, such as the IBM xSeries and Netfinity systems, clustering is a relatively recent development with the promise of high performance and availability at industry standard prices. Each platform in IBM's server family has its own unique strengths, and IBM xSeries and Netfinity servers benefit from the strategy of migrating key technologies to them from high-end systems. This strategy, and the continuing evolution of Intel-based systems, have now made clusters of these servers a feasible low-cost alternative to their more powerful cousins.

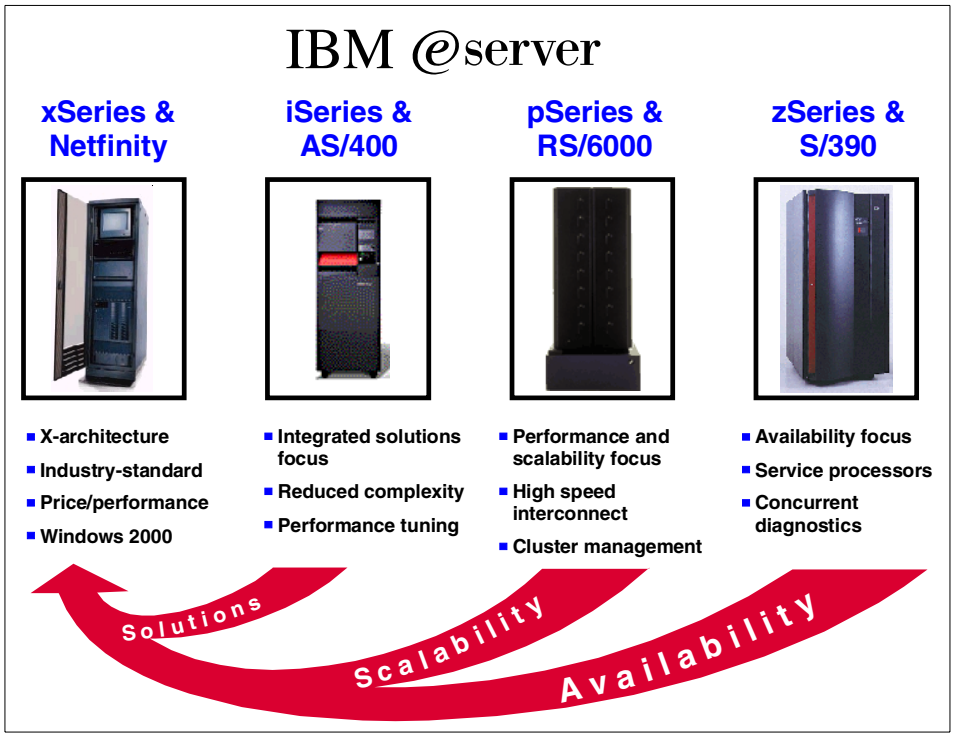

Figure 2. IBM @server family

1.1 How this book is structured

There are many facets to planning a cluster. You need to understand what clusters can do for you (and what they cannot do). Then you have to consider the different ways that a cluster can be implemented. For example, most major network operating systems offer clustering solutions and, as an alternative, certain applications offer their own clustering implementation independently of the underlying operating system. Hardware considerations also need to be taken into account before you can design a clustered solution.

Because clustering affects many areas of a system, there are several ways we could have organized this book. We chose to take a layered approach, so separate chapters cover operating systems, applications, management, hardware, and planning.

To examine different approaches to clustering, we describe the leading clustering products available for the Microsoft Windows 2000 and Novell

NetWare platforms and also several important cluster-aware applications. You will find these topics covered in Chapter 3, “Cluster at the operating system level” on page 19 and Chapter 4, “Application clustering” on page 95, respectively.

As you will discover, clusters can place stringent demands on servers and their subsystems. Information on Netfinity and xSeries servers and the key areas of disk subsystems and cluster interconnects is provided in Chapter 5, “Clustering hardware” on page 145. Some useful guidance on implementing uninterruptible power supplies (UPSs) for clustered systems is also provided in this chapter.

In Chapter 6, “Maintenance” on page 223, we discuss the aspects of maintaining your cluster and the issues associated with that.

Finally, Chapter 7, “Putting the pieces together” on page 239, summarizes the information spread throughout the rest of the redbook and highlights the most important areas you need to consider before implementing a clustered solution.

To plan a specific cluster, we expect that you will have to refer to different sections of the redbook. In this way, it is more of a reference book than a cookbook.

Chapter 2. Why clustering, why xSeries and Netfinity?

Why is there growing interest in the use of clustering techniques? And how do xSeries and Netfinity systems line up against the requirements for systems that are used in a clustered environment? These are topics we will now examine.

2.1 The promise of clustering

What exactly is a cluster?

In simple terms, a cluster is a group of computers that, together, provide a set of network resources to a client. If we take Microsoft Cluster Server (MSCS) as an example, the number of systems in a cluster is two but, in general, any number of systems could provide those resources. The key point is that the client has no knowledge of the underlying physical hardware of the cluster.

This means that the client is isolated and protected from changes to the physical hardware, which brings a number of benefits. Perhaps the most important of these benefits is high availability. Resources on clustered servers act as highly available versions of unclustered resources.

If a *node* (an individual computer) in the cluster is unavailable or too busy to respond to a request for a resource, the request is transparently passed to another node capable of processing it. Clients are therefore unaware of the exact locations of the resources they are using. For example, a client can request the use of an application without being concerned about either where the application resides or which physical server is processing the request. The user simply gains access to the application in a timely and reliable manner.

Another benefit is scalability. If you need to add users or applications to your system and want performance to be maintained at existing levels, additional systems can be incorporated into the cluster. A typical example would be a Web site that shows rapid growth in the number of demands for Web pages from browser clients. Running the site on a cluster would allow the growth in demand to be easily accommodated by adding servers to the cluster as needed.

Buying a large symmetric multiprocessing (SMP) machine and just adding central processing units (CPUs) and memory as demand increases is not a viable long-term solution for scalability. As you can see in Figure 3, an SMP machine scales very poorly when the number of CPUs increases beyond a

certain point that depends on the SMP implementation. The primary bottleneck is the bandwidth available to access the system's memory. As the CPU count increases, so does the amount of traffic on the memory bus, which eventually limits system throughput. In contrast, a well-implemented cluster can scale almost linearly.

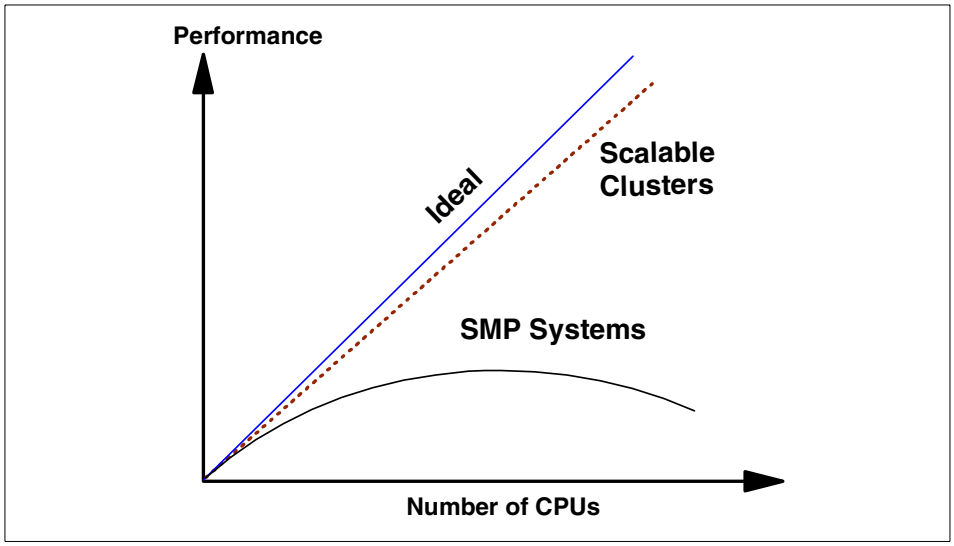

Figure 3. Scalable clusters versus SMP machines

In an ideal cluster, users would never notice node failures and administrators could add or change nodes at will. Unfortunately, this is not the case today. Current Intel-based clusters provide many of the features and functions of an idealized cluster but fall short in some areas as we will see in the coming chapters. IBM and others in the industry are working to get closer to the ideal.

2.1.1 Why consider a cluster?

There are several different reasons you might want to implement a cluster. We have already touched upon high availability and scalability and they are reiterated below along with other desirable characteristics of clusters:

- **High availability**

  If one node in a cluster fails, its workload is passed to one or more of the other servers in the cluster. Failure of a unclustered server means work comes to a halt.

- **Scalability**

  As the demands placed on a system grow, it will begin to suffer from overload. With clustering, if you outgrow a configuration, a new node can be added to the cluster with minimal or no downtime.

- **Performance**

  Using a cluster for load balancing could allow you to support a larger number of simultaneous users.

- **Price/performance**

  A clustered system can get leading-edge performance by linking inexpensive industry-standard systems together.

- **Manageability**

  Administrators can use a graphical console to move resources between the different nodes. This is used to manually balance workloads and to unload computers for planned maintenance without downtime (rolling upgrades).

- **Administrative clustering**

  For ease of administration, servers can be consolidated and clustered.

2.1.1.1 What is high availability?

Most companies are concerned about system availability or uptime. Mission-critical applications, such as e-business servers, cannot afford to suffer downtime due to unplanned outages. However, because computer systems are constructed using components that can wear out or fail, including software, system design must expect such failures and minimize their impact.

Traditionally, companies have used very reliable mainframes to host critical applications. Users and business managers have become used to this level of availability, which is not usually achieved in low-end PC systems. However, the cost/performance characteristics of Intel-based systems are compelling. Therefore, many system administrators now have a task to significantly improve the availability of their PC servers.

Before we discuss ways to increase system availability, we will try to define it. Simply stated, availability is the percentage of time that a system is running and available for access by its users. Availability is calculated only for the hours during which a system is supposed to be available. For example, if your business requires a system to be up from 6:00 a.m. to 11:00 p.m. each day, then downtime for system maintenance from 11:01 p.m. to 5:59 a.m. the next day does not count against your system availability. However, if you host an

online store that is open 24 hours a day, seven days a week, each second of downtime counts.

High availability is a relative characteristic: a highly available system will be operational for a higher percentage of the total time it is required to be available than it would be if no special system features or operational procedures were in place. As a reference, normal system availability in a mainframe environment has typically been measured at around 99.5%. For highly available systems, this improves to, perhaps, 99.99% or better. You can reach this level of availability only by eliminating or masking unplanned outages during scheduled periods of operations. To accomplish this, an advanced system design incorporating fault tolerance has to be used.

Advanced system design with fault tolerance enables a system to continue to deliver acceptable service in the event of a component failure. To achieve this, the proper configuration of system features and operational procedures have to be in place. The most common method of providing fault tolerance is to provide redundancy of critical resources, either in the same machine or elsewhere on the network, so that the backup can be made available in the event of a failing primary resource.

Some components are able to predict failures and employ preventive measures to avoid them or at least prevent these failures from affecting normal operation of the system. For instance, even in unclustered systems, hard drives using predictive failure analysis (PFA) can alert the system of an impending disk failure, allowing the disk controller to move the drive offline and to replace it with a hot spare without any manual intervention or downtime at all.

Clustering goes one step further. In clustered solutions, major components or subsystems belonging to a node may fail without users being affected. The clustering software detects the failure and makes another instance of the resource available from another system in the cluster. Users, at worst, see a brief interruption in availability of the resource. In many cases, they may be completely unaware that a problem has occurred.

When implementing a cluster solution for high availability, the classifications in Table 1 are often used. These levels are sometimes referred to by the number of nines in the Percent Available column. For example, a four 9s solution means you will only suffer a little under an hour's downtime per year; a five 9s solution reduces this to about five minutes. The more 9s you want, the more you will have to initially invest in your system. You will have to make a business judgment, balancing the cost of downtime against this investment.

Table 1. System availability classification

Percent Available	downtime/Year	Classification
99.5	3.7 days	Conventional
99.9	8.8 hours	Available
99.99	52.6 minutes	Highly Available
99.999	5.3 minutes	Fault Resilient
99.9999	32 seconds	Fault Tolerant

High availability is important for almost every industry in today's business world. Failure of a critical IT system can quickly bring business operations to a grinding halt, and every minute of downtime means lost revenue, productivity, or profit. While avoiding downtime is not a new requirement for businesses, its importance is emphasized by business strategies that are either based on or incorporate enterprise resource planning (ERP) and e-business applications.

There is a growing demand for solutions with increased availability that allow businesses to be up and running, 24 hours a day, 365 days a year without interruption. Without high availability, these businesses do not operate at their full potential. Worst case, the cost of downtime can be enough to put a company out of business. Table 2 indicates some estimated costs of downtime for different types of applications:

Table 2. Downtime costs by application

Application	Cost/Minute
Call Location	$27,000
ERP	$13,000
Supply Chain Management	$11,000
E-Commerce	$10,000
Customer Service Center	$3,700
ATM/POS/EFT	$3,500
Source: “Average cost per minute of downtime” The Standish Group - 1998	

As an example, take the price of downtime per minute for an e-commerce application from Table 2, which is $10,000 and total that against a 99.9% availability figure; 8.8 hours is 528 minutes of downtime. At $10,000 per

minute, this is a total of 5.28 million dollars, expensive for any company and potentially catastrophic.

IBM offers a 99.9% Guarantee Program. This program, which requires specific hardware configurations and specific installation and maintenance services, offers a very high level of availability of the physical and logical layers of the solution. See the following Web page for more information:

`http://www.pc.ibm.com/ww/netfinity/999guarantee.html`

2.1.1.2 Server consolidation

Server consolidation can be approached in three ways according to the Gartner Group. These are: logical consolidation, physical consolidation, and re-centralization.

- **Logical consolidation**

  Logical server consolidation normalizes the operational server environment for procedures such as backup and user maintenance. The benefits from this are a reduction in administrative staff or the local administrator's workload and, at the same time, having the lowest overall associated risk while providing a reasonable return on investment (ROI).

- **Physical consolidation**

  Servers are relocated into a centralized data center to be racked and stacked, allowing improved physical security and capacity planning across the servers and better sharing of peripherals. This in turn means reduced hardware, packaging, and cabling costs.

- **Re-centralization**

  A number of servers are collapsed into a single, more powerful and larger server. This process can be iterated to reduce the total number of servers in an organization by a significant factor.

  An obvious benefit of this is the possibility of reducing the total unused capacity of the replaced servers, but it also has a number of spinoffs:

  – Operating system consolidation
  – Reductions in the number and complexity of software licenses
  – Application instance consolidation
  – Reduction in the number of application versions supported

2.2 Types of clusters

There are several different ways you can categorize a cluster:

- Is the cluster technology software or hardware based?

- Does the cluster operate generally (as part of an operating system) or is it for a specific application?
- What kind of hardware approach to data clustering is used?

Today's Intel-based clusters utilize a number of different approaches. To help you understand the products available, this section discusses some useful ways to classify clustering technologies.

2.2.1 Software for clusters

Depending on what you are trying to accomplish and the availability of suitable products, there are different methods of implementing your cluster. From a software perspective, the primary types of clustering available are:

- At the operating system (OS) level, such as Microsoft Cluster Server
- At the application level, such as Lotus Domino clustering
- A combination of OS and application clustering

These approaches are covered in Chapter 3, "Cluster at the operating system level" on page 19 and Chapter 4, "Application clustering" on page 95.

2.2.2 Hardware for clusters

Hardware approaches to providing storage within a cluster also give us a way to classify clusters. The two most common cluster types are the *shared disk* cluster and the *shared nothing* (sometimes referred to as *partitioned*) cluster.

- **Shared disk**

  Disk storage is provided by a common disk subsystem that can be accessed by all cluster members. The clustering software manages disk accesses to prevent multiple systems from attempting to make changes to the same data simultaneously.

- **Shared nothing**

  Each cluster node has its own disk storage space. When a node in the cluster needs to access data owned by another cluster member, it must ask the owner. The owner performs the request and passes the result back to the requesting node. If a node fails, the data it owns is assigned to another node or another set of nodes in the cluster.

MSCS implements shared nothing

As described in 3.1, “Microsoft Cluster Server” on page 19, Microsoft Cluster Server implements a *shared nothing* cluster. With MSCS, even though the disks are physically connected to both systems, the disks are owned and accessible by only one system at a time.

It is confusing that the common disk subsystem used by MSCS is sometimes referred to as a shared disk subsystem. We will use the first term (*common disk subsystem*).

Symmetric multiprocessing (SMP) systems have overhead associated with managing communication between the individual CPUs in the system that eventually means that SMP machines do not scale well. In a similar way, adding nodes to a cluster produces overhead in managing resources within the cluster. Cluster management data has to be transferred between members of the cluster to maintain system integrity. Typically, cluster nodes are linked by a high-speed interconnect that carries a heartbeat signal for node failure detection and cluster-related data. However, careful design of clustering software, coupled with efficient intracluster communication, can minimize these overheads so that the linear scalability of an ideal cluster can be approached.

As already suggested, the disk subsystem and intracluster connections are two important elements of clustering. To date, these have generally been provided by extensions of mature technology. For example, a typical disk subsystem for clustering can be formed by having a common SCSI bus between two systems. Both systems are able to access disks on the common bus.

Similarly, the interconnect is typically implemented with a dedicated 100 Mbps Ethernet link. As the development of faster and more flexible systems continues, and the demand for clusters supporting more than two nodes grows, high-speed centralized disk subsystems (storage area networks or SANs) and switched interconnects will become increasingly common. You can read more about these topics in Chapter 5, “Clustering hardware” on page 145.

2.2.3 Active and passive servers

Nodes in a cluster can operate in different ways, depending on how they are set up. In an ideal two-node cluster, both servers are active concurrently. That is, you run applications on both nodes at the same time. In the event of a

node failure, the applications that were running on the failed node are transferred over to the surviving system. This does, of course, have implications on server performance, since the work of two nodes now is handled by a single machine.

A solution for this is to have one node passive during normal operation, stepping into action only when the active node fails. However, this is not a particularly cost-effective solution, since you have to buy two servers to do the work of one. Although performance in the failure mode is as good as before the failure, the price/performance ratio in normal operation is comparatively high.

We, therefore, have another way we can usefully classify clusters (particularly two-node clusters):

- **Active/active**

  This is the most common clustering model. It provides high availability and acceptable performance when only one node is online. The model also allows maximum utilization of your hardware resources.

  Each of the two nodes makes its resources available through the network to the network's clients. The capacity of each node is chosen so that its resources run at optimum performance, and so that either node can temporarily take on the added workload of the other node when failover occurs.

  All client services remain available after a failover, but performance is usually degraded.

- **Active/passive**

  Though providing maximum availability and minimum performance impact on your resources, the active/passive model requires a fully equipped node that performs no useful work during normal operation.

  The primary (active) node handles all client requests while the secondary (passive) node is idle. When the primary node fails, the secondary node restarts all resources and continues to service clients without any noticeable impact on performance (providing the nodes are themselves comparable in performance).

- **Hybrid**

  A hybrid model is a combination of the two previous models. By enabling failover only for critical applications, you can maintain high availability for those applications while having less critical, nonclustered applications conveniently running on the same server in normal operation.

In a failover situation, the less critical applications that were running on the failed server become unavailable and do not have any adverse impact on the performance of the surviving applications. You can therefore balance performance against the fault tolerance of your entire application suite.

2.3 IBM cluster strategy

There are technical challenges in implementing effective Intel CPU-based clusters. Hardware manufacturers have to develop high-speed interconnect methods, efficient storage subsystems and powerful processor complexes. Software designers need to provide clustering versions of operating systems, middleware layers (such as databases, online transaction processing (OLTP), and decision support), and applications. Importantly, this has to be achieved while conforming to industry standards and price points.

To address these challenges, IBM has developed a three-pronged clustering strategy:

1. Migration of established technologies from IBM's high-end clustering portfolio onto the Intel platform to drive the industry in the development and exploitation of the necessary technology.
2. Help establish and lead industry efforts to provide open, industry-standard cluster solutions.
3. Provide solutions to customers across major operating system and application platforms.

IBM Netfinity clusters offer key advantages to customers:

- High availability systems from cost-effective mainstream servers to high-performance enterprise class systems
- Support for Windows 2000 Advanced Server, Windows NT 4.0 Enterprise Edition, NetWare, and soon Windows 2000 Datacenter Server
- A wide choice of disk subsystems and connectivity options
- Industry-standard implementations
- Enhanced cluster system management capability
- Worldwide service and support

2.4 IBM xSeries and Netfinity clusters

IBM provides configuration aids to make assembling the necessary hardware for a cluster an easy task. A software tool, the *Netfinity Rack Configurator*,

helps you to put together rack-based systems (often used for clusters) and the detailed examples of server and cluster configurations in the *Netfinity Server Paper Configurator Guide* are an excellent starting point for your own clusters. You can find both of these tools by clicking **Configuration Tools** at:

`http://www.pc.ibm.com/us/netfinity/tech_library.html`

2.4.1 Choosing the right hardware

You have decided that clustering is the way to go to solve your availability problems. What type of servers are needed? Can you use the specifications for your current stand-alone server as a basis for the cluster?

The simple answer to this question is: yes. Neglecting the small overhead associated with a two-node cluster, an active/passive cluster will give the same level of performance as the equivalent stand-alone machine. This is not the most cost-effective solution though, for the reasons described in 2.2, “Types of clusters” on page 10.

By implementing an active/active or hybrid solution, you can boost the performance of your applications or provide additional network resources to your users without impacting performance.

Based on previous experience of running your applications in a standard server environment, and considering the number of active users you wish to accommodate on the cluster, you can estimate the type of server you are likely to need. If you are sizing an application without having this kind of experience behind you, you will have to refer to relevant application documentation for guidance.

2.4.2 IBM ServerProven program

IBM has a certification program called ServerProven that has taken the complexity out of configuring, installing, and setting up options and network operating systems. By testing the compatibility of various IBM and third-party hardware and software products with xSeries and Netfinity servers, IBM ensures that installation and operation is trouble-free.

The IBM ServerProven program gives commercial software and hardware developers the opportunity to test their business solutions on xSeries and Netfinity servers in real-world environments, reducing integration risks, and enabling smoother installations and reliable implementations.

What this means is that you do not have to worry whether IBM and these other companies' products will work together; they we been have tested to eliminate any incompatibilities.

You can find out more about IBM ServerProven at:

http://www.pc.ibm.com/us/netfinity/serverproven

More recently, we have expanded the ServerProven program to incorporate ServerProven Solutions, a commitment by IBM to work with independent software vendors and industry-leading hardware manufacturers to provide you with fully integrated solutions that meet your business needs.

2.4.3 IBM ClusterProven program

IBM's ClusterProven program offers solution developers the opportunity to optimize their clustering application offerings for operation on xSeries and Netfinity servers in a clustered environment. In this program, solution developers have their application solutions thoroughly tested and, if the solutions exhibit the characteristics required, they are registered as either ClusterProven or Advanced ClusterProven.

Application solutions meeting these requirements are allowed to display a mark to indicate that they meet the stringent requirements. These applications are also included in IBM's online Software Solutions Guide. This can be found at:

http://www.ibm.com/solutions/

The ClusterProven program encompasses all of the IBM platforms as described at:

http://www.ibm.com/servers/clusters/

For information specific to xSeries systems, please see:

http://www.pc.ibm.com/us/netfinity/clusterproven.html

2.4.4 IBM X-architecture

IBM X-architecture is a design blueprint that leverages existing, innovative IBM technologies to build the most powerful, scalable and reliable Intel processor-based servers for your business, whether you support ten or tens of thousands of users.

X-architecture is covered in detail at:

http://www.pc.ibm.com/us/eserver/xseries/xarchitecture.html

There are four principles in X-architecture:

- The OnForever initiative — designing xSeries and Netfinity servers with a goal of providing uninterrupted computing

- Bringing down the cost of enterprise computing
- Becoming a leader in establishing industry-wide collaboration
- Making servers easier to deploy and use

These are described below.

2.4.4.1 OnForever

OnForever is intended to extend the high availability feature of Active PCI and ChipKill memory to include hot-plug CPUs and RAM, and to provide online, real-time system diagnostics.

By using highly reliable parts and incorporating predictive failure analysis, the number of failures will be kept to an absolute minimum, striving to reach continuous availability, or “Double Zero”, as shown in Figure 4:

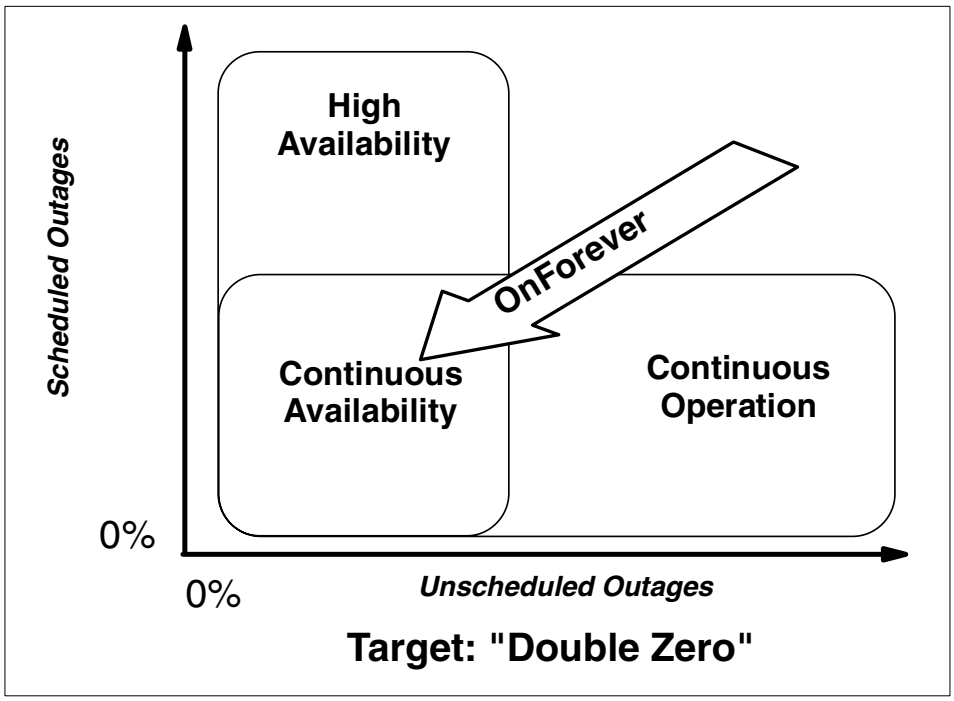

Figure 4. OnForever continuous availability

For more information about the OnForever initiative, see the white paper *IBM Netfinity OnForever Initiative* at:

`http://www.pc.ibm.com/us/techlink/wtpapers`

Table 3 shows the functions that IBM has delivered in Netfinity and xSeries systems to implement the X-architecture strategy:

Table 3. X-architecture implementation

OnForever Computing	Low-cost Enterprise Technologies	Driving Industry Standards	Making IT Easy
• Chipkill Memory • Active PCI • Common Diagnostics Model • Software Rejuvenation • Advanced System Management Processor • IBM Center for Microsoft Technologies • OPS Cluster Enabler • Predictive Failure Analysis • SAF-TE • Capacity Manager • Remote Mirroring	• Windows 2000 Datacenter • 8-node MSCS • Capacity Manager • Netfinity SP Switch • ESCON Channel Connection • Integrated Netfinity for AS/400 • FlashCopy • RAID 1E & 5E • Active Security • Logical Drive Migration	• PCI-X • InfiniBand I/O • Intel • Monterey • Linux • Microsoft Windows • Novell NetWare • SCO UnixWare • Netfinity Director • Fibre Channel	• ServerGuide • Light Path Diagnostics • Remote Connect • Update Connector • User-centered Design • LANClient Control Manager • System Migration Assistant • Software Delivery Assistant • FAStT Storage Manager • ServeRAID Manager • Tivoli Storage Manager

You can find more information about IBM Netfinity X-architecture at:

`http://www.pc.ibm.com/us/techlink/wtpapers/index.html`

Chapter 3. Cluster at the operating system level

This chapter discusses clustering technologies that operate at the level of the underlying operating system. Clustering software is either directly built into the operating system, or it is a middleware product that adds the function to a base operating system.

We will be examining:

- Microsoft Cluster Server
- Novell NetWare Cluster Services
- Linux clustering

Although these clusters often include an application programming interface to allow applications to take advantage of clustering features, an important aspect of these products is that many existing applications can also gain the benefits of clustering without any modification.

Chapter 4, “Application clustering” on page 95 looks at the alternative approach in which the clustering function is provided by specific applications.

3.1 Microsoft Cluster Server

Microsoft Cluster Server (MSCS) is part of the three enterprise-level Windows operating systems:

- Windows NT 4.0 Enterprise Edition
- Windows 2000 Advanced Server
- Windows 2000 Datacenter Server

MSCS (or Cluster Service, as its known in Windows 2000) is part of Microsoft's push into the enterprise computing arena. Providing the capability to link servers together to form a single computing resource is one way Microsoft is positioning Windows 2000 as a viable alternative to UNIX in large-scale business and technical environments.

MSCS is particularly important because it provides an industry-standard clustering platform for Windows NT and 2000 and it is tightly integrated into the base operating system. This provides the benefits of a consistent application programming interface (API) and a software development kit (SDK) that allow application vendors to create cluster-aware applications that are relatively simple to install.

The first release of MSCS was implemented in Windows NT 4.0 Enterprise Edition. This release linked two servers together to allow fault tolerance through server failover. Even before the release of MSCS, hardware manufacturers such as IBM provided redundancy for many server components, including power supplies, disks, and memory. This, however, would only protect you from component failure and not application failure.

The next release of MSCS was released with Windows 2000 Advanced Server. This introduced additional features, which include clustering services such as DHCP, WINS, SMTP, and NNTP.

Providing system redundancy means that a complete server can fail and yet client access to server resources is largely unaffected. MSCS extends this by also allowing for software failures at both operating system and application levels. If the operating system fails, all applications and services can be restarted on the other server. Failure of a single application is managed by MSCS individually. This, in effect, means that a failure can occur, but the cluster as a whole remains intact, still servicing its users' requests.

MSCS achieves this by continually monitoring services and applications. Any program that crashes or hangs can be immediately restarted on the same server or on the other server in the cluster.

If a failure does occur, the process of restarting the application on the other server is called *failover*. Failover can occur either automatically, such as when an application or a whole server crashes, or manually. By issuing a manual failover, the administrator is able to move all applications and resources onto one server and bring the first server down for maintenance. When the downed server is brought back online, applications can be transferred back to their original server either manually or automatically. Returning resources to their original server is often referred to as *failback*.

MSCS as implemented in Windows NT 4.0 Enterprise Edition and Windows 2000 Advanced Server allows only two servers, or *nodes*, to be connected together to form a cluster. Windows 2000 Datacenter Server will initially support four nodes.

The nodes are made available to client workstations through LAN connections. An additional independent network connection is used for internal housekeeping within the cluster. Both nodes have access to a common disk subsystem. Figure 5 shows the basic two-node hardware configuration to support MSCS:

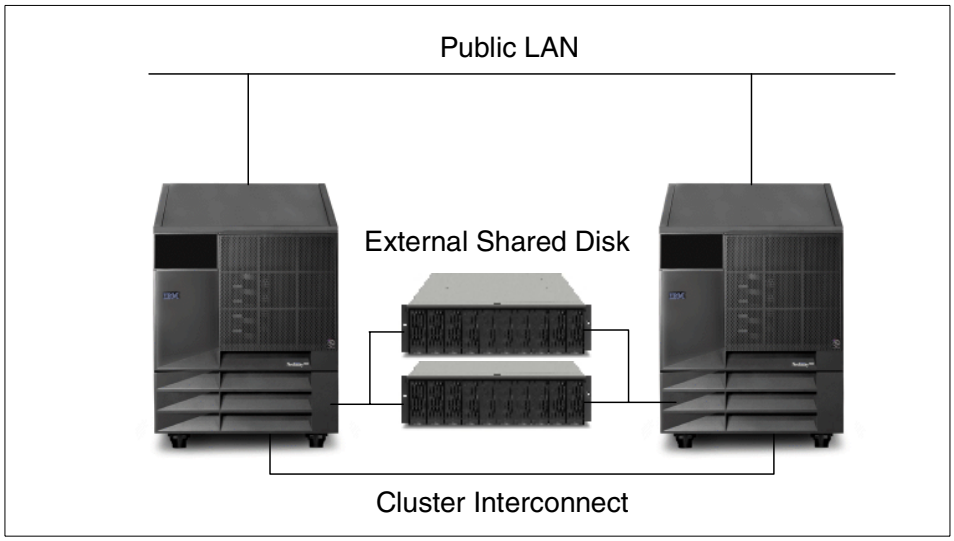

Figure 5. Basic cluster configuration

Common versus shared disk

As we saw in 2.2.2, “Hardware for clusters” on page 11, two common cluster topologies are *shared disk* and *shared nothing*. MSCS implements a shared nothing cluster. It is confusing that the common disk subsystem used by MSCS is sometimes referred to as a shared disk subsystem. We will use the first term (*common disk subsystem*).

Disks are only shared in that they are accessible by both systems at a hardware level. MSCS allocates ownership of the disks to one server or the other. In normal operation, each disk is accessed only by its owning machine. A system can access a disk belonging to the second system only after MSCS has transferred ownership of the disk to the first machine.

3.1.1 Resources

In MSCS terminology, the applications, data files, disks, IP addresses, and any other items known to the cluster are called *resources*. Cluster resources are organized into *groups*. A group can reside on either node, but only one node at any time, and it is the smallest unit that MSCS can fail over.

Resources are the applications, services, or other elements under the control of MSCS. The status of resources is supervised by a Resource Monitor.

Communication between the Resource Monitor and the resources is handled by resource dynamic link library (DLL) files. These resource DLLs, or resource modules, detect any change in state of their respective resources and notify the Resource Monitor, which, in turn, provides the information to the Cluster Service. Figure 6 shows this flow of status data between the Cluster Service and a cluster's resources:

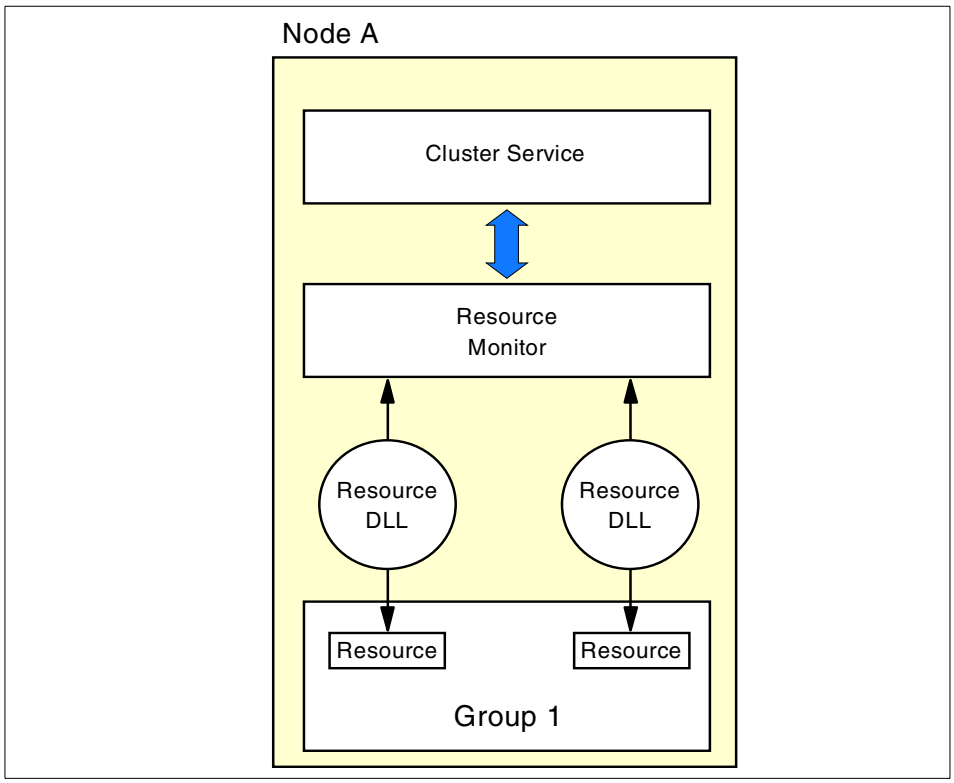

Figure 6. Communication between the Cluster Service and the resources

3.1.2 Resource Monitor

A Resource Monitor watches its assigned resources and notifies the Cluster Service if there is any change in their state. We have shown only a single Resource Monitor in Figure 6, but in fact, each node may run more than one of them. By default, the Cluster Service will start one Resource Monitor to service all resources on a node. You can choose to run a resource under its own separate Resource Monitor when the resource is defined during the cluster installation process. You would normally do this only for resources that are being debugged or if conflicts with other resources have occurred.

The Resource Monitor is separated from the Cluster Service to provide an extra level of security. The resource DLLs are running in the address space of the applications themselves. If the applications fail, the resource DLLs may malfunction, causing the Resource Monitor to fail as well. In these circumstances, the Cluster Service, however, should remain available to the cluster.

3.1.3 Dependencies

Dependencies are used within Microsoft Cluster Server to define how different resources relate to each other. Resource interdependencies control the sequence in which MSCS brings those resources online and takes them offline.

As an example, we look at a file share for Internet Information Server. A file share resource requires a physical disk drive to accommodate the data available through the share. To bind related resources together, they are placed within an MSCS group. Before the share can be made available to users, the physical disk must be available. However, physical disks and file shares initially are independent resources within the cluster and would both be brought online simultaneously. To make sure that resources in a group are brought online in the correct sequence, dependencies are assigned as part of the group definition.

Other items are required to make a fully functional file share, such as an IP address and a network name. These are included in the group, and so is an Internet Information Server (IIS) Virtual Root (see 3.1.4, “Resource types” on page 24 for more information on this resource). The group structure is shown in Figure 7:

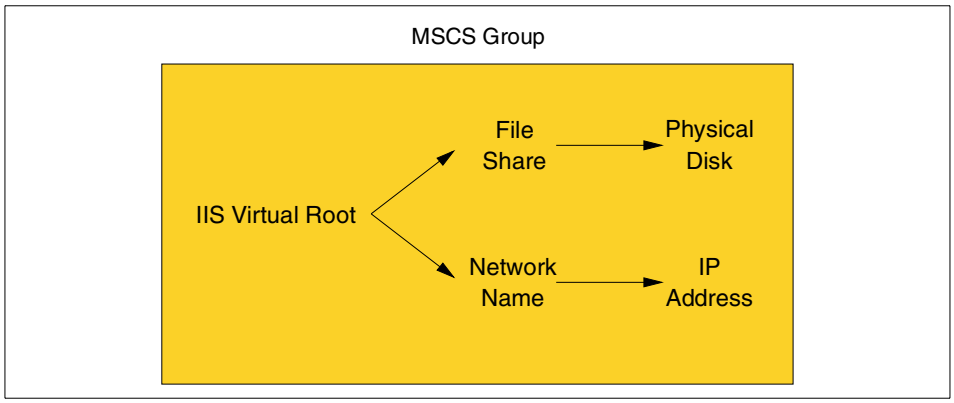

Figure 7. Resource dependencies

This diagram shows the hierarchy of dependencies within the group as a tree structure, where an arrow points from a resource to another resource upon which it depends. We see that the IIS Virtual Root is dependent on two other resources:

- A File Share resource that is itself dependent on a Physical Disk resource
- A Network Name resource that is itself dependent on an IP Address resource

When MSCS is requested to bring the IIS directory online, it now knows that it must use the following sequence of steps:

1. Bring the Physical Disk and the IP Address resources online

   then

2. When the Physical Disk becomes available bring the File Share online

   and, simultaneously

   When the IP Address becomes available bring the Network Name online

   then

3. When both the File Share and the Network Name become available, bring the IIS Virtual Root online.

As described in 3.1.6, “Resource groups” on page 27, all dependent resources must be placed together in a single group and a resource can only belong to one group.

3.1.4 Resource types

MSCS defines 12 standard resource types. Other resource types can be offered by third parties if they create suitable resource DLLs using the API in the Microsoft Platform SDK.

- DHCP Server

  This resource type provides Dynamic Host Configuration Protocol (DHCP) services from a MSCS cluster. The only parameter that is specific to the DHCP Server is the path to the database. DHCP Server resources have required dependencies on an IP Address resource and a storage class resource (typically a Physical Disk resource).

  Note: Although support for this resource type is discussed in the *MSCS Administrator's Guide*, a DHCP server resource is not supported in Windows NT 4.0 Enterprise Edition. Refer to Microsoft Knowledge Base article Q178273.

- Distributed Transaction Coordinator

  This resource type allows you to use Microsoft Distributed Transaction Coordinator (MSDTC) in MSCS. Two dependencies are required for this resource: a Physical Disk resource and a Network Name resource.

- File Share

  The File Share resource type lets you share a directory on one of the clustered disks in your configuration to give access to that directory to network clients. You will be asked to enter the name of the share, the network path, a comment, and the maximum number of users that can connect to the share at the same time.

  The configuration of a File Share resource type is identical to the configuration of a file share in Windows NT Explorer. File Shares require a Physical Disk resource and a Network Name resource.

- Generic Application

  The Generic Application resource type allows existing applications that are otherwise not cluster-aware to operate under the control of MSCS. These applications can then fail over and be restarted if a problem occurs. There are no mandatory resource dependencies.

  MSCS is often demonstrated using the Windows NT clock program. To do so, the clock.exe program is defined to the cluster as a Generic Application.

- Generic Service

  This resource type can be used for services running on Windows NT. You must enter the exact name of the service at the creation of the resource. Just as for Generic Applications, the Generic Service resource does not have any resource dependencies.

- IIS Virtual Root

  The IIS Virtual Root resource type provides failover capabilities for Microsoft Internet Information Server Version 3.0 or later. It has three resource dependencies: an IP Address resource, a Physical Disk resource, and a Network Name resource.

- IP Address

  An IP Address resource type can be used to assign a static IP address and subnet mask to the network interface selected in the *Network to Use* option during the definition of the resource. IP Address resources do not have any dependencies.

- Microsoft Message Queue Server

  This resource type supports clustered installations of Microsoft Message Queue Server (MSMQ) and is dependent on a Distributed Transaction Coordinator resource, a Physical Disk resource, and a Network Name resource.

- Network Name

  The Network Name resource type gives an identity to a group, allowing client workstations to see the group as a single server. The only dependency for a Network Name resource is an IP Address resource.

  For example, if you create a group with a Network Name resource called FORTRESS1 and you have a file share resource with the name UTIL, you can access it from a client desktop entering the path \\FORTRESS1\UTIL. This will give access to the directory on the share regardless of which cluster node actually owns the disk at the time.

- Physical Disk

  When you first install MSCS on your nodes, you are asked to select the available disks on the common subsystem. Each disk will be configured as a Physical Disk resource. If you find it necessary to add more disk after the installation, you would use the Physical Disk resource. This resource does not have any dependencies.

- Print Spooler

  The Print Spooler resource type allows you to create a directory on a common storage disk in which print jobs will be spooled. Two resources are needed to create a Print Spooler resource: a Physical Disk resource and a Network Name resource.

- Time Service

  This is a special resource type that maintains date and time consistency between the two nodes. It does not have any dependencies. The cluster must not have more than one Time Service resource.

  Note: With Windows 2000 Advanced Server and Windows 2000 Datacenter Server, the Time Service resource type is not longer required because Windows 2000 cluster nodes in the same domain automatically adjust their clocks. The resource type is provided for compatibility and migration purposes only.

Note

Do not create more than one resource of each of the following:

- DHCP Server
- Distributed Transaction Coordinator
- Microsoft Message Queue Server
- Time Service

You can create one, but no more than one, of each of these. If more than one of each resource is made, then the new resource created will fail to come online and this could possibly take the first resource offline.

3.1.5 Resource states

Resources can exist in one of five states:

1. Offline — the resource is not available for use by any other resource or client.
2. Offline Pending — this is a transitional state; the resource is being taken offline.
3. Online — the resource is available.
4. Online Pending — the resource is being brought online.
5. Failed — there is a problem with the resource that MSCS cannot resolve. You can specify the amount of time that MSCS allows for specific resources to go online or offline. If the resource cannot be brought online or offline within this time, the resource will go into the failed state.

3.1.6 Resource groups

The smallest unit that MSCS can fail over from one node to the other is a group. Related resources that have been defined from the palette of available resource types are collected together into groups. Dependencies between the resources in a group are then assigned as already described in 3.1.3, “Dependencies” on page 23.

Dependent resources

Dependent resources must be grouped together.

When one resource is listed as a dependency for another resource, then the two resources *must* be placed in the same group. If all resources are ultimately dependent on the one resource (for example, a single physical disk), then all resources must be in the same group. This means that all cluster resources would have to be on a single node, which is not ideal.

Any cluster operation on a group is performed on all resources within that group. For example, if a resource needs to be moved from node A to node B, all other resources defined in the same group will be moved. Figure 8 depicts how MSCS groups might be distributed between the nodes:

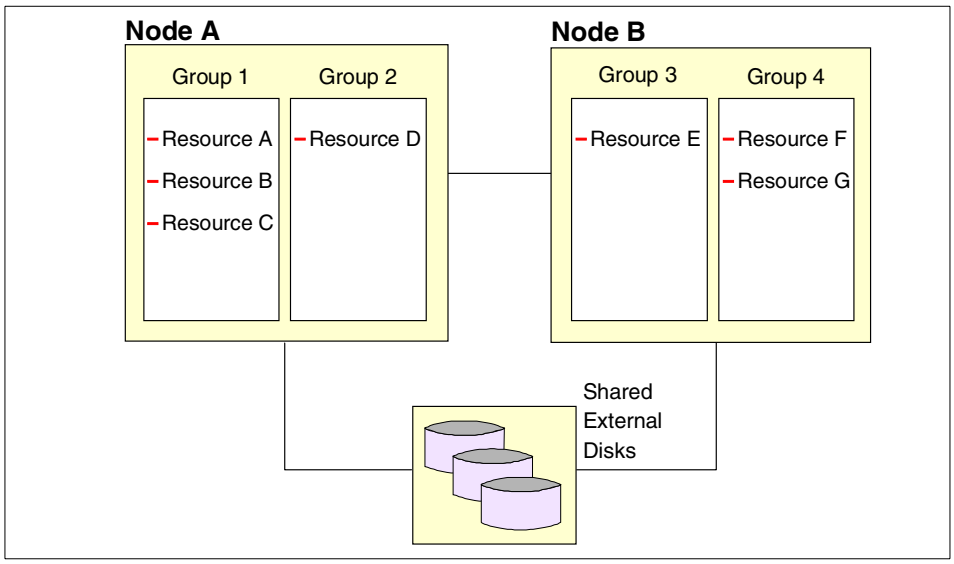

Figure 8. Example of MSCS Groups

Group states

A resource group can be in any one of the following states:

- Online — all resources in the group are online.
- Offline — all resources in the group are offline.

- Partially Online — some resources in the group are offline and some are online.

Virtual servers

Groups that contain at least an IP Address resource and a Network Name resource appear on the network as servers. They appear in Network Neighborhood on Windows clients and are indistinguishable from real servers as far as a client is concerned. These groups are, therefore, sometimes referred to as *virtual servers*.

To gain the benefits of clustering, your network clients must connect to virtual servers and not the physical node servers. For example, if you create a group with a Network Name resource called IIS_Server and then browse your network, you will see an entry (a virtual server) called IIS_Server in the same domain as the physical servers.

Browsing using physical server names

Although you can browse for the physical server names, you should not use them for connections, because this will circumvent the cluster failover functionality. You should use the virtual server name instead.

3.1.7 Quorum resource

One physical disk that is accessible from both nodes is used to store information about the cluster and is called the *quorum disk*. This resource maintains logged data that is essential to maintain cluster integrity and to keep both nodes in synchronization, particularly when the nodes fail to communicate with each other.

The quorum disk can be owned by only one node at a time and is used to determine which node will take ownership of cluster resources in certain situations. For example, if the nodes lose contact with each other, the node that cannot contact the quorum disk will withdraw from the cluster while the other node assumes ownership of all cluster resources.

You specify an initial quorum resource when installing the first MSCS node. It must be located on a drive in the common disk subsystem. Therefore, the physical drive used when defining the quorum logical drive must reside in a storage expansion enclosure with the other drives.

The drive containing the quorum resource may also contain other applications and data, but this is not recommended.

Note: In the case of IBM's ServeRAID adapters, the quorum resource is temporarily defined on a local disk during installation. By completion, however, the quorum resource has been migrated to one of the common subsystem drives.

3.1.8 TCP/IP

MSCS uses TCP/IP to communicate with network applications and resources. Cluster IP addresses cannot be assigned from a Dynamic Host Configuration Protocol (DHCP) server. These include IP Address resources, the cluster administration address (registered at the installation of MSCS), and the addresses used by the nodes themselves for intracluster communication.

Note that each node will usually have at least two network adapter cards installed. In Windows 2000 clusters, two network connections are *required*, and in Windows NT 4.0 Enterprise Edition clusters, two network connections are *recommended*.

Although a single network connection can be used with Windows NT 4.0 Enterprise Edition, Microsoft recommends using a private network for cluster traffic. One adapter is used to allow communication over the external network for administration and management of the cluster and for user access to cluster resources.

The physical server IP addresses assigned to these adapters could be obtained through DHCP, but it is important that users attach to clustered addresses. We recommend the use of static IP addresses for all adapters in your cluster; otherwise, if a DHCP leased address expires and cannot be renewed, the ability to access the cluster may be compromised (see Microsoft Knowledge Base article Q170771).

The second adapter in each machine is for intracluster communication, and will typically have one of the TCP/IP addresses that conform to those reserved for private intranets. Table 4 shows the allocated ranges for private IP addresses:

Table 4. Private IP address ranges

IP address range	Description
10.0.0.0 through 10.255.255.255	A single Class A network
172.16.0.0 through 172.31.255.255	16 contiguous Class B Networks
192.168.0.0 through 192.168.255.255	256 contiguous Class C Networks

For more information refer to *TCP/IP Tutorial and Technical Overview*, GG24-3376, and Chapter 3 of *Microsoft Cluster Server Administration Guide*.

Use of NetBEUI and IPX

You must have TCP/IP installed on both servers in order to use MSCS. Applications that use only NetBEUI or IPX will not work with the failover ability of MSCS. However, NetBIOS over TCP/IP will work.

3.1.9 Additional comments about networking with MSCS

- **Clusterable applications**

  Not all Microsoft software require shared hardware clustering for high availability:

  – Microsoft SNA Server uses SNA connection pooling
  – Proxy Server can cluster at the application level
  – WINS can use replication partners to achieve maximum availability

  Some functions simply cannot work in a clustering environment. With Windows NT 4.0 Enterprise Edition, DHCP is an example of this (DHCP can be clustered with Windows 2000 Advanced Server).

- **Implementing WINS in a cluster**

  Windows NT 4.0 Enterprise Edition requires a WINS server that is accessible to both clients and the cluster servers. This server should *not* be a member of the cluster, but a third machine somewhere on the domain. This requirement may not seem necessary if the clients and the cluster are on the same segment, but will become readily apparent when the clients and the cluster are separated on different segments. WINS servers will also help prevent clients from accidentally obtaining the wrong IP address for any clustered resource.

  Windows 2000 Advanced Server is able to fail over WINS, so a third server on the network is not necessary. It is not recommended that you use WINS because failing over WINS can create problems with the tombstone record. If you still have a Windows NT 4.0 domain controller, then you should still use WINS on a separate server on the domain. If you are in a Windows 2000 environment, WINS is not necessary because DNS takes over the function of a WINS server.

- **The Computer Browser and NetBIOS**

  The Computer Browser service uses NetBIOS to distribute and maintain the browse list. However, NetBIOS cannot be failed over in a cluster and therefore should not be enabled.

  Clients cannot browse or search for cluster virtual servers in a non-NetBIOS environment.

  To configure a non-NetBIOS environment, the client must have the WINS client property set or inherited from the DHCP server. You can determine this by making sure the setting on each network adapter is NetBIOS-less or inherited from the DHCP server, by viewing the settings in Network and Dial-up Connections. In the TCP/IP properties for each network adapter, click **Advanced** then go to the WINS tab. This shows three options:

  – Enable NetBIOS over TCP/IP
  – Disable NetBIOS over TCP/IP
  – Use NetBIOS setting from the DHCP server

  Make sure that **Use NetBIOS setting from the DHCP server** checkbox is selected.

  For information about the Computer Browser service in Windows 2000, see Knowledge Base article Q188001.

  For information about heartbeat configurations in Windows 2000 and Windows NT 4.0 Enterprise Edition, see Knowledge Base article Q258750.

- **Using multiple network cards**

  Clustered servers can be connected to multiple subnets through the use of multiple network cards. The normal rules of Microsoft networking apply, however, and only one default gateway can exist for all physical adapters. The use of multiple adapters, however, only provides for bandwidth distribution and not failover.

- **Failed network connections**

  MSCS in Windows NT 4.0 Enterprise Edition does not monitor for a link failure, and cannot route clients from one subnet to another. MSCS monitors only adapter failures. For instance, Ethernet cards do inform the operating system they can no longer have link status, so MSCS has no way to truly understand that the network has failed.

  Consequently, it will *not* move resources to the node that has a correctly functioning network. IP addresses cannot be used to detect this failure either, since they only require a bound network driver to function.

In Windows 2000 clustering, as described in Knowledge Base article, Q242600 “Network Failure Detection and Recovery in a Two-Node Server Cluster”, link failures can be detected in some situations.

The next bullet discusses some alternatives to help you avoid failed network connections.

- **Redundant network adapters**

  IBM has made available a wide variety of adapter failover solutions that provide redundant link access to the network. For example, the Adaptec ANA-6911 and ANA-6944 with Dual-Link Option are on the IBM ServerProven list.

  For even greater flexibility, the IBM Netfinity 10/100 Fault Tolerant Adapter supports multiple fault-tolerant pairs and adapter load balancing. IBM also offers fault-tolerant token-ring and Gigabit Ethernet adapters. The combination of Microsoft MSCS node protection and adapter link protection can significantly increase availability of cluster resources, and is recommended in any mission-critical environment.

  Note: MSCS cannot use a second network card as a hot backup for the client access. This means that the card may be a critical failure point.

- **Use of Kerberos**

  Clients cannot use Kerberos to authenticate a connection to a cluster virtual server. Since Kerberos cannot be used, the clients will attempt to authenticate with NTLM authentication.

- **Administering the cluster**

  Administrators of the cluster cannot use the browse functionality of Cluster Administrator to find the cluster name on a network adapter that has NetBIOS over TCP/IP disabled, even if the two computers are on the same subnet.

- **Multihomed environments**

  Cluster problems may arise due to the priority of network adapters in a multihomed Windows NT 4.0 Enterprise Edition. Since each cluster node has adapters connected to at least two different networks (the cluster’s private link and the public LAN), each cluster node is also a multihomed host. On such systems, the question, “What is my IP address?” is answered by a list of IP addresses assigned to all network cards installed in the machine. The Windows Sockets API call `gethostbyname()` is used to obtain these addresses.

  Some cluster applications (for example, Oracle FailSafe and SAP R/3) are sensitive about the IP address order in the list returned by

gethostbyname(). They require that the IP address of the adapter to which their cluster virtual address will be bound appears on top. This means that the address of the adapter connected to the public LAN must be listed before the address of the cluster's private link adapter. If not, it may be impossible to connect to the application using the virtual address after a failover.

To avoid such problems, you should check the order of assigned IP addresses in the address list before you install MSCS (to ensure that the network assignments are right from the beginning). The two simplest ways to do this are:

– Ping each node from itself. For example, on Node_A you type in a command window:

  ping node_a

  The address from which the ping is answered must be the address assigned to the adapter card in the public LAN.

– The ipconfig command shows all addresses in the order of the gethostbyname() list.

If the result you get from either of the above methods is in the wrong order, correct it before installing MSCS. When the network adapters are of the same type, then it is sufficient to simply exchange their outgoing cable connections. If you have different adapter types (for example, 10/100 EtherJet for the cluster-private link and redundant FDDI for the public LAN), then you need a way to control the internal IP address order.

Assuming that your two network adapter cards have the driver names Netcard1 and Netcard2 (the IBM Netfinity 10/100 EtherJet Adapter has the driver name IBMFE), PING and IPCONFIG show you Netcard1 first, but your public LAN is on Netcard2. To change the IP address order so that Netcard2's address is listed first, you must edit the Windows Registry as follows (remember that this can have serious consequences if not done carefully):

a. Start REGEDT32 (REGEDIT will not work since it cannot add complex value types) and select the following subkey:

   HKEY_LOCAL_MACHINE\SYSTEM\CurrentControlSet\Services\Netcard1

b. Add a new value with the following specifications:

   Value Name:DependOnService
   Value Type:REG_MULTI_SZ
   Data:Netcard2

c. Exit the Registry Editor and reboot the machine. Entering `ping` and `ipconfig` should now show you the addresses in the required order.

DependOnService may reset to the default

Note that the new registry value, `DependOnService`, will be deleted whenever Windows rebuilds the networks bindings. Thus after each modification of network parameters you should verify that the order is still correct. If you change the IP settings frequently, you will save time by exporting the value to a .REG file for convenient registry merging.

3.1.10 Domains

The following information specifies the criteria for clustered MSCS servers in regard to domains:

- The two servers must be members of the same domain.
- A server can only be a member of one cluster.
- The following are the only valid domain relationships between cluster nodes:
  - Two domain controllers (Windows 2000)
  - A PDC and a BDC (Windows NT)
  - Two BDCs (Windows NT)
  - Two stand-alone servers

In general, we recommend that the nodes are set up as stand-alone servers. This will remove the additional workload generated by the authentication tasks and the domain master browser role performed by domain controllers. However, there are situations, such as when domain size is small, when it may be appropriate for nodes also to be domain controllers.

If you change a member server to a domain controller or vice versa after you install the Cluster service, the service may not start and you receive the following error messages in the system event log:

Event ID: 7013
Source: Service Control Manager

Description: `Logon attempt with current password failed with the following error: Logon failure: the user has not been granted the requested logon type at this computer.`

Event ID: 7000
Source: Service Control Manager

Description: `The Cluster service failed to start due to the following error: The service did not start due to a logon failure.`

This problem can occur if the account used to install the Cluster service does not have explicit rights that are needed to run the Cluster service.

Refer to Knowledge Base article Q247720 for recovery procedures in Windows 2000 clustering.

3.1.11 Failover

Failover is the relocation of resources from a failed node to the surviving node. The Resource Monitor assigned to a resource is responsible for detecting its failure. When a resource failure occurs, the Resource Monitor notifies the Cluster Service, which then triggers the actions defined in the failover policy for that resource. Although individual resource failures are detected, remember that only whole groups can fail over.

By default, MSCS configures all the resources in the cluster to share a single Resource Monitor. This means that any time any resource has failed, the single instance of the Resource Monitor must check *all* resources and take appropriate actions. For obvious reasons, this can cause resources to take extended amounts of time to come online, since they are waiting for the single resource monitor to service it.

To avoid this problem, IBM recommends setting all IP Address resources and Physical Disk (ServeRAID Logical Disk) resources to use their own resource monitors. This will maximize the availability of the core MSCS resources, and eliminate 99% of any lag in bringing resources online.

Note: Each Resource Monitor is its own application consuming both memory and CPU, so it is *not* recommended that you put every resource into its own monitor.

Failovers occur in three different circumstances: manually (that is, at the request of an administrator), automatically, or at a specific time as set by IBM Cluster Manager.

Automatic failovers have three phases:

1. Failure detection
2. Resource relocation
3. Application restart (usually the longest part of the failover process)

An automatic failover is triggered when the group failover threshold is reached within the group failover period. These are configuration settings, defined by the administrator.

Group and resource failover properties

Both groups and resources have failover threshold and period properties associated with them. The functions these properties control, however, depend on whether they are associated with a group or a resource.

- Resource failover settings

  Failover threshold is the number of times in the specified period that MSCS allows the resource to be restarted on the *same* node. If the threshold count is exceeded, the resource and all other resources in that group will fail over to the other node in the cluster.

  Failover period is the time (in seconds) during which the specified number of attempts to restart the resource must occur before the group fails over.

  After exceeding the threshold count of restart attempts, MSCS fails over the group that contains the failing resource and every resource in that group will be brought online according to the startup sequence defined by the dependencies.

- Group failover settings

  Failover threshold is the maximum number of times that the group is allowed to fail over within the specified period. If the group exceeds this number of failovers in the period, MSCS will leave it offline or partially online, depending on the state of the resources in the group.

  Failover period is the length of time (in hours) during which the group will be allowed to fail over only the number of times specified in Threshold.

For example, consider the application Domino in group Lotus Domino. Other resources in the group include a File Share resource and a Physical Disk resource, as shown in Figure 9.

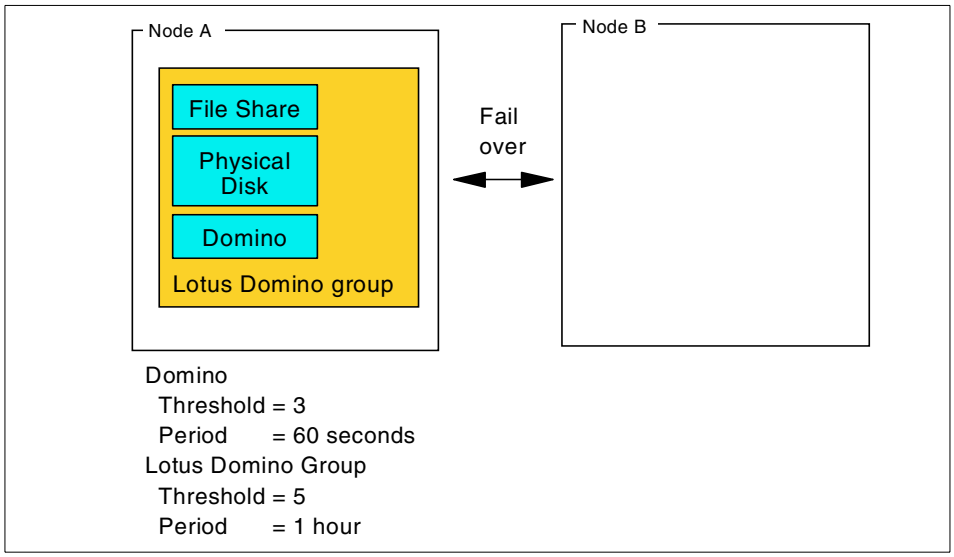

Figure 9. Failover example

The administrator that set up this cluster has assigned a failover threshold of 3 with a failover period of 60 seconds to the Domino resource and a failover threshold of 5 with a failover period of 1 hour to the Lotus Domino Group.

Consider now the situation when Domino continually fails. The program (a Generic Application resource type) will be restarted on Node A three times. On the fourth failure within one minute, it and its group, Lotus Domino Group, will fail over to Node B. This counts as one Lotus Domino Group failover. When Domino fails four times (that is, one more than the resource threshold) on Node B, it will fail over to node A. This counts as the second Lotus Domino failover.

After the fifth Lotus Domino Group failover within one hour (Node A→B→A→B→A→B), MSCS will not attempt a restart of Domino, nor will it fail over Lotus Domino Group. Instead, it will leave Domino in the failed state and Lotus Domino Group will be placed in the partially online state. The other resources in the group will be placed in the failed state if they are dependent on Domino; they will remain online if they are not dependent on Domino.

With Windows 2000 Datacenter Server when there are three or four nodes in the cluster, the failed resource group is failed over to a surviving node in the preferred owners list, in the order in which they are listed. If there are no surviving preferred owners, then the group is failed over to one of the nodes in the possible owner list (which may or may not be all the nodes). It is

therefore important to carefully select the preferred and possible owners in a four-node cluster.

3.1.12 Failback

Failback is a special case of failover and is the process of moving back some or all groups to their *preferred owner(s)* after a failover has occurred.

A group's preferred owner is the node in the cluster that you have declared as the one upon which you prefer the group of resources to run. With Windows 2000 Datacenter Server clusters with three or four nodes, there could be multiple preferred owners.

If the preferred owner fails, all of its clustered resources will be transferred to the surviving node. When the failed node comes back online, groups that have the restored node as their preferred owner will automatically transfer back to it. Groups that have no preferred owner defined will remain where they are.

You can use the preferred owner settings to set up a simple load-balancing configuration. When both servers are running with failback enabled, the applications and resources will move to their preferred owner, thereby balancing out the workload on the cluster according to your specifications.

When you create a group, its default failback policy is set to disabled. In other words, when a failover occurs, the resources will be transferred to the other node and will remain there, regardless of whether the preferred node is online. If you want failback to occur automatically, you have the choice of setting the group to fail back as soon as its preferred node becomes available, or you can set limits so the failback occurs during a specific period, such as outside of business hours.

Automatic failback not recommended

IBM does not recommend setting failback to automatic, as some cluster applications can take a considerable amount of time to become available (some large Exchange clusters can take up to 5 hours to come back online!), which can significantly impact high availability of the clustered resource data.

If administrators do not want to manually move resources at off hours when downtime is scheduled, they can use the IBM Cluster Manager to schedule a move when users can tolerate the potential loss of resources.

3.1.13 LooksAlive and IsAlive

To determine if resources are available, the Resource Monitor *polls* (requests status information from) the resource DLLs for which it is responsible. Two levels of polling are supported by the Resource Monitor and you can adjust how frequently each of them occurs for each resource. The two levels are called *LooksAlive* and *IsAlive* polling. They are defined as follows:

- LooksAlive polling

  The Resource Monitor makes a superficial check to determine if the resource is available.

  If a resource fails to respond to a LooksAlive poll, then the Resource Monitor will notify the Cluster Service. When you create a new resource, you define the interval (in milliseconds) between LooksAlive polling requests. The default interval is 5,000 milliseconds (5 seconds).

- IsAlive polling

  The Resource Monitor performs a complete check of the resource to verify that it is fully operational. If a failure is returned, the Cluster Service is immediately notified and, depending on the configuration defined for the resource, the Resource Manager will either terminate the resource or try to bring it back online on the same node or on the other node (as part of a group failover). The default interval is 60,000 milliseconds (1 minute).

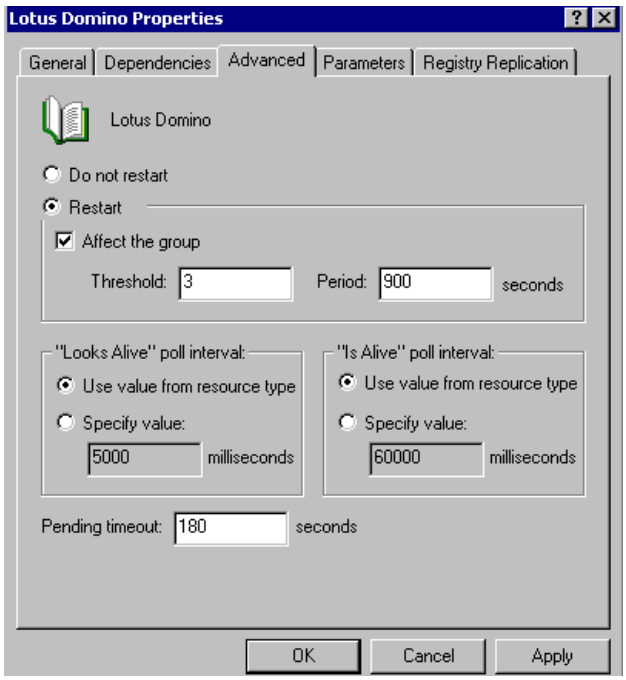

Figure 10. LooksAlive / IsAlive poll interval example

Consider again the Domino example in Figure 9 on page 38. Assuming that the Generic Application resource is created with the default parameters, the Resource Monitor calls the LooksAlive function in the resource DLL every five seconds to make a cursory check that Domino is functioning. Every 60 seconds, the IsAlive function is called to perform a more rigorous test to check that Domino is operating correctly.

.

Superficial versus complete checks

Exactly what constitutes a superficial check or a complete check in the descriptions above is determined by the programmer who wrote the resource DLL. For generic applications such as the Domino program, the two tests may be identical. More sophisticated resources such as database elements will usually implement a different test for each entry point.

3.2 Windows 2000 Advanced Server

With Windows 2000 Advanced Server, Microsoft incorporates two clustering technologies that can be used independently or they can coexist to provide organizations with a complete set of clustered solutions that can be used based on the requirements of a given application or service. The two clustering technologies are:

- **Cluster Service**

  This service, which offers a similar function to that of MSCS in Windows NT 4.0 Enterprise Edition, is intended primarily to provide failover support for applications such as databases, messaging systems, and file and print services. Windows 2000 Advanced Server supports two-node failover clusters only. Cluster Service will help ensure the availability of critical line-of-business and other back-end systems.

  Name change

  In Windows 2000 Advanced Server, MSCS is built into the operating system and is now referred to as Cluster Service.

  Cluster Service ensures that critical business applications are online when needed, by removing the physical server as a single point-of-failure. In the event of a hardware or software failure in either node, the applications currently running on that node are migrated by Cluster Service to the surviving node and restarted. No data is lost during failover because Cluster Service uses a shared-disk configuration such as ServeRAID or Fibre Channel. Another benefit of Cluster Services is that it can significantly reduce the amount of unplanned application downtime caused by unexpected failures.

  The Setup wizard has been greatly improved. The setup requires fewer entries and less time to install and configure than Windows NT 4.0 Enterprise Edition. The Cluster Administrator is now a Microsoft Management Console (MMC) snap-in.

  Refer to the following Web site for more information about the improvements and features:

  `http://www.microsoft.com/windows2000/guide/server/features/`

The following scenarios can be implemented using Cluster Service technology:

  – File and print servers
  – Database and messaging
  – E-Commerce sites

- **Network Load Balancing (NLB)**

  This service load balances incoming TCP/IP traffic across clusters of up to 32 nodes. The practical limit to the number of nodes varies depending on your implementation. The availability and scalability of Internet server-based programs such as Web servers, streaming media servers, and Windows Terminal Services are enhanced by Network Load Balancing.

  NLB acts as the load balancing infrastructure and provides controlled information to the management applications built on top of Windows Management Instrumentation (WMI). This allows NLB to integrate seamlessly into an existing Web server farm infrastructure.

  NLB provides an integrated infrastructure for building critical and in-demand Web sites in a distributed, load-balanced manner. When distributed application features of Component Services and the enhanced scalability of Microsoft Internet Information Services (IIS) are combined, NLB ensures that Web services can handle heavy traffic loads, while it also protects against planned and unplanned server downtime.

  NLB uses a statistical load-balancing model to distribute incoming IP request across a cluster of up to 32 servers. You can add capacity to Web-based applications because NLB is integrated into Windows 2000 networking infrastructure.

  NLB can fail over clustered Web servers in less than 10 seconds. This means that routine maintenance or planned upgrades, and even unplanned server downtime will cause little interruption of Web services.

  The following scenarios can be implemented by Network Load Balancing technology:

  – Web server farm
  – Terminal services
  – e-commerce sites

Figure 11 shows an example of both technologies coexisting with each other. It shows a client making a request to a Web page. When the client connects to the Web page, he is unaware he is connecting to the NLB Web farm.

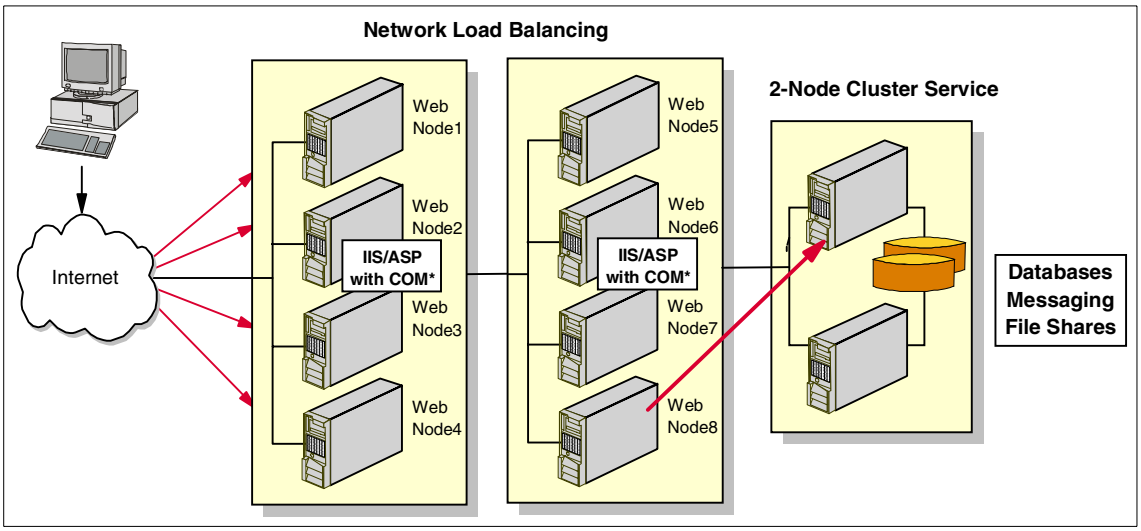

Figure 11. Example of both technologies being deployed

NLB distributes IP traffic to multiple copies (or *instances*) of a TCP/IP service such as the Web server in Figure 11. Each instance is running on a host within the cluster. NLB transparently partitions the client requests among the hosts and lets the client access the cluster using one or more virtual IP addresses. Again, from the client's point of view, the cluster appears to be a single server that answers the client's request.

Each server is running Microsoft Internet Information Services 5.0 and NLB distributes the networking workload among them. This means that the eight-host cluster works as a single virtual server to handle network traffic.

This speeds up the normal processing so that the Internet client can see a faster turnaround on its requests. The two-node cluster running Cluster Service houses a back-end application database that fulfills any request's from the clients. The two-node cluster is an option administrators can add when a back-end database is needed in an NLB clustered environment.

3.2.1 Using Active Directory in a clustered environment

Virtual servers defined in a server cluster running the Cluster service as Network Name resources are not integrated into the Active Directory structure and therefore cannot take advantage of many of its new features and functionality.

When you integrate the Cluster service with the directory service, the nodes that are Windows 2000-based computers that belong to a cluster and are members of a Windows 2000 domain are integrated in the directory as computer objects noted by their computer name. The server cluster name and all defined virtual servers in the cluster are not represented by computer objects in Active Directory. This means that the network names that are seen by clients as computers on the network cannot be found by using the Active Directory Find Computers option in Active Directory Users and Computers.

Note: The option to search for a computer on the network in My Network Places uses more methods than Active Directory to find computers and can find cluster virtual servers.

This is discussed in Microsoft Articles Q235529 and Q247720.

For information about Active Directory, review the book *Creating Active Directory Infrastructures*, by Curt Simmons. Chapter 14 “Understanding Active Directory Replication” can be viewed online via Microsoft TechNet at:

`http://www.microsoft.com/TechNet/win2000/understa.asp`

3.3 Windows 2000 Datacenter Server

Windows 2000 Datacenter Server adds several important technical and support features to Windows 2000 Advanced Server. It follows the same basic model, but supports four-node clustering, and up to 64 GB of memory and 32 CPUs. It also includes Winsock Direct, a new high-speed communications technology.

The biggest changes in Datacenter Server relate to configuration management and support of the product. Datacenter Server cannot be purchased directly from Microsoft; it must be purchased from a Microsoft Certified Datacenter Partner such as IBM.

The IBM solution is to provide a controlled environment for Datacenter installations. IBM’s initial offering will be to integrate the Netfinity 8500R with solutions from IBM Global Services and will include services, warranties and support for Netfinity, which are derived from IBM’s enterprise servers.

IBM has developed a set of standard services around the Datacenter operating system that will raise its reliability, availability and serviceability to a level consistent with traditional enterprise computing architectures. This includes factory integration, on-site installation, solutions assurance review, 24x7 technical support, and a single point of contact for the hardware and operating system.

The complete IBM offering, including systems, options, drivers, and applications, must successfully complete a rigorous testing process. Users cannot add any piece of software that touches the kernel (for example virus scanners and open file agents) without first having IBM test and certify it. The benefit of this controlled environment will be maximum system uptime for the customer.

The Windows Datacenter Program is designed to support longer system life cycles through certified hardware and software configurations. IBM and Microsoft are committed to supporting a change management process that ensures customers the highest level of service. Updates made to the hardware or operating system will be handled through a formal process to minimize the disruption of service and customer interaction.

For more information about Windows 2000 Datacenter Server, see

`http://www.microsoft.com/windows2000/guide/datacenter/overview/default.asp`

3.4 Netfinity Availability Extensions for MSCS

Formerly codenamed *Cornhusker*, IBM Netfinity Availability Extensions for MSCS (NAE) is a clustering technology developed by IBM. It extends the capabilities of MSCS from its present two-node solution to an six-node solution on the IBM Netfinity platform. NAE's origins are based on the IBM High Availability Cluster Multi-Processing (HACMP) solution for the AIX platform.

This extension is a result of the strategic alliance of IBM with Microsoft for API selection and validation for industry users of Windows clustering products. Enterprises with established MSCS clusters can plan to expand their clustered systems.

Providing a virtual multinode cluster that manages a collection of MSCS clusters as if they were one, Netfinity Availability Extensions for MSCS enables failover between clusters. It provides MSCS-like failover for up to 6 nodes using file and print, Lotus Domino and DB/2 failover resources.

This multi-node solution is currently used only with Windows NT 4.0 Enterprise Edition and NAE Service Pack 3. For a multi-node (more than two nodes) Windows 2000 solution, IBM is offering Windows 2000 Datacenter Server.

For a more detailed explanation of IBM Netfinity Availability Extensions for MSCS, refer to *IBM Netfinity Availability Extensions for Microsoft Cluster Server* at

`http://www.pc.ibm.com/ww/netfinity/clustering/mscs.html`

3.5 Cluster Management

IBM offers two management products for Microsoft clusters, in addition to the Microsoft Cluster Administrator that is part of Cluster Service:

- IBM Cluster Systems Management
- Netfinity Director

3.5.1 IBM Cluster Systems Management

IBM Cluster Systems Management (ICSM), is an administration tool for MSCS that is integrated into IBM Netfinity Manager. It can also integrate smoothly with Intel LANDesk and Microsoft SMS for a clear view of clustered resource components.

IBM Netfinity and xSeries servers running MSCS will provide additional features that promote ease of use and increased productivity, as well as event and problem notification for a clustered server configuration—all from a single console. ICSM is supported only on Windows NT 4.0.

IBM Cluster Systems Management allows a systems administrator to:

- Discover and display individual clusters and, using a GUI, set up and manage those clusters
- Schedule manual load balancing of MSCS resources
- Set up and manage multiple clusters from one GUI

Netfinity Availability Extensions for MSCS requires the use of ICSM. You should not use Microsoft Cluster Administrator with NAE because it will eventually corrupt your shared clustered disks.

Netfinity Director, the replacement for Netfinity Manager, also includes ICSM as the task Cluster Systems Management as described in 3.5.2.2, “Cluster Systems Management” on page 49.

3.5.2 Netfinity Director

Netfinity Director is the next generation systems manageability software from IBM, a powerful, highly-integrated, systems management solution built upon industry standards and designed for ease of use.

To enable Netfinity Director's cluster support you must meet certain prerequisites:

- Cluster nodes must be running Windows NT 4.0 with Service Pack 3 or higher or Windows 2000 Advanced Server, and have MSCS installed.
- You must also install the UM Server Extensions plug-in for Netfinity Director. You can download this plug-in from:

  `http://www.pc.ibm.com/ww/netfinity/systems_management/nfdir/serverext.html`

Figure 12 shows the Netfinity Director Console after the cluster has been discovered. Selecting **Clusters** in the Groups panel, then right-clicking the specific cluster in the Group Contents panel displays a pop-up menu as shown.

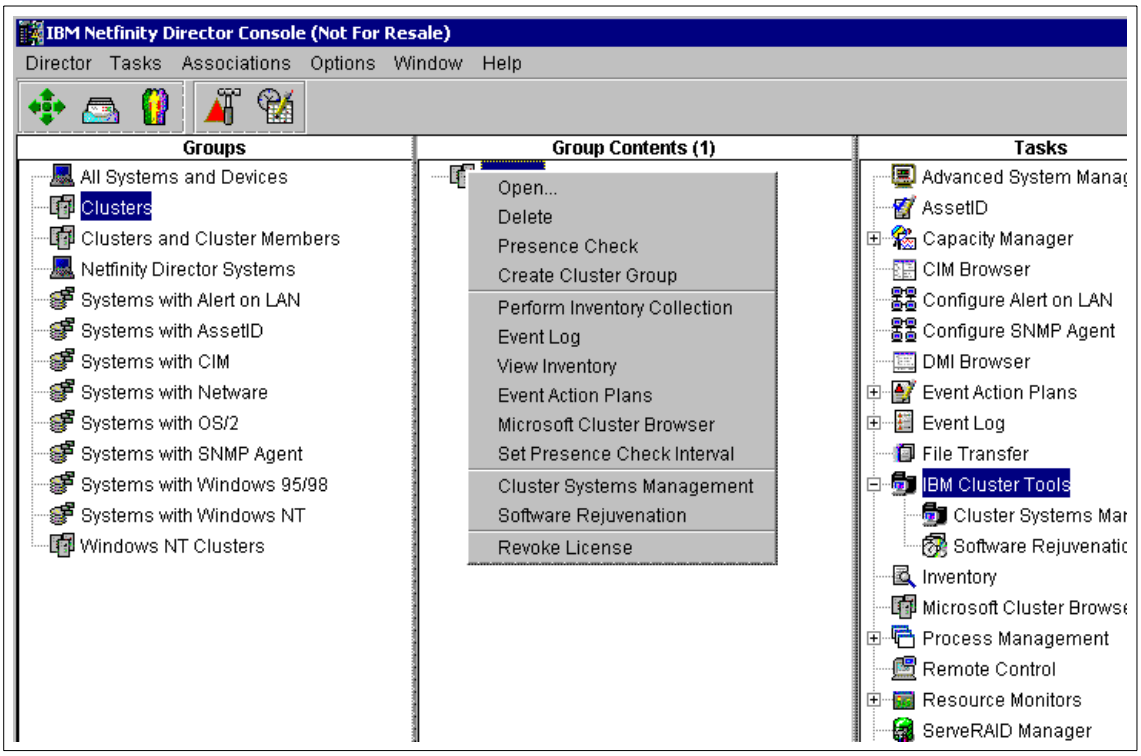

Figure 12. IBM Netfinity Director Console

This pop-up menu shows all the tasks that you can perform on the cluster. The three tasks specific to clusters are:

- Microsoft Cluster Browser
- Cluster Systems Management

- Software Rejuvenation

Note: As well as right-clicking the entry in the Group Contents panel, you can also drag tasks from the Tasks panel to the cluster to perform them.

3.5.2.1 Microsoft Cluster Browser

This browser allows you to view information only — you cannot perform data-changing operations. It lets you determine the structure, nodes, and resources associated with a cluster, which means that you can determine the status of a cluster resource and view the associated properties of the cluster resources.

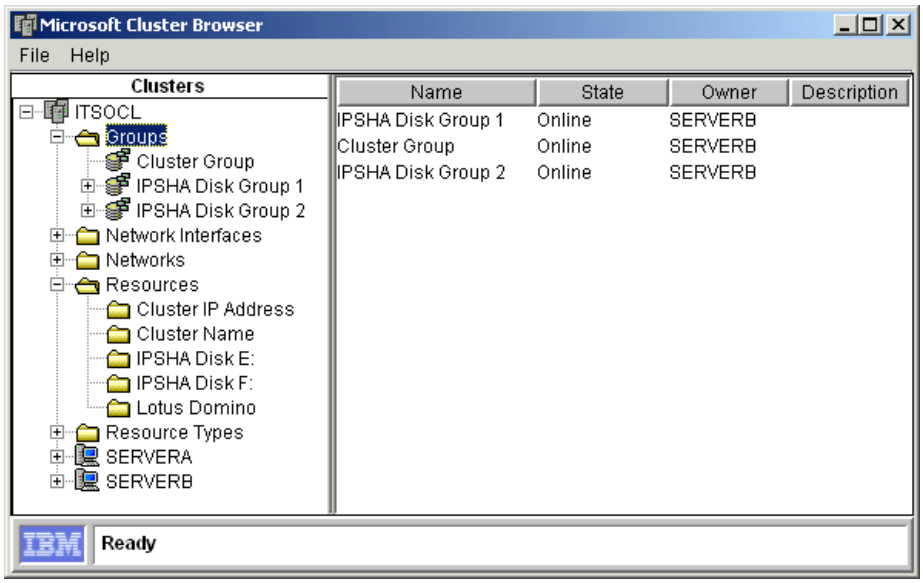

Figure 13. Microsoft Cluster Browser

3.5.2.2 Cluster Systems Management

This tool is similar to the IBM Cluster Systems Manager in Netfinity Manager. It lets you perform management tasks on your clusters from any Netfinity Director Console. This task is part of the UM Server Extension plug-in.

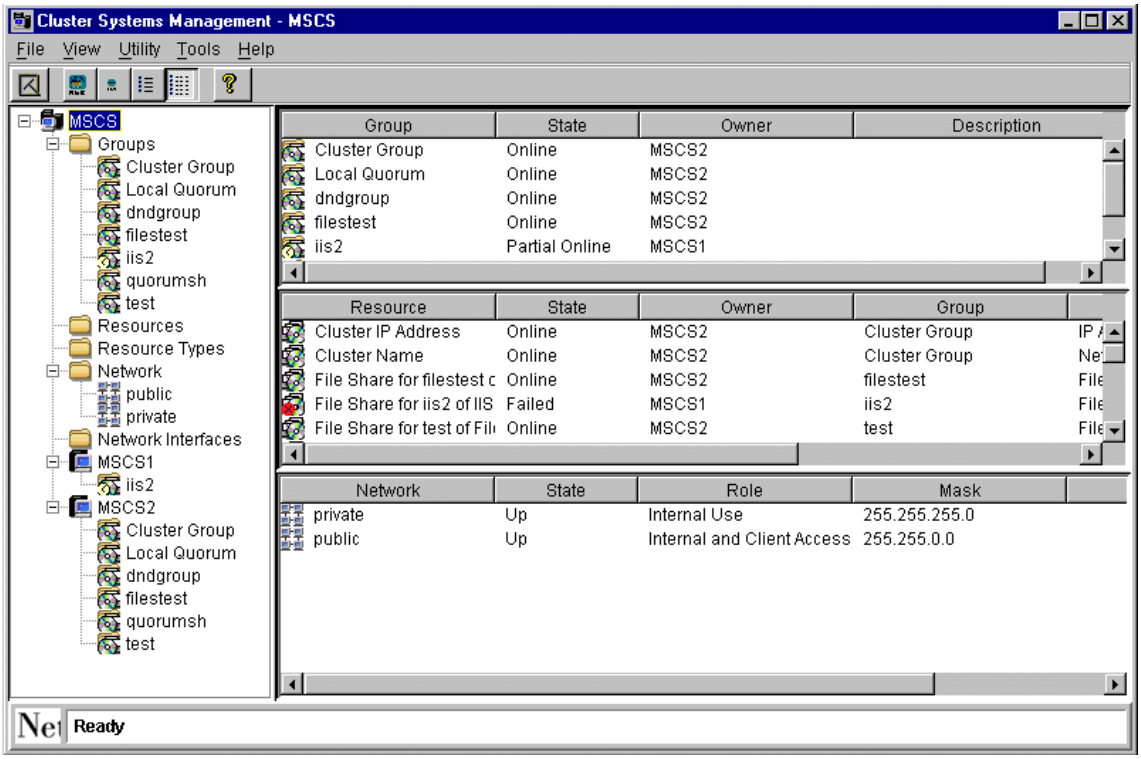

Figure 14. Cluster Systems Management

3.5.2.3 Software Rejuvenation

The IBM Software Rejuvenation program allows you to selectively restart nodes in a cluster. In restarting a node, you rejuvenate, or refresh, software resources. Software Rejuvenation not only restarts a node in a selected cluster, but also allows for schedule restarts. This includes scheduling of multiple nodes on various dates.

To schedule the restarting of a node, simply drag the node onto the calendar. You will then be prompted to enter the details of the schedule including repeat rate and time of day. You can also change other clusterwide options by clicking the **Option** button.

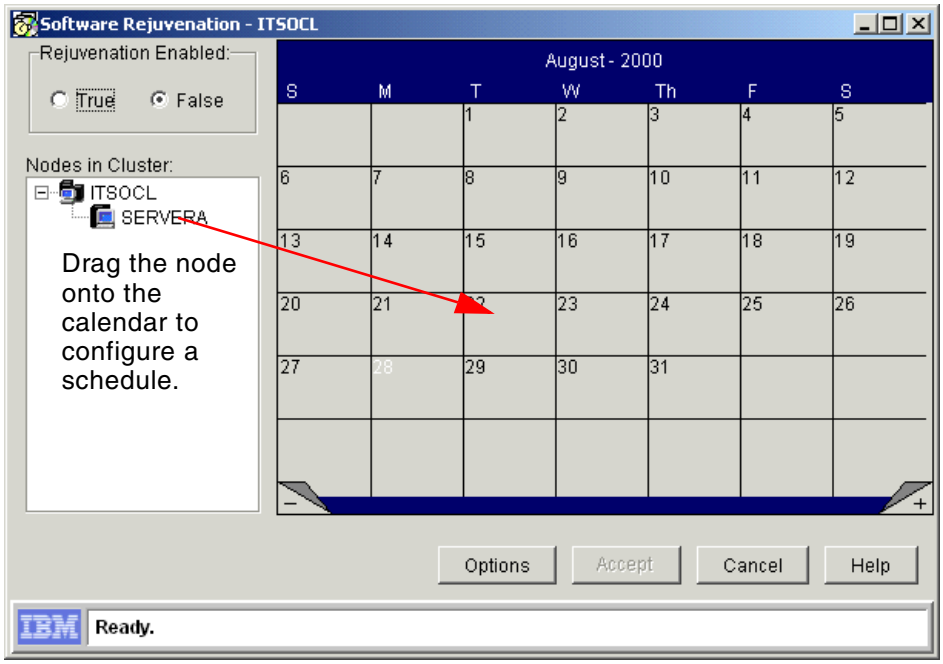

Figure 15. Software Rejuvenation

3.6 Legato Co-StandbyServer for Windows NT

Important note

Legato does not offer a Windows 2000 Co-StandbyServer solution. However, they do offer Legato Cluster Enterprise and eCluster, which is supported with Windows 2000. Legato also offers Mirroring Extensions for Microsoft Cluster Service, which lets you implement a mirror configuration rather than a common disk configuration and separate the nodes by up to 10 km.

We will not be discussing these products in this redbook. For more information about these products, refer to:

`http://www.legato.com/products`

Legato Co-StandbyServer for Windows NT is a server-clustering solution that offers high availability to critical network resources including data, applications, shares, IP addresses, and printers. The product allows two

servers to support each other, both of which can be fully functional network servers running applications. In the event of a failure, Co-StandbyServer transfers clustered resources from the failed server to the surviving server. This is an active/active configuration as discussed in 2.2.3, “Active and passive servers” on page 12. Co-StandbyServer is based on a shared-nothing architecture (see 2.2.2, “Hardware for clusters” on page 11).

The single greatest benefit of the Co-StandbyServer approach is the elimination of all single points of failure within the cluster. In this configuration, each server within the cluster maintains its own copy of the clustered resources and its own Windows NT registry database.

3.6.1 Requirements

Implementing a Legato Co-StandbyServer for Windows NT solution places a number of prerequisites on the two server systems to be used:

- Servers

  The two servers must be Intel-based and compatible with Microsoft Windows NT Server 4.0. They do not have to be identical. For example, one server might be a four-way SMP system while the other is a uniprocessor machine.

  Note: IBM requires both systems to be the same.

  The clustered servers must be in the same domain or workgroup, and in one of the following relationships to each other:

  – A primary domain controller and a backup domain controller
  – Two backup domain controllers
  – Two stand-alone servers

- Network interface cards

  Industry-standard network interface cards are used to connect network clients to the servers. All network protocols supported by Windows NT can be used to make up the network backbone.

WAN capability

A major benefit of Co-StandbyServer is that there are no limitations on the distance between the servers imposed by a common disk subsystem. A wide area network (WAN) connection can be used as long as a low-latency, high-speed link is used. The use of bridges and routers on this link should be minimized.

Implementation of disaster-recovery solutions is therefore relatively simple, provided that you must be careful to eliminate the possibility of the two servers losing complete contact. That is, you should run the private and public connections between the servers along physically separate routes. If separate routes cannot be ensured, we recommend you do not allow automatic failover.

- High-speed interconnect

  A high-speed interconnect is used to carry mirroring traffic between the two clustered servers. The dedicated link also acts as a second heartbeat path for checking the status of each server within the cluster. This eliminates the possibility of a false failover based on a partial network failure. The high-speed interconnect is nothing more than another network segment through which Co-StandbyServer directs all its mirroring traffic. The high-speed interconnect can be configured with any industry-standard card with a network driver interface specification (NDIS) driver. The Microsoft TCP/IP protocol must be bound to this network segment.

- Boot drive and mirrored disks

  It is important to install Windows NT on a physical disk (or logical drive within a RAID array) that is independent of the drive where data to be mirrored resides. Regardless of partitioning, the drive housing the Windows NT system partition cannot be mirrored. Co-StandbyServer supports all hard-disk controllers and storage devices compatible with Microsoft Windows NT. These include SCSI, IDE, SSA, RAID, and others.

  The implication of this is that a full active/active configuration requires a minimum of three disks in each server. These disks can be either physical or RAID logical drives.

A typical Legato Co-StandbyServer for Windows NT system configuration is shown in Figure 16:

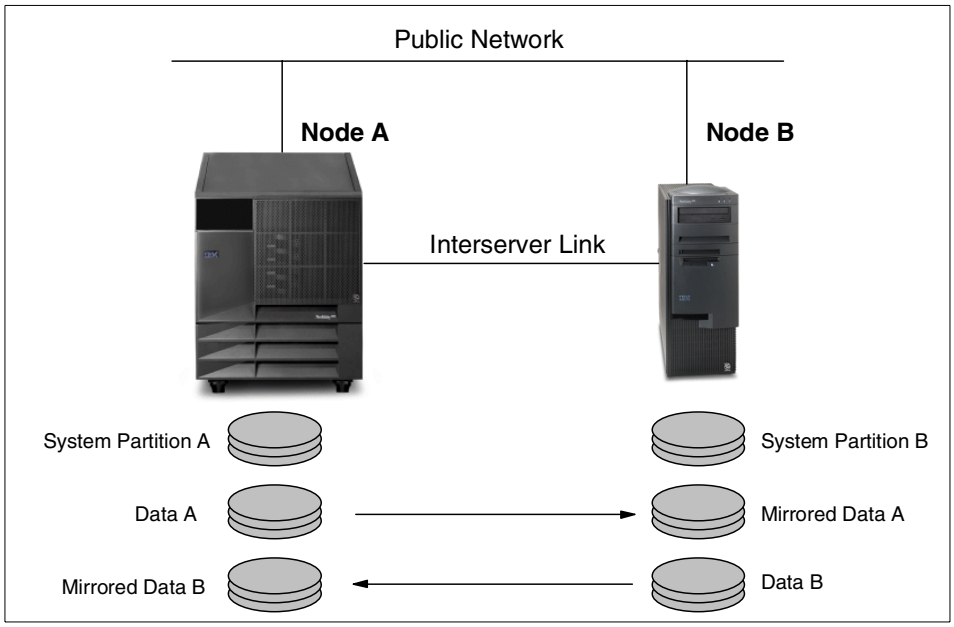

Figure 16. Legato Co-StandbyServer for Windows NT

Co-StandbyServer has its own disk-mirroring engine. The mirroring engine is invoked when a Windows NT volume is configured as a clustered resource. Each block of data residing on the clustered volume is copied to its partner disk device inside the mirror set. Open files can cause difficulties when file-level copying schemes are used. The block-level mirroring of Co-StandbyServer allows open files to be mirrored without any problem.

It is arguably more accurate to say the disks in a Co-StandbyServer mirror set are duplexed, since they are effectively on separate controllers. The process is much like duplexing disk drives internally on a server, except the disks in a mirrored pair each reside on a different server.

To avoid data corruption, the Co-StandbyServer software prevents each server from accessing any drive that is hosting a mirror copy of data belonging to the other server.

3.6.2 Resources

Cluster resources are the applications, services, or components running under the control of Legato Co-StandbyServer for Windows NT. Five resource types are supported:

- Disks, volumes, and partitions
- Shares
- Network cards and IP addresses
- Printers
- Applications

Co-StandbyServer clusters a resource by creating an inactive duplicate of an existing resource on the server that does not own that resource. Each resource in the system can be in one of the four following states:

- Not Clusterable - The resource cannot be clustered.
- Not Clustered - The resource can be clustered but is not at the moment.
- Clustered Active - The resource is clustered and active.
- Clustered Inactive - The resource clustered but inactive.

3.6.2.1 Clustering a volume with Legato

A cluster volume protects important data from disk device failures, controller failures, and server failures. A cluster volume can ensure continuous availability of important files, database tables, print queues, and application data. Cluster volumes are the foundation for building a successful high-availability configuration. Cluster shares, printers, and applications all depend upon information that will be stored on cluster volumes. For that reason, at least one volume must be clustered before you cluster these resources.

3.6.2.2 Clustering a share with Legato

A share is a shortcut name that network clients use to gain access to server data. By using a name to hide the actual path to the server data, system administrators can simplify clients' use of server resources and benefit from greater flexibility in administering those resources without exposing the details to users.

3.6.2.3 Clustering an IP address with Legato

Both shares and IP addresses connect network clients to resources on the server, but whereas shares provide file system connection points, IP addresses provide connections to applications through server names, which then resolve to an IP address.

3.6.2.4 Clustering a printer with Legato

A print server provides a centralized printing service to network-connected clients. To create a clustered printer, you must have a physical printer set up locally on each server.

3.6.2.5 Clustering an application with Legato

Co-StandbyServer maintains two application images in the system, one on each clustered server. Only one of the images will be active at a time. When the failover group that contains the clustered application moves from one server to another, either manually or as a result of a server failure, the active application image is deactivated while the inactive image is activated. During the failover process, Co-StandbyServer follows the directions in the application script to stop or start services and update registry information.

3.6.3 Failover and recovery within Legato

To understand the failover and recovery process of Co-StandbyServer you must first understand the relationship between failover groups and server names. You also need to know that Microsoft Windows NT Server Version 4.0 has the ability to create alias computer names. Aliases allow you to associate multiple computer names to one physical server.

Each server within the cluster maintains its own resources. When a resource is clustered using Co-StandbyServer, it is associated with an alias NetBIOS computer name that belongs to a failover group. The failover group can be activated on either server within the cluster. Activating a failover group on a server allows users to see the alias computer names on the network that are pointing to the server hosting the clustered resources.

Before installing Co-StandbyServer, the two servers to be clustered have their own unique NetBIOS names. These are the names of the servers as they appear when browsing the network. During the installation of Co-StandbyServer, the original NetBIOS names of the two servers are replaced and the original server names become alias NetBIOS names. Each of these aliases is applied to one of the two failover groups created during the installation.

For example, if the original names were SVRA and SVRB, the physical servers might be renamed SVRA-0 and SVRB-0 while the failover groups are given the original names, SVRA and SVRB. A client on the external network will now see individual servers using these four names in its browser list as soon as the failover groups are activated. This is shown in Figure 17.

Failover groups are equivalent to the resource groups or virtual servers discussed in 3.1.6, “Resource groups” on page 27. From a network resource planning perspective, the most significant difference between them is that Co-StandbyServer limits you to two failover groups, whereas MSCS will allow you to define as many resource groups as you wish.

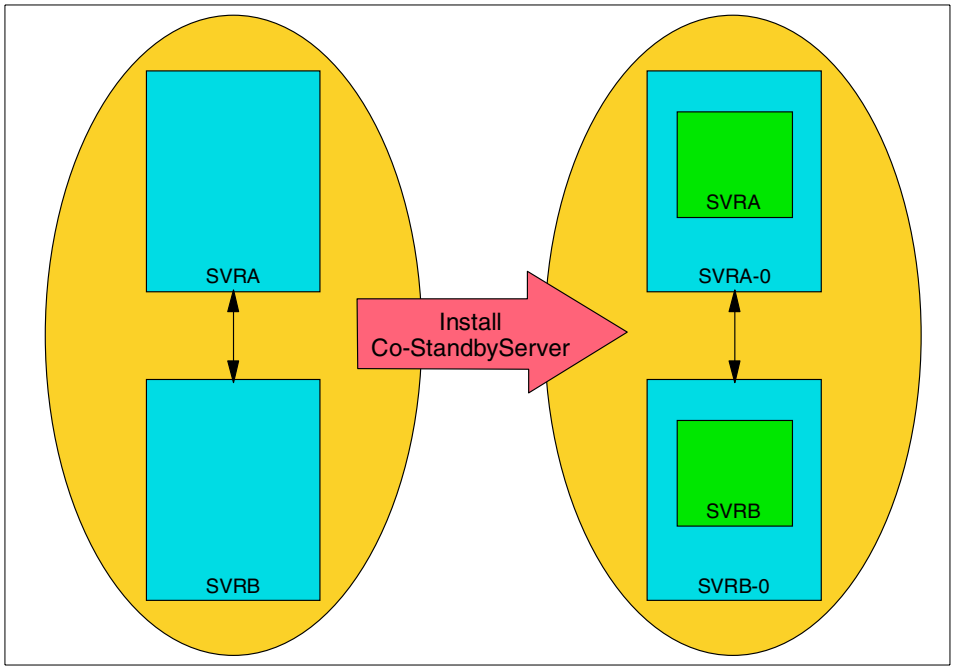

Figure 17. Effect of Co-StandbyServer on network names

If SVRA-0 encounters a failover, the SVRA failover group can automatically be activated on SVRB-0. Following this failover, you will only see three servers with the network browser, namely SVRB-0, SVRA, and SVRB. The failover groups contain all the clustered resources. SVRB-0 is now hosting both failover groups and all their clustered resources.

Just as with MSCS, your users must connect to resources owned by the failover groups. Connections to resources on the physical servers will be lost should a failure occur.

3.6.4 Failover groups

Failover groups transfer from one physical server to the other either manually or automatically. System administrators initiate manual transfers by issuing commands through the graphical management console. Automatic transfers

take place in response to a failure. The precise criteria for automatic failover are configured using the management console. Although the net effect of both failover types is to move resources and client connections between servers, there are some differences between manual and automatic failover that are worth noting.

3.6.4.1 Manual failover

There are at least two reasons why a system administrator might want to move resources and clients between servers:

- Scheduled maintenance

  Every server requires maintenance from time to time: disk drives need to be added, network interface cards replaced, software upgrades installed, and so on. For a mission-critical server, scheduled maintenance presents a problem for the system administrator. Either users do without important functions during the maintenance downtime or the maintenance occurs at a time of relatively little network activity, usually at a time most inconvenient for the system administrator. Manually moving both failover groups to the other server in the cluster allows the system administrator to perform maintenance when it is most convenient to him or her, without loss of critical network resources.

- Load balancing

  With only two failover groups, the scope for load balancing is limited, but this can be useful if your servers are supporting resources that are not clustered in addition to those that are clustered. To optimize performance for a specific set of workload conditions, the system administrator might decide to move applications and clients from one server to the other.

  An additional feature of manual, as opposed to automatic, failover is that Co-StandbyServer is able to set selected resources to the inactive state on the server that the failover group is leaving.

3.6.4.2 Automatic failover

Co-StandbyServer constantly monitors the connections between the cluster servers looking for failures. When a failure is detected, Co-StandbyServer checks the properties of the failover groups. If there are any failover groups currently active on the failed server and they are configured to automatically fail over, Co-StandbyServer prepares the surviving server to receive the failover group from the failed server. The necessary resources are then activated on the surviving server and the failover process is complete.

Automatic failover occurs only in the event of a failure of one of the cluster servers. After an automatic failover the cluster is no longer in a protected state and the cluster resources are at risk should the second server fail.

This is in contrast to manually moving a failover group, which does not alter the availability state of the cluster.

3.6.5 Cluster management

Co-StandbyServer provides a graphical management console that is installed on either a Windows NT client or one of the cluster servers. The console provides simple drag-and-drop controls to make cluster administration tasks easy to perform. Administrators view clustered resources and their properties by right-clicking the relevant objects. Manual failovers are performed by dragging failover group icons and dropping them onto the icon for the physical server on which you wish them to reside.

A number of dialogs are provided to allow you to set up alerts and command files to be executed when failovers take place. Alerts are issued when a server failure is detected and the command files allow you to fine-tune the way the cluster behaves during and after a failover. Additional command files are executed when a manual failover occurs, allowing a different set of actions to be initiated.

More information, including trial versions of the Legato Co-StandbyServer for Windows NT product and online documentation, is available from:

`http://www.legato.com`

3.7 Marathon Endurance array

Marathon Technologies is a ServerProven partner with their Marathon Endurance array solution. This Windows NT configuration offers a fault-tolerant solution with no single point of failure.

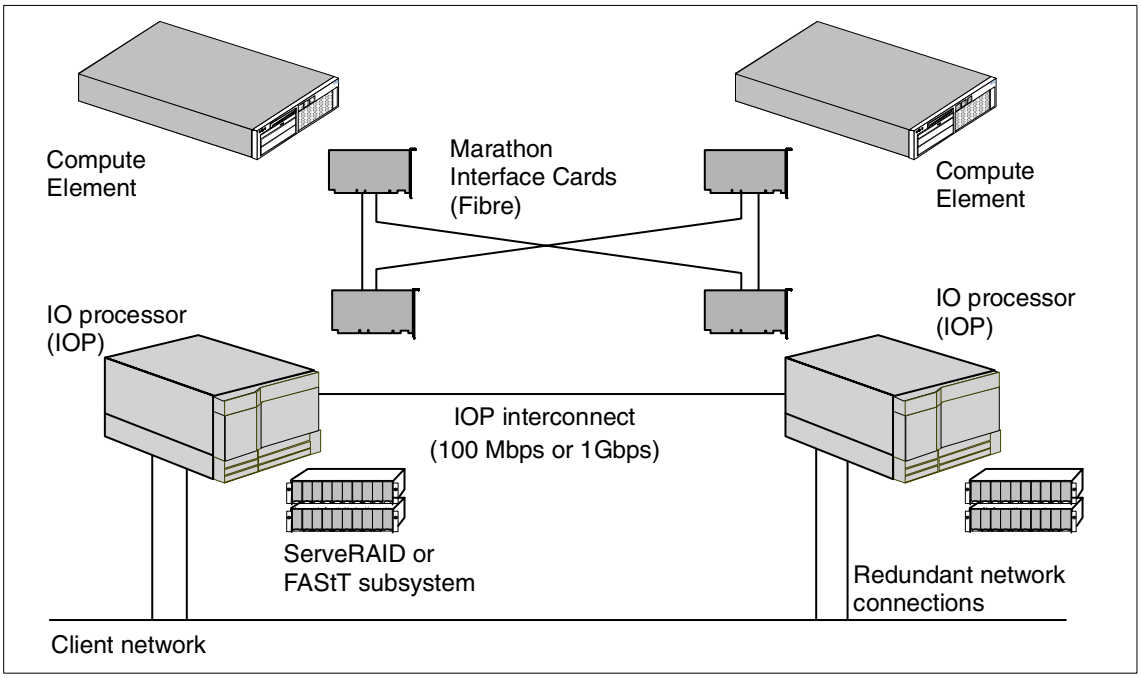

Figure 18. Marathon Endurance array — example configuration

This solution is designed to provide fault tolerance at both the hardware and operating system levels to hide many of the complexities traditionally associated with redundant, high availability systems. It runs on Windows NT 4.0 Server, Windows NT 4.0 Enterprise Edition, or Windows NT 4.0 Terminal Server with unmodified applications such as Microsoft Exchange.

Unlike a standard MSCS configuration, the Marathon Endurance array prevents users from experiencing server failures including loss of in-flight transactions. The Marathon system uses an implementation of redundant server components that ensures no single point of failure protecting applications, user sessions, in-flight transactions, continuous access to data, and network connectivity. All data is maintained in two mirrored locations.

With the Endurance array, the transfer of services to another node is unnecessary because redundant synchronized components are online continuously. Failed components can be repaired or replaced while the system is running, and are automatically reconfigured into the array and re-synchronized with the surviving components.

3.7.1 The Endurance array

The Endurance array offers active redundancy in the form of two logical synchronous working servers called *tuples* that perform the same tasks at the same time. The Marathon fault-management software manages both tuples transparently to applications that are running. When a failure occurs in one tuple, its counterpart in the other tuple continues to service applications and users without interruption. Thus, application context, user session context, in-flight transactions, access to data, and network connectivity run without interruption.

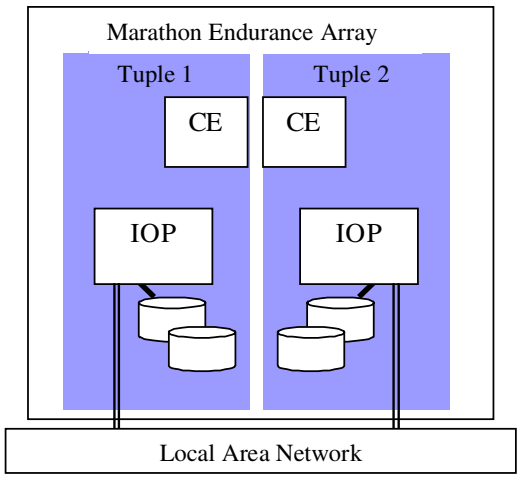

Figure 19. Marathon Endurance array — logical view

Each tuple is a single logical working server comprised of a Compute Element (CE) and an I/O Processor (IOP).

- The CE contains the CPU and memory and performs the synchronous processing, running a copy of the operating system and a copy of the application.
- The IOP contains the video, keyboard and mouse, mirrored SCSI devices, redundant network adapters, and other I/O devices and performs all of the asynchronous I/O.

Two tuples are cabled together with fiber so that there is no single point of failure in the system. Marathon software monitors and manages both tuples. The CEs of each tuple are forced to run in lockstep, while the two IOPs perform as a loosely synchronized redundant pair.

Separating the synchronous application processing from the more variable asynchronous I/O processing eliminates many potential operating system

failures typically associated with interrupts and I/O device interfaces. Separating tuples across two geographically dispersed sites eliminates many potential hazards that would render a single computer out of service.

Marathon fault-management software manages all of the active redundant components within the array. When an array component has failed or an administrator manually disables it, the array is in a vulnerable state because a single point of failure exists until the failed component is repaired.

3.7.2 Marathon mirrored disks

Mirrored volumes with dual simultaneous writes protect important data from disk device failures, controller failures, and server failures. A mirrored volume can ensure continuous availability of important files, database tables, print queues, and application data. Simultaneously mirrored volumes are a primary component for building a successful fault-tolerant configuration.

All I/O devices in the Endurance array physically reside on the IOPs, and because there are two IOPs, the mirrored SCSI devices (one on each IOP) provide fault tolerance. The Marathon fault-management software coordinates all I/O processing so that each pair of redundant components appears as one logical device. If a disk failure occurs, the disk operations can continue operating without affecting applications or users.

In the mirrored disk set, the two physical disks appear as one logical disk and provide RAID-1 capability. Standard RAID-5 or RAID-0 products can be added to provide even greater performance and data availability. Marathon software manages disk I/O within the mirrored disk set so that disk writes are performed simultaneously on each of the two physical disks. There is no single point of failure so there is no window of vulnerability for data failure or corruption. Traditional mirroring techniques create a single point of failure between the time a primary write occurs and the time when the mirror copy is written.

The Endurance array initiates a mirroring process only after a disparity exists between the two mirrored disks because of downtime for failure or maintenance. The administrator can permit the Marathon software to automatically re-configure the disk back into the array and initiate a mirror copy of the surviving disk to the newly joined disk. Another administrative option is to manually manage the re-synchronization and mirroring process. The mirroring process operates as a background task that does not interrupt applications or users.

3.7.3 IP and MAC addresses with Marathon

The Endurance array supports multiple Ethernet connections on one to four networks to provide network redundancy and network traffic control. It maintains continuous network connectivity by detecting any network adapter or connectivity failures. If a failure occurs, the redundant network connection continues processing all network traffic.

The Endurance array presents itself to the network as a single server with a single MAC address and a single IP address, even though there are multiple connections to the network. The Marathon fault management software manages the Network Interface Cards (NIC) so that two or more may listen on the network, but only one is activated to transmit.

3.7.4 Disaster tolerance

The Endurance array implements SplitSite technology that enables the two tuples to be deployed across different geographic locations up to 500 meters apart. This real-time hot-site capability provides a complete range of data protection and application availability from minimum business interruptions (such as a power failure) to maximum disaster protection (a building disaster). If one tuple is destroyed, the other continues the required work without loss of data or interruption of application processing.

The Endurance array maintains two application images in the system, one on each tuple. Both images are active and simultaneously perform lockstep execution of each instruction. When a failure occurs in one server, the other server simply continues to process the next instruction and maintains the application context, user session context, in-flight transactions, access to data, and network connections.

Standard off-the-shelf applications become fault tolerant when installed on an Endurance array. The application is loaded once, no scripting is required, and the application remains unmodified.

3.7.5 Information

For more information about the Endurance array, see:

- Endurance product information:
  `http://www.marathontechnologies.com/productinfo/Endurance_6200.html`
- Whitepaper: *A Technical Overview of Marathon Assured Availability Endurance*: `http://www.tagteam.com/ttserverroot/Download/78887_MarathonAAEnduranceTechOverview.doc`

- Marathon ServerProven program participants: `http://www.pc.ibm.com/us/compat/serverproven/index.htm` (click the Marathon logo)

3.8 Linux clustering

WIth the adoption of Linux as a mature server operating system by IBM in early 2000, what had been a relatively obscure “hacking” project, suddenly became the talk of the IT world. The approach taken by IBM towards Linux has “legitimized” Linux in the eyes of IBM’s more traditional customers, and has caused these customers to think seriously about Linux for the first time.

Linux now offers an alternative server operating system and is an ideal match for the IBM range of Intel-based servers - the xSeries and Netfinity systems.

In this section we study specific Linux solutions for creating clusters of machines to provide high-availability configurations: software solutions using Linux to provide higher availability with multiple machines than with single server solutions. The combination of the Linux operating system, sophisticated and reliable software clustering, and xSeries and Netfinity hardware offers high availability at a low price. Even in an enterprise environment, where an obvious choice for a highly reliable back-end database server would be the zSeries Parallel Sysplex environment, for example, Linux high-availability clustering solutions can provide a reliable front-end Web server.

The two primary benefits of a Linux high-availability cluster are:

- Fault tolerance — if a single Linux server in a cluster should fail, then the server function of the total cluster solution is not impacted.
- Scalability — as workload demands grow it should be possible to add machines to an existing cluster to handle the load. This compares with a single-box solution in which at some point a total hardware replacement is required to upgrade the server function.

The three typical ways clustering is used in a Linux environment are:

- **High performance computing or scientific computing**

  The most commonly known implementation of Linux clustering is *Beowulf*. The Beowulf Project implements high performance computing (HPC) using message-passing parallel programs. To really use the Beowulf concept, your application has to be written (or rewritten) using parallel virtual machine (PVM) or message passing interface (MPI). At least you should

be able to run the processing in parallel using shell script front ends, so that each node works at a specific range of the whole task.

Beowulf is not a single software package — instead, it consists of different parts (PVM, MPI, Linux kernel, some kernel patches, etc.). You can get more information about Beowulf at

`http://www.beowulf.org/`

- **Load balancing or scalability**

  This is a very important topic for any fast growing business, which most of today's e-business sites are. Most of these sites start small with only a few Web servers and a back-end database. So when they grow, they have to change their hardware more often, as their number of customers and the number or level of services they provide increases. Changing or upgrading your hardware means outages, downtimes, and lost money and it doesn't look professional nor does it provide the kind of service your business needs to grow.

  With a load-balancing cluster, you can just add another box into the cluster if the demand or load you get increases. If one server fails, just change the cluster configuration automatically and take the broken server out for service. Later you can reintegrate this server or a replacement box back into the cluster again.

  Most of today's Linux load-balancing cluster solutions are based on the Linux Virtual Server (LVS) project and one of the major products implementing LVS is TurboLinux Cluster Server. For more information on LVS, see:

  `http://www.linuxvirtualserver.org/`

  Another approach to load sharing and distributed computing is called *MOSIX* (The Multicomputer OS for UNIX). It allows you to run processes distributed on a collection of clustered nodes transparently. MOSIX migrates processes from very loaded nodes to other less loaded nodes dynamically and scales very well. No special support from the application is necessary. It simply looks like normal SMP but with more than one physical box.

  Actually, MOSIX doesn't exactly fit in any of these categories. It's something between HPC and load sharing, but currently doesn't provide improved additional availability. For more information see:

  `http://www.mosix.org/`

- **High availability and failover**

  High availability is also part of load balancing as discussed above and is known as an active/active configuration (where all nodes are doing real or active work). However, high availability can also be configured as active/passive — in Linux, this concept is called the Fail Over Service (FOS).

  With two-node FOS systems, you have one master and one standby system. In normal operation the master is running your service or application and the second system is just watching the master. If the master fails, the second system takes over the service immediately and shuts the master down (if this has not already happened). This provides you with a highly available system.

  FOS is also provided by the Linux Virtual Server project. One currently available commercial product is Red Hat HA Server and the next release of TurboLinux Cluster Server will provide FOS as well.

  At the time of publication, there was also another project going on to port SGI's FailSafe HA solution to Linux. SuSE and SGI are the major contributors.

3.8.1 Implementing Linux clustering

In the following sections, we examine the latter two aspects of Linux clustering as they are implemented in the Linux Virtual Server:

- High availability and failover
- Load balancing

At the time of writing, the products that implement LVS are Red Hat HA Server and TurboLinux Cluster Server.

Note: While both Fail Over Service (FOS) and load balancing are provided by the Linux Virtual Server project, most people refer to LVS when talking about load balancing, and FOS when they mean real high availability, and so do we.

As with other clustering implementations such as Microsoft Cluster Server, a Linux cluster has the following components and characteristics:

- A heartbeat connection between the nodes.

  With a two-node cluster this can be a simple crossover Ethernet connection, but with three or more nodes, a private switched Ethernet network is recommended, either 100-BaseT or Gigabit Ethernet.

- Separate network connections to the network for normal data traffic

- The cluster as a whole is virtualized as one single server complete with one or more virtual IP addresses.

3.8.2 Fail Over Service

Failover means that we have two servers, one primary or master node and one secondary or backup node. Both know about each other via heartbeat and are attached to the client network as well, as shown in Figure 20:

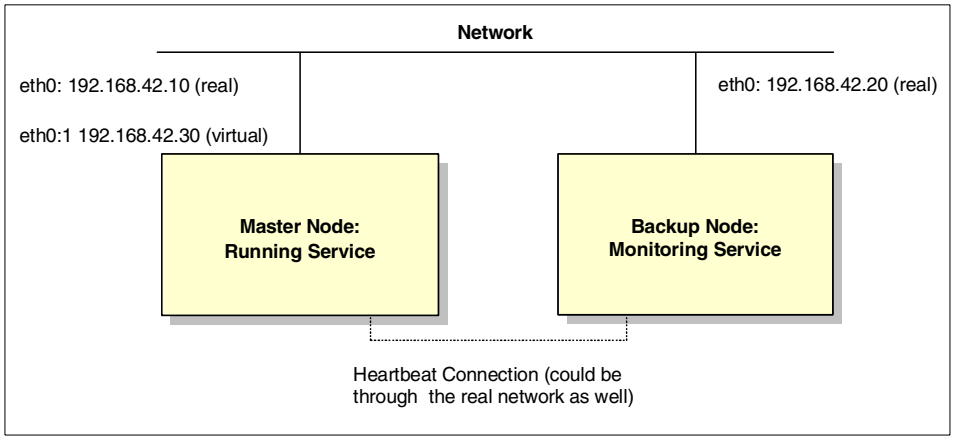

Figure 20. Failover service in normal operation

In normal operation, the master server is running and providing the service, a Web server for example. The backup node monitors the master such as by trying to connect to the master server's HTTP port (80) every 10 seconds and retrieve a Web page. Heartbeats are exchanged by both servers. As the picture implies, both servers have a real IP address assigned (192.168.42.10 for the master and 192.168.42.20 for the backup in this case) to their real interfaces (eth0). As the master node is active, it gets a second IP address, the virtual cluster IP address. In Linux terms, this IP address is an alias address (eth0:1) defined on top of the real network interface (eth0).

Both real and virtual interfaces can be seen via ARP (Address Resolution Protocol), responsible for the IP to MAC address mapping. Actually, both eth0 and eth0:1 share the same MAC address, which is why eth0:1 is called an alias.

What happens if the service on the master server becomes unavailable or the server itself goes down? This situation will be noticed via monitoring (if only the service fails) or via heartbeat (if the complete machine goes down). Figure 21 shows what will happen then:

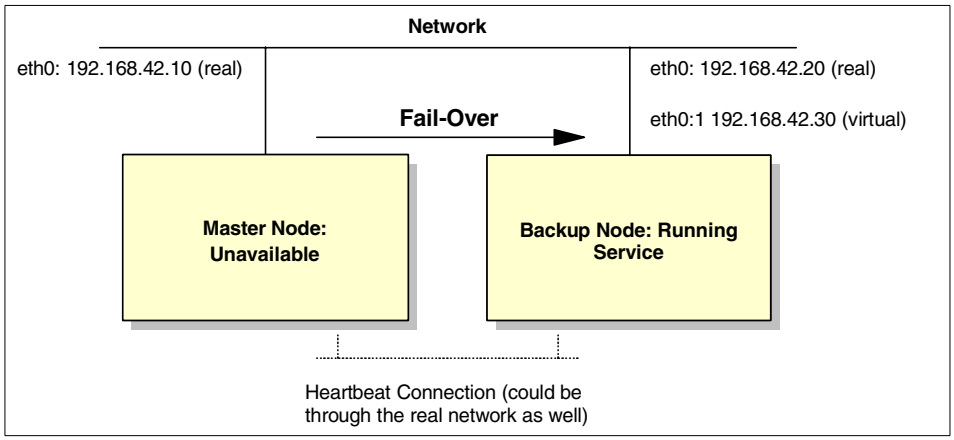

Figure 21. Failover service showing actual fail-over operation

The backup node takes over the virtual cluster IP address from the master node and gets its aliased eth1:0 up and running. After that it starts the service that was originally available on the master node and everything is fine again. This process is called failover.

As the virtual IP address is transferred to another real network interface, its associated MAC address changes too. To get this change reflected to all other computers on the network, the new active (and former backup) node broadcasts an ARP message for the IP address of the cluster containing the new MAC address. This process is known as *gratuitous ARP* or *courtesy ARP* and enables the other machines on the network to update their ARP tables with the new MAC address of the cluster.

If now the master becomes available again (observed via heartbeat), also called *failback* or *fallback* can take place (see Figure 22). The backup node stops running the service, the master node takes over the virtual IP address, issues the gratuitous ARP broadcast and starts its service. At this time everything looks like no failover had happened at all.

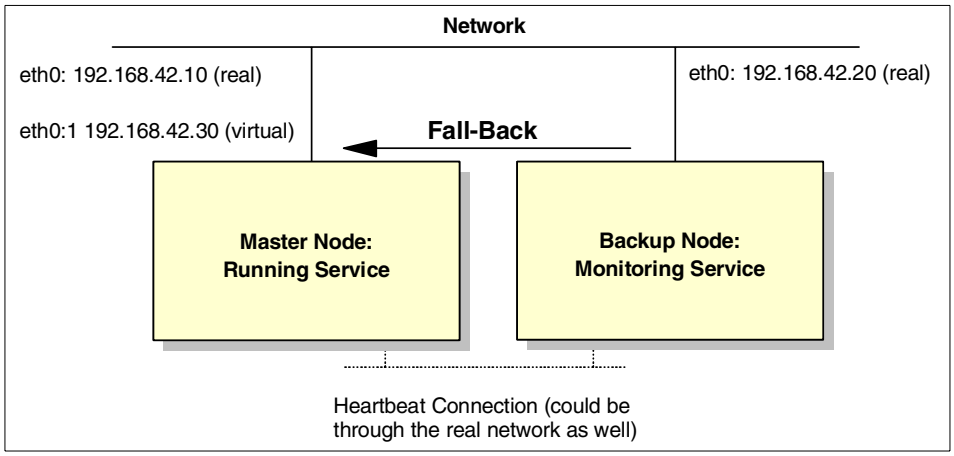

Figure 22. Failover service, resumption of normal operation

With the current Linux product implementations, some restrictions apply:

- Only two node FOS configurations are supported.
- No selective failovers (for individual services) are possible; all services are monitored and failover as a group.

3.8.3 Load balancing

Load balancing works similar to Fail Over Service, but aims at scalability and reducing system outages. It spreads incoming traffic to more than one server and lets all these servers look like one large server. It uses heartbeats like FOS, but implements another concept, unique to load balancing: traffic monitors or managers. A very simple LVS setup is shown in Figure 23. There's no dedicated, internal cluster network; all machines are connected to the same physical network.

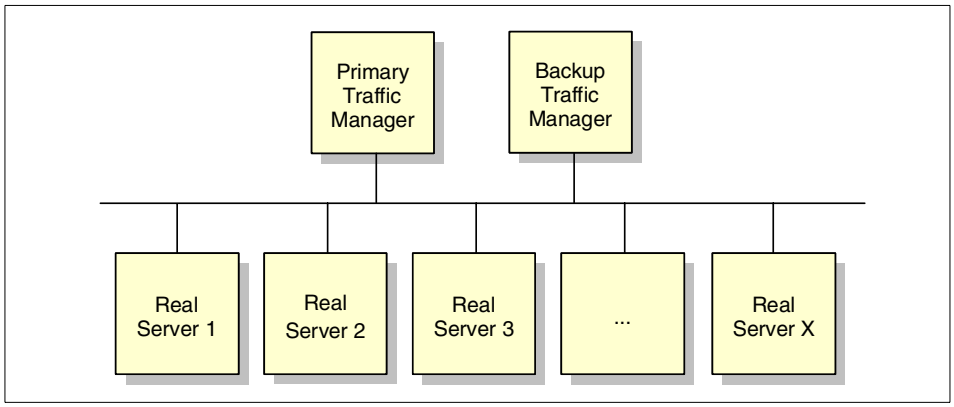

Figure 23. Simple Linux Virtual Server setup

As with FOS, there's a virtual server formed out of individual boxes. The primary and backup traffic manager behave like a FOS cluster concerning network connection and heartbeat service. The active traffic manager gets the virtual IP address assigned and redirects the incoming traffic to the real servers, based on the chosen load balancing and routing scheme. The traffic manager monitors the real servers for heartbeat, service, and load (if supported).

Scheduling mechanisms for distributing the incoming traffic can be one of the following, depending on the product:

- Round robin — all traffic is equally distributed to all real servers.
- Least connections — more traffic is distributed to real servers with fewer active connections.
- Weighted round robin — more traffic gets distributed to the more powerful servers (as specified by the user) and dynamic load information is taken into account.
- Weighted least connections — more traffic is spread to the servers with fewer active connections (based on a user-configured capacity) and dynamic load information is taken into account.

The next steps are to get requests from the traffic manager to the cluster nodes and then respond to the clients.There are three options, depending on the product you use:

- Direct routing
- Network address translation
- Tunneling

3.8.3.1 Direct routing

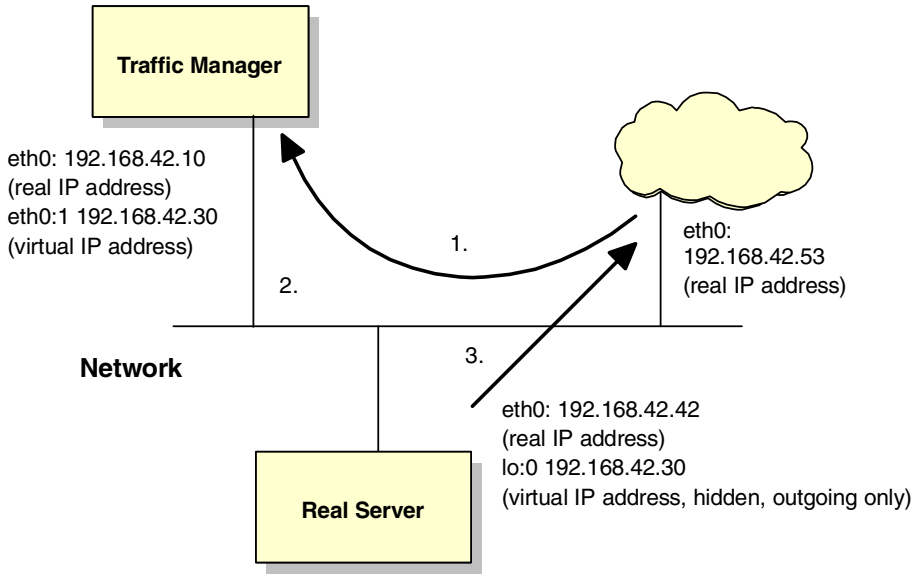

Figure 24. Direct routing of returned packets

In Figure 24, the client accesses the virtual server (192.168.42.30). Its traffic gets routed to the traffic manager, which redirects it to the real server by simply changing the MAC address of the data frame and retransmitting on the LAN. The real server itself has a physical network interface (eth0) for the incoming traffic and one aliased, ARP-hidden network interface (lo:0) for the outgoing traffic.

So the real server sends the response back directly to the requesting Client 1 using lo:0 as its source address, thus using the virtual IP address. From the perspective of the client, an IP packet has been sent to the virtual server's address and a response has been received from the same address. The client never sees any response to its request as coming from the server's "real" eth0 address. It only sees the virtual IP address.

The lo:0 address in Figure 24 is called a "hidden" address because it must be configured in such a way that the server owning this network interface will not respond to ARP requests for the IP address. The only network device that should respond to ARP requests is the traffic manager. The traffic manager determines which actual server is to be used for the received packet and forwards the packet to the server by re-transmitting the received packet onto

the network, but now with the destination Layer 2 MAC address of the packet now being the MAC address of the desired server.

The server will receive the packet, because it is now destined to its hardware MAC address, and will examine the packet and discover that it contains an IP packet destined for an IP address known to the server as its internal “hidden” IP address. It will then pass the packet to the IP application (such as a Sockets application) bound to this IP address. The application will respond and the same IP address will be used as the source address in the response, and the response packet will be sent out over the network directly to the client. The response does not pass through the traffic manager.

In our cluster implementations we examined all-Linux environments, in which the traffic managers and the servers themselves are running the same distribution of Linux code; this certainly eases implementation of the clusters but it should be noted that other operating system environments such as Windows 2000 or even OS/390 can be used for the server environments in a Linux cluster. The only requirement is that the servers themselves must be configured with both “real” and “hidden” IP addresses in a similar manner to Linux servers.

Because only the traffic manager responds to ARP requests for the IP address of the cluster, a full implementation of a load-balancing cluster environment will include a backup traffic manager as shown in Figure 23 on page 70. There will now be an additional requirement for the backup traffic manager to maintain cluster state information such as information on the state of open TCP connections into the cluster, and this information will allow the backup traffic manager to take over operation of the cluster without disrupting existing connections.

Although not shown explicitly in Figure 23 on page 70, the traffic manager function can also reside on the same physical server as one of the “real” servers. The function can be “co-located” with the server itself. This reduces the total number of machines required to implement a load-balancing cluster if this is an issue. A cluster could be implemented on only two machines, with one machine acting as the primary traffic manager and the other as the backup traffic manager and with the server functions themselves residing on the same machines.

This basic configuration is the easiest and fastest solution to implement, but has one major disadvantage: the traffic manager and the real servers must have interfaces to the same physical LAN segment. As traffic to the cluster increases, this may lead to congestion. Each packet inbound to the cluster appears on the network twice (once to the traffic manager from outside and

once from the traffic manager to the actual server) and then each response packet also crosses the same network.

It's a good idea to have a separate internal cluster network where possible like the one shown in Figure 25. In this network traffic between the traffic managers and the servers flows over the private network, and this network could also be used for the flow of heartbeat information, meaning that all intracluster network traffic is isolated from the external client network environment.

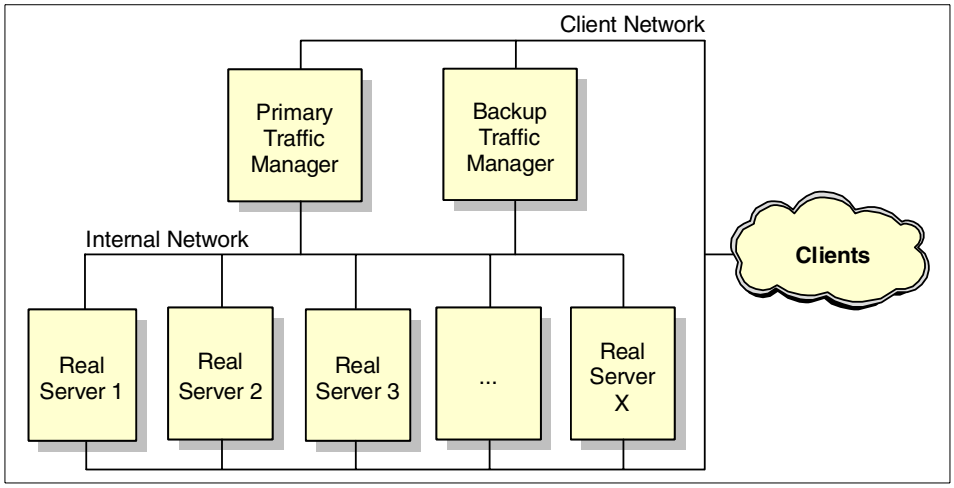

Figure 25. More sophisticated setup using an internal cluster network

3.8.3.2 Network Address Translation

Another option for hiding the internal cluster network is called Network Address Translation (NAT). NAT requires the traffic managers to take on one more job role; they have to translate the IP addresses of incoming traffic to direct it to one of the real servers and on the way back they have to re-translate the IP addresses of the outgoing traffic. Unlike the previous configurations, this requires that both inbound and outbound traffic have to flow through the traffic manager. Figure 26 shows this process:

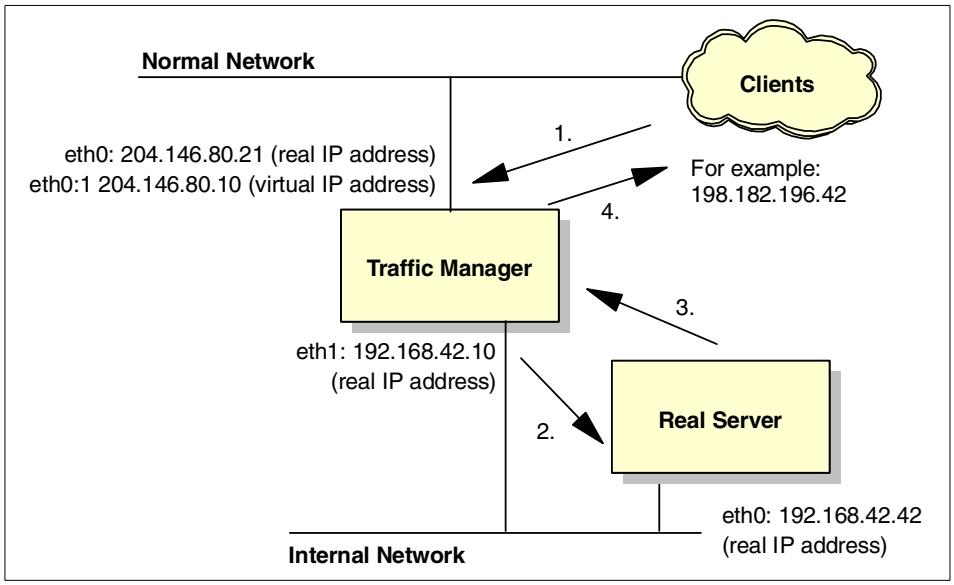

Figure 26. Network Address Translation

When the client talks to the virtual server represented by the traffic manager, its traffic looks like:

Source = 198.182.196.56
Destination = 204.146.80.10

Now the traffic manager selects a real server for this traffic and after translating the addresses passes on the traffic:

Source = 198.182.196.56
Destination = 192.168.42.42

After the real server does its job, it sends back the response using:

Source = 192.168.42.42
Destination = 198.182.196.56

Finally the traffic manager forwards the traffic to the outside world after a retranslation:

Source = 204.146.80.10
Destination = 198.182.196.56

The translation is done inside the traffic manager using a hash table and IP address-port mappings.

This is a very convenient way to implement a cluster because it only requires a single external IP address and all the destination servers can be defined on an internal private IP network. It does have one really significant disadvantage, mentioned above: all outgoing traffic has to pass through the traffic manager now. One of the justifications for permitting inbound traffic to pass through the traffic manager in a basic cluster is that outgoing traffic is usually much more significant in volume than the incoming traffic. This is because incoming requests such as HTTP requests are small in comparison to the volume of traffic sent back in response. Your traffic manager finally becomes the bottleneck of your cluster, and so NAT is suitable for a smaller cluster environment with not too much expected traffic.

And, in any case, what's to stop the servers in Figure 24 on page 71 from being configured with "real" IP addresses in a private IP network? There is no reason why even a simple cluster environment with a single network infrastructure can implement multiple IP networks over the same physical infrastructure. So NAT may not be required even in cases where multiple external IP addresses are not possible.

However, NAT has one other major attraction in that the destination servers themselves do not need to be configured with a "hidden" IP address at all. In the early days of clustering this was a problem on certain operating systems, and certainly adds complexity to the server configuration process even today. The NAT solution means that absolutely any IP server platform can be used as the target servers in a cluster without having to consider the quirks of IP addressing using "hidden" IP addresses configured on loopback interfaces.

3.8.3.3 Tunneling

Another interesting option for building up a LVS cluster is to use IP tunneling. It allows you to cluster real servers spread around the world, being part of different networks. But it needs the support of IP tunneling on each server of the cluster. Figure 27 shows the setup:

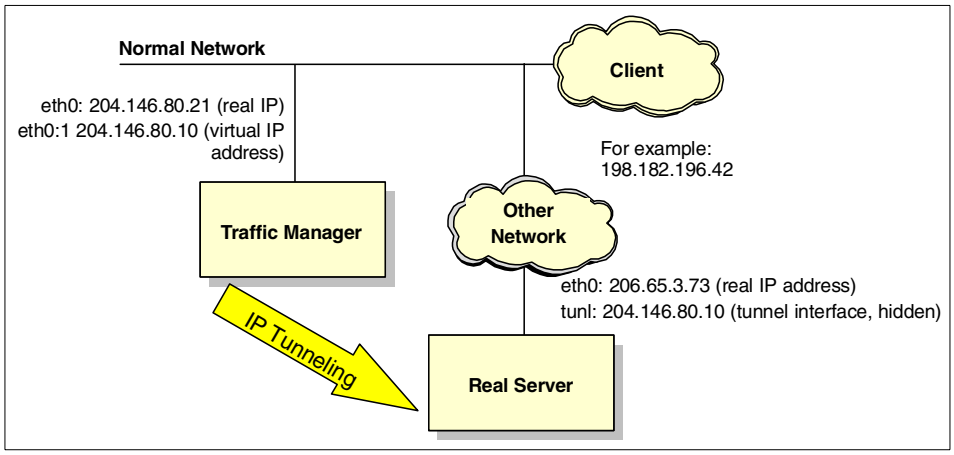

Figure 27. IP tunneling

Here, when a client accesses the virtual server the client sends a packet to the traffic manager, which advertises the IP address of the cluster and responds to ARP requests for it. Having received a packet from a client, the traffic manager encapsulates the packet into an IP datagram addressed to the real server, forwards it and stores this connection information in its hash table. All subsequent IP packets belonging to this connection end up at the same real server over the same IP tunnel. The real server itself de-encapsulates the packet and responds to the client directly using the virtual IP address as its source address.

IP tunneling is a very flexible way to build up a widespread cluster solution, but depends on the IP encapsulation protocol support of all participating cluster servers/nodes. In current implementations, this requires that all the servers be Linux servers, whereas the other solutions we discussed can use a mix of server operating systems in a single cluster.

3.8.4 Services supported

All IP services using a direct socket connection can be implemented with the current Linux clustering solutions. Here are some examples:

- HTTP
- FTP (INETD)
- SMTP
- POP
- IMAP
- LDAP

- NNTP
- SSH
- Telnet

Services depending on a secondary port connection besides the listening port are not supported.

3.8.5 Sharing the data between nodes

One of the most important aspects of the cluster is the availability to the nodes and the consistency of that data between the nodes. There are, as always, multiple solutions, mostly depending on the frequency of changes to your data and the amount of data involved. We cover the following, starting with the simplest:

- rsync
- Network File System
- Global File System
- Intermezzo
- Back-end databases

3.8.5.1 rsync

If your content is primarily static Web pages (contact information, for example) or a reasonably small FTP site, you can store all the data locally on each of the actual servers. Then to keep the data synchronized you can simply use a mirroring tool such as rsync that runs periodically, say twice an hour. With this solution you get good availability, since all data is stored on each server individually. It doesn't matter if one server goes down for some reason, nor do you rely on a central storage server. Figure 28 shows this solution:

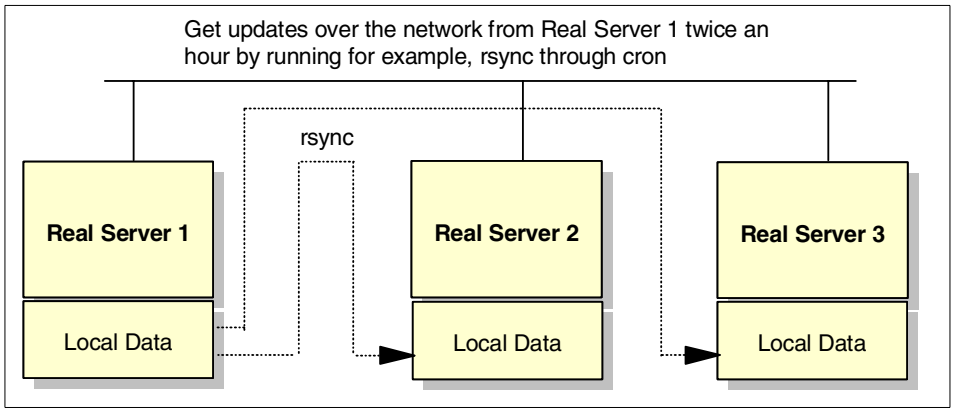

Figure 28. Using rsync for local server data synchronization

But this solution will not be suitable if you have really large amounts of data (more than a few gigabytes) changing more often (more than a few times in a week) and you have to keep it synchronized. Now network or distributed file systems come into the picture.

3.8.5.2 Network File System

Again, starting with the simplest approach, we can use NFS, the widely used, commonly known and stable network file system. It's easy to use and requires a central NFS server that exports the shared data. This data is "mounted" by the real servers across the network. This approach looks like the one shown in Figure 28:

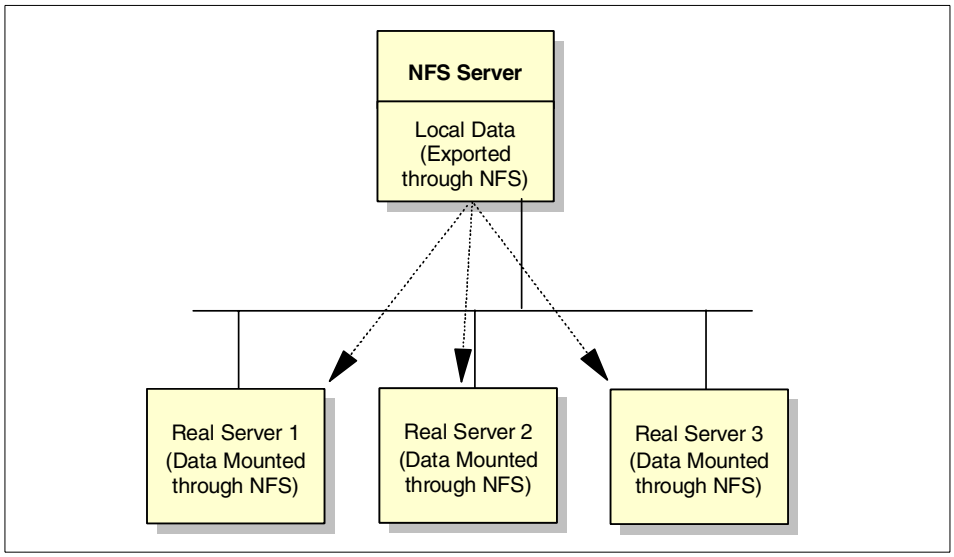

Figure 29. Using NFS for central data storing

Although NFS is simple to implement and use (on Linux), it has two major drawbacks:

- Slow performance
- Single point of failure

Although the performance may be acceptable for a small cluster solution, you should always be aware that if the NFS server dies then the real servers will no longer be able to get to your data and therefore will not be able to provide their service. This might make you think about setting up a redundant, highly available NFS server, but that's no trivial thing to attempt; you have to take care of the clients' file handles, keep the clustered NFS servers synchronized and there is no "out of the box" solution here. So after all, NFS is no real solution for a cluster environment.

That's why there are real cluster-capable file systems, such as the Global File System (GFS) and Intermezzo, which offers different approaches to a cluster file system.

3.8.5.3 Global File System

GFS implements the sharing of storage devices over a network. This includes Shared SCSI, Fibre Channel (FC), and Network Block Device (NBD). The Global File System sitting on top of these storage devices appears as a local file system for each box.

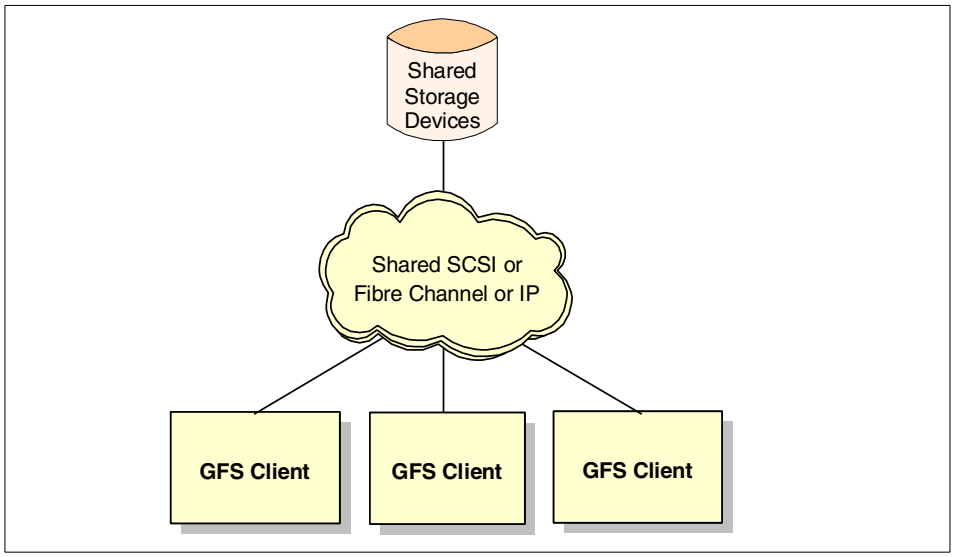

Figure 30. GFS

The Global File System is a 64-bit, shared disk file system focusing on:

- Availability — if one of the clients goes offline, the data can still be accessed by all the other GFS clients
- Scalability — which means it doesn't suffer from concepts based on a central file server, as NFS does

Furthermore GFS is able to pool separate storage devices into one large volume and to load balance between the workload generated by all the clients.

To set up GFS, you need to decide on the transport medium to use:

- Shared SCSI (although typically you are limited to clusters of two nodes).
- Fibre Channel
- IP (akin to using tunneling to attach to your client over a traditional network), not yet a widely used option and limited by the network bandwidth but allows you to attach any client without direct FC or SCSI connection to your storage pool

GFS itself implements the storage pooling, the file locking, and the real file system. It's still under development, but should be quite usable already.

3.8.5.4 Intermezzo

Unlike GFS, which is a shared file system, Intermezzo is an implementation of a distributed file system. This means that there's no central storage pool, but each machine has its own kind of storage locally. The storage gets synchronized via traditional TCP/IP networks.

Intermezzo features a client/server model. The server holds the authoritative data, while the clients only have a locally cached version of the data, which is kept synchronized. Intermezzo even supports disconnected operation and is able to reintegrate when connected again. Figure 31 shows a simple Intermezzo configuration:

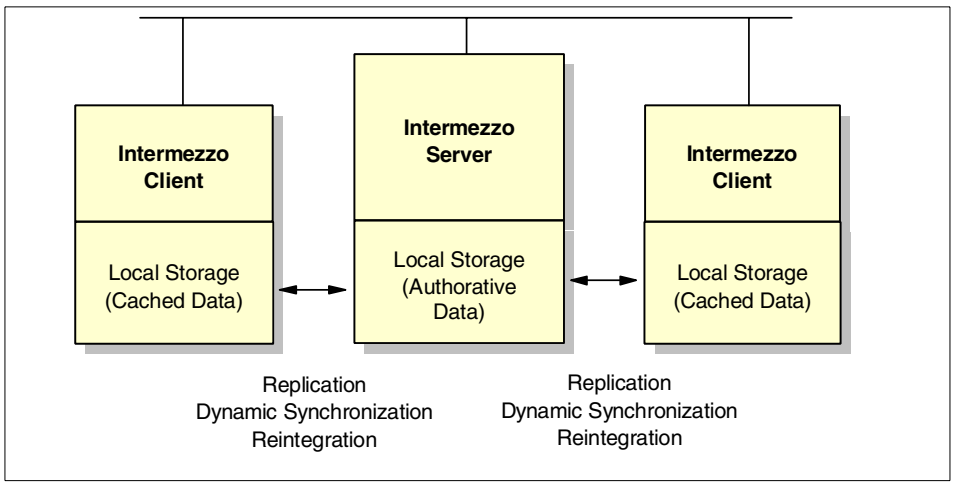

Figure 31. Sample Intermezzo setup

Intermezzo uses a traditional file system such as ext2 to hold its data and puts a layer in between that is responsible for journaling updates and keeping the data synchronized. Intermezzo, like GFS, is still under development, but already usable (it requires a Linux kernel recompilation to implement it today).

3.8.5.5 Back-end database

Another option for storing and accessing your data, and one that you might already have in place, is a back-end database, such as DB2. This database itself can be highly available, but that's not part of the Linux clustering solution. Your real servers simply have to be capable of connecting to this database and putting data into it or getting data from it using, for example, remote SQL queries inside PHP featuring dynamic Web pages. This is a very convenient and widely used option. An example of such a configuration is shown in Figure 32. Consider the back-end database as the existing

enterprise database server running on S/390 or other UNIX platforms, for example:

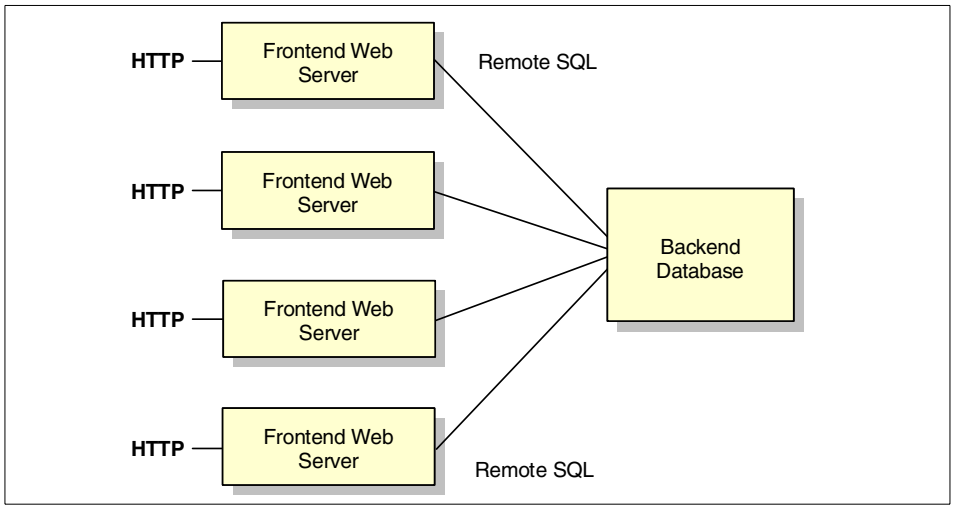

Figure 32. A cluster of front-end servers in conjunction with a fault-tolerant back-end database

3.8.6 Putting it all together

After discussing the different aspects of clustering, we now put all these things together to get a complete picture. The first question to ask is:

Why do we want to do clustering? The possible answers are:

- We want a scalable solution. So we go for Load Balancing.
- We want a highly available solution. So we go for Fail Over Service.

The important things about load balancing are:

- Make sure your services are able to run in a cluster environment (single listening TCP/IP port) and can be balanced (servers can act in parallel).
- Keep your system scalable from the point of network technology and cluster implementation.
- Think about a backup traffic manager. Otherwise, all your real servers are useless if the traffic manager dies.
- Based on the amount of data and change frequency, select an appropriate method to access and store your data.

Talking about high availability consider the following thoughts:

- Make sure you can monitor the services accordingly.

- Think about storing and accessing your data safely and available. There's little sense in building up a high availability Web server if the database it connects to does not offer comparable high availability.
- Be paranoid (up to a certain point). For example, consider a second, backup Internet provider if you want to offer Internet services. Otherwise you may end up with a really highly available internet service locally that is not accessible if your provider goes offline.
- Think about high availability from the hardware side (UPS).
- Think about disaster prevention, such as putting nodes in separate buildings.
- Finally, don't forget to test your setup on a regular basis.

For more information, see the following:

- Linux HA project: `http://linux-ha.org/`
- Linux Virtual Server (LVS) project: `http://www.linuxvirtualserver.org/`
- Red Hat HA server project: `http://people.redhat.com/kbarrett/HA/`
- Global File System (GFS): `http://www.globalfilesystem.org/`
- Intermezzo: `http://www.inter-mezzo.org/`

For details on implementing Linux clusters on Netfinity hardware, see Chapter 1 "Linux high availability cluster solutions" of the redbook *Linux on IBM Netfinity servers - A collection of papers*, SG24-5994.

3.9 NetWare Cluster Services

Novell NetWare remains a widely used and popular network operating system. High availability can be as important in a NetWare environment as it is for Windows. In this section we take a look at a product that provides clustering features for NetWare.

NetWare Cluster Services v1.01 for NetWare 5.x is a server-clustering system you can use to provide high availability and manageability of critical network resources, including data (volumes), applications, server licenses, and services. It is a multi-node, Novell Directory Services (NDS)-enabled clustering product for NetWare 5.x that supports failover, failback, and migration (load balancing) of individually managed cluster resources.

You can configure up to 32 NetWare 5.x servers into a high-availability cluster using NetWare Cluster Services.This means resources can be dynamically moved to any server in the cluster.

There are special considerations when using Fibre Channel and NetWare. Consult Chapter 2 “Preparing for installation” in *IBM Netfinity FAStT Storage Manager Version 7.02 for Windows Installation and Support Guide* for more information.

Figure 33 displays three nodes connected to a Fibre Channel switch using a common disk subsystem. Each server has its own local SYS volume.

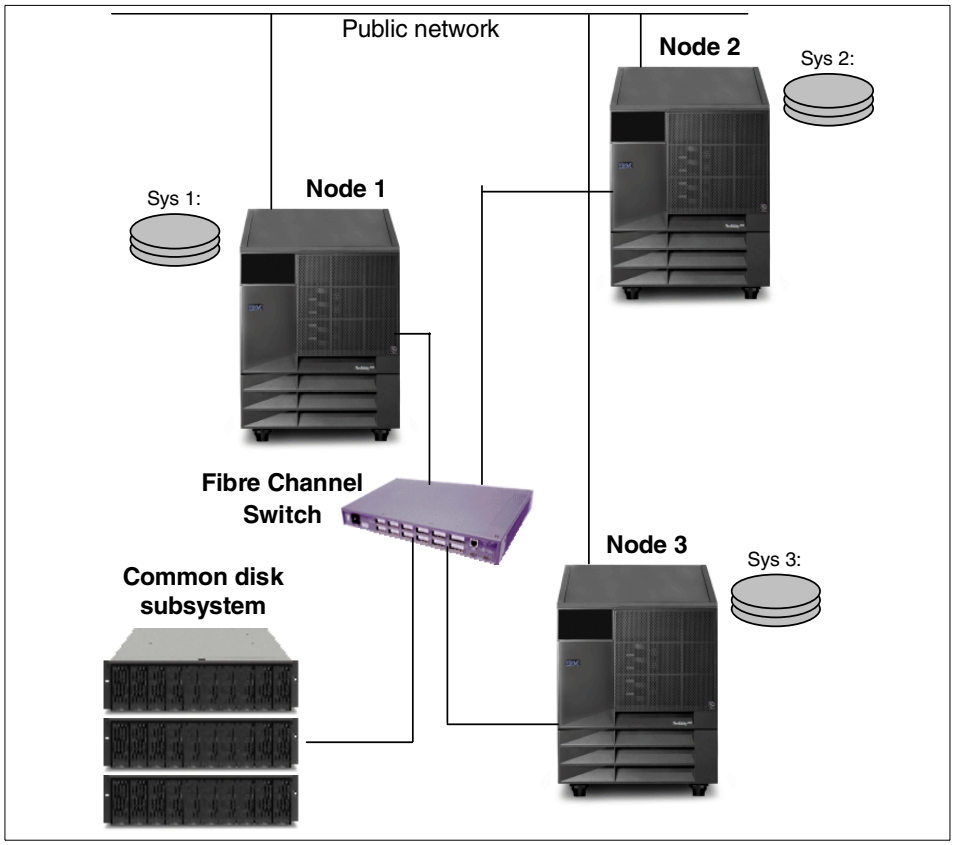

Figure 33. NetWare Cluster Services

3.9.1 Requirements for NCS

Before installing Novell Cluster Services, there are several requirements that must be reviewed and implemented. Understanding these requirements will help the installation process flow smoothly. Failing to comply with these requirements could give you undesirable results.

The following must be met:

- Each server should have a minimum of 64 MB of memory, but 128 MB is recommended.
- The common disk subsystem must be a Fibre Channel configuration. ServeRAID is currently not supported.
- All servers in a cluster must be in the same NDS tree and must be replicated on at least two servers.
- Each server must have at least one local disk device for a SYS volume.
- The common disk system volumes must be configured to use Novell Storage Services.
- RAID-5 or RAID-1 should be used to protect the system against failure in the common disk subsystem.
- NetWare 5.0 requires Support Pack 4 or greater. NetWare 5.1 requires Support Pack 1or greater.
- NCS requires that all servers in the cluster be configured with the IP protocol and be on the same subnet.
- On the management workstation, Novell Client Version 4.50.819 or higher for Windows NT or Version 3.0.0.0 or higher for Windows 95/98 must be installed. The workstation should have at least a 300 MHz processor and 90 MB of memory.
- All nodes must be running the same version of NetWare.

3.9.2 Failover and failback

You have a great deal of flexibility in defining how the resources in the cluster behave when problems occur. When a node fails, the clustered resources it was hosting are migrated to the remaining nodes in the cluster. The target node for each affected resource depends on the failover order defined in the properties of the individual resources.

In the event of a server failure, you can configure resources to be moved automatically or you can move them manually when you need to troubleshoot hardware or balance the workload.

Although automatic failover is appealing, manual failover allows you to make decisions based on the circumstances at hand. For example, it may be that some servers are heavily loaded at particular times and you would prefer not to add extra workload to them during those periods. Using manual failover, you can decide how best to handle a failure, perhaps allowing less important

failed resources to remain unavailable until the workload on the cluster falls back to normal levels.

Failback can also be handled either manually or automatically. Automatic failback ensures cluster resources automatically move back to their preferred node when the system becomes available. Resources do not automatically fail back to any other server in the cluster apart from their preferred node.

Manual failback is fundamentally the same as manual failover; you choose exactly when you want a resource to return to its preferred node.

3.9.2.1 Resources and resource scripts

Resources are defined through the ConsoleOne graphical user interface (GUI). Templates are provided for selected resource types, which makes setting up those resources a quick and easy task. Resources without templates require some extra work to define, load, and unload scripts that control how resources are started and stopped. Templates include predefined load and unload scripts.

The commands required to start a service or application or to mount a volume are contained in a load script, one of which is required by NetWare Cluster Services for each resource, service, or volume in the cluster. Certain types of resources also require an unload script that gracefully shuts down the resource. Unload scripts may be required when a resource is manually failed over to ensure it is stopped on one server before being restarted on another. Commands used in the script files consist of the same commands you could use in an NCF file, executed from the server console.

As resources are defined, they are automatically distributed to members of the cluster. You can redistribute them, however, defining as many nodes as you wish as potential owners of the resource. A preferred node is defined for each resource.

3.10 Novell StandbyServer

Novell StandbyServer for NetWare is a high availability solution whose operation is very similar to an active/passive implementation of Legato Co-StandbyServer for Windows NT (see 2.2.3, “Active and passive servers” on page 12 and 3.6, “Legato Co-StandbyServer for Windows NT” on page 51).

In this configuration, one server (the primary system) is protected by a secondary machine that performs no other task than to monitor the primary’s

status in normal operation. If the primary fails, however, the secondary assumes the role of the primary so that users receive minimal loss of service.

Although perhaps not as cost effective as an active/active solution, it has the following useful features:

- Using identical servers means that performance in a failed-over state is the same as in normal operation.
- There is no restriction on server type (as long as NetWare is supported). An older, less powerful system that might otherwise be decommissioned can be used as the standby system. Although performance will not be as good in the failed-over state, your users can keep working.
- If you have critical applications, there is the option to configure multiple secondary machines. Although only one secondary system can be nominated for automatic failover, in the unfortunate case where the failover machine has problems itself, one of the other secondaries can be brought online manually, thus reducing the impact on users.
- Although StandbyServer is an active/passive solution as far as clustering is concerned, you can run other server tasks on the secondary machine in what Novell calls a *Utility Server* configuration. Examples of suitable tasks are file shares, print spoolers, fax servers, and so on. The important point to remember is that these tasks will not survive a failure of the secondary machine. In addition, Utility Server tasks will be interrupted and have to be restarted when a failover occurs.

3.10.1 StandbyServer requirements

StandbyServer is a software only product. You can take any two servers running NetWare today and implement StandbyServer clustering as long as you have sufficient physical disk and memory configured in the machines. It is recommended, however, that an extra network adapter be added to each machine to form a private link between the systems for mirrored data traffic.

Here, is a list of the specific requirements for StandbyServer:

- Two servers, each capable of running NetWare (both systems must be running the same version). All current versions of NetWare are supported by StandbyServer. Remember to check that your Netfinity servers support the selected version of NetWare by checking the IBM Netfinity ServerProven Web site:

  `http://www.pc.ibm.com/us/netfinity/serverproven`

  If you do not plan to run a Utility Server (that is, the secondary server really is doing nothing besides monitoring the primary), a two-person

run-time version of NetWare is shipped with StandbyServer for use on the standby machine, saving the cost of a license.

- Sufficient memory in each machine. StandbyServer requires a maximum of 1 MB more than the amount needed to run your NetWare environment. The standby machine needs the same amount of memory (or more), since it has to take over the full role of the primary when a failover occurs.
- The two systems must have at least one LAN connection between them. A second, private link is recommended for redundancy. You must use NetWare-supported adapters that are compatible with IP, IPX, or VIPX (VIPX cannot be routed) for both links. A high-speed wide area network (WAN) connection, such as 100 Mbps Ethernet, can be used for the private link as long as it has low latency, making the implementation of disaster-recovery solutions relatively simple.
- The disk configurations of the two machines must be similar. The secondary machine needs its own boot partition and sufficient disk space to hold mirror copies of the partitions on the primary machine. An extra physical disk is required if you wish to run the secondary machine as a Utility Server. More information on disk configuration is given in the following section.

3.10.2 Disk configuration

StandbyServer for NetWare duplicates data that resides on the primary machine using standard NetWare mirroring. Because NetWare mirroring is used, the restrictions are primarily those imposed by NetWare itself:

- The hard drives in the primary machine and standby machine may be different sizes, but partition sizes must be identical.
- Disk devices can contain both a DOS partition and a NetWare partition. Normally only the first drive contains a DOS partition as this is used to start the machine. NetWare mirrors only NetWare partitions, not DOS partitions.
- Only one NetWare partition can exist per disk device. That partition may contain more than one volume, or may contain only part of a volume. NetWare allows up to eight volumes on a single NetWare partition, or a single volume can span up to 32 different disk devices with NetWare partitions.

NetWare 5

NetWare 5 supports multiple NetWare partitions on a single drive.

- Physically, a NetWare partition consists of the data area and a small hot fix spare area. NetWare mirrors the data area only. In this way the hot fix area is maintained separately for each part of the mirrored set.

3.10.3 Novell StandbyServer operation

In normal operation, the primary server behaves as a standard NetWare server with mirrored disks. Any data written to the server is transferred to the disks physically attached to the primary server and a mirror copy is written to the equivalent partition on the standby machine by way of the dedicated link. As well as writing the mirrored data to its own disks, the secondary machine is waiting, checking for the presence of the primary. A typical configuration would look similar to that shown in Figure 34:

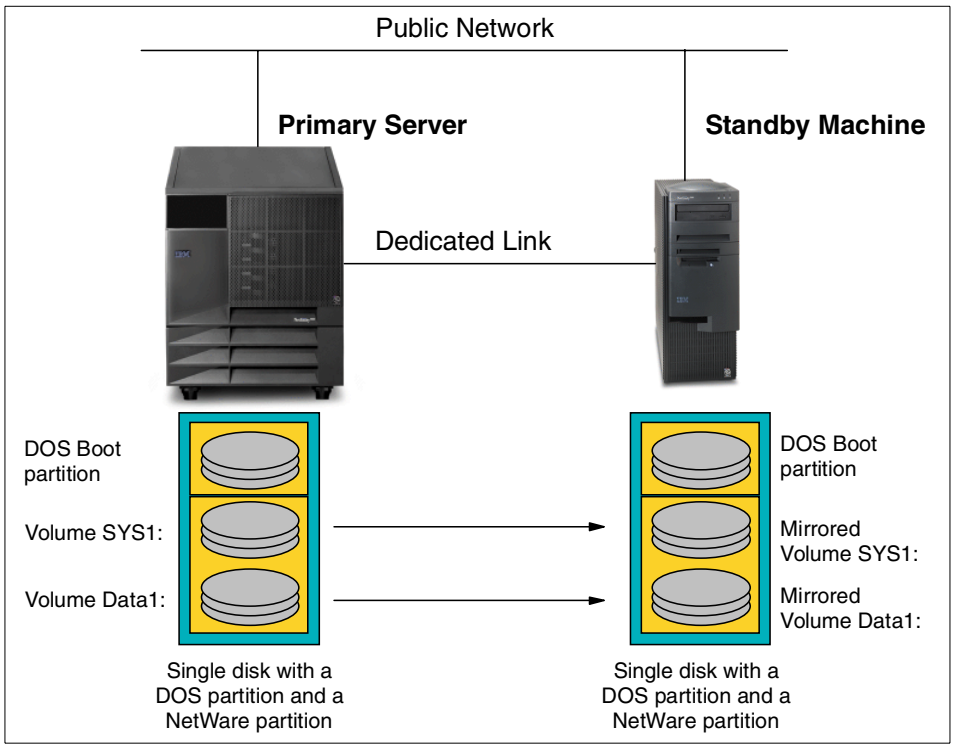

Figure 34. Novell StandbyServer configuration

When a fault occurs on the primary server, StandbyServer detects the problem and initiates a failover. A fault on the primary server is indicated when the standby system has not received IPX `ping` packets from the primary through either the dedicated link or the LAN link for 40 seconds (this is the default). A NetWare Loadable Module (NLM), called AutoSwitch, is executed

on the standby server to commence a failover. Failover is achieved by bringing down the standby server (either Utility Server or the run-time version) and then bringing up the server that was running on the primary machine, using the mirror copy of the primary server's volumes. The time to complete the switchover depends on exactly what software has to be loaded.

When the failover occurs, the standby server loads SBWARN.NLM that broadcasts a message to all users. If you wish to perform more complex alerting in conjunction with failover events, IBM Netfinity Manager provides sophisticated options for notification, including paging and contacting other LANs to notify administrators. A copy of Netfinity Director is shipped with each xSeries and Netfinity system.

3.10.3.1 Effect on clients

If your client machines are running older NetWare shell (NETX) or DOS Virtual Loadable Module (VLM) client software, they will be disconnected from the network when a failover occurs. To regain access to network resources, they must log in again when the standby machine has restarted. Because the standby has assumed the role of the primary, clients log in normally. Clients using newer 32-bit software will reconnect to the server without user intervention.

3.10.4 Recovery

Following a failover, you will want to bring the failed machine back into service as soon as possible after it has been repaired. The Novell StandbyServer product manuals describe several different scenarios for handling recovery. The most straightforward is to bring the primary server up in standby mode, have the disks re-establish a good mirror and then shut both servers down. Restart the standby server in standby mode and then the primary server as the primary. Both sets of disks will have all of the data mirrored, so no data will have been lost.

Important

It is important to bring down the server that is currently in the primary role before bringing down the standby machine. If the standby system is brought down first, the mirror will be broken. Similarly, it is important that the standby machine is brought up before the primary system.

3.10.5 Installation and maintenance considerations

- The latest NetWare Service Pack (3a or up) is required to be able to remove the system (SYS) volume from the standby server. If you are unable to remove SYS during installation, you will have to reinstall NetWare.
- StandbyServer is not compatible with the INETCFG utility. LAN drivers must be loaded before the SYS volume is loaded.
- Be sure to move any commands that are placed in the AUTOEXEC.NCF on the primary server to the AUTOEXEC.NCF on the standby server. Also, make sure that any products you install reference the AUTOEXEC.NCF in the NWSERVER directory, not in SYS:SYSTEM.
- StandbyServer installation disables the automatic ABEND recovery in NetWare 4.11. This prevents the standby server from coming up at the same time the primary server is still running.
- If StandbyServer is installed during the installation of NetWare, make sure that the primary server holds the master copy of partitions and that it is the single reference time server. Do not allow the standby server to hold a master replica.
- Mirroring times can be long with the large disks often used in servers. Running the dedicated link at less than maximum speed can significantly lengthen mirror times and affect server performance under heavy data write loads. To ensure the dedicated link you use is running at the fastest speed, check your system and adapter documentation. Some Ethernet adapters, for example, require command-line options to make them run at full duplex and at 100 Mbps.
- NetWare Storage Services (NSS), introduced with NetWare 5, allows much more flexibility in the way you organize your disk subsystems. Unfortunately, NSS does not support mirroring at present, so NetWare 5 servers cannot use NSS if you wish to implement StandbyServer. Novell has said that mirroring for NSS will be available in the future. For more information about NSS go to the following Web site:

  `http://www.novell.com/documentation/lg/nw5/docui/index.html#../usfile/nss__enu/data/hcf8v0n5.html`

3.10.6 Comparing clustering solutions for NetWare

In Table 5 we compare the two high availability solutions available for NetWare 5.x.

Note: The NetWare clustering solutions that do not support NetWare 5.x are not covered here. These include:

- Novell SFT III for NetWare
- Novell StandbyServer Many-to-One
- Novell High Availability Server

Table 5. Comparison of NetWare 5.x high availability products

Feature	NetWare Cluster Services (page 83)	Novell Standby Server (page 86)
Full application failover support	Yes	Yes
Protects data in memory not yet written to disk	No	No
Back up open files	Yes	Yes
Active/Active configuration	Yes	No
Failover detection delay	Complete failover can take place in a few seconds	Up to two minutes + volume remount time
Maximum number of nodes	32	2
Identical hardware required	No	No
Common disk storage supported	Yes	Yes
Single-system image client view	Yes	No
NetWare versions supported	5, 5.1	3.12, 3.2, 4.11, 4.2, 5, 5.1

3.11 Disks: common subsystem or mirroring?

In 2.2, “Types of clusters” on page 10, we discussed a number of ways in which clusters may be classified. Now, in this chapter, we have looked at several specific solutions that provide clustering functionality through the operating system. Essentially, they have all been shared nothing systems, implemented either by utilizing a common disk subsystem or by mirroring data to a second set of disks in a system that will take over if the server holding the primary data disk fails. The mirror disk is not normally available to the standby system, only becoming so when the primary system fails.

Each approach has its own particular strengths. The following sentence offers some thoughts for your consideration.

3.11.1 Common disk subsystem

Clusters implemented with a common disk subsystem usually offer excellent performance and availability.

- Performance

  Clusters with common disk subsystems are often true active/active clusters, so both servers execute their own workload. Each system can take over the load of the other in the event of a failure. In contrast, many mirroring solutions assume one-way failover, from a primary server to a standby or backup server.

- Availability

  Common disk clustering systems can monitor and recover individual applications. Most mirroring solutions only monitor for total server failure. Also, the ability to transfer ownership of common SCSI disks ensures that applications always start up with exactly the same disk-based data as at the moment of failure. Mirrored solutions have a finite amount of time during each mirroring operation in which the local and remote disks are out of synchronization.

Cost can also be an important factor. For large disk installations, the additional cost of an external enclosure can be more than compensated for by savings in the number of disk drives required for a common RAID-5 array disk subsystem in comparison with a mirrored configuration of an equivalent capacity.

3.11.2 Mirrored disks

The main strengths of mirrored-disk failover solutions are the ability to do live backup and the flexibility of location of the cluster members.

- Live backup

  Mirrored solutions create a near real-time second copy of data, possibly at a remote disaster recovery site. Common disk subsystem solutions can resolve the single point of failure issue by implementing RAID and other redundant technologies in the disk subsystem. Some mirrored systems, however, offer the additional advantage of being able to back up open files from the mirrored copy.

- Distance

  Mirrored-disk failover solutions have virtually no distance limitation between the primary server and the recovery server. In a shared disk cluster, the two servers can be no further apart than allowed by the common disk connection, which can be as little as a few feet unless relatively expensive disk subsystems, such as Fibre Channel, are used.

3.12 Summary

The number of Intel-based clustering solutions available is constantly growing as the need to support mission-critical applications on these systems becomes prevalent. As we have seen in this chapter, many of these solutions provide only a subset of true clustering.

High availability is the dominant characteristic found in today's operating system clustering offerings. Although a vital function, high availability by itself is not a full clustering solution. The other clustering benefits of scalability and workload balancing are generally not provided or are present only in rudimentary form. Microsoft Cluster Server and NetWare Cluster Services can provide some limited workload balancing at the operating system level. This is changing, and new product announcements in the coming months will begin to address these important areas.

Chapter 4, "Application clustering" on page 95 discusses some clustering solutions that do offer scalability, load balancing or both for specific applications today.

Chapter 4. Application clustering

Most applications are written to run on a single machine. Some, however, particularly those intended for execution on a server, are written to take advantage of the multiple processors available in a symmetric multiprocessing (SMP) machine. An SMP-aware application divides its tasks into separate threads that can be executed in parallel with each other. The SMP machine's operating system then distributes the application threads among the system's processors.

The problem is, SMP machines eventually run into performance bottlenecks that prevent them from scaling as processors are added (see 2.1, "The promise of clustering" on page 5). Clustering is regarded as the way to improve performance beyond that attainable through SMP. As we have seen in Chapter 3, "Cluster at the operating system level" on page 19, however, today's Intel-based servers offer little in the way of scalability when clusters are implemented in the operating system.

As a way of providing the advantages of clustering in this marketplace, several server application vendors have implemented proprietary forms of clustering within their applications.

This chapter looks at some of the leading applications that offer clustering solutions for Intel-based servers.

4.1 Lotus Domino clustering

The Domino Server family of products offers an integrated messaging and Web application software platform. It has particular strengths for companies that wish to implement workflow processes and collaborative working to improve responsiveness to customers and streamline business processes. Domino is also effective as a Web server and, in its latest version (R5.0), integrates well with Internet standards for mail and other services. Applications based on Domino can serve a variety of users, including Notes clients, standard Internet e-mail clients, and Web browsers.

Domino is the world's most widely deployed messaging and application platform, based on a proven architecture that has been in production for more than 10 years. It is the foundation for many of the largest networks in use, some supporting more than 150,000 users.

For more detailed information on Domino and Domino clustering, refer to *Lotus Domino R5 Clustering with IBM @server xSeries and Netfinity Servers*,

SG24-5141, and *Netfinity and Domino R5.0 Integration Guide*, SG24-5313, from which much of the information in this section was obtained.

4.1.1 Introduction to Domino clustering

Because of its central role in many businesses, the cost of Domino server downtime can be great. In common with other cluster solutions, Domino's clustering capability enhances the availability of resources, in this case Domino-based business applications. In addition, however, Domino clustering can also be used to provide scalability and load balancing.

Clustering has been a feature of Domino for several years. Lotus first released Domino clustering as a Notes for Public Networks offering for telecommunication and Internet service providers as part of Domino R4.0. Several of the larger telecommunications and Internet service providers are using clusters of as many as six Domino servers to provide Domino failover support to their customers.

General customers gained the benefits of Domino clustering (with the Domino Advanced Services license) as part of Domino R4.5. Now, with Domino R5.0, clustering will ship with the Domino Enterprise Server. In this latest release, Domino clustering has been enhanced to support failover and load balancing of Web browsers (both Netscape 3.x and later and Microsoft IE 3.x and later) as well as Notes clients.

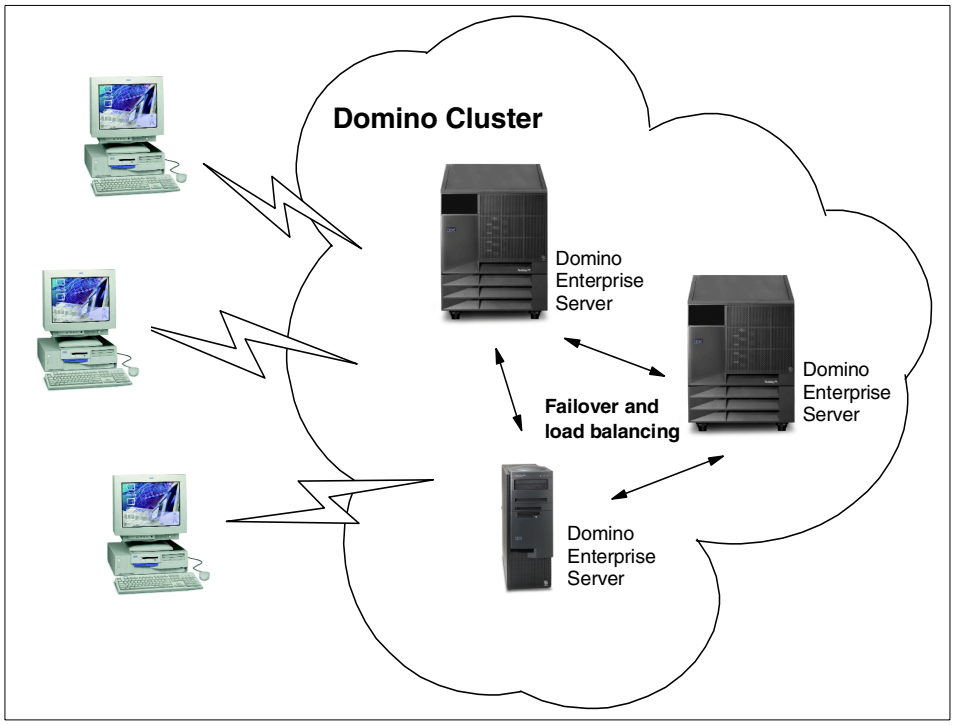

Figure 35. Domino clustering

Domino clustering provides content-based routing, that is, intelligent failover and load balancing of client requests are performed based on the content requested by the user. Cluster replication maintains synchronized replicas of databases that are critical to a business and which must be highly available. When a user requests data from a server in a Domino cluster, information about the whereabouts of replicas and server availability is used to direct users to an appropriate server that contains the content desired by the user. This is a dynamic process that can ensure that no single server is overloaded with requests for frequently accessed data. Load balancing in this way is an integral part of the operation of a Domino cluster.

Cluster replication

If you are familiar with Domino, you will know that replication is an important function that allows replicas of databases to be distributed among servers that may be physically separated by great distances. Cluster replication is a special form of Domino replication, unique to clusters.

The basic difference between the two forms of replication is:

- *Conventional replication* is scheduled. At predetermined intervals, changes to a database are transmitted to its replicas.
- *Cluster replication* is event driven. Each time data is written to a database, all cluster replicas are updated.

No specific certification is necessary for hardware that you wish to use in a Domino cluster. Any server that supports Domino can participate in a cluster. Clustering with Domino is particularly unusual in that you can cluster heterogeneous operating system and hardware platforms. A single cluster, for example, can comprise servers running Domino on Windows NT/2000, OS/400, UNIX and Linux; these operating systems could be executing on a Netfinity server, an AS/400, RS/6000 and a System/390.

A Domino cluster can have between two and six member servers (nodes).

4.1.2 Who uses Domino clustering and why?

Today, customers use Domino clustering for the following reasons:

- High availability of critical applications
- Increased scalability of workloads
- Migrations of platforms, OS, or Domino versions
- Disaster recovery processes

Domino applications are organized around databases stored on servers. For customers who have critical Domino applications with a target of 24 hours a day, seven days a week availability, Domino clustering can help. Multiple replicas of the application databases are distributed among the cluster members. Users accessing servers that are too busy or otherwise unavailable are redirected to a server with a replica copy of the requested database.

This means that, as the workload on a server within a cluster increases, at some predetermined point additional load is passed to another machine. If necessary, an additional server can be introduced to the cluster (up to a

maximum of six systems can participate in a single cluster), and additional replicas of application databases created on the new server to handle the extra load. In this way, Domino clustering provides real scalability.

Domino clustering can also assist in migrations of hardware and software. One customer was able to move a Domino server farm, which serviced several thousand users, to another geographic location. The customer introduced additional servers into an existing cluster, siting the machines at the new location. By restricting the servers in their old location the users were automatically failed over to servers in the new location and the original servers were then taken out of the cluster. Users saw no interruption in services.

Other customers have completed server consolidations and platform migrations by combining smaller servers into one large Domino server through clustering. All servers were incorporated into a Domino cluster and users were then redirected from the smaller servers to the one larger Domino server with no disruptions.

Some customers are implementing disaster recovery plans by using Domino to cluster servers in geographically dispersed locations across a wide area network (WAN). With the appropriate network connectivity, this approach can provide a very cost-effective solution to protect you against natural hazards or intentional damage.

Note

Clustering across a WAN requires a guaranteed high-speed, low-latency connection between sites.

4.1.3 System requirements

To successfully set up a Domino cluster, a number of system requirements must be met. Here is the list of prerequisites for your servers:

- All servers in a cluster must be running Domino Release 5.0 or 4.62 Enterprise Server license, or the Domino Release 4.5 or 4.6 Advanced Services license.
- All servers in a cluster must be connected using a high-speed LAN. You can also set up a private LAN just to carry cluster traffic.
- All servers in a cluster must use TCP/IP, be on the same Notes named network and use the same set of network protocols.

- All servers in a cluster must be in the same Domino domain and share a common Domino Directory.
- You must specify an administration server for the Domino Directory in the domain that contains the cluster. If you do not specify an administration server, the Administration Process cannot change cluster membership. The administration server does not have to be a member of the cluster or be running the Enterprise Server license.
- Each server in the cluster must have a hierarchical server ID. If any servers have flat IDs, you must convert them to hierarchical IDs to use them in a cluster.
- A server can be a member of only one cluster at a time.
- Each server must have adequate disk space to function as a cluster member. Servers in clusters typically require more disk capacity than unclustered Domino servers because clustered systems normally have more database replicas.
- Each server must have adequate processing power and memory capacity. In general, clustered servers require more power than unclustered servers.

Clients, too, have some requirements to work correctly with clustered servers:

- Notes clients must run Notes Release 4.5 or higher to take advantage of the cluster failover feature.
- Clients accessing a server in a cluster should be using TCP/IP.

4.1.3.1 Mixed Domino cluster environments

In a mixed Release 4.6/Release 5 cluster, users cannot have an R5 mail file on both R4 servers and R5 servers. The R5 mail template does not work properly on an R4 server. In this situation, use the R4 design for mail files or make sure that you place users' mail files only on R5 servers if you want to use the R5 design.

The R5 mail template will not function correctly on R4 servers because cluster replication ignores selective replication formulas and you cannot prevent the mail file design from replicating to other clustered servers.

To avoid potential problems with your mail files, we recommend all servers in a cluster are running the same release of Domino.

4.1.4 Cluster planning considerations

The planning process can be broken down into a number of steps:

1. Define the services required.
2. Define database characteristics.
3. Define user characteristics.
4. Define your hardware.

4.1.4.1 Which services are required?

The first step to take when planning to use Domino clustering is to define which server services are required. Some environments will need only base mailing and application database services. Others may require, for example, the Domino HTTP server or other services.

Look carefully at your application set to decide which are essential. For example, you may decide your users cannot do without access to their mail, but that the company intranet can withstand some downtime. In this case, the HTTP load can be placed on a server outside the cluster holding the mail databases.

At this point, it is also worth considering what operating system services are not required. Any unwanted services should be disabled to avoid unnecessary loads on the servers.

4.1.4.2 Defining database characteristics

The next step is to define your database characteristics. Taking the time to understand database characteristics will enable you to make the correct decision when estimating the hardware required to provide a particular level of service.

The important characteristics are:

- Number and type of databases in your environment
- Size of the databases
- Expected volume of new data, records added, updated, or deleted per day, week, or year
- Time sensitivity of data
- Number and distribution of database users
- Location of database replicas across Domino servers
- Network connections between servers

To make the definition process easier, create a spreadsheet table containing all the databases in your company and define the characteristics that are relevant for planning clustering. The following table illustrates the information that should be collected:

Table 6. Database characteristics example

Database name	Maximum number of concurrent users	Transaction rate	# replicas (standard, not cluster)	Database size	Database growth	Need for high availability
Help Desk application	10	Medium	1	1.2 GB	Medium	Medium
Discussion database	500	High	2	2.5 GB	Fast	High
Business application	350	Medium	2	800 MB	Fast	Very High
Web application	200	Medium	1	1 GB	Medium	Medium (Only available in Domino 5)
Link library	300	Medium	1	600 MB	Slow	Low

After creating the table it is easier to judge which databases need high availability and what the expected disk consumption will be. In this example, there are two databases that are growing fast and need high availability. One of the databases is crucial for the business.

Although up to six servers may be included in a single cluster, it is unusual for a database to be replicated to that number of servers. For most databases, two replicas is sufficient. You might want to have three replicas of an important database such as the business application in Table 6. However, increasing the number of replicas also increases cluster replication traffic, so this needs to be taken into account when making these decisions.

Using mail as an example, an efficient way to use a six-node cluster would be to split your users into three groups, each of which has its own mail database. Place each mail database on a separate server, then replicate each database to another cluster member that does not yet hold a mail database or a copy of one. Each server now has a single mail database and each user is protected against a server failure. The mail workload has been distributed evenly among the cluster members.

4.1.4.3 Defining user characteristics

The workload a user creates on a server depends a lot on the user's familiarity with Lotus Domino. Beginners tend to use simple operations in comparison with more advanced users, who may use functions that require

more system resources such as sending mail with large file attachments, and performing full text searches in databases.

Users can be roughly divided into four main groups depending on the type of operations they perform, although often a user's workload is represented by a combination of several categories. The four categories of user workload that we have defined for this example can be described as follows:

- Mail only

  Users who send and receive mail but do not access any other databases, or use Notes calendar and scheduling functions.

- Mail with calendar and scheduling

  Mail users who also use calendar and scheduling functions.

- Database

  Users who are only performing heavy shared database operations. They perform view operations, navigate unread documents, make additions and updates to documents, and run full text searches in a shared database.

- Web user

  Users who are accessing the Domino server using a Web browser. Web clients can use HTTP, POP3, LDAP, or IMAP protocols to perform functions. The corresponding task must be running on the Domino server to enable the protocol. Web user workload on the Domino server is different from a typical Notes client workload, since an HTTP client typically opens a session to a server to transfer data, and closes the session when the operation has been completed. A Notes client opens a session to a server once, and keeps it open until it is explicitly closed, or if the timeout value is exceeded. The actual workload difference depends greatly on the protocol being used, and the type of database and application accessed on the server.

The following table presents what is considered a typical distribution by user type. Web users have not been included because it is difficult to quantify them. Depending on the amount of server processing required and whether external users can access your Web site, it may easily provide a greater load than your internal Notes users:

Table 7. User characteristic breakdown by type

User type	Percentage of total users
Mail only	10%
Mail with calendaring and scheduling	60%

User type	Percentage of total users
Database	30%

4.1.5 Hardware for Domino

When planning hardware for Domino clustering, the same principles apply as for a standard Domino server. The main differences are that you must reserve some extra memory and CPU cycles for the cluster tasks, and allowance must be made for the potential increased workload that may occur if clients are redirected from other servers.

Servers within the cluster do not have to be identical. For example, in a two-server cluster, the first system could be a Netfinity 7000 M10 with a RAID disk subsystem, ECC memory, and four Pentium III Xeon processors installed, while the second server could be a Netfinity 3000 uniprocessor machine with standard SCSI disks. The servers could even be running different operating systems.

Hardware sizing

This section gives some guidelines for configuration of your server. For detailed information on sizing and tuning IBM Netfinity servers for a Domino environment see *Netfinity and Domino R5.0 Integration Guide*, SG24-5313.

The essential question is whether the hardware is capable of offering the required level of performance. Hardware bottlenecks are most often one or more of the following:

- Memory
- CPU
- Disk space
- Networking

4.1.5.1 Memory

Memory can become a bottleneck in a Domino environment, especially when users perform more advanced functions. The operating systems that support Domino provide virtual memory that uses a page file on a hard disk to extend the capacity of the system beyond that of the physically installed memory. Accessing the disk is much slower than accessing memory, so performance suffers. Avoiding excessive paging is of paramount importance in achieving good performance.

You can estimate the memory requirement for active users by using the following formula:

```
Installed memory = Recommended basic memory + (Number of concurrent
users / 3) MB
```

where the recommended basic memory when running Lotus Domino server on Windows NT 4.0 is 64 MB or more. This algorithm is appropriate for mail and application servers, and mail hubs. Users in this case are connected, active users, and not those users whose mail files are located on the server. There can be 1,000 users with mail files on a given server, but usually there are significantly fewer users active at any one time. A useful approximation, based on analysis by IBM, is:

```
Number of actively connected users = Number of registered users / 3
```

Additional memory may be required for performance reasons (for example, if you run additional replicators). Information on sizing your IBM Netfinity server for performance can be found in *Netfinity and Domino R5.0 Integration Guide*, SG24-5313.

4.1.5.2 CPU

In practice, Domino clustering requires at least a Pentium level processor or better. The decision to use a faster CPU, or to use SMP, will be influenced by the server and cluster workload. If the cluster is used as a mail or database server, increasing CPU power will probably not provide a significant improvement in response time because a bottleneck is likely to be due either to memory or disk I/O bandwidth limitations. Alternatively, if the server is an application server, where LotusScript or Java applications are heavily used, additional CPU capacity will probably be of benefit.

Clustering in itself does not have an impact on the performance of the individual servers. The amount of server-to-server communication that is used to communicate status among the cluster's members is inconsequential. After incorporating servers into a cluster, given that the workload remains the same, the CPU utilization will also remain the same. Additional server CPU workload will be applied when replicas of databases are created.

In the case where two existing mail servers are put into a new cluster of two, and all the mail databases are replicated on the other server, the CPU utilization will increase approximately additively. As an example, assume we have two identical servers with CPU utilizations of 35% and 40% respectively. When these systems are combined in a cluster, with event replication enabled on all databases, you can expect CPU utilization to increase to approximately

75%. In practice, utilization may be around 65% because not all of the CPU load on each server is related to updating the replicated databases.

When servers are clustered, cluster replication keeps the information in the databases synchronized in all replicas. In the case of our example, if one server becomes unavailable, the remaining server, which has up-to-date copies of all databases, has to support all users. While the failed server is down, the other server will handle all user requests, and cluster replication will not take place. Each server has to be able to handle this full load.

Often the only way to see if the CPU is the bottleneck is to set up a test environment and monitor the CPU load under real circumstances. In Windows NT, monitoring is made easy by using the Performance Monitor program.

If the processor utilization is constantly close to 100%, consider either spreading the server workload within the cluster or adding another processor. Domino server can take full advantage of the SMP environment. High CPU utilization can also be a result of excessive paging, a malfunctioning adapter card, or simply too many applications running on the server. Windows NT's Performance Monitor can help you determine how much CPU each application is consuming and the paging rate. If the paging rate is high, adding more memory will improve performance. See 4.1.5.1, "Memory" on page 104 for more information.

4.1.5.3 Disk space

When you are configuring hard disk space for Domino clustering, the following areas need to be considered:

- Operating system requirements
- Paging file location and size
- Lotus Domino program files
- Lotus Domino databases
- Databases replicated from other servers in the cluster

Using Windows NT 4.0 Server as an example, a typical installation will require up to about 150 MB of disk space, depending on the installation options selected. You might need to reserve some extra disk space for other system applications, such as backup tools. It is recommended that you not run any other application servers or Windows NT file and print services alongside Domino server.

When Windows NT Server starts, it reserves space on disk for a paging file to implement a virtual memory system. Virtual memory effectively increases the memory space available to the system beyond the amount installed as

physical components. As memory is used by the system, a point is reached where no more physical memory is available. To service an additional request for memory, the system temporarily writes the contents of a block of memory not currently in use to the paging file, thus freeing it for use. When the contents previously written to disk are needed, another memory block is freed up in the same way and the required data is retrieved from the disk. Disk reads and writes are much slower than direct access to memory, so paging in this way is not good if it happens constantly. Virtual memory does, however, allow the server to cope gracefully with peak demands that would otherwise overload the system.

We recommend a paging file size that is 10% greater than the amount of physical memory installed in the machine:

```
Paging file = Physical memory installed * 1.1
```

Lotus Domino with clustering requires about 200 MB of disk space for the program code and standard databases. You can reduce this somewhat by not installing the Notes help databases on every server. After defining the database characteristics, you should be able to define the disk space requirements for the databases. Remember to reserve enough disk space for the replicas from other servers in the cluster as well.

For mail databases, we recommend reserving around 50-100 MB of disk space for each Domino mail user. This will accommodate typical usage of mail with liberal use of attachments and calendaring and scheduling.

Performance tips

- Make sure that paging is minimized. Memory utilization should generally stay below, say, 70%.
- Place the Notes program files on a separate disk from the Notes database files.
- Try to avoid placing databases with high transaction rates on the same physical disk as each other.
- Place the transaction logging files on a separate disk from the Notes database files (Domino R5.0).

More performance tips can be found in *Netfinity and Domino R5.0 Integration Guide*, SG24-5313.

4.1.5.4 Networking

Clients can connect to a cluster with any Domino-supported protocol, such as TCP/IP, SPX/IPX, or NetBEUI, but the cluster replication tasks require TCP/IP. The servers in a cluster must all have the same set of protocols installed. For example, if two existing servers are configured as a cluster, and clients have been using SPX/IPX to connect to one system and NetBEUI to connect to the other, both servers must install SPX/IPX and NetBEUI for client traffic and TCP/IP for the cluster's internal communications.

Lotus recommends that TCP/IP be the only protocol used in a clustered environment. Failover is optimized for TCP/IP and therefore takes place faster than when using, for example, the NetBEUI protocol. Because of this, we will focus on the TCP/IP protocol from here on.

Domino clustering does not limit your choice of network. Domino clustering supports all technologies supported by the underlying hardware and operating system platform. Windows NT supports all the major networking offerings including token-ring, Ethernet, FDDI, and ATM. As long as the network adapter is configured and functioning, and the required protocol settings are complete, Domino clustering can operate.

Cluster replication is constantly updating replicated databases over your network. Although these transactions are primarily small records, the additional traffic may be a performance concern. We recommend that you dedicate a private LAN for the cluster's internal traffic. A WAN cluster link can be used as long as it is a high-speed, low-latency connection.

4.1.6 Lotus Server.Planner

The best way to determine the optimum hardware is to set up a pilot system and tune the server by adding memory, upgrading the processor, or changing server hardware parameters. This is time-consuming and is not always an option. To help you decide on the hardware you will need, Lotus has developed a tool called Lotus Server.Planner.

Lotus Server.Planner is a set of Domino database templates and hardware vendor-provided databases. These databases are:

- DSPV.NTF (Vendor database template)
- DSPA.NTF (Analyst database template)
- DSPD.NTF (Decision Maker template)
- CPxxxxx.NSF (Vendor-provided database, where xxxxx is the vendor name)

You can download the newest release of the template files from the URL:

`http://www.notesbench.org`

These database templates are used to create new databases to help select the best hardware for the environment. Vendor databases contain the results from Domino tests.

Lotus Server.Planner allows you to enter the user characteristics and distribution, the server role and operating system, and provides a matrix of servers fulfilling the requirements. Although Lotus Server.Planner does not specifically support cluster planning, you can use it to plan the server hardware after defining its characteristics.

4.1.7 Scalability, failover, and load balancing in a cluster

Scalability, failover, and load balancing are the three primary cluster functions.

Scalability means being able to support more resources and more clients by adding additional hardware, ideally in a linear relationship so that moving from two servers to three lets you support 50% more clients. Domino offers good scalability, supporting up to six servers in a cluster. A cluster can start small, with just two servers, and be expanded as the user population grows.

We now look more closely at failover and load balancing.

4.1.7.1 How failover works in a Domino cluster

Failover provides protection for business-critical databases and servers. When a client requests data from a server that has failed or is otherwise unavailable, the cluster services the request from an alternative source.

Figure 36 illustrates the situation when a Notes client tries to open its mail file on a clustered server that is unreachable at that moment:

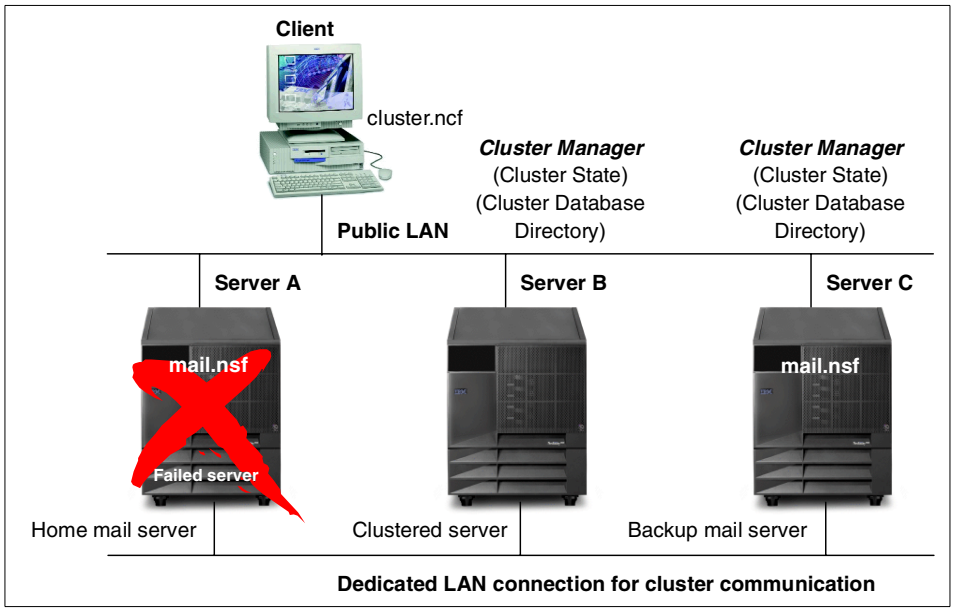

Figure 36. Example of a failover event within a cluster

The home mail server for the client is Server A, which is clustered with Servers B and C. The user's mail database has a cluster replica on Server C. Server A is down and the Cluster Managers in Servers B and C know the status of all the servers in the cluster and where database replicas are stored.

The steps involved in failover from one server to another are described below:

1. The first time a client accesses a server that is a cluster member, details of all servers in the cluster are stored in the client's server cache.

   So, assuming that the client has successfully accessed its mail at some time in the past, the client is aware that Server A is a member of a cluster along with Servers B and C.

2. The user double-clicks the desktop icon to open his or her mail database.

3. When the request to Server A fails, a `Server not responding` message is returned to the client. If Server A were not a cluster member, the user would be presented with this message. In the clustered case, however, the client searches its cluster cache to identify the other members of the cluster. The client then accesses the next server in the list.

4. The Cluster Manager of the selected server determines where the replica of this user's mail file is located. It may or may not be on the server that the client is currently accessing.

5. If the user's mail file is not on this server, the Cluster Manager returns the name of the server that contains the replica to the client, who then accesses that server.
6. The user sees the mail file opening on his or her desktop. (There are some server-opening messages displayed on the message line as usual.) An icon pointing to the replica mail file is added to the desktop.

An alert user may notice the new icon on his or her desktop, but apart from this, the failover is transparent to the user. Note that failover only occurs during a request to open a database. If the server fails after the database has been opened, the user will have to exit from the database and attempt to reopen it to get failover.

4.1.7.2 Load balancing

Load balancing is used to distribute the total server workload among the available servers in a way that maximizes the utilization of the cluster resources. Heavily used databases are replicated in the cluster, and user sessions are directed to another server in the cluster when the load on a particular server exceeds a defined threshold.

When running as part of a cluster, a Domino server constantly monitors its own workload. The Cluster Manager process automatically computes the server's availability index statistic every minute. All servers in a cluster know about the availability of all other servers in the cluster.

When the server availability index drops below the server availability threshold, the server becomes busy and, if a client tries to open a session with a request that triggers failover, he will be redirected to an available clustered server. (Note that the index has to drop *below* the threshold because it is an *availability* index, inversely related to the server's load.)

Workload balancing appears as virtually transparent to the user. If users are switched to a different server, they will just see an additional icon on the desktop, or a stacked icon for the database on the second server. They will not receive an error message or other indication that workload balancing has occurred.

4.1.8 Domino R5.0 Web clustering features

The Internet Cluster Manager (ICM) is a new Domino R5.0 server task that supports failover and workload balancing of HTTP and HTTPS client access (Web browsers) to the Domino HTTP servers running in a Domino cluster. The ICM serves as an intermediary between the HTTP clients and the HTTP servers in a Domino cluster.

A Web browser directs a request for a database to the ICM by way of a Uniform Resource Locator (URL), as for any other Web page. Meanwhile, the ICM periodically sends inquiries to the Web servers in the cluster determining their status and availability. When a client request is received, the ICM looks at the information in the Cluster Database Directory to find a server that contains the requested database. The ICM determines the most available server that contains the requested database, and then redirects the client to that server. The client then closes the session with the ICM and opens a new one with the nominated server. The user may see this as a change in the host name in the URL.

If the page that a Web server presents to a client includes links to other databases on the same server or to other databases in the cluster, the Web server includes the host name of the ICM in the link URLs to ensure that users accessing them go through the ICM.

Note that cluster replication applies only to Domino databases. This means that if there are pages in your Web site that are not generated from Domino databases, such as standard HTML data, they must be replicated by other means.

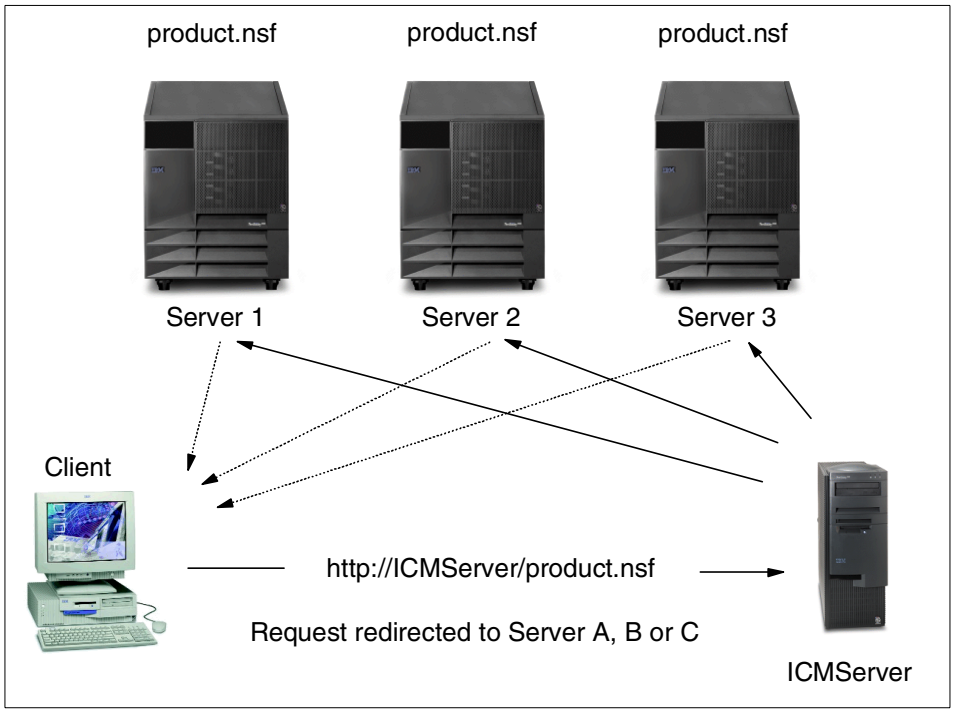

Figure 37. Lotus Domino Internet Cluster Manager

The ICM can perform the following functions:

- Monitor back-end Domino servers for availability
- Check the Domino HTTP Web service for availability
- Disallow any new connections to servers that are out of service
- Provide failover (redirect clients) to the best available server
- Provide load balancing by setting availability thresholds for your Domino servers
- Support virtual IP addresses and map ports
- Provide content routing for your clients

Internet Cluster Manager does not require any additional hardware. You can install it on your Domino server and it runs on all operating system and hardware platforms supported by the Domino server. The ICM supports client connections through TCP/IP only.

4.1.8.1 Configuration of the ICM

An ICM is dedicated to a single cluster. If you have two clusters, you must have separate ICMs for each cluster. The ICM needs to be in the same

domain as the Domino cluster because the ICM will always use the local copy of the Domino Directory. You can have multiple ICMs on a single physical machine by using Domino partitioned servers.

There are several ways you can implement the ICM:

- Dedicated server outside the cluster

  One option is to have the Domino server running the ICM dedicated to that purpose and configured to run outside of the cluster. In this environment, the ICM server would not contain any databases other than those necessary for server operation, and would run only the basic set of server tasks. Executing the ICM outside the cluster can improve the reliability of the ICM, especially if the server is dedicated to the ICM function, because the system has a lighter load than it would have as a cluster member.

  You can improve the ICM availability by configuring a second ICM for the cluster. Typically, the two ICMs will be configured in the DNS with the same host name. This way, if one of the ICMs fails, Web clients can fail over to another ICM while continuing to use the same host name.

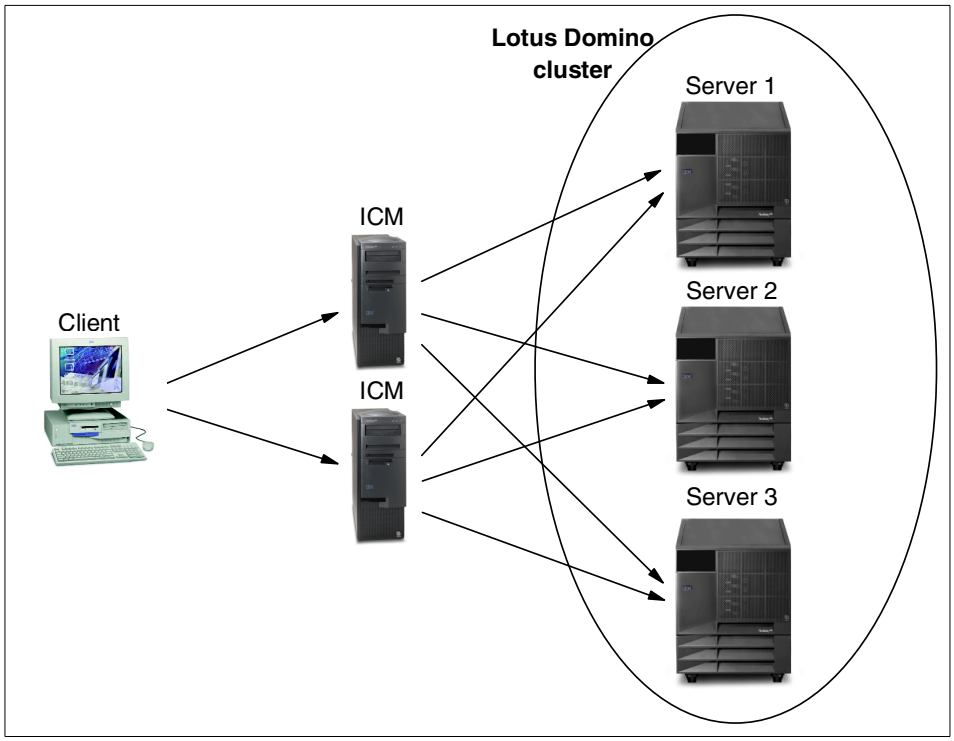

Figure 38. Multiple ICMs outside the cluster

Note

You can use operating system cluster products to provide failover support for the ICM running outside of the cluster.

- Run ICM on servers in the cluster

  Another option is to configure the ICM to run on one or more of the Domino servers in the cluster. If you choose this implementation, you should run the ICM on the most powerful server or servers in the cluster or the least loaded systems. Make sure that systems hosting ICM can handle the additional traffic.

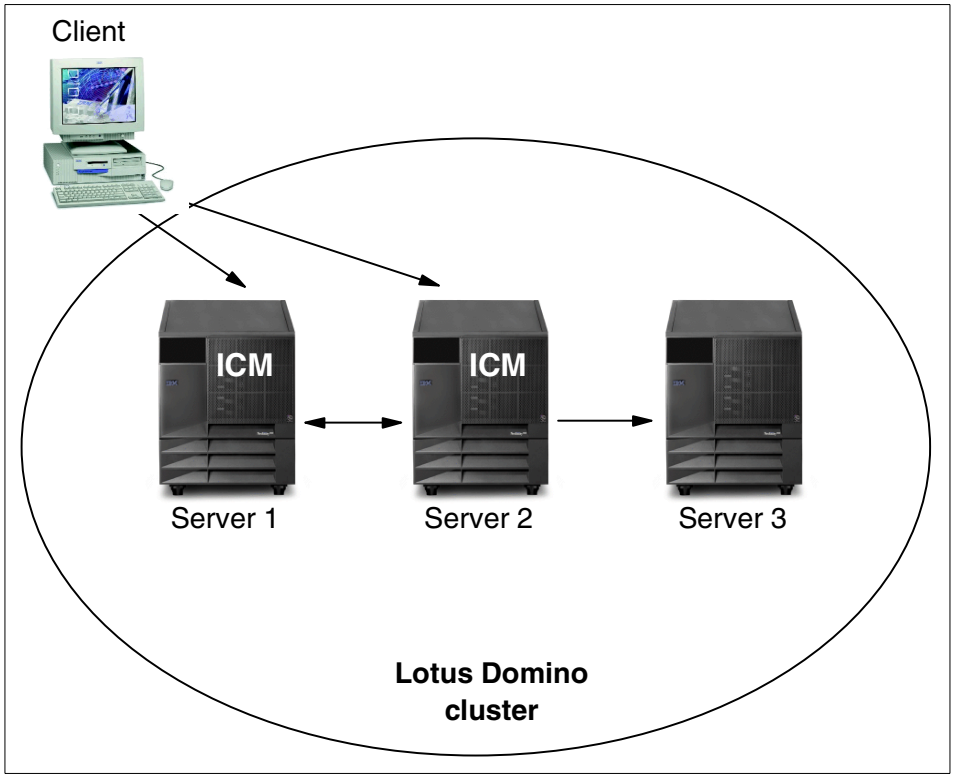

Figure 39. Multiple ICMs inside the cluster

4.1.9 Lotus Domino on Microsoft Cluster Server

Domino clusters operate in a very different way from Microsoft Cluster Server. Each approach offers unique benefits. In some situations it may be worth considering underpinning a Domino cluster with MSCS. By comparing the

features of both clusters, you can decide if one or the other offers the availability you are seeking, or whether a combined implementation is appropriate.

4.1.9.1 Comparison and considerations

Lotus Domino offers a platform-independent solution for clustering Domino servers. The Lotus solution is implemented at the application level and provides fault tolerance and load balancing for databases.

Microsoft offers Cluster Server as part of Windows NT 4.0 Enterprise Edition, Windows 2000 Advanced Server, and Windows 2000 Datacenter Server. Microsoft Cluster Server is implemented at the operating system level and is largely transparent to applications running on top of it.

The Domino cluster implementation relies on database replication between the nodes in the cluster. Replication is event driven, and ensures the integrity of data in the databases within the cluster. Microsoft Cluster Server is based on disk device sharing.

Since Domino clustering and Microsoft clustering are implemented in very different ways and based on different premises, straightforward comparison would not be fair to either of the products. The two systems have been developed for different uses. Microsoft Cluster Server offers a platform for general application servers to gain fault tolerance, but does not offer workload balancing.

Implementing MSCS clustering constrains your server platform choice, because it requires hardware that is specifically certified for MSCS.

There is nothing to prevent you from implementing both an MSCS cluster and a Domino cluster to work in conjunction with each other. This solution can be attractive, for example, if there are several geographical sites in the Domino network, and when high availability is required at each site.

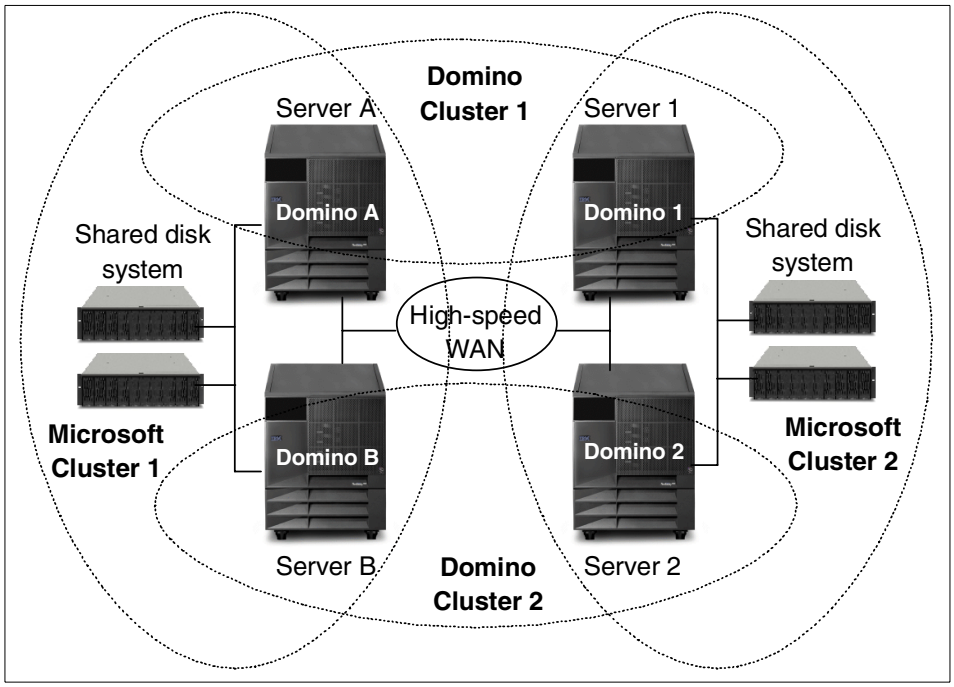

Figure 40. Combining Domino clusters with MSCS clusters

In Figure 40, loss of the high-speed WAN link does not compromise the high availability of the servers at either site.

4.1.10 Managing Domino

Managing a Lotus Domino Cluster is, in many ways, no different from managing stand-alone Domino servers. As we saw in 4.1.3, “System requirements” on page 99, there are no specific hardware requirements to implement a Domino cluster. The clustering functionality is built into the Domino R5 or R4.62 Enterprise Server license and the Domino R4.5 or R4.6 Advanced Services license. Only Domino R5 supports failover for Web-based clients.

Cluster management is through the standard Notes administrative client.

4.1.10.1 Strategies

After reviewing your Notes environment, you should plan a cluster strategy that tries to create a balance between your users’ requirements for data availability and the physical ability of each computer in your cluster to manage additional workload.

There are different strategies for optimizing database and server access:

- **Failover strategy**

  This strategy attempts to keep important databases always available to clients. Obvious candidates to consider for this strategy are the users' mail databases. You can make mail databases highly available by creating a replica of the databases on another server in the cluster. If the client accesses the mail database and the server is unavailable, the user fails over to the replica on another server.

  If you have databases for which you must provide maximum protection, create more than two replicas within the cluster.

- **Workload strategy**

  Domino clusters have an optional workload balancing feature that lets you distribute the workload of heavily used databases across multiple servers in the cluster. Again we should consider mail databases in balancing workload across servers in the cluster. It is recommended that you distribute the user mail files equally across the cluster so that no single server has a significantly larger number of mail files than any other server.

  There are two notes.ini parameters that assist in workload balancing:

  a. Server_Availability_Threshold

     This parameter specifies the level of system resources below that you do not wish to fall. By setting this value for each server in a cluster, you determine how the workload is distributed among cluster members. Domino compares this value against a server's availability index. When the availability index falls below the Server_Availability_Threshold value, the server becomes BUSY and further requests for access will be failed over to a database replica on another server.

  b. Server_MaxUsers

     This parameter specifies the maximum number of active user sessions allowed on a server. When this limit is reached, the server goes into a MAXUSERS state. The Cluster Manager then attempts to redirect new user requests to other servers in the cluster. Note that the Server_MaxUsers parameter is a valid setting for any Domino server; it is not restricted to clustered systems.

- **Mixed strategy**

  The mixed strategy is a combination of high failover and active workload balancing. When you have a database that users access continuously, such as a special discussion database, not only do you need high

availability, but you also need to distribute the workload caused by users accessing the database.

4.1.10.2 Management tools

There are a number of tools that can help you manage Domino clusters:

- **Cluster Administration process**

  It is responsible for the correct operation of all cluster components. On clustered servers, the process runs automatically at server startup and whenever the cluster membership changes. When a server detects a change to its Address Book, the Cluster Manager and the Cluster Administration Process (cladmin) are executed.

- **The Cluster Database Directory Manager**

  The Cluster Database Directory Manager (cldbdir) task keeps the Cluster Database Directory (cldbdir.nsf) up to date with the most current database information. It also manages databases with cluster-specific attributes such as databases marked out of service or pending deletion. A replica of the Cluster Database Directory database resides on each server in a cluster and contains information about all the databases and replicas within a cluster.

- **The Public Name and Address Book**

  This database is at the heart of any Domino implementation. All objects within Domino, such as users, servers, and connections, and their settings, are contained in the Public Name and Address Book (names.nsf). Every server in a cluster has a replica of this database.

  In Domino R5, names.nsf is called the Domino Directory.

- **notes.ini file**

  Many of the settings in names.nsf are reflected in the notes.ini file. After the server is added to a cluster, its notes.ini file has new entries inserted to reflect its new status. The next time the server starts, two new server tasks are automatically started, the Cluster Database Directory Manager and the Cluster Replicator tasks. The ServerTask statement in the notes.ini file of the server is appended with these two new task parameters:

  – cldbdir

    All Domino servers within a cluster run a Cluster Database Directory Manager (cldbdir) task that creates and maintains the cluster database directory. When you add or remove a database on a server, the cldbdir

task immediately changes the information in the cluster database directory to reflect this.

All members of a cluster share a common cluster database directory. When one server updates its list of databases, the Cluster Replicator replicates the changes to the other servers in the cluster. In this way each cluster member has an up-to-date directory of databases in the cluster.

– clrepl

The Cluster Replicator (clrepl) task is responsible for the tight synchronization of data among replicas in a cluster. Each server in a cluster runs one Cluster Replicator by default although administrators may choose to increase the number of clrepl tasks on a server.

Whenever a change occurs to a clustered database, the Cluster Replicator pushes the change to the other replicas in the cluster. This behavior ensures that each time you access a database, you are looking at the most up-to-date version.

The Cluster Replicator task only pushes replications to other servers in the cluster. The standard Notes replicator task (replica) is still responsible for replicating changes to and from servers outside of the cluster.

Lotus recommends that you also have standard replication running within a cluster to update any databases that cluster replication may have been unable to complete for unforeseen reasons.

- **MailClusterFailover parameter**

In addition to the parameters already mentioned, there is the MailClusterFailover parameter that is important if you want MailRouter failover. This parameter enables or disables MailRouter request failover. If users have replicas of mail files located on multiple servers, you can set this variable in the notes.ini file of all Notes Release 4.x servers in the domain to enable users to receive mail from servers within and outside the cluster when their home servers are down. The format of the parameter is:

```
MailClusterFailover=value
```

There is no default value for this parameter. Use the following values to set this variable:

0 - Disables Mail Router request failover
1 - Enables Mail Router request failover

4.1.10.3 Notes log

As the cluster administrator, you should closely monitor your system for usage patterns and adjust cluster resources accordingly. Modifying the number of cluster replicas, changing parameters to control server workload and, if necessary, increasing the number of servers in your cluster are all ways you can fine-tune the performance of your Domino system.

A number of tools are provided to help you in the management task, such as the Domino Statistics & Events Database, and by using the Windows NT Performance Monitor, you can identify and minimize the effect of any bottlenecks in your system.

For detailed information on using these tools, we recommend the redbook *Netfinity and Domino R5.0 Integration Guide*, SG24-5313.

4.1.11 Summary

Lotus Domino's clustering implementation combines the important basic functions of high availability, scalability, and workload balancing. Domino currently allows you to group up to six servers to form a cluster. Since it is an application clustering solution, it provides you with a degree of operating system independence. This flexibility of support for multiple operating systems includes mixing operating systems within a cluster.

High availability is achieved by allowing clients to fail over to another server whenever their data is not available on its usual server, due either to failure or overloading. Your users can keep working.

Domino offers good scalability. It is easy to add servers to a cluster, or to remove them, as the workload changes over time.

Load balancing lets the system distribute workload as demand fluctuates or in the event of a server failure. As an administrator you do not want to have to continually monitor which server is getting overloaded. Workload balancing lets the clustering technology distribute the workload dynamically, based on performance thresholds set in each server.

New to Domino Release 5.0 clustering is a major enhancement called the Internet Cluster Manager (ICM). This enables Domino clusters to provide the benefits of failover, scalability, and workload balancing to Web browser clients.

Domino clustering is also easy to implement. There are no special hardware requirements and no need to rewrite your Domino applications.

For more information about Domino clustering see the following redbooks:

- *Lotus Domino R5 Clustering with IBM @server xSeries and Netfinity Servers*, SG24-5141
- *Netfinity and Domino R5.0 Integration Guide*, SG24-5313

4.2 Oracle Parallel Server and high availability

Oracle Corporation, one of today's leading database vendors, offers a high availability option for its base product (currently Oracle8*i*) called Oracle Parallel Server (OPS). OPS provides scalability for non-partitioned applications, such as ERP applications and Internet commerce applications. Oracle Parallel Server permits users and applications running on any node of a cluster to access all data in a database and if one cluster node fails over to surviving node.

4.2.1 Oracle8i and OPS

In this section we introduce the key components of an OPS solution on Netfinity servers. Comprehensive information about configuration, installation, tuning and how OPS works can be found in the redbook *Oracle Parallel Server and Windows 2000 Advanced Server on IBM Netfinity*, SG24-5449.

Oracle offers two clustering products:

- Oracle Fail Safe (OFS), which offers high availability on Windows NT and Windows 2000 servers.

  Oracle Fail Safe is layered over Microsoft Cluster Server and tightly integrates with the shared nothing cluster environment. In a shared nothing cluster such as Microsoft Cluster Server, resources (for example, disks) are owned and accessed by only one node at a time. If one server fails, the other server of the cluster can handle the user workload. With OFS, however, there is no performance scalability; the maximum computing power is a single system's computing power.

- Oracle Parallel Server (OPS), which offers high availability and scalability on Windows NT and Windows 2000 servers.

  By contrast with OFS, Oracle Parallel Server implements a "shared data" model (see 2.2.2, "Hardware for clusters" on page 11), allowing multiple servers to simultaneously access the disk storage. Each server hold a database instance. This allow a single workload to scale across multiple systems and allows load-balancing capabilities. OPS does not use Microsoft Cluster Server to implement its clustering.

4.2.2 IBM OPS configurations

The IBM solutions for OPS that have been certified by Oracle are listed at:

`http://www.oracle.com/database/options/parallel/certification/oracle8i_ibm.html`

The major components of the IBM-certified configuration are:

- Up to eight Netfinity 7000 M10 servers
- A Netfinity Fibre Channel controller unit with two RAID controllers
- Dual Vixel 7-port Fibre Channel Hubs and loops
- Netfinity Fibre Channel PCI Host Adapter
- Optical cabling, GBICs
- Netfinity EXP15
- Nways 8271-712 Ethernet Switch or a Netfinity SP Switch and associated adapter

This is shown in Figure 41.

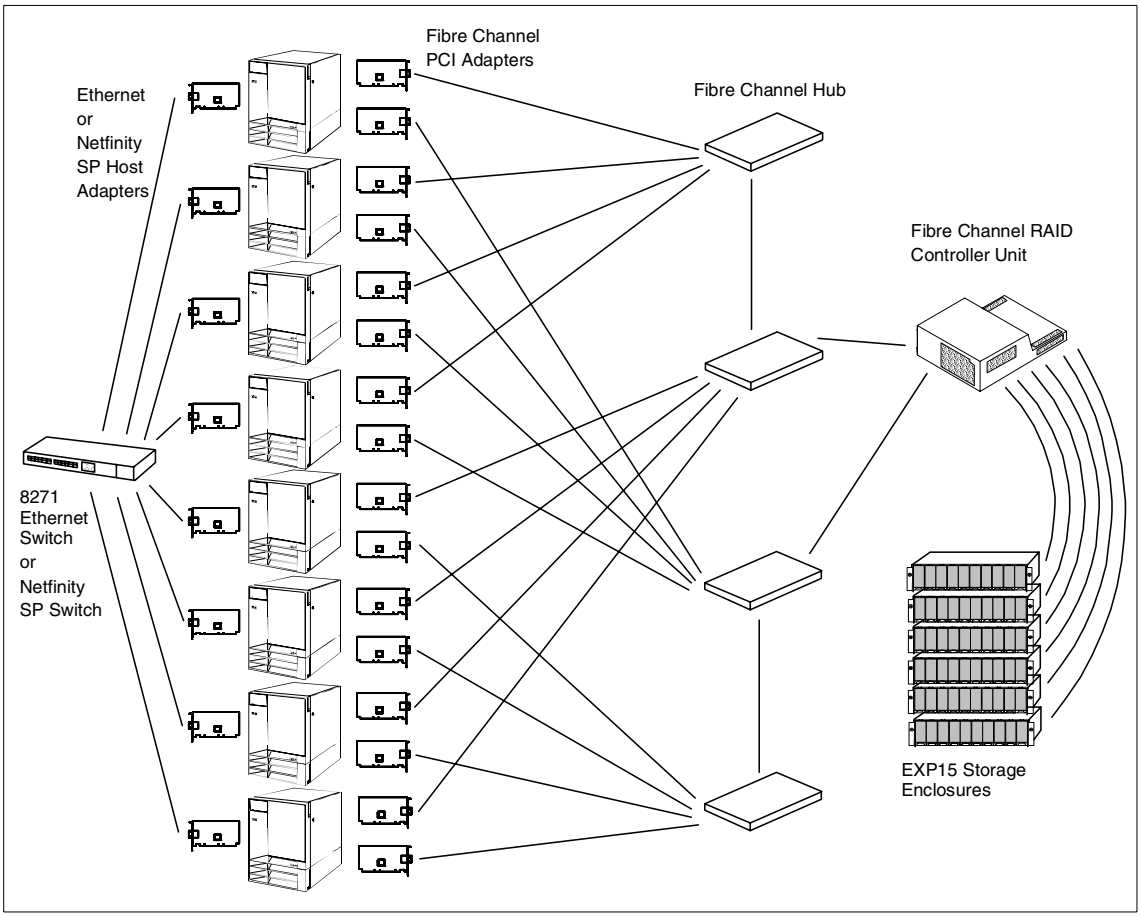

Figure 41. OPS eight-node configuration

IBM also intends to support OPS and Windows 2000 and plans to have a certified solution in 2000. The certification process is likely to include the following components:

- Netfinity 8500R cluster
- Netfinity FAStT Fibre Channel hardware
- Windows 2000 Advanced Server
- Oracle Parallel Server 8i R2 (8.1.6)
- Netfinity Advanced Cluster Enabler for OPS

A similar configuration was implemented to write *Oracle Parallel Server and Windows 2000 Advanced Server on IBM Netfinity*, SG24-5449. This configuration included:

- Two Netfinity 8500R systems with 8 CPUs and 4 GB RAM each
- Netfinity Fibre Channel RAID controller unit with two RAID controllers
- Two Vixel 7-port hubs and loops
- Netfinity Fibre Channel PCI Adapters
- Optical cabling, GBICS
- Netfinity EXP200
- Windows 2000 Advanced Server

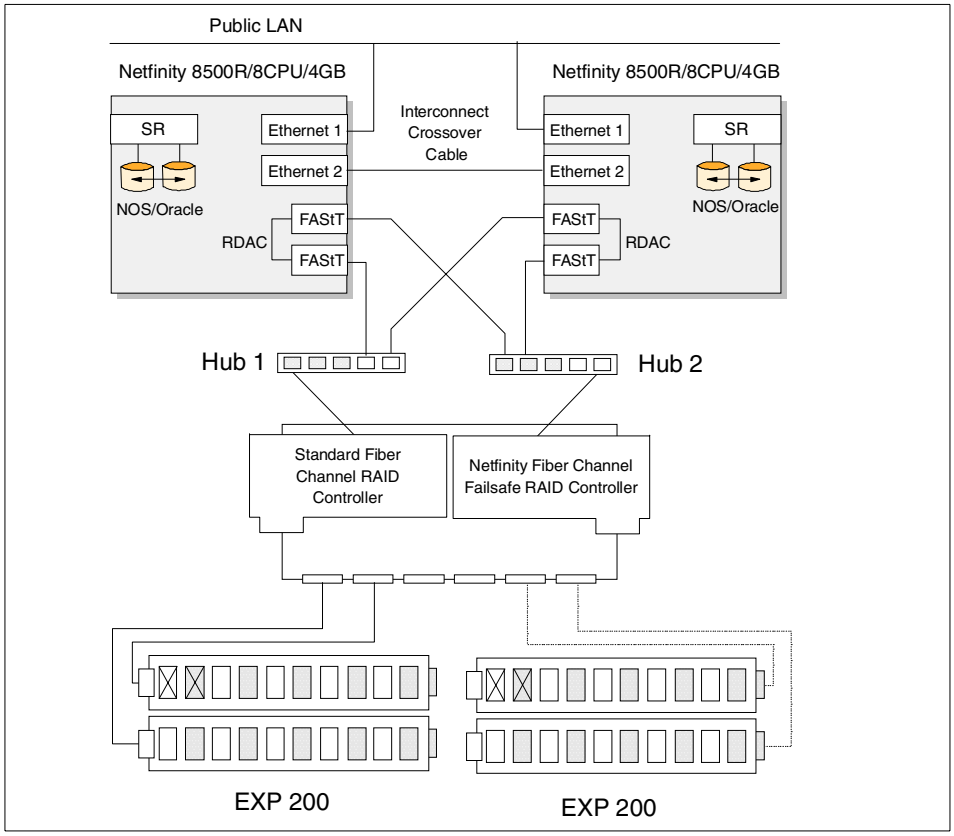

Figure 42. OPS two-node configuration

IBM has developed the Netfinity Advanced Cluster Enabler for OPS for integration with Oracle Parallel Server v8i. This component implements the operating system dependent (OSD) layer for OPS.

The OSD layer consists of several software components distinct from Oracle Parallel Server. These components provide key services required for proper operation of the Oracle Parallel Server options. In total, the OSD components provide what can be viewed as the Oracle Parallel Server interface to the cluster and its distributed services. The OSD layer's most important components are:

- Cluster Manager (CM)
- Interprocess Communication (IPC)
- I/O (input/output)

All of these components are the interface between Oracle and platform-specific cluster services that provide the architected functions necessary for the operation of OPS.

The OSD was developed by Oracle, and IBM has written the installation code for it. You can download it from IBM at:

`http://www.pc.ibm.com/us/netfinity/parallel_server.html`

4.2.3 Scalability

Oracle8i is designed to support extremely high user populations by using Oracle Multithreaded Server (MTS) configuration. MTS is based on a database resource-sharing architecture where database listeners route user connections to a group of dispatchers that interact with server processes to handle database connections. Oracle Parallel Server environments can be configured with MTS, where each node in the cluster is configured with a group of dispatchers.

OPS offers a number of functions that aid scalability:

- **Cluster Load Balancing**

  Oracle8i for connections to the database use Oracle standard Net8 protocol or industry-standard IIOP protocol for HTML or Java clients. Figure 43 explains the connection and cluster load-balancing schemes.

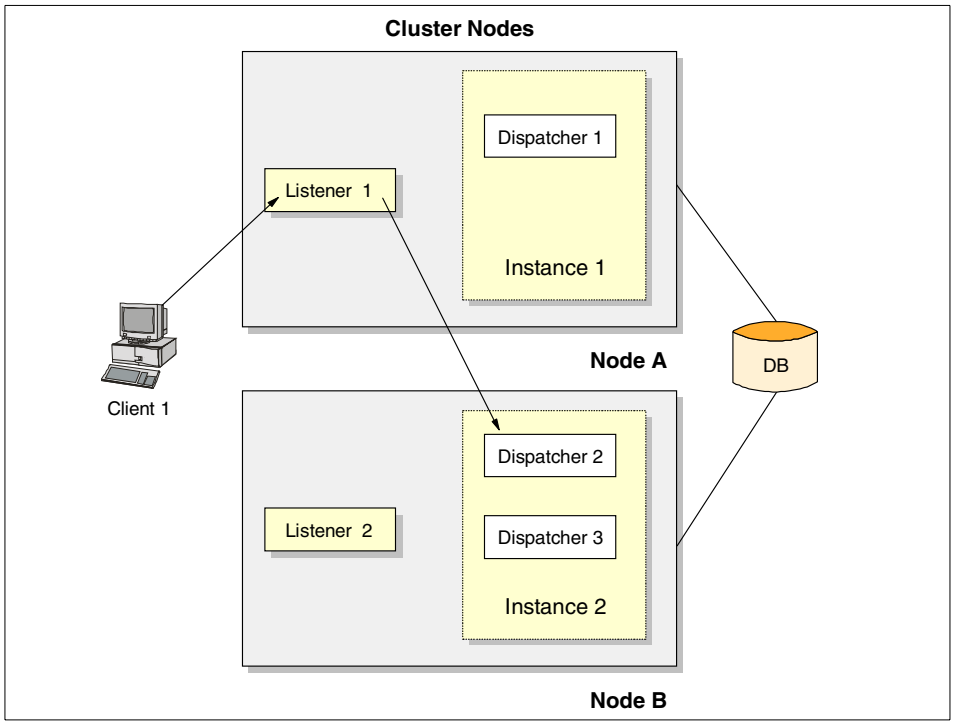

Figure 43. Connection load balancing scheme

Our sample OPS configuration has two nodes A and B running Instance1 and Instance2 to serve database requests to same database service. Listener 1 and Listener 2 run on nodes A and B respectively. Listener is a service responsible for listening for clients connections and directing them to dispatcher that controls connections. The connections are routed via a dispatcher. To handle more connections we can have more than one dispatcher for a particular instance. In our example we have Dispatcher 1 on Instance 1 and Dispatchers 2 and 3 on Instance 2.

Every client has a configuration file with a list of available listeners (L1 and L2). To spread connections efficiently across available nodes, clients use random connections to available listeners.

An Instance monitors resource utilizations and reports that information to all Listeners. Listener 1 compares load information on Node A and Node B. If the CPU load is lower on Node B, then Listener 1 will direct the client to Node B. Listener 1 also checks the active connection number on Dispatcher 2 and Dispatcher 3. If Dispatcher 2 has a less active

connection than Dispatcher 3, Listener 1 will establish a client connection to Dispatcher 2 on Instance 2.

- **Parallel Cache Management**

  Parallel Cache Management is the technology that provides concurrent multi-instance access to the database. Synchronization of access is required to maintain the integrity of the database when one instance needs to access or update data that has been modified by another instance and is in the buffer cache of that instance. Cluster nodes accessing the data in changed blocks are not able to read a current form of that data from disk. Parallel Cache Management allows cluster nodes to read changed data residing in the buffer cache of a remote node and ensures necessary communications and synchronization.

- **Cache fusion**

  In Oracle8i, a new improved parallel cache management method is used known as *cache fusion*.

  In the previous releases of Oracle, the block had to be written on the disk, before the block could be read by another instance. This mechanism avoids the need to write the block on the disk — it is transmitted to the requesting instance through the interconnect link.

  Cache fusion reduces the access time for a remotely cached block.

4.2.4 High availability

The high availability features of IBM's OPS configuration include:

- **Netfinity Fibre Channel**

  OPS is built on a shared disk cluster architecture (see 2.2.2, "Hardware for clusters" on page 11). Cluster nodes uses their own CPU and memory to run the operating system, database and application, but they share the disk subsystem with a single/common database.

  On the hardware level, high availability is made possible by a fully redundant Fibre Channel storage subsystem with no single point of failure. The path (host adapters, cables, hubs, and RAID controllers) from each server to the RAID array is duplicated. To avoid two-fold presence of the same disk space, a Redundant Dual Active Controller (RDAC) multi-path driver is used to access and control the subsystem.

  RDAC is an I/O path failover driver installed on the host computers that access the storage subsystem. Usually, a pair of active controllers is located in a storage subsystem. Each logical drive in the storage subsystem is "owned" by one controller, and it controls the I/O between

the logical drive and the application host along the I/O path. When a component in the I/O path fails, such as a cable or the controller itself, the RDAC multi-path driver transfers ownership of the logical drives assigned to that controller to the other controller in the pair.

The RDAC multi-path driver manages the I/O data path failover process for storage subsystems with redundant controllers. If a component (for example, a cable, controller, or host adapter) fails along the I/O data path, the RDAC multi-path driver automatically reroutes all I/O operations to the other controller.

- **Transparent application failover**

  Comparing OPS high availability with generic application failover (such as that implemented in MSCS), we see much faster and easier reconnections for clients. OPS is designed for high availability and has its own client.

  The Oracle client knows all cluster nodes and there is no need to move the IP address or logical drives to the surviving node. Disks are shared and all cluster nodes in OPS see the shared database. Clients automatically connect to another node and simply redo the transaction.

  Clients can also be configured to preconnect to eliminate the reconnection procedure and delay. This feature is called *transparent application failover* (TAF) and works with applications that are written to use the Oracle Call Interface.

4.2.5 Manageability

The new and improved Java-based tools help manage and administer Oracle databases in a cluster:

- **Oracle Universal Installer**

  The installer automatically discovers cluster resources and cluster configuration parameters. The installer wizard enables users to select the appropriate nodes of the cluster and have the install propagate to all nodes of the cluster. The installer prompts the user to indicate if the database is a single-instance regular Oracle database or a multi-instance Oracle Parallel Server database. If the database is multi-instance, the installer automatically discovers the cluster nodes and enables a cluster-wide install.

- **Database Configuration Assistant**

  The Oracle Parallel Server-enabled Database Configuration Assistant (DBCA) also delivers a configuration wizard for cluster-wide database configuration. The Oracle8i DBCA extends the database configuration tools and capabilities that are part of DBCA to include Oracle Parallel

Server. DBCA automatically discovers the cluster nodes and provides facilities for installing a variety of seed databases or sample databases.

- **Oracle Enterprise Manager**

  Oracle Enterprise Manager allows you, from a single workstation, to manage, administer and monitor multiple complex databases. It is a set of centralized services and management applications that provide a database administrator (DBA) with all the tools to manage multiple instances and parallel databases.

  Oracle Enterprise Manager has a three-tiered architecture (Figure 44). The first tier is the client or administration console, which consists of a Java-based console and integrated applications. The second tier is the management server, which provides control between the clients (first tier) and the different components that will be managed (third tier). The third tier can, for example, be an OPS solution running, for example, a parallel database, but it can also be normal databases or other services.

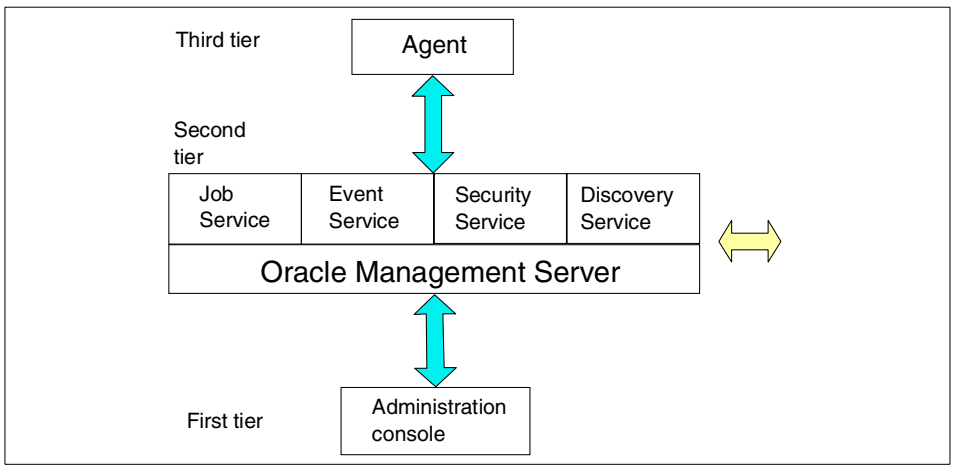

Figure 44. Oracle Enterprise Manager architecture

- **Cluster Manager**

  Cluster Manager is an element of the IBM version of the OSD layer, the Netfinity Advanced Cluster Enabler for OPS. This service controls and manages membership information for systems in the cluster. Its primary task is to interface with OPS to control the process of nodes either joining (attaching) or leaving (detaching) the cluster. Another major task of this service is to manage failover of instances within the cluster.

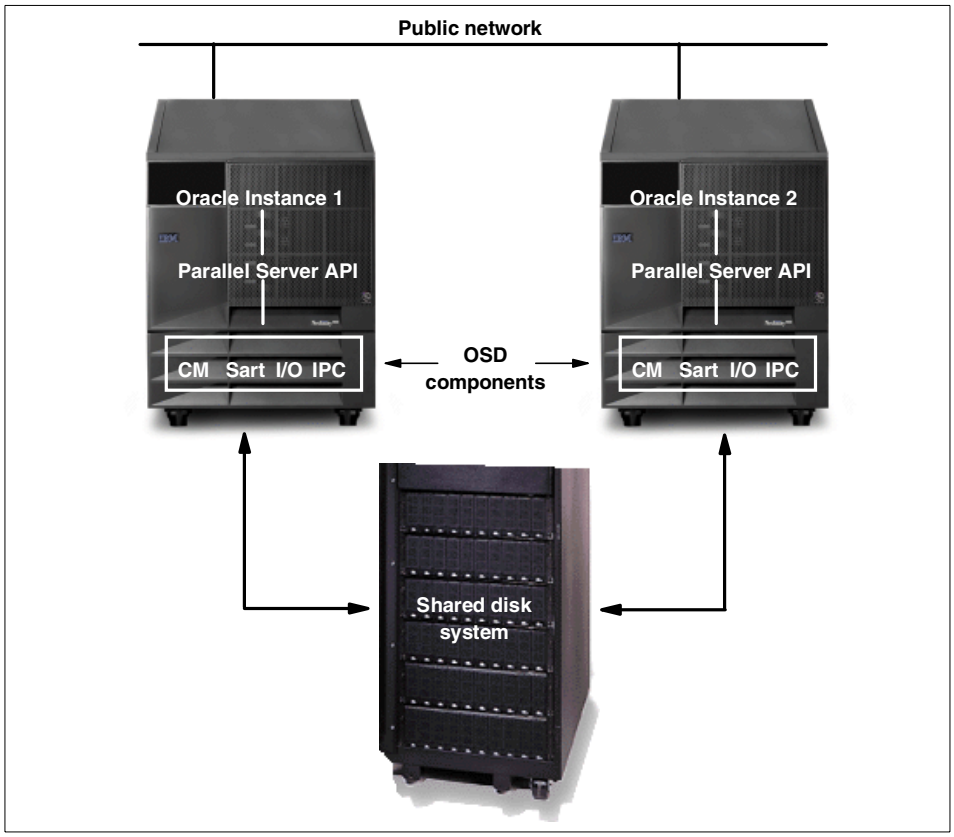

Figure 45. OSD components

In a failover situation it is critical that all remaining instances are notified of the failure and that relevant OPS recovery operations can be started. This allows the processes that were running on the failed node to move to a new instance and restart. Cluster Manager reconfigures the cluster to isolate the failed node. This is transparent to user applications.

Cluster Manager has two separate parts:

– The *Node Monitor* component manages access to the shared disks and checks the status of the nodes, network, and the Cluster Management component.

– The *Cluster Management* (CM) component monitors the topology of the cluster and maintains current membership status for the nodes.

The Cluster Management component manages instance members. Each instance registers with its database-specific group, which is managed by

the CM component. This information is passed to the Node Monitor and then through the CM to OPS. Further information can be found in *Implementing Oracle Parallel Server on Netfinity Servers*, SG24-5449.

4.2.6 Planning for OPS

Your priorities will decide the characteristics of your database. Is transaction speed more important than data safety? Will the database be used for a large number of short transactions or, perhaps, a smaller number of large transactions? Is the database to be accessed by a single application or several?

These, and other factors, will influence your implementation. To compare characteristics, we can conveniently define four basic types of database applications: online transaction processing (OLTP), decision support, development and mixed database.

We can give only general guidelines here, since each application will exhibit its own characteristics in the way the database is accessed. The following sections should be considered in this light.

- **OLTP database**

  An OLTP database, such as a bank automated teller machines (ATMs), has a very high volume of transactions (measured in transactions per second). When planning a database for such purposes, you usually need to:

  – Split database files so that disk accesses are spread across many disk subsystems.
  – Implement redo log archiving.
  – Be available 24 hours a day, 7 days a week.
  – Use raw disk partitions.
  – Use high availability hardware (mirrored disks, for example).
  – Institute performance tuning.
  – Set database parameters to accommodate more users and increase performance.

- **Decision support database**

  A decision support database, such as an inventory system, has a relatively low number of database updates (measured in transactions per hour). The users tend to make few queries, and they look at the result of

these queries for many minutes at a time. They may prepare reports from these queries. A decision support database:

  – Has less need to distribute input/output across multiple disks and controllers.
  – Has less need to split the logical database design.
  – Has less need to implement redo log archiving.
  – Can usually be brought down for maintenance or backups.
  – Table partitioning and materialized views will increase the performance.

- **Development database**

  Application development require the presence of live database. The development team probably does not need the database to have the same level of robustness as a production database (unless testing under stress conditions). Typically development databases:

  – Do not need to split input/output to the disks.
  – Can usually be brought down for maintenance or backups.
  – Have little need for backup more than once a day.
  – Have little need for performance tuning.

- **Mixed database**

  A mixed database combines various functions, such as decision support and translation processing. Most databases fall within this category. Such a database:

  – Has an even mixture of queries and updates.
  – Has some performance tuning needs, but not as much as an OLTP database.

4.3 DB2 Universal Database

To satisfy the demand for complex decision support and data warehousing applications on Netfinity servers, IBM has extended the rich feature set of DB2 Universal Database Enterprise Extended Edition (EEE) to the Windows 2000/NT platform. The database's shared nothing architecture (see 2.2.2, "Hardware for clusters" on page 11) allows parallel queries to be made with minimal data transfer among nodes.

There are four different versions of DB2 available to meet different market segment requirements:

- Personal Edition, suitable for individual applications that do not require remote access.

- Personal & Workgroup Edition, suitable for individual or departmental applications on uniprocessor Intel-based hardware platforms.
- Enterprise Edition (UDB EE), suitable for UNIX or Intel-based applications on SMP machines.
- Enterprise Extended Edition (UDB EEE), suitable for clustered or large Intel or UNIX servers.

In a clustered environment, DB2 EEE supports separate applications and databases on each server of the cluster and will be able to use other servers in the cluster for failover processing, gaining high availability of the applications.

The main difference between DB2 EE and DB2 EEE is the partitioning feature included in the latter. Using the hash-based partitioning architecture from DB2 Parallel Edition, databases are partitioned to enable interquery parallelism and scaling across clusters up to 512 nodes.

Currently DB2 runs on the following platforms: OS/2, Windows NT/2000, AIX, HP-UX, Solaris, NUMA-Q and Linux.

4.3.1 DB2 scalability

Earlier this year, IBM published impressive TPC-C benchmark results using a 32-node Netfinity 8500R database cluster running a large DB2 database on Windows 2000 Advanced Server, which demonstrated that the combination of IBM Netfinity servers and DB2 Universal Database Enterprise Extended Edition delivers outstanding power and multiuser throughput performance for companies running business intelligence applications.

All the machines are connected by a communications facility. This environment is referred to by many different names, including cluster, cluster of uniprocessors, massively parallel processing (MPP) environment, and shared-nothing configuration. The latter name accurately reflects the arrangement of resources in this environment. Unlike an SMP environment, an MPP environment has no shared memory or disks. The MPP environment removes the limitations introduced through the sharing of memory and disks.

A partitioned database environment allows a database to remain a logical whole, despite being physically divided across more than one partition. The fact that data is partitioned remains transparent to most users. Work can be divided among the database managers; each database manager in each partition works against its own part of the database.

32 Netfinity servers running Windows 2000 were used for the test as shown in Figure 46, producing the TPC-C results.

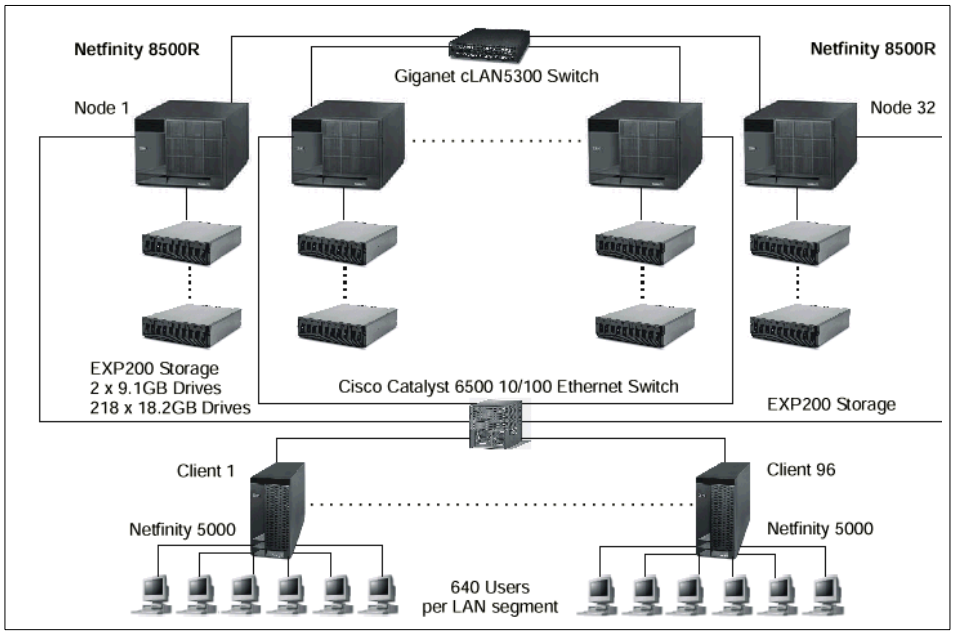

Figure 46. 32-node DB2 cluster on Netfinity 8500R servers

Test results are published on the Web at:

`http://www.tpc.org/results/individual_results/IBM/ibm.8500r.00070302.es.pdf`

The real benefit of partitioning, as implemented in UDB EEE, is that more than one server in a cluster can be used to execute a single application. Because of this, the application is not limited to the capacity constraints of a single SMP system. Partitioning also enables faster insert, update, and delete processing, since these operations will execute across all partitions in parallel. Manageability is also increased, because partitioning provides more granularity for database operations.

As an aside, in some cases it may make sense to deploy UDB EEE on a single SMP machine to take advantage of the granularity of operations, especially in large SMP environments.

As your data volume increases, adding more capacity is as easy as installing another node in your cluster to add another partitioned DB2 EEE server, as shown in Figure 47:

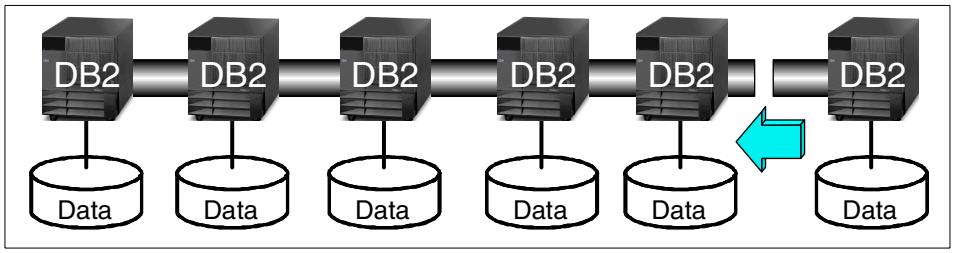

Figure 47. DB2 UDB EEE data partitioning

For more detailed information about partitioned databases systems and DB2 EEE see *Quick Beginnings of DB2 Universal Databases Enterprise -Extended Edition for Windows*. You can read this document online on URL:

`ftp://ftp.software.ibm.com/ps/products/db2/info/vr7/pdf/letter/db2v6e70.pdf`

The DB2 Universal Database architecture is a thread-based implementation. DB2 engine threads are mapped to the Windows operating system threads providing a high degree of parallelism, load balancing and throughput, resulting in high performance for users. Several features are incorporated to enhance input/output (I/O) performance. Examples include:

- Prefetch Data Pages - one or more pages are retrieved in anticipation of their use.
- Large Block Reads - several pages are read at a time with a single I/O operation.
- Parallel I/O - many I/O operations can be performed in parallel for a single query.

The thread-based DB2 implementation, coupled with Windows NT kernel-based thread model, and SMP support results in a high-performance combination.

DB2 is designed for parallel operation and hence very suitable for multiprocessor or multinode systems. Figure 48 shows how a parallel query and a parallel transaction can be spread across multiple CPUs (or nodes):

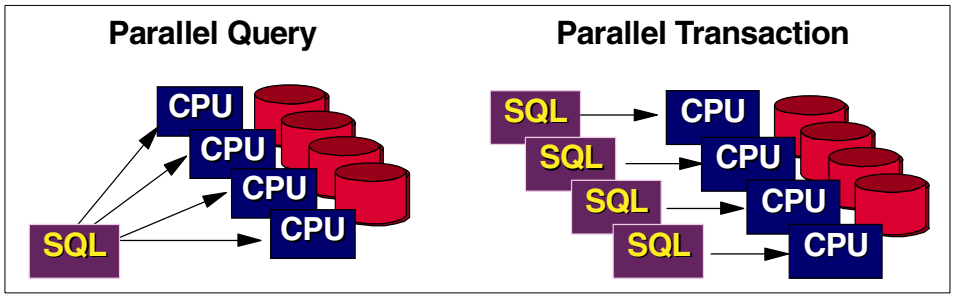

Figure 48. DB2 parallel query versus parallel transaction

4.3.2 DB2 high availability with MSCS

This section provides information about implementing DB2 in a Microsoft Cluster Server environment. If you are not familiar with MSCS, please review 3.1, “Microsoft Cluster Server” on page 19.

DB2 UDB installation provides a new resource type *DB2* for MSCS to monitor the cluster (see 3.1.4, “Resource types” on page 24 for more information). It represents a DB2 database partition server (also known as a DB2 node). You define a resource of type DB2 for each node that requires failover support. The instance owning the DB2 node has dependencies on the instance directory file share and the DB2 network name with an associated IP address for DB2. Each DB2 node will also have dependencies on disk resources that will be required to store user data.

4.3.2.1 DB2 UDB failover support

DB2 UDB for Windows NT/2000 provides support for MSCS, allowing DB2 to be defined and managed as an MSCS-aware server application, using the following:

- A resource module, which allows a DB2 instance to be defined and managed as an MSCS resource. This allows MSCS to start and stop DB2 UDB for Windows NT/2000.
- Pre- and post-online script support, allowing additional configuration, before and after DB2 is brought online in a clustering scenario.
- Utilities to manage DB2 instances defined in MSCS clusters.

To set up, configure and maintain DB2 UDB in an MSCS environment you have utilities such as:

- db2mscs

  When you run the db2mscs utility, it creates the infrastructure for failover support for all machines in the MSCS cluster. To remove support from a machine, use the `db2iclus` command with the `drop` option. To re-enable support for a machine, use the `add` option.

- db2drvmp

  When you create a database in a partitioned database environment, you can specify a drive letter to indicate where the database is to be created. On an MSCS cluster, you cannot share common disks from both nodes. For proper drive mapping on MSCS cluster nodes during installation of a partitioned database system, we use the db2drvmp utility to set up the drive mapping.

Client application behavior is unchanged in a clustered environment except for the following sequence of events:

1. During DB2 UDB failover, client application will receive a communications failure.
2. DB2 UDB will be brought online on the other node of the cluster, at which time client applications can reconnect to DB2.
3. In-flight transactions that were not completed during the failover are rolled back by DB2. The application must resubmit the transaction.

4.3.2.2 DB2 UDB EEE failover configurations

There are two possible configurations for DB2 UDB EEE in an MSCS environment:

- **Hot standby**

  Also known as active/passive. In this configuration, an MSCS clustered DB2 instance is defined with one or more database partition servers running on one server, while the other server is inactive. Upon failover, DB2 will be started on the inactive server, the standby system. The standby system is simply used to provide high availability.

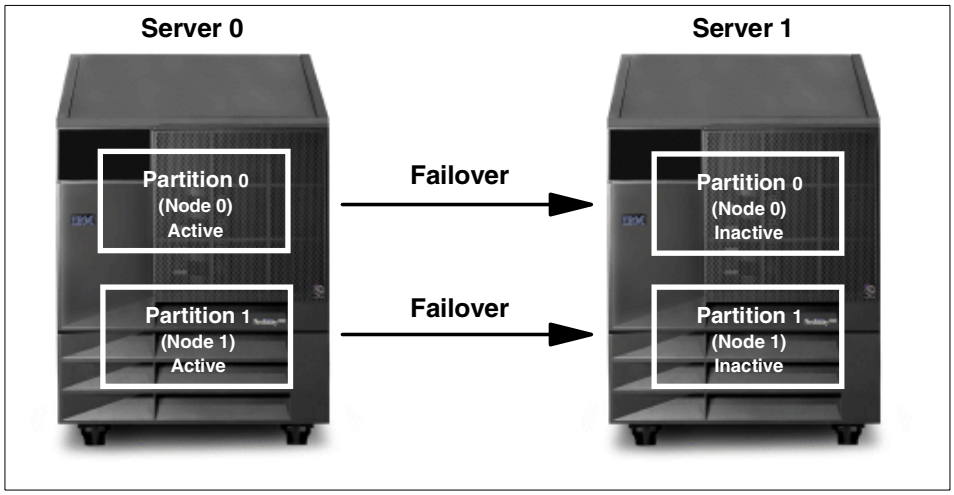

Figure 49. DB2 UDB EEE and MSCS: hot standby failover configuration

- **Mutual takeover**

  This is also known as active/active. Within the MSCS-enabled DB2 instance, database partition servers are defined on the two servers in the MSCS cluster, each with its own set of MSCS resources. During failover, database partition servers from the failed system will be started on the remaining server.

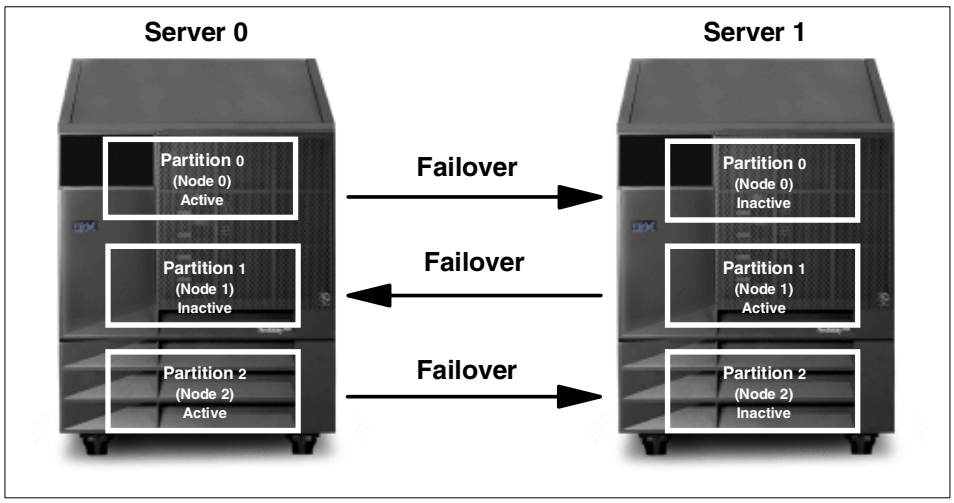

Figure 50. DB2 UDB EEE and MSCS: mutual takeover failover configuration

Because DB2 UDB EEE instances may be defined over more than two machines, you may have a mix of hot-standby and mutual takeover MSCS clusters within the same DB2 instance.

Comprehensive information about high availability in the Windows NT environment and DB2 configuration for MSCS can be found in “High Availability in the Windows NT Environment” in *Administration Guide: Planning*, available from:

`http://www-4.ibm.com/cgi-bin/db2www/data/db2/udb/winos2unix/support/v7pubs.d2w/en_admin`

For more detailed information about implementing DB2 databases on Windows NT, we recommend *The Universal Guide to DB2 for Windows NT*, SC09-2800.

4.4 Microsoft SQL Server

Microsoft SQL Server is a prime application for use with clustering. In this section, we describe several issues we faced before and after we implemented clustering with SQL Server. Unlike the other applications described in this chapter, SQL Server does not handle clustering itself—it is, however, tightly integrated with Microsoft Cluster Services (MSCS).

The three major areas of concerns we cover here are:

- Installation
- Performance
- Support

It is very important to design a well-thought-out plan that will meet your business requirements. Understanding what your line-of-business needs are is critical if your company is to be successful.

In this section we used Windows 2000 Advanced Server with Cluster Services (MSCS) and apply SQL Server 7.0 as our database. We used two Netfinity 7000 M10s with four 500 MHz processors and 4 GB of RAM in each system.

4.4.1 Installation concerns

We chose SQL Server V7.0 over Version 6.5 is because of the problems with v6.5 uninstalling. V6.5 does not clean out the registry, so if you have to uninstall and then reinstall you need Microsoft’s RegClean for SQL. This is available in the Windows NT or Windows 2000 Resource Kits.

When installing SQL Server, you need to make sure that you have all program and data files on the common disk subsystem. All the virtual drives in a SQL Server configuration, along with the Network Name and IP Address resources, should be kept in the same MSCS cluster group. This is mainly for failover purposes.

We plan to have an active/active setup. This means that basically we will be running two separate SQL database, with one normally on each node. They will be using different application binaries and more importantly have two different sets of data on the common disks.

4.4.2 Performance and sizing concerns

We opted to put our transaction logs on RAID-1 volumes. This is for performance reasons, because a write takes two actions to the log for each write to the data partition. The data is stored on RAID-5 arrays.

We made sure that we did not mix the SQL and the quorum devices because too much I/O against the quorum partition may cause unwanted failover of the cluster.

Because we have 4 GB of RAM, we put the TempDB in RAM. This will greatly increase our performance. If you do not have a lot of RAM, you should consider moving TempDB out of MasterDB and putting it on its own RAID-1 volume to help with performance.

Since we put TempDB in RAM, the RAM allocated to TempDB had to be committed to physical memory. This is very important when running active/active configurations; otherwise performance will suffer. We committed our TempDB to 2 GB of memory because if we stacked two instances on top of each other in a failed-over cluster, we would have enough RAM to support the failover.

We decided to go with 16 KB stripe size because this is usually optimal for an SQL database. If you are using this database to store large images, you might need to experiment with different stripe sizes to achieve optimum performance. The Windows NTFS file system stripe size is not important (all SQL Server data is essentially stored in one NTFS file), but the RAID stripe is going to be key here.

Since we have decided to use active/active, we have set a baseline for maximum performance of the hardware in terms of SQL throughput. Once we had the number, we then sized the amount of the resources properly so that each virtual server consumed a portion of the available resources. Keeping in mind that our implementation shares half the total resources in each SQL

Server instance, we had to make sure that the SQL Server did not grow greater than 50% of available resources. If this happens then we will have performance degradation when it shares that resource with our other server.

For more information about SQL Server performance, see Chapter 21 of the redbook *Tuning Netfinity Servers for Performance: Getting the Most Out of Windows 2000 and Windows NT 4.0*, SG24-5287.

4.4.3 Support concerns

SQL Server clustering provides *no* protection from data loss. This is why it is very important to implement a solid tape backup solution. Do not rely on RAID-5 alone to provide protection against a disk failure — this is not a valid backup solution. If there are enough bad parity bits, there is no way to restore your data. Also, RAID doesn't protect you against deleted or corrupted data.

We carefully thought out future plans as well. Sometimes it is necessary to remove SQL Server from the cluster before upgrading the operating system and applying service packs because each operating system and service pack requires certain DLL files to be available or to be removed. As a result, we had to make sure that we had reviewed all the readme files for any future updates.

Make sure that the administrators understand that if they are using an active/passive configuration, the passive node is not a test server or test cluster. Do not apply service packs or any other general update to the passive node in the cluster while in production. Remember, what you do to one node can adversely affect the other node.

We also avoided *Open file agents* as a backup method because many of them do not work with clustering. We recommend Tivoli Storage Manager because it can back up the cluster while it is online through an agent. See *Using TSM in a Clustered Windows NT Environment*, SG24-5742 for more information.

As you can see, there is a lot of thought that goes into planning a high availability cluster solution using SQL Server. Some solutions, such as Microsoft Exchange, may not require as much thought when planning and implementing. Planning and implementing your cluster with the help of professional IT architects can improve your chances of building a successful cluster solution.

4.5 Citrix MetaFrame

Citrix MetaFrame is an extension to Windows NT Server 4.0, Terminal Server Edition and Windows 2000 Terminal Services.

It adds support for additional client types by leveraging Citrix's Independent Computing Architecture (ICA). Even though MetaFrame does not itself offer full clustering support, such as application failover, it does let you install your applications in a way that allows MetaFrame to decide which server should process your request, thus providing the ability to load balance among systems.

Note: MetaFrame does not fail over clients in the event of a node failure. As a result, user sessions hang and they get a message saying their session is disconnected. They can then reconnect and get automatically redirected to a surviving node.

Load balancing gives you the ability to group multiple MetaFrame servers into a virtual cluster, also known as a unified server farm, to be able to serve a growing number of connections. There is no maximum number of MetaFrame servers that can be members of a server farm beyond the limits imposed by network addressing constraints. A server farm can reside anywhere on the network, and can be load balanced across a LAN or a WAN, provided enough bandwidth is available. Exactly how much bandwidth is *enough* depends on application requirements.

To fully secure your MetaFrame environment you should combine a MetaFrame server farm with an MSCS cluster, as illustrated in Figure 51:

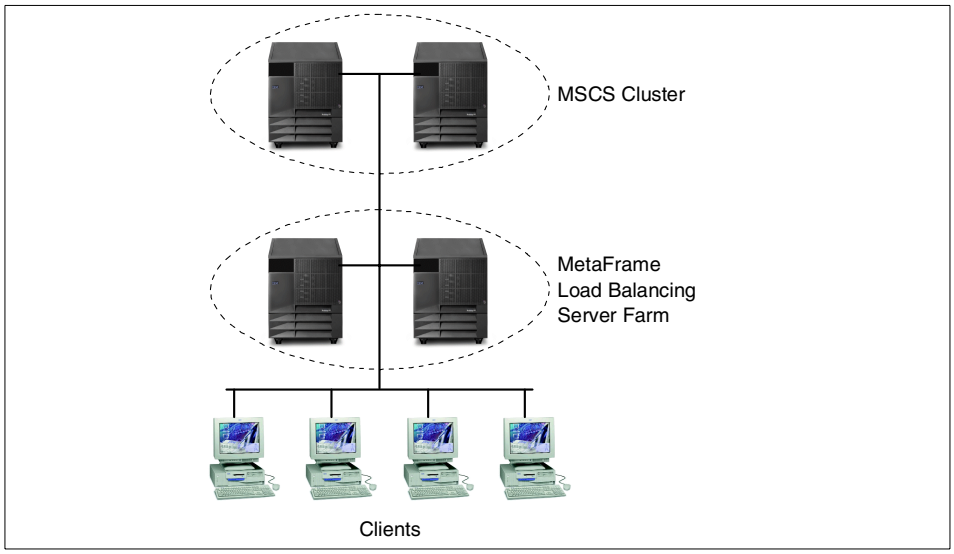

Figure 51. Citrix MetaFrame high availability

The MSCS cluster contains all user data, such as home directories, and the MetaFrame load balancing server farm contains your published applications.

Clients will access their home directories and data, hosted by the MSCS cluster (and thus highly available), while at the same time using the load balancing features of MetaFrame to access applications. In the event of a MetaFrame server failure, client requests are handled by the surviving MetaFrame server, but all sessions will be terminated, resulting in the need of an application restart from the client. Any data not saved may be lost.

Load-balanced MetaFrame servers connected to LANs or WANs can be managed from one location as a single entity, simplifying application and user management. Load balancing also ensures a level of fault tolerance, by enabling vital applications and data on a failed server to be accessed from another server in the server farm through replication.

Load Balancing Services for MetaFrame is a separate product and has to be purchased separately.

Details of Citrix products can be found at:

`http://www.citrix.com/`

Chapter 5. Clustering hardware

In this chapter we look at the xSeries and Netfinity server hardware with particular focus on the requirements for clustering. After reviewing the server systems, we examine the disk subsystem and cluster interconnect hardware in some detail because of their importance in cluster implementation. At the end of the chapter we consider the implications clustering has for implementing uninterruptible power supply protection.

Note: Many clustering products, such as Microsoft Cluster Service, require certified hardware configurations. You should ensure your configuration is certified by visiting ClusterProven at:

`http://www.pc.ibm.com/ww/netfinity/clustering/clusterproven.html`

5.1 The IBM @server xSeries server family

Power, control, scalability, reliability, and serviceability are terms used by many manufacturers when describing their server offerings. IBM has designed the xSeries and Netfinity servers to make them more than just words. Leading benchmark results, award-winning management software, and innovative X-architecture technical solutions such as Advanced System Management, Light Path Diagnostics and Active PCI combine to differentiate Netfinity systems from Intel-based systems offered by other vendors.

In addition to the basic server hardware, IBM offers a selection of disk subsystems giving customers flexibility and choice not available elsewhere. Technology leadership products, such as ServeRAID, FAStT Fibre Channel and Serial Storage Architecture (SSA) storage solutions, provide the capacity and performance to meet the needs of the most demanding application.

The xSeries and Netfinity servers are a family of Intel-based, industry-standard server products. They cover a whole range of price/performance points from entry-level, easy-to-use, low-cost servers to enterprise-level, high-performance systems.

Business-critical applications tend to need powerful systems and, it must be remembered, when a system fails, the load on the surviving systems increases. For the predominant two-node clusters, this means that one server has to be capable of carrying the cluster's entire workload. So, although clusters could be based on the smaller machines in the family, in practice, customers are likely to need the power and capacity of large systems.

As new servers are added to the family, it is the more powerful systems that are tested for clustering first. Smaller systems will be tested later if they are deemed likely to be in demand as a platform for a given clustering technology.

Review the clustering compatibility matrix for the latest configurations:

`http://www.pc.ibm.com/ww/netfinity/clustering/matrix.html`

When planning a cluster, it is important to determine what hardware is critical for particular clustering software. For instance, for MSCS it is important to have certified common disk subsystem hardware. Other solutions such as Legato StandbyServer require a supported network interface adapter and enough capacity to run the application.

More detailed information about supported operating systems and clustering software is located at:

`http://www.pc.ibm.com/us/compat/nos/matrix.shtml`

To simplify replication of tested clustering solutions, IBM has developed the ClusterProven Program. ClusterProven is designed to help leverage customers' investments by encouraging the development of solutions that provide meaningful availability and scalability benefits, as well as meeting carefully defined technical and functional requirements. Implemented cluster solutions go through the certification process and later are published at the ClusterProven Web site:

`http://www.pc.ibm.com/ww/netfinity/clustering/clusterproven.html`

5.1.1 xSeries and Netfinity systems technology features

The IBM X-architecture design blueprint has the following features:

- **Predictive Failure Analysis**

  IBM Netfinity and xSeries servers come with Predictive Failure Analysis (PFA), an IBM-developed technology that periodically measures selected attributes of a component or its activity. If a predefined threshold is exceeded, PFA sends a warning message that enables timely replacement of the failing component before the failure actually occurs. Netfinity Director can be set up to alert an administrator to this impending failure so that corrective action can be taken—often from a remote location. Specific PFA-enabled components on Netfinity servers include hard-disk drives, power supplies, VRMs, cooling fans, processors and memory.

- **Active PCI**

  Active PCI lets you upgrade your hardware and software, replace hardware and make other changes to your servers without having to shut

down your Netfinity servers. Active PCI features are designed to increase total server availability and can be described in three major categories:

  – Hot Add, allowing you to add adapters to the server to expand capacity

  – Hot Replace, allowing you to replace an adapter in the system that is no longer operating correctly

  – Failover, so that if the first adapter fails, a second adapter can pick up the workload

- **Chipkill Error Correction Code**

  Chipkill ECC memory and automatic server restart features work to minimize server downtime. IBM Chipkill memory, initially developed for NASA's space mission, is an excellent example of IBM's commitment to provide systems that remain highly available. Chipkill memory is more reliable than standard ECC DIMMs at preventing certain system memory errors. With the latest Chipkill memory technology, select xSeries and Netfinity servers will be protected from any single memory chip that fails and any number of multi-bit errors from any portion of a single memory chip.

- **Light Path Diagnostics**

  Light Path Diagnostics contribute to advanced manageability. xSeries and Netfinity servers are designed with quick problem isolation as a goal: implementing a light-path service panel in conjunction with the component LEDs. Lights are attached to specific components on selected servers. These lights illuminate at the time of a failure. Components include memory, processors, VRMs, hard-disk drives, power supplies and cooling fans. Service personnel can quickly and easily identify a failing component, potentially without even running diagnostics.

- **Advanced System Management**

  Now you can diagnose and resolve many problems without having to send an expert to your remote locations. The IBM Advanced System Management processor provides you with remote management capabilities so that problems can be diagnosed and corrected quickly, saving both time and money. It enables functions such as:

  – Auto recovery

  – Remote setup and diagnostics via the remote POST console

  – Remote access to the event log

  – Monitoring and alerting of critical server events:

    - Internal temperature and voltage thresholds

- Power status
- Power on Self Test (POST) progress
- Operating system load progress
- Operating system heartbeat
- Fans, power supplies, power subsystem
- PFA events for processors, VRMs, memory, fans, power supplies
- System/POST events and errors

– Automatic alerting capabilities include:

- Paging to numerical or alphanumerical pagers
- Sending or forwarding messages over the modem or LAN to a Netfinity Director console
- Sending messages to Netfinity Director on a local console or forwarding them over the enterprise network to a Netfinity Director console

The system management processor is standard on all midrange and high-end Netfinity servers and optional on entry-level models. Selected Netfinity servers feature a dedicated advanced system management bus on the planar.

- **FlashCopy**

  FlashCopy is a high-availability tool that minimizes application downtime associated with performing data backups, as well as increases performance by offloading host resources. This tool takes a snapshot of the source drive and places it on the target drive, which then can be extracted and used in another server or placed on tape. Previously, this type of technology was available only on high-end enterprise storage platforms.

We now examine each of the product types in the IBM xSeries and Netfinity server family. From the entry level Netfinity 1000 and 3000 systems through to the top-of-the-range Netfinity 8500R, the IBM Intel-based servers offer increasing levels of power, capacity, control and manageability. In addition, as you move up the range, more sophisticated high-availability characteristics such as Active PCI with hot-swap PCI slots, advanced ECC memory, and redundancy features are built in to make choosing a system to match your requirements a simple task.

Not all systems are supported for clustering

The servers in the following sections are the complete range of xSeries and Netfinity servers. However, not all of them are supported for clustering. For the latest support matrix, visit:

`http://www.pc.ibm.com/ww/netfinity/clustering/matrix.html`

5.1.2 xSeries 200

The xSeries 200 is an excellent value for the most cost sensitive small business - and they are easy to use and set up. These affordable uniprocessor xSeries servers are offered in several models with either Intel Celeron or Pentium III processors.

Figure 52. xSeries 200 server

Table 8 lists the specifications for the xSeries 200 server:

Table 8. xSeries 200

Component	Details
Form factor	Mini tower
CPU	One Pentium III or Celeron processor, PGA370 socket Processor speed varies by model 256 KB ECC full-speed cache with ATC on Pentium III models 128 KB cache on Celeron processor models 133 MHz front side bus (processor-to-memory bus)
PCI chipset	VIA Apollo Pro133A chipset
Memory	ECC 133 MHz unbuffered SDRAM, maximum of 1.5 GB Three DIMM sockets Installed amount varies by model
SCSI (non-RAID) controller	Adaptec 29160LP Wide Ultra160 SCSI PCI adapter on SCSI models One 16-bit SCSI channel (internal port only) Dual integrated EIDE controllers
RAID controller	None standard (ServeRAID supported)
Disk bays	Two 5.25” HH (1 for CD-ROM) Five 3.5” SL bays (1 for diskette) No hot-swap support
Adapter slots	Five full-length, 33 MHz PCI slots
Ethernet	Intel 82559 100/10 controller on planar
System management	Monitoring of processor temperature, voltage
Video	S3 Savage 4 LT adapter installed in the AGP slot 8 MB SDRAM
Power	Single 330 W power supply

5.1.3 xSeries 220

The xSeries 220 is aimed at entry-business and workgroup applications to deliver value and ease of use. Now, powerful two-way, SMP-capable 133 MHz FSB Pentium III processors, high-speed memory, high-bandwidth PCI buses, and Ultra160 SCSI data storage let you optimize your growing business network applications — all at an attractively low price.

Figure 53. xSeries 220

Table 9 lists the specifications for the xSeries 220:

Table 9. xSeries 220

Component	Details
Form factor	Mini tower
CPU	1-2 Pentium III processors, PGA370 sockets Processor speed varies by model 256 KB ECC full-speed cache with Advanced Transfer Cache 133 MHz front-side bus (processor-to-memory bus)
PCI chipset	ServerWorks ServerSet III LE, 2 PCI buses (32-bit, 64-bit), PCI 2.1 33 MHz
Memory	ECC 133 MHz SDRAM, maximum of 4 GB Four DIMM sockets Installed amount varies by model
SCSI (non-RAID) controller	Adaptec AHA-7892 Wide Ultra160 SCSI (160 MBps) on planar One channel (internal port only)
RAID controller	None standard (ServeRAID supported)
Disk bays	Two 5.25” HH (1 for CD-ROM) Five 3.5” SL bays (1 for diskette) No hot-swap support
Adapter slots	Three full-length 64-bit PCI 2.1 slots Two full-length 32-bit PCI 2.1 slots
Ethernet	Intel 82559 100/10 controller on planar
System management	Two ADM1024 chips (processor temperature, fans, voltage) Optional IBM Remote Supervisor Adapter supported
Video	S3 Savage 4 on planar 8 MB SDRAM
Power	Single 330 W power supply

5.1.4 xSeries 230

The xSeries 230 brings performance, power, and function to mainstream business applications. These servers use two-way SMP-capable, 133 MHz front-side bus Pentium III processors coupled with a 64-bit PCI bus and Ultra160 SCSI, and are packaged in a compact 5U case with ample bays to support general-purpose database, file, or print serving business applications.

The xSeries 230 is a follow-on to the Netfinity 5100 which is described on page 172.

Figure 54. xSeries 230

Table 10 lists the specifications for the xSeries 230:

Table 10. xSeries 230 specifications

Component	Details
Form factor	Rack device 5U or tower (tower-to-rack conversion kit available)
CPU	1-2 CPUs, Pentium III, Slot 1 connector 256 KB ECC cache full speed (Advanced Transfer Cache) 133 MHz front side bus Processor speeds vary by model
PCI chipset	ServerWorks ServerSet III LE chipset Two PCI buses (one 32-bit, one 64-bit), PCI 2.1 33 MHz
Memory	ECC 133 MHz registered SDRAM Four DIMM sockets 4 GB maximum, installed amount varies by model
SCSI (non-RAID)	Adaptec AHA-7899 Wide Ultra160 SCSI (160 MBps) on planar 64-bit PCI, two channels (internal, external)
RAID controller	None standard (ServeRAID supported)
Disk bays	Six 3.5” SL hot-swappable Three 5.25” HH (1 for CD-ROM), one 3.5” for diskette
Adapter slots	Three full-length 64-bit PCI 2.1 (non-hot-swap) Two full-length 32-bit PCI 2.1 slots (non-hot-swap)
Ethernet	AMD Am79C975 (32-bit PCI bus) on planar, 100/10 Mbps
System management	IBM Advanced System Management Processor on planar Light Path Diagnostics
Video	S3 Savage4 on planar 8 MB SDRAM
Power	One 250 W hot-swap power supply standard Two additional 250 W supplies optional Redundant with two supplies for power requirements <250 W Redundant with three supplies for power requirements >250 W

5.1.5 xSeries 240

The xSeries 240 is a powerful two-way SMP-capable, high-availability server packaged in a compact 5U case. CPU speeds start at 1 GHz. It is a follow-on to the Netfinity 5600, which is described on page 174.

Figure 55. xSeries 240

Table 11 lists the specifications for the xSeries 240:

Table 11. xSeries 240 specifications

Component	Details
Form factor	Rack device 5U or tower (tower-to-rack conversion kit available)
CPU	1-2 CPUs, Pentium III Slot 1 256 KB ECC cache full speed (Advanced Transfer Cache) 133 MHz front-side bus Processor speeds vary by model
PCI chipset	ServerWorks ServerSet III LE chipset Two PCI buses (one 32-bit, one 64-bit), PCI 2.1 33 MHz
Memory	ECC 133 MHz registered SDRAM Four DIMM sockets 4 GB maximum, installed amount varies by model
SCSI (non-RAID)	Adaptec AHA-7897 Wide Ultra2 SCSI (80 MBps) 64-bit PCI, two channels (internal, external)
RAID controller	None standard (ServeRAID supported)
Disk bays	Six 3.5” SL hot-swappable Three 5.25” HH (1 for CD-ROM), one 3.5” for diskette
Adapter slots	Three full-length 64-bit PCI 2.1 hot-swap slots Two full-length 32-bit PCI 2.1 slots (not hot-swap)
Ethernet	AMD Am79C975 (32-bit PCI bus) on planar, 100/10 Mbps
System management	IBM Advanced System Management processor on planar Light Path Diagnostics
Video	S3 Savage4 on planar 8 MB SDRAM
Power	Two 250 W hot-swap power supplies, redundant at <250 W Optional 250 W hot-swap for >250 W redundancy

5.1.6 xSeries 330

The xSeries 330 is a new 1U-high rack-drawer footprint. This rack-optimized platform features two-way, SMP-capable power, high availability, scalability, and a large internal data storage capacity. It is ideal for compute-intensive Web-based or enterprise network applications where space is of primary importance.

Figure 56. xSeries 330

Figure 57. 42 xSeries 330s in a rack

Table 12 lists the specifications for the xSeries 330:

Table 12. xSeries 330

Component	Details
Form factor	Rack device 1U
CPU	One or two CPUs, Pentium III, PGA370 socket Processor speed varies by model 256 KB ECC full-speed cache with Advanced Transfer Cache 133 MHz front-side bus
PCI chipset	ServerWorks ServerSet III LE 2 PCI 2.1 buses: one 32-bit 33 MHz primary, one 64-bit 33 MHz secondary
Memory	ECC 133 MHz SDRAM, maximum of 4 GB Four DIMM sockets Installed amount varies by model
SCSI (non-RAID) controller	Adaptec AHA-7892 Wide Ultra160 SCSI (160 MBps) Integrated on planar, one internal channel
RAID controller	None. (ServeRAID-4L supported)
Disk bays	One 5.25” SL (for CD-ROM) One 3.5” SL (for diskette) Two 3.5” SL hot-swap bays
Adapter slots	One full-length 64-bit PCI 2.1 slot One half-length 64-bit PCI 2.1 slot
Ethernet	Two Intel 82559 controllers (PCI bus) on planar, 100/10 Mbps
System management	Integrated Advanced System Management Processor
Video	S3 Savage4 Graphics Accelerator with 8 MB SDRAM
Power	Single 200 W power supply

5.1.7 xSeries 340

The xSeries 340 is a very high-density, two-way SMP-capable 3U rack server with power, scalability, control, and serviceability to handle your networked business-critical applications. With its 1 GHz or faster CPUs, It is ideal for compute-intensive Web-based or enterprise network applications where space is a primary consideration.

The xSeries 340 is a follow-on to the Netfinity 4500R, which is described on page 170.

Figure 58. xSeries 340

Table 13 lists the specifications for the xSeries 340:

Table 13. xSeries 340 specifications

Component	Details
Form factor	Rack device 3U
CPU	One or two CPUs, Pentium III Slot 1 Processor speed varies by model 256 KB ECC cache, full speed of the CPU 133 MHz front-side bus
PCI chipset	ServerWorks ServerSet III LE 2 PCI buses, 1x64-bit, 1x32-bit, PCI 2.1 33 MHz
Memory	ECC 133 MHz SDRAM, maximum of 4 GB Four DIMM sockets Installed amount varies by model
SCSI (non-RAID) controller	Adaptec AHA-7899 Wide Ultra160 SCSI (160 MBps) on planar Two internal channels (no external port)
RAID controller	None standard (ServeRAID supported)
Disk bays	Two 3.5” Ultraslim (US - 0.8”) (for CD-ROM and diskette) One 3.5” SL (for hot-swap drives) Two 5.25” HH bays (can be converted to three hot-swap SL drive bays)
Adapter slots	Two full-length 32-bit PCI 2.1 slot Three full-length 64-bit PCI 2.1 slot
Ethernet	AMD Am79C975 100/10 controller on planar
System management	IBM Advanced System Management Processor, on planar Light Path Diagnostics
Video	S3 Savage4 on planar 8 MB SDRAM
Power	Single 270 W power supply Second 270 W optional supply for redundancy

5.1.8 Netfinity 1000

The Netfinity 1000 is aimed at the most cost-sensitive small business. This affordable server, using an Intel Pentium III processor, gives you the power to make small business applications run faster and handle more complex networking requirements.

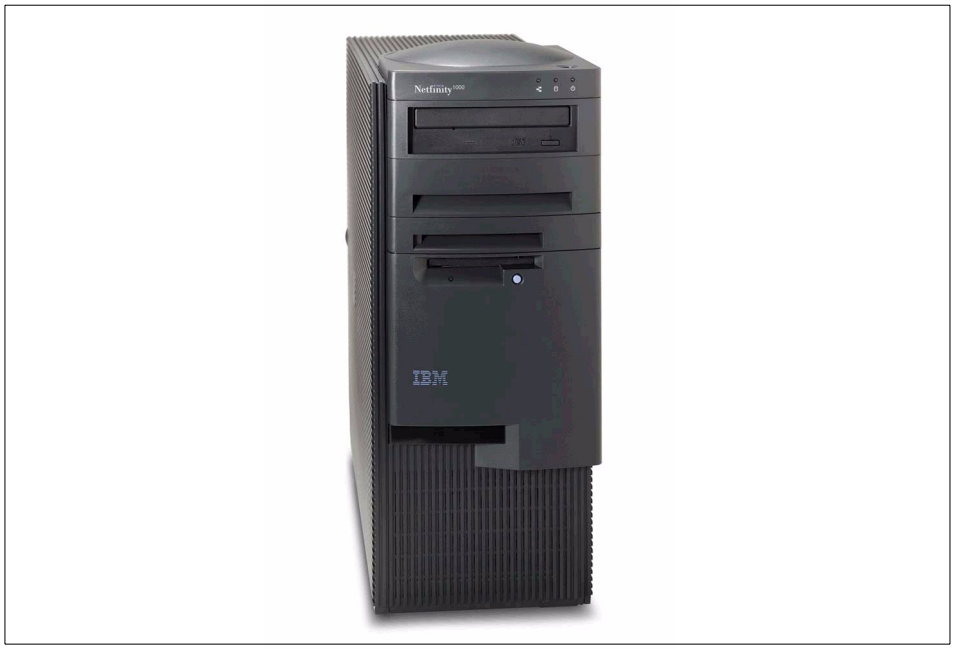

Figure 59. Netfinity 1000

Table 14 lists the specifications for the Netfinity 1000:

Table 14. Netfinity 1000

Component	Details
Form factor	Mini tower
CPU	Single CPU, Pentium III Slot 1, speed varies by model L2 cache size and speed varies by model 100 MHz front-side bus
PCI chipset	Intel 440 BX, 1 PCI bus, PCI 2.1 33 MHz
Memory	ECC 100 MHz SDRAM, maximum of 768 MB Three DIMM sockets Installed amount varies by model
Disk controller	Enhanced IDE on planar
Disk bays	Two 5.25” HH (1 for CD-ROM) Four 3.5” SL bays (1 for diskette) No hot-swap support
Adapter slots	Three full-length 32-bit PCI 2.1 slots Three full-length ISA slots One 32-bit AGP
Ethernet	Intel EtherExpress PRO/100B (PCI bus) PCI 82558 chip on planar
System management	LM80 compatible Advanced Systems Management Adapter (ISA) supported
Video	S3 Trio3D, 4 MB 100 MHz SGRAM
Power	Single 330 W power supply

5.1.9 Netfinity 3000

The Netfinity 3000 delivers excellent price performance and excellent functionality to the entry server marketplace. The affordable Netfinity 3000, using the Intel Pentium III processor technology, can make your business applications run faster or handle more complex networking requirements.

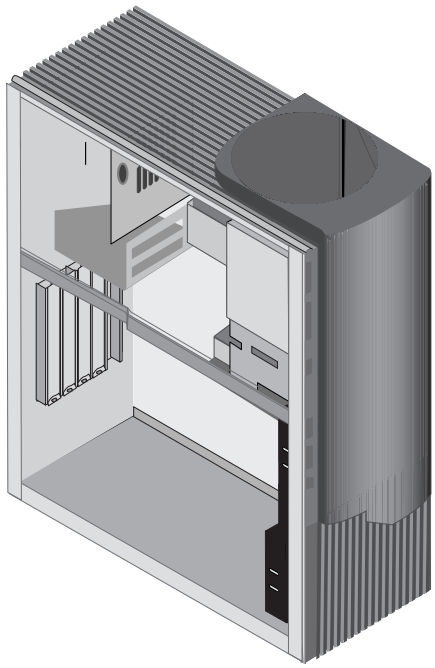

Figure 60. Netfinity 3000

Table 15 lists the specifications for the Netfinity 3000:

Table 15. Netfinity 3000

Component	Details
Form factor	Mini tower
CPU	Single CPU, Pentium III Slot 1, speed varies by model 256 KB ECC cache, full speed of the CPU 100 MHz front-side bus
PCI chipset	Intel 440 BX, 1 PCI bus, PCI 2.1 33 MHz
Memory	ECC 100 MHz SDRAM, maximum of 768 MB Three DIMM sockets Installed amount varies by model
SCSI (non-RAID) controller	Adaptec AHA-2940U2W Wide Ultra2 SCSI (80 MBps) on planar Single channel (internal and external port)
RAID controller	None standard (ServeRAID supported)
Disk bays	Two 5.25” HH (1 for CD-ROM) Four 3.5” SL bays (1 for diskette) No hot-swap support
Adapter slots	Three full-length 32-bit PCI 2.1 slots Three full-length ISA slots One 32-bit AGP
Ethernet	Intel EtherExpress PRO/100B (PCI bus) PCI 82558 chip on planar
System management	LM80 compatible Advanced Systems Management Adapter (ISA) supported
Video	S3 Trio3D, 4 MB 100 MHz SGRAM
Power	Single 330 W power supply

5.1.10 Netfinity 3500 M20

The Netfinity 3500 M20 is a powerful, SMP-capable server that offers great functionality to the entry-server marketplace. This affordable Pentium III-based server has the muscle to make your business applications run faster while providing two-way SMP scalability for future growth.

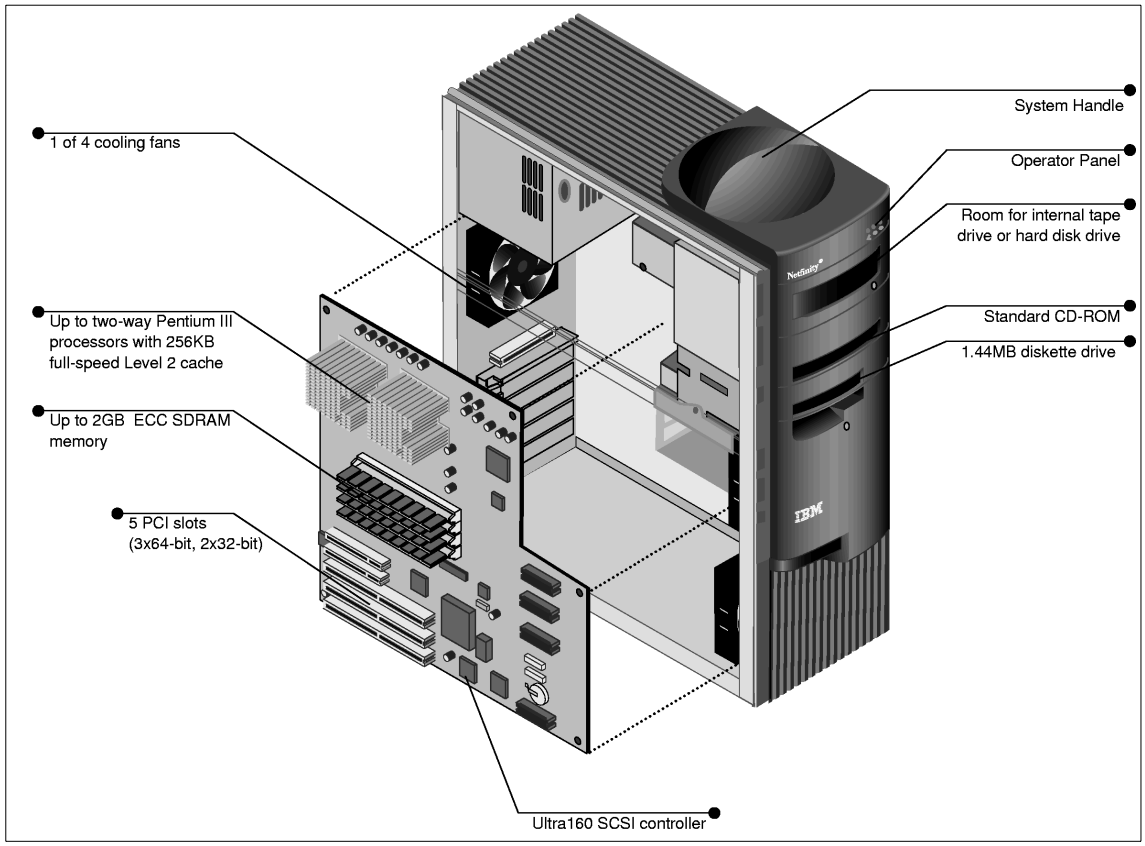

Figure 61. Netfinity 3500 M20

Table 16 lists the specifications for the Netfinity 3500 M20:

Table 16. Netfinity 3500 M20

Component	Details
Form factor	Mini tower
CPU	1-2 Pentium III processors, PGA370 sockets Processor speed varies by model 256 KB ECC cache, some models use Advanced Transfer Cache 133 MHz front-side bus (processor-to-memory bus)
PCI chipset	ServerWorks ServerSet III LE, 2 PCI buses (32-bit, 64-bit), PCI 2.1 33 MHz
Memory	ECC 100 MHz SDRAM, maximum of 2GB Four DIMM sockets Installed amount varies by model
SCSI (non-RAID) controller	Adaptec AHA-7892 Wide Ultra160 SCSI (160 MBps) on planar One channel (internal port only)
RAID controller	None standard (ServeRAID supported)
Disk bays	Two 5.25” HH (1 for CD-ROM) Five 3.5” SL bays (1 for diskette) No hot-swap support
Adapter slots	Two full-length 32-bit PCI 2.1 slots Three full-length 64-bit PCI 2.1 slots
Ethernet	Intel 82559 controller on planar
System management	Two ADM1024 chips (processor temperature, fans, voltage)
Video	S3 Savage4 on planar 8 MB SDRAM
Power	Single 330 W power supply

5.1.11 Netfinity 4000R

The IBM Netfinity 4000R is a powerful, ultrathin, rack-mount server designed specifically for high density, Web server environments. This SMP-capable, Pentium III-based server packs a tremendous amount of power and function into a space-saving, high-density 1U rack drawer.

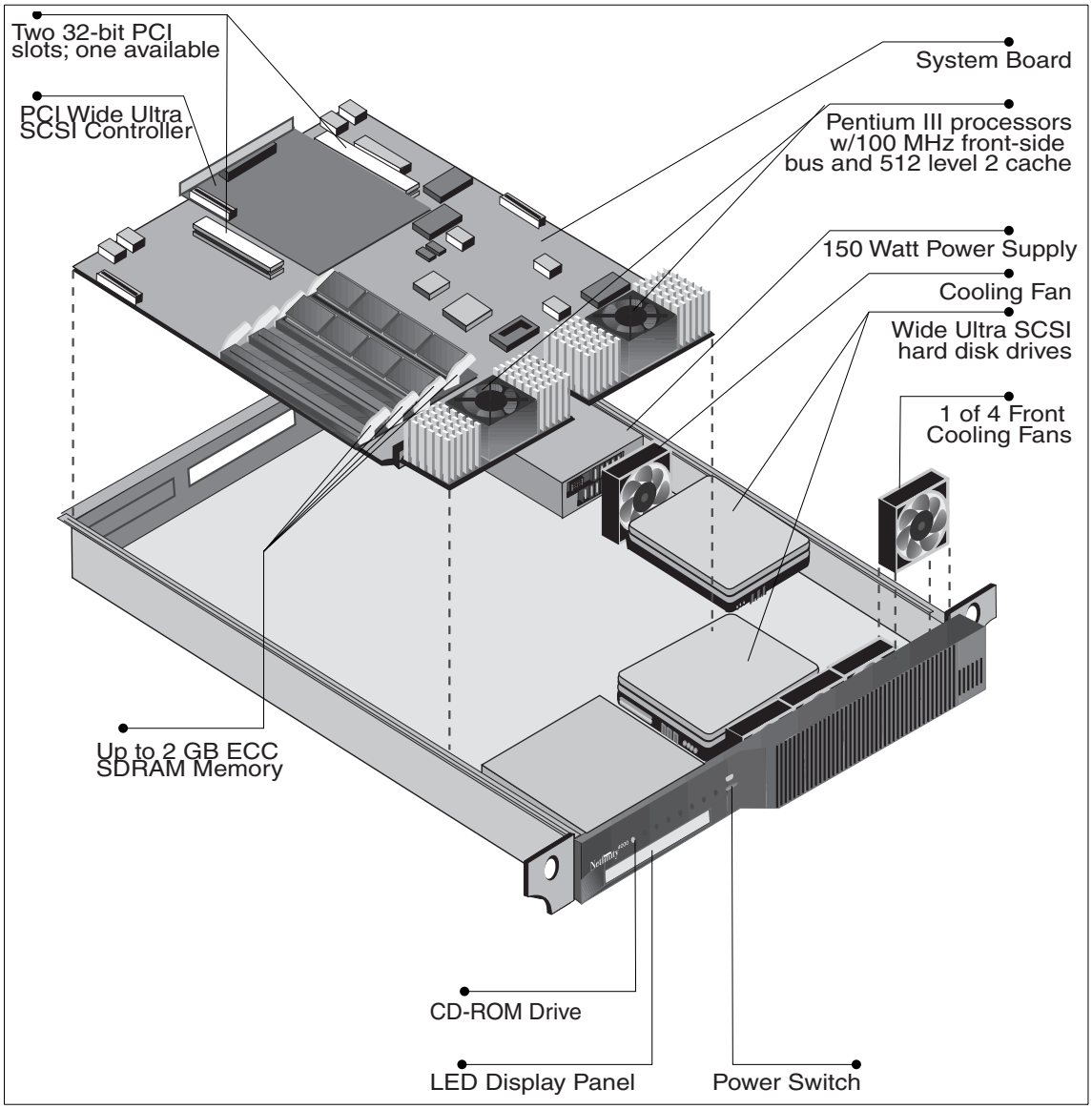

Figure 62. Netfinity 4000R

Table 17 lists the specifications for the Netfinity 4000R:

Table 17. Netfinity 4000R

Component	Details
Form factor	Rack device 1U
CPU	One or two CPUs, Pentium III Slot 1 Processor speed varies by model 256 KB ECC cache, full speed of the CPU 100 MHz front side bus
PCI chipset	Intel 440 GX, 1 PCI bus, PCI 2.1 33 MHz 32-bit
Memory	ECC 100 MHz SDRAM, maximum of 2 GB Four DIMM sockets Installed amount varies by model
SCSI (non-RAID) controller	Adaptec AHA-2940UW Wide Ultra SCSI (40 MBps) Full-length PCI adapter, one channel (internal and external port)
RAID controller	None standard (ServeRAID supported)
Disk bays	One 5.25” SL (for CD-ROM) Two 3.5” SL bays No hot-swap support, no diskette drive
Adapter slots	One full-length 32-bit PCI 2.1 slot One half-length 32-bit PCI 2.1 slot
Ethernet	Two Intel 82559 controllers (PCI bus) on planar, 100/10 Mbps
System management	ST Micro ST72251 voltage/temperature monitor
Video	Chips and Technologies B69000 HiQVideo 8 MB 110 MHz SDRAM
Power	Single 150 W power supply

5.1.12 Netfinity 4500R

Netfinity 4500R is a very high-density, two-way SMP-capable 3U rack server with power, scalability, control, and serviceability to handle your networked business-critical applications. It is ideal for compute-intensive Web-based or enterprise network applications where space is a primary consideration

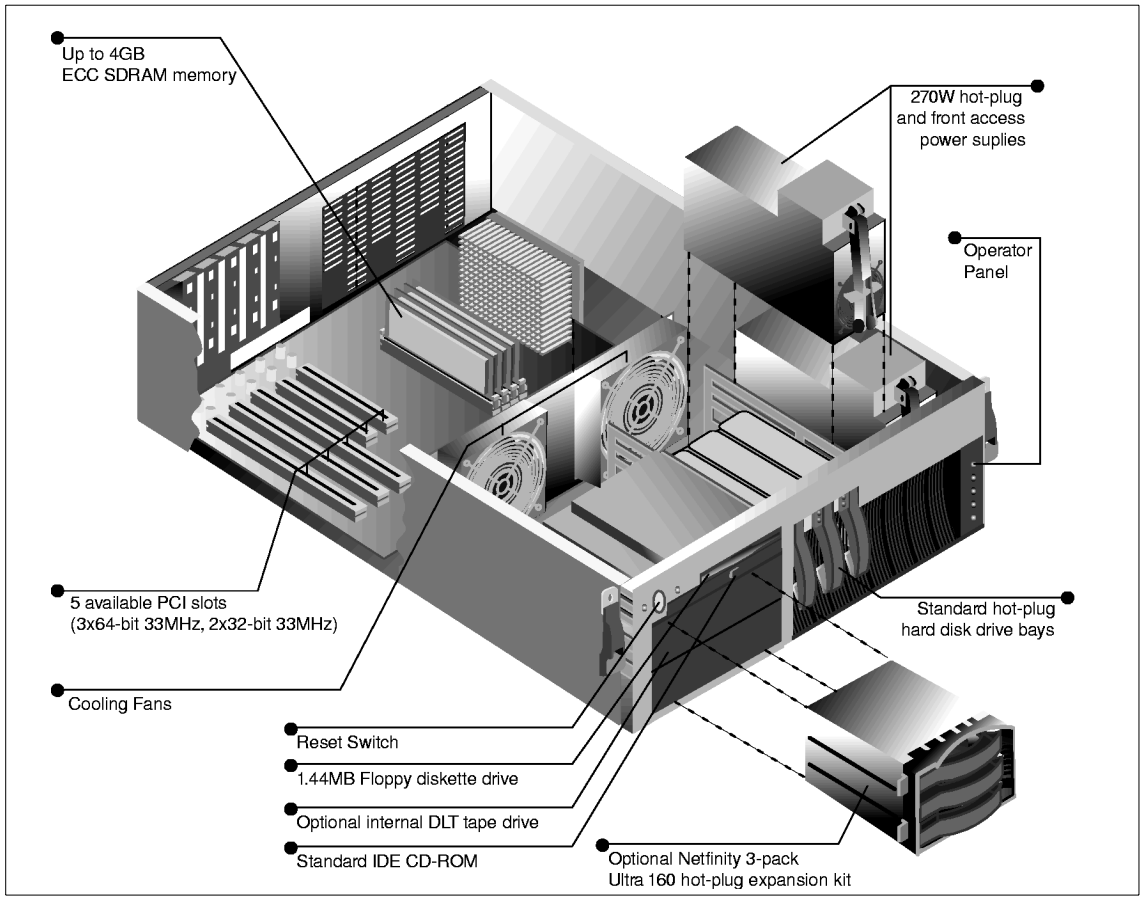

Figure 63. Netfinity 4500R

Table 18 lists the specifications for the Netfinity 4000R:

Table 18. Netfinity 4500R specifications

Component	Details
Form factor	Rack device 3U
CPU	One or two CPUs, Pentium III Slot 1 Processor speed varies by model 256 KB ECC cache, full speed of the CPU 133 MHz front-side bus
PCI chipset	ServerWorks ServerSet III LE 2 PCI buses, 1x64-bit, 1x32-bit, PCI 2.1 33 MHz
Memory	ECC 133 MHz SDRAM, maximum of 4GB Four DIMM sockets Installed amount varies by model
SCSI (non-RAID) controller	Adaptec AHA-7899 Wide Ultra160 SCSI (160 MBps) on planar Two internal channels (no external port)
RAID controller	None standard (ServeRAID supported)
Disk bays	Two 3.5” Ultraslim (US - 0.8”) (for CD-ROM and diskette) One 3.5” SL (for hot-swap drives) Two 5.25” HH bays (can be converted to three hot-swap SL drive bays)
Adapter slots	Two full-length 32-bit PCI 2.1 slot Three full-length 64-bit PCI 2.1 slot
Ethernet	AMD Am79C975 100/10 controller on planar
System management	IBM Advanced System Management Processor, on planar Light Path Diagnostics
Video	S3 Savage4 on planar 8 MB SDRAM
Power	Single 270 W power supply Second 270 W optional supply for redundancy

5.1.13 Netfinity 5100

The Netfinity 5100 brings performance, power, and function to mainstream business applications. These servers use two-way SMP-capable, 133 MHz front-side bus Pentium III processors coupled with a 64-bit PCI bus and Ultra160 SCSI, and are packaged in a compact 5U mechanical with ample bays to support general-purpose database, file, or print serving business applications.

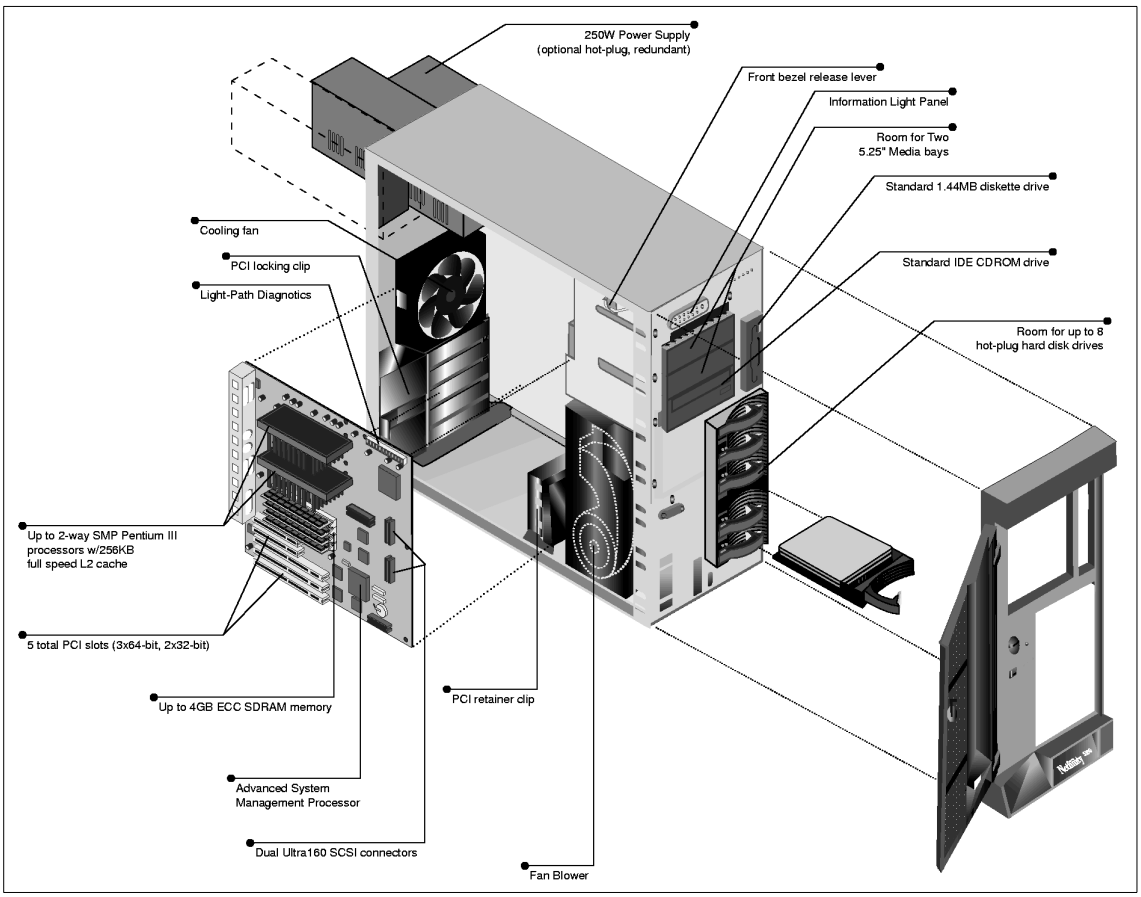

Figure 64. xSeries 230 / Netfinity 5100

Table 19 lists the specifications for the Netfinity 5100:

Table 19. Netfinity 5100 specifications

Component	Details
Form factor	Rack device 5U or tower (tower-to-rack conversion kit available)
CPU	1-2 CPUs, Pentium III Slot 1 256 KB ECC cache full speed (Advanced Transfer Cache) 133 MHz front side bus Processor speeds vary by model
PCI chipset	ServerWorks ServerSet III LE chipset Two PCI buses (one 32-bit, one 64-bit), PCI 2.1 33 MHz
Memory	ECC 133 MHz registered SDRAM Four DIMM sockets 4 GB maximum, installed amount varies by model
SCSI (non-RAID)	Adaptec AHA-7899 Wide Ultra160 SCSI (160 MBps) 64-bit PCI, two channels (internal, external)
RAID controller	None standard (ServeRAID supported)
Disk bays	Six 3.5” SL hot-swappable Three 5.25” HH (1 for CD-ROM), one 3.5” for diskette
Adapter slots	Three full-length 64-bit PCI 2.1 (non-hot-swap) Two full-length 32-bit PCI 2.1 slots (non-hot-swap)
Ethernet	AMD Am79C975 (32-bit PCI bus) on planar, 100/10 Mbps
System management	IBM Advanced System Management Processor on planar Light Path Diagnostics
Video	S3 Savage4 on planar 8 MB SDRAM
Power	One 250 W hot-swap power supply standard Two additional 250 W supplies optional Redundant with two supplies for power requirements <250 W Redundant with three supplies for power requirements >250 W

5.1.14 Netfinity 5600

The Netfinity 5600 is a powerful two-way SMP-capable, high-availability servers packaged in a compact 5U mechanical. This new platform is perfect for business-critical applications spanning customer sets from small to large business.

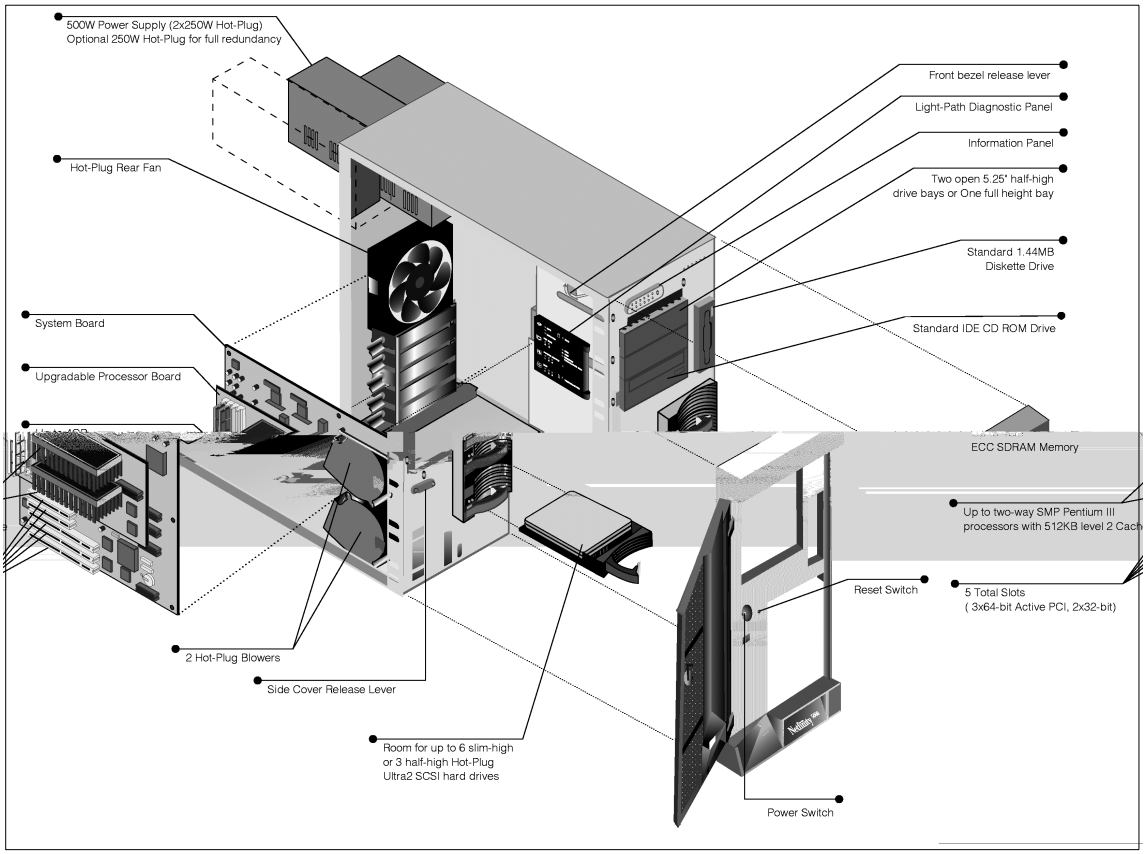

Figure 65. Netfinity 5600

Table 20 lists the specifications for the Netfinity 5600:

Table 20. Netfinity 5600 specifications

Component	Details
Form factor	Rack device 5U or tower (tower-to-rack conversion kit available)
CPU	1-2 CPUs, Pentium III Slot 1 256 KB ECC cache full speed (Advanced Transfer Cache) 133 MHz front-side bus Processor speeds vary by model
PCI chipset	ServerWorks ServerSet III LE chipset Two PCI buses (one 32-bit, one 64-bit), PCI 2.1 33 MHz
Memory	ECC 133 MHz registered SDRAM Four DIMM sockets 4 GB maximum, installed amount varies by model
SCSI (non-RAID)	Adaptec AHA-7897 Wide Ultra2 SCSI (80 MBps) 64-bit PCI, two channels (internal, external)
RAID controller	None standard (ServeRAID supported)
Disk bays	Six 3.5” SL hot-swappable Three 5.25” HH (1 for CD-ROM), one 3.5” for diskette
Adapter slots	Three full-length 64-bit PCI 2.1 hot-swap slots Two full-length 32-bit PCI 2.1 slots (not hot-swap)
Ethernet	AMD Am79C975 (32-bit PCI bus) on planar, 100/10 Mbps
System management	IBM Advanced System Management processor on planar Light Path Diagnostics
Video	S3 Savage4 on planar 8 MB SDRAM
Power	Two 250 W hot-swap power supplies, redundant at <250 W Optional 250 W hot-swap for >250 W redundancy

5.1.15 Netfinity 6000R

The Netfinity 6000R is a 4U rack drawer that provide four-way SMP-capable power, advanced high availability, scalability, and a surprisingly large internal data storage capacity. It is ideal for compute-intensive Web-based or enterprise network applications where space is a primary consideration.

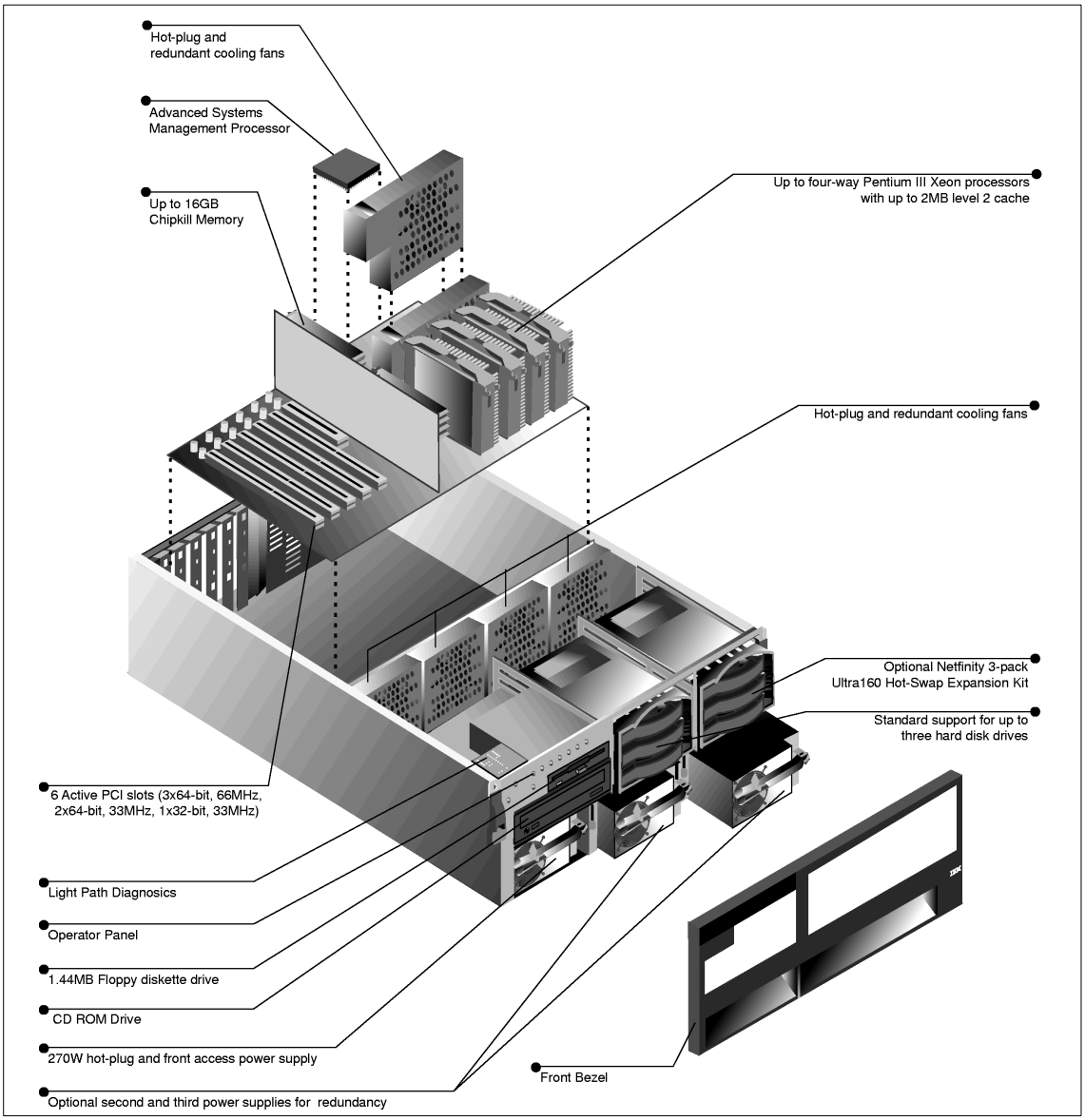

Figure 66. Netfinity 6000R

Table 21 lists the specifications for the Netfinity 6000R:

Table 21. Netfinity 6000R

Component	Details
Form factor	Rack device 4U
CPU	1-4 CPUs, Pentium III Xeon Slot 2 Processor speeds vary by model ECC cache full speed, size varies by processor 100 MHz front-side bus
PCI chipset	ServerWorks ServerSet II HE 3 PCI 2.2 buses (32-bit 33 MHz, 64-bit 66 MHz, 64-bit 33 MHz)
Memory	ECC 100 MHz registered SDRAM Installed amount varies by model 32 DIMM sockets, 16 GB maximum Chipkill multi-bit error correction
SCSI (non-RAID)	Adaptec AHA-7899 Wide Ultra160 SCSI (160 MBps) 64-bit PCI, two channels (internal, external)
RAID controller	None standard (ServeRAID supported)
Disk bays	Three 3.5” SL hot-swap drive bays Three additional 3.5” SL hot-swap bays optional One 5.25” HH for CD-ROM, one 3.5” for diskette
Adapter slots	Six total PCI 2.2 hot-swap slots: • One full-length 32-bit 33 MHz • Three full-length 64-bit 66 MHz • Two full-length 64-bit 33 MHz
Ethernet	AMD Am79C975 on planar, 100/10 Mbps
System management	IBM Advanced System Management Processor Light Path Diagnostics
Video	S3 Savage4 on planar 8 MB SDRAM
Power	One 270 W hot-swap power supply standard, three maximum Two redundant with power consumption <270 W

5.1.16 Netfinity 7100

The Netfinity 7100 offers four-way SMP processing power and scalability to handle advanced enterprise network server applications. With high-performance Pentium III Xeon processors, these servers deliver excellent performance while giving you high-availability and remote systems management features you need to handle business-critical applications.

Figure 67. Netfinity 7100

Table 22 lists the specifications for the Netfinity 7100:

Table 22. Netfinity 7100

Component	Details
Form factor	Rack device 8U
CPU	1-4 CPUs, Pentium III Xeon Slot 2 Processor speeds vary by model ECC cache full speed, size varies by processor 100 MHz front-side bus
PCI chipset	ServerWorks ServerSet III HE chipset 3 PCI 2.1 buses: (64-bit 66 MHz, 64-bit 33 MHz, 32-bit 33 MHz)
Memory	ECC 100 MHz registered SDRAM Installed amount varies by model 16 DIMM sockets, 16 GB maximum Optional Chipkill multi-bit error correction
SCSI (non-RAID)	Adaptec AHA-7896 Wide Ultra2 SCSI (80 MBps) 64-bit PCI, two channels (internal, external)
RAID controller	None standard (ServeRAID supported)
Disk bays	Ten 3.5” SL hot-swap drive bays (or seven HH hot-swap drive bays) Three 5.25” HH (1 for CD-ROM), one 3.5” for diskette
Adapter slots	Four full-length 64-bit 33 MHz PCI 2.1 slots (optional hot-swap upgrade) Two full-length 64-bit 66 MHz PCI 2.1 slots
Ethernet	AMD Am79C975 on planar, 100/10 Mbps
System management	IBM Advanced System Management Processor Light Path Diagnostics
Video	S3 Trio3D on I/O function card, 4 MB SGRAM
Power	Two 250 W hot-swap supplies; four maximum Need a minimum of three for redundancy, depending on consumption

5.1.17 Netfinity 7600

The Netfinity 7600 is a four-way SMP system that is similar in configuration to the Netfinity 7100. However, it offers standard features such as hot-swap Active PCI slots, Chipkill multibit memory error correction and a standard three-channel ServeRAID controller.

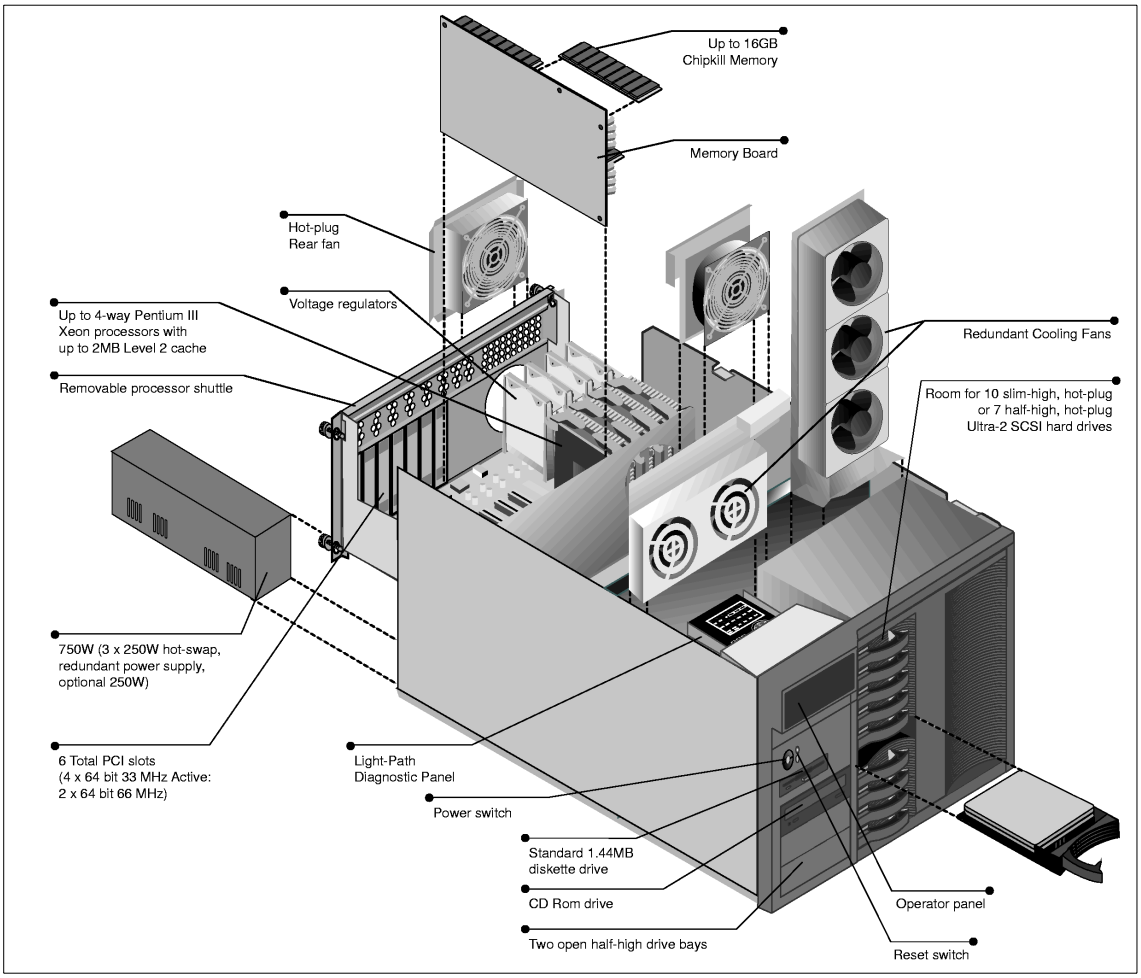

Figure 68. Netfinity 7600

Table 23 lists the specifications for the Netfinity 7600:

Table 23. Netfinity 7600

Component	Details
Form factor	Rack device 8U
CPU	1-4 CPUs, Pentium III Xeon Slot 2 Processor speeds vary by model ECC cache full speed, size varies by processor 100 MHz front-side bus
PCI chipset	ServerWorks ServerSet III HE chipset 3 PCI 2.1 buses: (64-bit 66 MHz, 64-bit 33 MHz, 32-bit 33 MHz)
Memory	ECC 100 MHz registered SDRAM Installed amount varies by model 16 DIMM sockets, 16 GB maximum Chipkill multi-bit error correction
SCSI (non-RAID)	Adaptec AHA-7896 Wide Ultra2 SCSI (80 MBps) 64-bit PCI, two channels (internal, external)
RAID controller	ServeRAID-3HB standard
Disk bays	Ten 3.5” SL hot-swap drive bays (or seven HH hot-swap drive bays) Three 5.25” HH (1 for CD-ROM), one 3.5” for diskette
Adapter slots	Four full-length 64-bit 33 MHz PCI 2.1 hot-swap slots Two full-length 64-bit 66 MHz PCI 2.1 slots
Ethernet	AMD Am79C975 on planar, 100/10 Mbps
System management	IBM Advanced System Management Processor Light Path Diagnostics
Video	S3 Trio3D on I/O function card, 4 MB SGRAM
Power	Two 250 W hot-swap supplies; four maximum Need a minimum of three for redundancy, depending on consumption

5.1.18 Netfinity 8500R

IBM Netfinity 8500R advanced, eight-way SMP-capable enterprise servers are optimized for advanced clustering and storage area network (SAN) environments. Using high-speed Pentium III Xeon processors, they pack incredible performance and scalability into a dense 8U rack-mountable package.

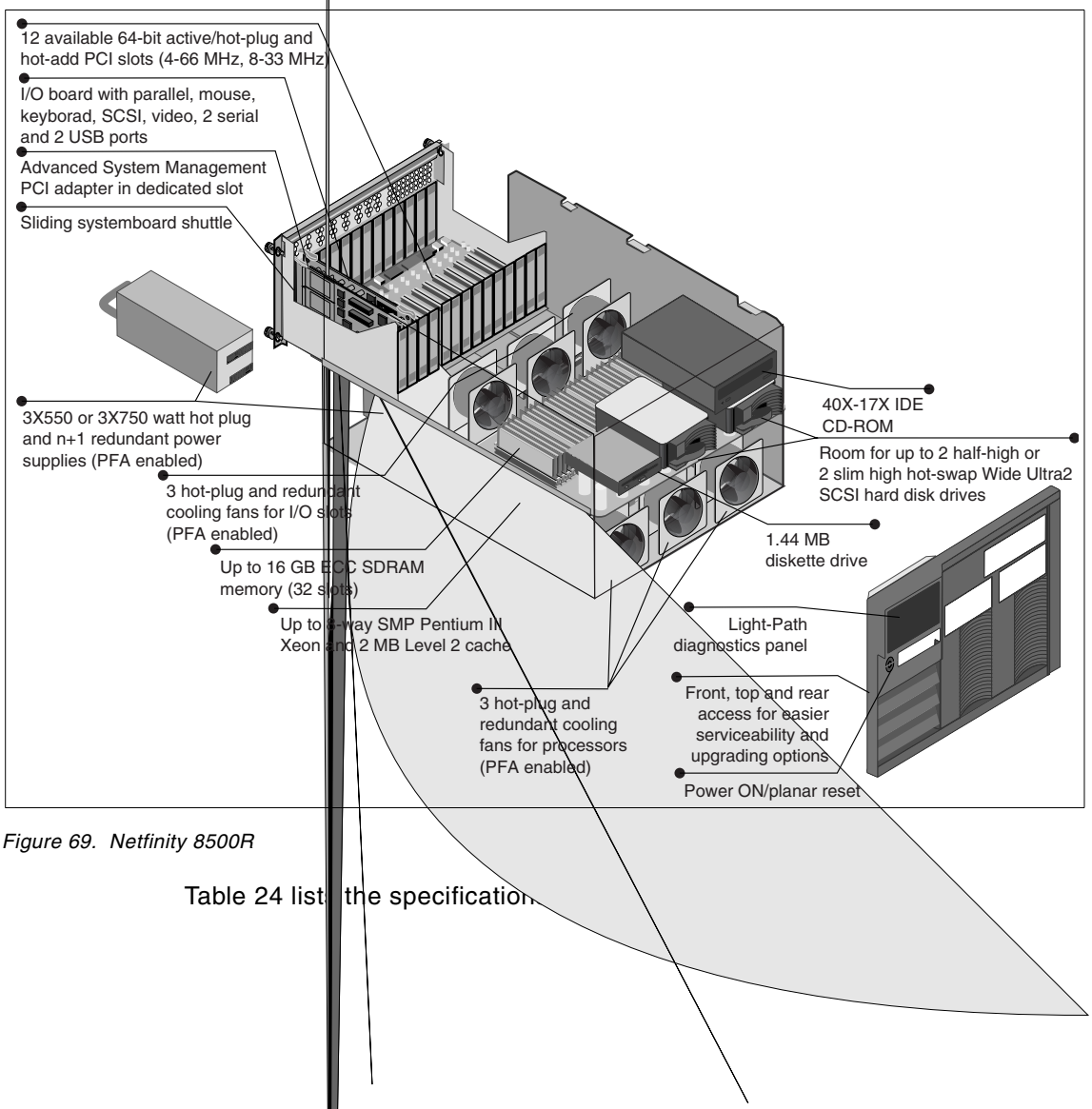

Figure 69. Netfinity 8500R

Table 24 lists the specification

Table 24. Netfinity 8500R

Component	Details
Form factor	Rack device 8U
CPU	1-8 CPUs, Pentium III Xeon Slot 2 Processor speeds vary by model ECC cache full speed, size varies by processor 100 MHz front-side bus
PCI chipset	Intel Profusion chipset 4 PCI buses, 64-bit PCI 2.2 (two 33 MHz, two 66 MHz)
Memory	ECC 100 MHz registered SDRAM Installed amount varies by model 32 DIMM sockets, 32 GB maximum
SCSI (non-RAID)	Adaptec AHA-7897 Wide Ultra2 SCSI (LVD 80 MBps) 64-bit PCI, two channels (internal, external)
RAID controller	None standard (ServeRAID supported)
Disk bays	Two 3.5” HH hot-swappable One 5.25” HH for CD-ROM, one 3.5” for diskette
Adapter slots	Four full-length 64-bit 66 MHz PCI 2.2 hot-swap slots Eight full-length 64-bit 33 MHz PCI 2.2 hot-swap slots
Ethernet	None standard
System management	IBM Advanced System Management PCI adapter Light Path Diagnostics LED panel with two 16-character lines for status
Video	S3 Trio3D on I/O function card, 4 MB SGRAM
Power	Three hot-swap redundant supplies (each 750 W at 220 Volts, 550 W at 110 Volts); for 220 Volts connections, only two required for redundancy

5.2 Disk subsystems

The disk subsystem is one of the potential bottlenecks in your cluster. If you examine benchmark test configurations in detail, you will discover that the systems that deliver the leading results always have very large disk configurations. In some cases, there may be 80 or more drives attached to a single server to get sufficient data to keep the processors busy. Implementing a fast and flexible storage solution from the start will save you time and money later.

The IBM Redbook *Netfinity Server Disk Subsystems*, SG24-2098 is the definitive guide to xSeries and Netfinity disk subsystems. It is aimed at technical staff within IBM, customers, and business partners who wish to understand the range of available storage options for the IBM Intel-processor servers. Reading it will provide you with sufficient information to be able to make informed decisions when selecting disk subsystems for your servers. It will also prove invaluable to anyone involved in the purchase, support, sale, and use of these leading-edge storage solutions.

When building a cluster, in most cases a common disk subsystem is shared by a number of servers. The disk subsystem should be reliable and backed up by RAID implementation. More information about RAID levels can be found in the white paper *IBM Netfinity RAID Technology* available on Web at:

`http://www.pc.ibm.com/us/netfinity/tech_library.html`

5.2.1 IBM ServeRAID adapters

There are currently three ServeRAID adapters available from IBM:

- The IBM ServeRAID-4H Ultra3 SCSI Adapter
- The IBM ServeRAID-4M Ultra3 SCSI Adapter
- The IBM ServeRAID-4L Ultra3 SCSI Adapter

These ServeRAID-4 Ultra160 SCSI controllers are the fourth generation of the IBM ServeRAID SCSI controller family and will replace the ServeRAID 3 family. These 64-bit, Active PCI (hot-swap/hot-add) controllers support full RAID functions, FlashCopy, and clustering failover.

- Up to 14 HDDs supported per channel
- ServeRAID Manager RAID configuration and monitoring software for powerful, easy-to-use RAID management across your storage assets
- Support for Active PCI (hot-swap and hot-add of controllers) for Windows NT, Windows 2000 and NetWare environments
- Support for adapter failover in Windows NT environments

- Support for clustering with Microsoft Cluster Services and Netware High-Availability Services

For a feature comparison of these adapters, see Table 25:

Table 25. ServeRAID features

Features	ServeRAID-3HB	ServeRAID-4H	ServeRAID-4M	ServeRAID-4L
Internal Connectors	1	2	2	1
External Connectors (standard / maximum)	2/3	2 / 4		1 / 1
Cache	32 MB	128 MB	64 MB	16 MB
Battery-backup cache	Yes	Yes	Yes	No
RAID levels	0, 1, 1E, 5, 5E	0, 1, 1E, 5, 5E, 00, 10, 1E0, 50	0, 1, 1E, 5, 5E, 00, 10, 1E0, 50	0, 1, 1E, 5, 5E, 00, 10, 1E0, 50
Autosync	Yes	Yes	Yes	Yes
PCI bus width	64- and 32-bit	64- and 32-bit	64- and 32-bit	32-bit
SCSI Interface	Ultra 2 SCSI	Ultra 160 SCSI	Ultra 160 SCSI	Ultra 160 SCSI
I_2O enabled	Yes	Yes	Yes	Yes
Hot-Swap PCI support	Yes	Yes	Yes	Yes
Fault-Tolerant Adapter Pair	Yes	Yes	Yes	Yes
Clustering support	Yes	Yes	Yes	Yes
Automatic Hot Swap Rebuild	Yes	Yes	Yes	Yes

ServeRAID adapters supports a wide range of RAID levels. To build a fault tolerant disk subsystem you need to configure your disk arrays with the appropriate RAID level. Table 26 will help you to choose a RAID level with fault tolerance.

Table 26. RAID levels at a glance

RAID level	0	1	1E	5	5E	00	10	1E0	50
Also known as	Striping	Mirroring	Mirroring with an odd number	Striping with distributed parity	Striping with distributed hot spare and parity	Striping across multiple RAID-0 arrays	Striping across multiple RAID-1 arrays	Striping across multiple RAID-1E arrays	Striping across multiple RAID-5 arrays
Fault tolerance	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
Redundancy type	None	Duplicate	Dup.	Parity	Parity	None	Dup.	Dup.	Parity
Hot spare option	No	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
Disks required	1+	2	3+	3+	4+	2+	4+	6+	6+

Assess your storage and expansion needs, then select the ServeRAID controller that provides the best match to your requirements.

5.2.1.1 ServeRAID considerations for MSCS

It is important to read product documentation (such as the *IBM ServeRAID SCSI Adapter User's Guide*) so that you understand any configuration constraints and functional limitations before implementing a cluster solution. Doing so will allow you to implement a technically sound solution. For a detailed description of clustering solutions and installation procedures in a ServeRAID environment, see *Installing the IBM ServeRAID Cluster Solution*, available for download at:

`ftp://ftp.pc.ibm.com/pub/pccbbs/pc_servers/00n9132a.pdf`

ServeRAID adapters, and SCSI subsystems in general, support a two-node shared disk cluster environment. Limitations on bus-length and device addressing make SCSI impractical for clusters with more nodes. If you do wish to use more than two nodes in a cluster, Fibre Channel is required.

A mixture of ServeRAID adapters are supported in clustering. Clustering was first supported on the ServeRAID II adapter, so the ServeRAID I adapter is not supported. Identical adapters must be connected together (ServeRAID II to ServeRAID II). If the ServeRAID II contains an MSCS quorum disk, a SCSI quorum cable must be connected to channel 3 of the ServeRAID II adapter pair (even the latest firmware does not remove this requirement from the ServeRAID II adapter as it does for the ServeRAID IIIs and 4s).

In a ServeRAID-based cluster, the shared disk drives have to reside in external storage enclosures, and are connected to a ServeRAID adapter in each node, as shown in Figure 70. The configuration is similar to that of a fault-tolerant pair.

A fault-tolerant adapter pair is used to increase external SCSi disk subsystem availability by duplicating RAID controllers in the server. With this feature, you can configure a ServeRAID adapter pair and connect both adapters to the same storage enclosure in order to provide access to the disk drives even after one of the adapters has failed.

You can use this feature on ServeRAID-4, ServeRAID-3 and ServeRAID II adapters. The original ServeRAID adapter and the ServeRAID controllers integrated onto the planar of some Netfinity systems are not supported.

You can only use a fault-tolerant pair of ServeRAID adapters with disk drives installed in external disk drive enclosures. Disks connected to internal server backplanes are not supported.

In both cases you connect the two ServeRAID adapters to a common set of disks. There are differences, however:

- When using a fault-tolerant pair, one adapter is active and the other is passive. That is, only one adapter has access to the drives.
- When using clustering, both adapters can actively access the disks. The clustering software allocates ownership of logical drives to one node or the other and only the owning node can access a specific drive.
- A fault-tolerant pair cannot be part of an MSCS shared cluster.

The MSCS high-availability cluster configuration gives higher availability by duplicating adapters and whole server machines, but still leaves the clustered system dependent on the common disk subsystem.

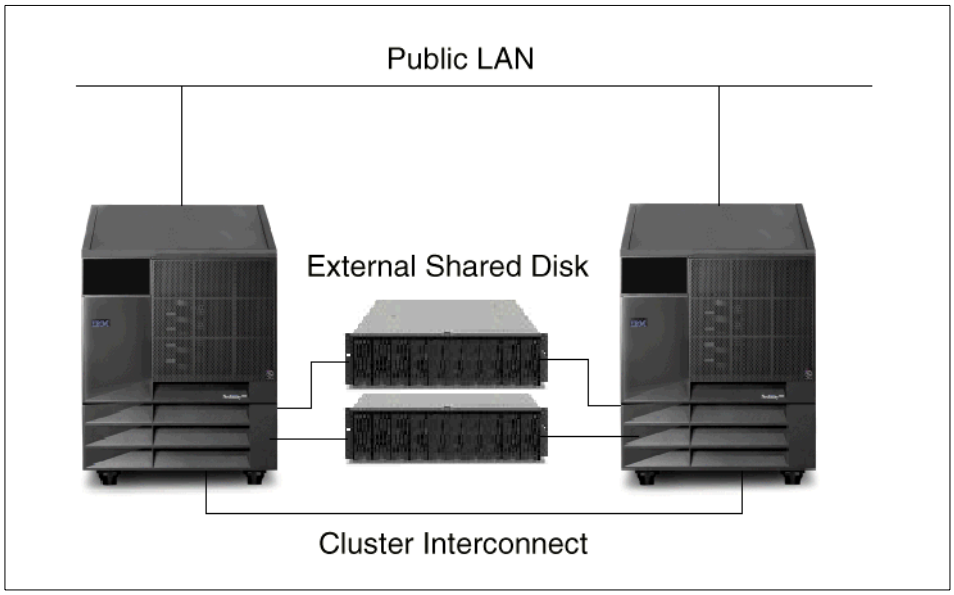

Figure 70. Two-node cluster with ServeRAID adapters

The following are important considerations for ServeRAID clustering:

- **Operating system**

  Each node has the operating system installed on its own local disk drives. These disk drives are not a part of clustered storage and may not even be attached to a ServeRAID adapter. However, if they are attached to the ServeRAID adapter that is also connected to the shared storage, you must assign non-shared merge IDs to the operating system logical drives.

- **SCSI Initiator IDs on the shared SCSI bus**

  The ServeRAID adapters will be connected to either an EXP200 or EXP300 external storage enclosure, using external SCSI cables. You should make sure that the two backplanes in the storage enclosure are configured as a single SCSI bus. It is also mandatory to connect the enclosure to the same SCSI channel on both adapters.

  Taking a closer look at a single, shared SCSI channel, the devices on the SCSI bus are:

  – The two ServeRAID adapters
  – The backplane in the external storage enclosure
  – The disk drives in the external storage enclosure

For correct operation, all devices on a SCSI bus must have unique SCSI IDs, so the ServeRAID adapters have to use different SCSI initiator IDs. Typically, we set the SCSI ID of the first adapter to 6 and leave the second adapter's ID at the default setting of 7.

- **Unattended mode**

  You must enable unattended mode on both adapters. The active node that currently owns the disk drives will detect them as *online* and the standby node will not be able to access them at all; it will mark them as *defunct*. If unattended mode is disabled (the default), the standby node will not boot into the operating system, but rather stop booting at a ServeRAID POST message and wait for user input.

- **Quorum disk drive**

  When using Microsoft Cluster Server, the quorum log should reside on a RAID-1 logical drive. Locating the quorum on a RAID-5 logical drive can cause problems in case of a disk failure, since the quorum resource will not be able to fail over until the failed disk drive is rebuilt to a hot spare or the drive is replaced and rebuilt. This could potentially disable the cluster if the active node failed.

- **Write-through cache policy**

  Do not use a write-back cache policy for the shared logical drives. Doing so will cause data loss when failover occurs. There is no way to transfer dirty data in the ServeRAID adapter cache from the failing node to the surviving one. Therefore, all shared logical drives should use a write-through cache policy.

- **Clustering parameters**

  You must specify the clustering parameters:

  – Specify the host adapter name and the partner adapter name.

  – Assign shared merge IDs for all shared logical drives.

  – Assign non-shared merge IDs for all non-shared logical drives on a ServeRAID adapter pair. For example, these could be the logical drives local to the server, which contain the operating system.

  Note: You only create arrays and logical drives on one node. Also, the shared merge IDs are only assigned on the first node. The second node will pick the configuration and merge IDs up from the first node. But you must still create non-shared merge IDs for local logical drives on the second node.

- **IBM ServeRAID Windows NT Cluster Solution Diskette**

  In the Microsoft Cluster Server environment, after the Microsoft Cluster Server is installed on both nodes, you must apply the IBM ServeRAID Windows NT Cluster Solution diskette. Do not forget to do it on both nodes. This will add the new ServeRAID disk resource into the MSCS environment. It will also create a group and a resource of this type for each shared logical drive.

  When installing Microsoft Cluster Server on Windows NT in a ServeRAID environment, it is important to use the `/localquorum` switch. The quorum resource can be moved to the shared logical drives only after installation completes, because the correct resource type (ServeRAID disk resource) is available only at the end of the installation procedure.

- **Hardware considerations**
  - Do not connect non-disk devices such as tape drives to common channels. They will be accessible to both servers and conflicts could arise.
  - The SCSI channels in an adapter pair must be cabled in pairs so that connected channel numbers match (1-1, 2-2, 3-3, 4-4).
  - You must connect all disks that will be shared to the shared channels. This includes hot-spare drives as well as data drives.
  - All shared physical drives must be part of a logical drive or defined as a hot spare to one or the other servers. You must remove all other physical drives from the shared disk enclosure to ensure proper operation.
- **Controller considerations**
  - Each ServeRAID controller must have a unique controller name assigned.
  - The quorum disk can be placed on any ServeRAID channel shared by the two servers. The quorum logical drive may also be used to store customer programs and data.
  - Do not change the controller name on both ServeRAID controllers in a pair at the same time. Doing so can cause a problem, such as a server lockup.
  - The stripe unit size of each ServeRAID adapter in a pair must be set to the same value (8 KB, 16 KB, 32 KB, or 64 KB).
  - You cannot share hot-spare drives between controllers in a cluster pairing. If you want hot-spare protection, each ServeRAID controller must have a hot-spare drive defined.

- **Array or logical drive**
  - RAID-5 logical drives will not fail over if they are in a critical state. Therefore, it is highly recommended to implement a hot spare disk drive in a clustering environment. A hot spare in a clustered environment is allocated to a specific adapter. This means that you must therefore assign a hot spare drive for each node.
  - Only one logical drive must be created for each physical array because only one server at a time can own a physical drive. Using ServeRAID nomenclature, only logical drives A0, B0, C0 and so on are supported, not A1, B1, and so on.
  - Merge ID numbers in the range of 1 to 8 must be assigned to each logical drive that will be common. Merge ID numbers must be unique for each shared logical drive in an adapter pair. Merge ID numbers in the range of 200 to 215, where the last two numbers are the SCSI initiator of a common channel, must be assigned to the local logical drives.
  - If you are starting (booting) the operating system from a common bus adapter, define the first logical drive as the startup drive and assign a local Merge ID, for example, 206 for Node A.
  - The total number of logical drives allowed per adapter pair is a maximum of eight. This includes any drives transferred to an adapter because of a failover. A failover will not complete if this number is exceeded.
  - Logical drives that are currently undergoing Logical Drive Migration (LDM) operations, or RAID 5E logical drives undergoing compression or decompression, will not be allowed to fail over. However, all other drives will be able to fail over if necessary.
  - If a failover occurs while a critical RAID-1, RAID-1E, RAID-10 or RAID-1E0 logical drive is rebuilding to a spare disk, the rebuild automatically starts a few seconds after the failover is completed.
  - The cluster support software will initiate a synchronization of RAID-1 and RAID-5 logical drives immediately after a failover. If a drive fails before this synchronization is complete, logical drive access is placed in the blocked state.
  - When a logical drive spans multiple channels and a failure within the drive subsystem occurs that is unique to a channel (for example, a disconnected cable), then the entire physical array will be marked as bad even though access from the surviving server can occur. Therefore, make sure that drives do not span multiple channels.

5.2.2 Fibre Channel hardware

Fibre Channel is a high-speed serial data transport technology used for mass storage and networking. It consists of an integrated set of standards that define new protocols for flexible information transfer using several interconnection topologies.

For information about Fibre Channel specifications, see:

`http://www.t11.org/`

This is the home page for Technical Committee T11, which is the committee within NCITS (National Committee for Information Technology Standards) responsible for device level interfaces.

Note

Fibre Channel is a term for an architecture and does not represent a specific device or a network topology. Fiber (note the different spelling) is a general term used to cover all physical media types supported by Fibre Channel, such as optical fiber or electrical copper cable.

For more general information about Fibre Channel and also about Fibre Channel disk subsystems implementation see the redbook *Netfinity Server Disk Subsystems*, SG24-2098.

IBM's FAStT Fibre Channel storage subsystem gives you the ability to design and create extremely flexible and expandable configurations. To implement a Fibre Channel solution, you interconnect the following components:

- **IBM FAStT Host Adapter**

  This adapter fits into a PCI slot within a server to allow it to be connected to a Fibre Channel network. Two adapters can be installed in a single server as a fault-tolerant pair. The total number of adapters in a server is limited only by the number of available PCI slots. It contains a powerful RISC processor, fibre protocol module with one-gigabit transceivers, and 66 MHz, 64/32-bit PCI local bus interface.

- **IBM SAN Fibre Channel Managed Hub**

  The IBM SAN (storage area network) Fibre Channel Managed Hub is a 1U high, rack-mounted, entry-level device for connecting together components in a Fibre Channel fabric. It is designed to support a homogeneous cluster of host servers and storage systems. An option is available to allow configuration as a stand-alone device.

The hub has eight FC-AL ports. Seven of them support fixed short-wave optical media. The eighth port is a gigabit interface converter (GBIC) slot that can be configured for either short-wave or long-wave optical media.

An arbitrated loop is logically formed by connecting all eight ports on the hub into a single loop, or the ports can be zoned into several independent arbitrated loops. Each port supports 100 MBps full duplex data transfer. Two Managed Hubs can be cascaded, providing a loop of up to 14 ports, and the hub can also be attached to the IBM SAN Fibre Channel Switch, providing loop attachment of storage devices.

The StorWatch FC Managed Hub Specialist, included with the Managed Hub, enables you to configure, manage, and service the hub.

- **IBM FAStT500 Mini Hub**

  Each Netfinity FAStT500 Mini Hub contains two ports and supports either short- or long-wave Fibre Channel GBICs to support optical fiber cabling to devices. The Netfinity FAStT500 RAID Controller supports up to four host and four drive Netfinity FAStT500 Mini Hubs (two hosts and two drive Netfinity FAStT500 Mini Hubs are included standard).

- **IBM SAN Fibre Channel Switch 8/16-port**

  The switch used in the Netfinity Fibre Array solutions is the IBM SAN Fibre Channel Switch as described in 5.3.1, “SAN components” on page 208.

- **Netfinity Fibre Channel RAID Controller Unit**

  The Netfinity Fibre Channel RAID Controller Unit provides Fibre Channel optical attachment to a host server, and six SCSI channels are available for connection to external disk enclosures. The RAID controller converts the incoming Fibre Channel optical data to an electrical signal, performs RAID calculations, and then directs the appropriate SCSI commands to the low-voltage differential SCSI (LVDS) channels.

  The controller unit is a rack-mounted device that connects to and controls disks installed in EXP300 and EXP200 storage enclosures.

  – Six high-speed, Ultra2 LVDS channels

  – Supports up to 15 devices per channel (Netfinity EXP storage units support up to 10)

  – RAID levels 0, 1, 3, and 5

  – 128 MB write-back cache with battery backup

  The unit contains one controller as shipped. A second controller, called the Netfinity Failsafe RAID Controller, can be added to form a redundant

pair of controllers. The controllers provide the RAID function within the disk subsystem.

- **IBM FAStT500 RAID Controller**

  This device provides the Fibre Channel-to-Fibre Channel interface between the host and storage devices. Its 4U mechanical is designed for maximum high availability and serviceability. Customer Replaceable Units (CRUs) and component redundancy are used throughout to minimize downtime and service costs. The Netfinity FAStT500 RAID Controller supports dual host channels and four drive channels to enable configuration flexibility and redundant dual-loop drive storage. Key features:

  – Dual high-performance, RAID controller cards - supports up to 100 MBps data transfer rate per controller

  – Support for RAID 0, 1, 3, 5, and 10

  – 256 MB write-back cache with battery backup

  – Optional Netfinity FAStT500 256 MB Cache available for complex configurations

  – Two 175 W auto-ranging, hot-swap, redundant power supplies

  – Two RS-232 ports and two 10BaseT and 100BaseT ports for diagnostics and remote system management of the array through the Netfinity FAStT Storage Manager

  – Two, two-port mini hubs on both drive and host sides are standard - remaining mini hubs are optional

- IBM FAStT EXP500 Storage Expansion Unit

  The FAStT EXP500 is a highly available data storage unit that supports high-performance, hot-swap Fibre Channel HDDs. This 3U storage enclosure supports four Fibre Channel short-wave GBICs that provide the optical fiber connection to FAStT500 RAID Controllers. Both single and dual loop configurations are supported.

  Up to 22 FAStT EXP500s can be attached to a single FAStT500 RAID Controller by adding an appropriate number of short-or long-wave GBICs and supported optical cabling options.

  – Two hot-swap, 350 W auto-ranging redundant power supplies

  – Redundant fans - two hot-swap, dual-fan units

  – LED indicators on all critical components warn of faults, excessive temperature, and other abnormalities

  – Ten drive bays - support slim-high or half-high FC hot-swap HDDs

- **IBM FAStT200R and FAStT200 RAID/Storage Units**

  These new compact units are ideal for creating a cost-effective SAN using Fibre Channel-to-Fibre Channel technology. Configurations support disk and tape pooling connected to xSeries and Netfinity servers in redundant loops to create a highly available SAN.

  These new FAStT200 RAID/storage units combine RAID controller technology and hot-swap HDD storage into a highly integrated 3U mechanical package. They contain:

  – Ten 3.5-inch, hot-swap HDD bays, supporting the new fourth-generation 10,000 rpm FC hot-swap HDD options.

  – One or two 350 W, auto-ranging, hot-swap power supplies (model dependent).

  – One or two RAID controllers (model dependent) performance optimized for up to 30 drives.

  – Two redundant fan canisters — each contain two blowers.

  – Midplane — provides dual FC loops.

  – LED indicators on all critical components warn of faults, over temperature, and other abnormalities.

  FAStT200 RAID/Storage Units are connected to a supported xSeries and Netfinity server through a FAStT Host Adapter installed in the server and a GBIC option connected to the RAID controller of the RAID/storage unit. Alternatively, they can be connected to the host server through:

  – IBM 3534 Managed Fibre Channel Hub
  – IBM 8-port or 16-port FC Switches

  The RAID/storage units are designed for easy service. Hot-plug components, Light Path Diagnostics, and Customer Replaceable Units (CRUs) are used throughout to minimize downtime and service costs.

  Two RAID/storage units are available:

  – FAStT200R RAID/Storage Unit - high-availability FC RAID storage solution. This RAID/storage unit is designed for high-availability applications requiring a high degree of component redundancy. It features two hot-plug RAID controllers and two hot-plug power supplies.

  – FAStT200 RAID/storage unit - entry level FC RAID storage solution. This RAID/storage unit is an entry level version that is identical to the FAStT200R RAID/storage unit except it contains a single RAID controller and one power supply. This unit can be upgraded for

high-availability applications by installing an optional FAStT200 Failsafe RAID Controller and EXP200 350 W redundant power supply

- **IBM EXP300 Storage Enclosure**

  Fully redundant 3U rack-mount, base unit — includes hot-plug redundant power supplies, fans, and drive bays supporting:

  – 14 hot-swap slim-high bays

  – Both Ultra160 and Ultra2 technology HDDs with ServeRAID-4 controllers

  – Accommodates single or dual SCSI bus configurations

  – Dual hot-swap, 500 W redundant power supplies with integrated fan assemblies

  – PFA for fans and HDDs

  – Tower capability through optional Rack-to-Tower Conversion Kit

- **Netfinity EXP200 Storage Enclosure**

  External, rack-mountable data storage units with 10 hot-swap drive bays that support both half-high and slimline 7200 RPM or 10,000 RPM Ultra2 SCSI disk drives. These units are fitted with dual hot-swap power supplies and dual hot-swap fans for high availability.

- **GigaBit Interface Converters (GBICs)**

  A device used to connect fiber optic cables to the electrical interface in the Fibre Channel hub.

- **Cabling**

  Optical fiber cables are used to connect all the components together. There are two basic connections supported by the cable: short-wave (multimode) and long-wave (single mode). Short-wave connections are supported by all devices and can be up to 500 m apart. Long-wave connections are supported only by the Fibre Channel hub (that is, hub-to-hub connections) and can be up to 10 km apart.

In addition to these fundamental building blocks for a Fibre Channel-based storage subsystem, IBM has recently announced a number of products aimed at customers who wish to implement storage area networks (SANs). Information about the IBM SAN Fibre Channel Switch, the IBM SAN Data Gateway Router, Vicom Fibre Channel SLIC Router, McDATA Enterprise Fibre Channel Director, and IBM Fibre Channel RAID Storage Server can be found in 5.3.1, “SAN components” on page 208.

5.2.3 Disk configurations

5.2.3.1 A basic MSCS or NCS cluster configuration

Figure 71 illustrates a basic clustered configuration using Fibre Channel disk subsystems and offering data protection and RAID redundancy. It uses two Netfinity 5500 M10 servers for failover redundancy, connected to the FAStT500 RAID Controller with its redundant control units.

If one of the servers, cables or controllers fails, the system will remain operational. In addition, another Fibre Channel host adapter may be added to each of the servers for additional redundancy. This configuration is typical for implementations of Microsoft Cluster Server (MSCS) and Novell Cluster Services (NCS).

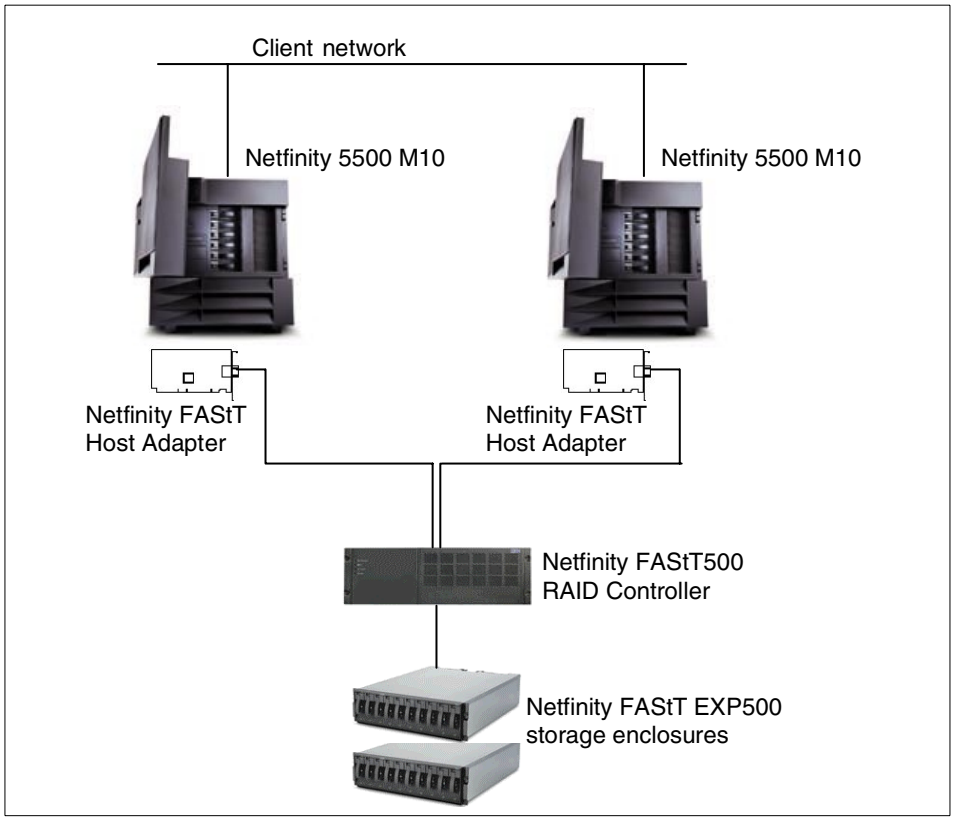

Figure 71. IBM FAStT disk subsystem for two-node clusters

5.2.3.2 Netfinity high-availability solutions using MSCS

The next scenario illustrates a clustering environment that allows server-based applications to be made highly available by linking two servers or nodes running Microsoft Windows NT 4, Enterprise Edition or Microsoft Windows 2000 Advanced Server using their clustering technology called Microsoft Cluster Server (Services in Windows 2000) (MSCS). This configuration provides high availability, and reduced planned or unplanned downtime.

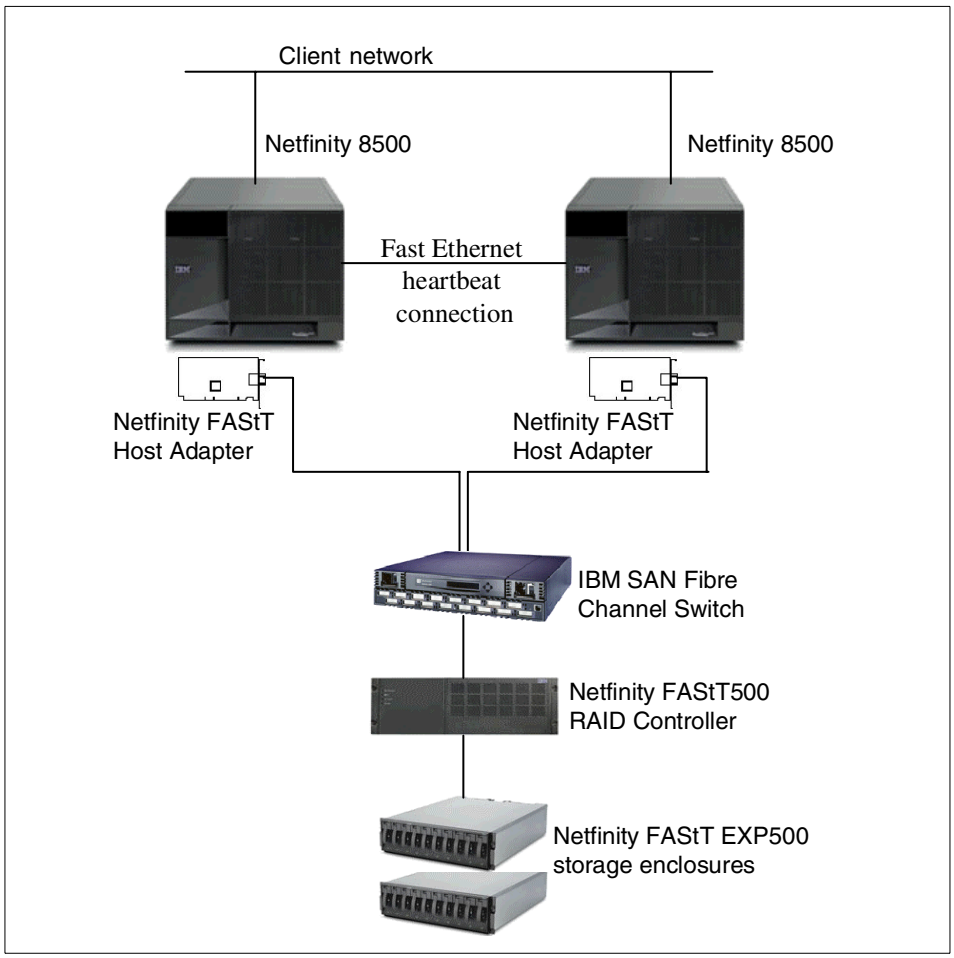

Figure 72. Two-node MSCS cluster using Fibre Channel disk subsystems

The specifics of this configuration are shown in Table 27:

Table 27. Hardware/software requirements using Fibre Channel subsystems

Configuration	Requirements / Recommendations
Software	Microsoft Windows NT 4.0 Enterprise Edition with Service Pack 5 or greater, or Microsoft Windows 2000 Advanced Server
Server	Two identical Netfinity servers certified for clustering. Each server has two identical local hard disks for RAID-1 mirroring of the network operating system.
Interconnects	• Four 10/100 Ethernet network adapters to provide connections to the public LAN and for the MSCS heartbeat (other approved interconnects could be used for the heartbeat connection) • Two IBM ServeRAID-4L Adapters for internal disks • Two IBM FAStT Host Adapters • One IBM SAN Fibre Channel Switch 2109-S16 • Fibre Channel cables
Storage	• One IBM FAStT500 RAID Controller • Two IBM Fibre Channel EXP500 Storage Expansion Units • 10 36.4 GB Fibre Channel Hard Disk Drives per storage expansion unit

Heartbeat connection

The heartbeat connection used for Microsoft Cluster Service no longer has to be certified as part of a certified cluster configuration The requirement is now that it be on the Windows NT/2000 HCL and NDIS capable.

This means that the integrated Ethernet adapter is now supported as a heartbeat connection, as well as adapters such as the Netfinity 10/100 Ethernet Network Adapter or Netfinity 100/10 EtherJet PCI Adapter.

Also note that the MSCS heartbeat connection should be point to point — the use of hubs is not recommended.

Although the configuration shown in Figure 72 increases the availability of server resources to your clients, there are several single points of hardware failure that can still cause loss of service that could be avoided. These include:

- The IBM SAN Fibre Channel Switch
- The IBM FAStT500 RAID Controller
- The IBM FAStT EXP500 Storage Expansion Units

- Cable connections between the switch and the controller, and between the controller and the storage expansion units

Any failure in these devices or connections will cause both of the servers to lose access to the storage system. This can be avoided by implementing a fully redundant Fibre Channel disk subsystem. A fully redundant Fibre Channel storage configuration offers the highest levels of data protection and availability and access to data.

Note

Fully redundant configuration is available only for MSCS, because RDAC (Redundant Disk Array Controller) software is available only on Windows. RDAC is a software component that comprises a multipath driver and a hot-add utility. This software is installed on the host system and provides redundancy in the case of component failure.

The FAStT500 RAID Controller contains a pair of redundant disk controllers as standard and can be used to connect an alternate path to the storage units and to the switch. A second switch is then added to protect against a switch failure. Finally, we have also added a second host adapter to each server to avoid failover in the event of an adapter failure. This configuration is shown in the following diagram:

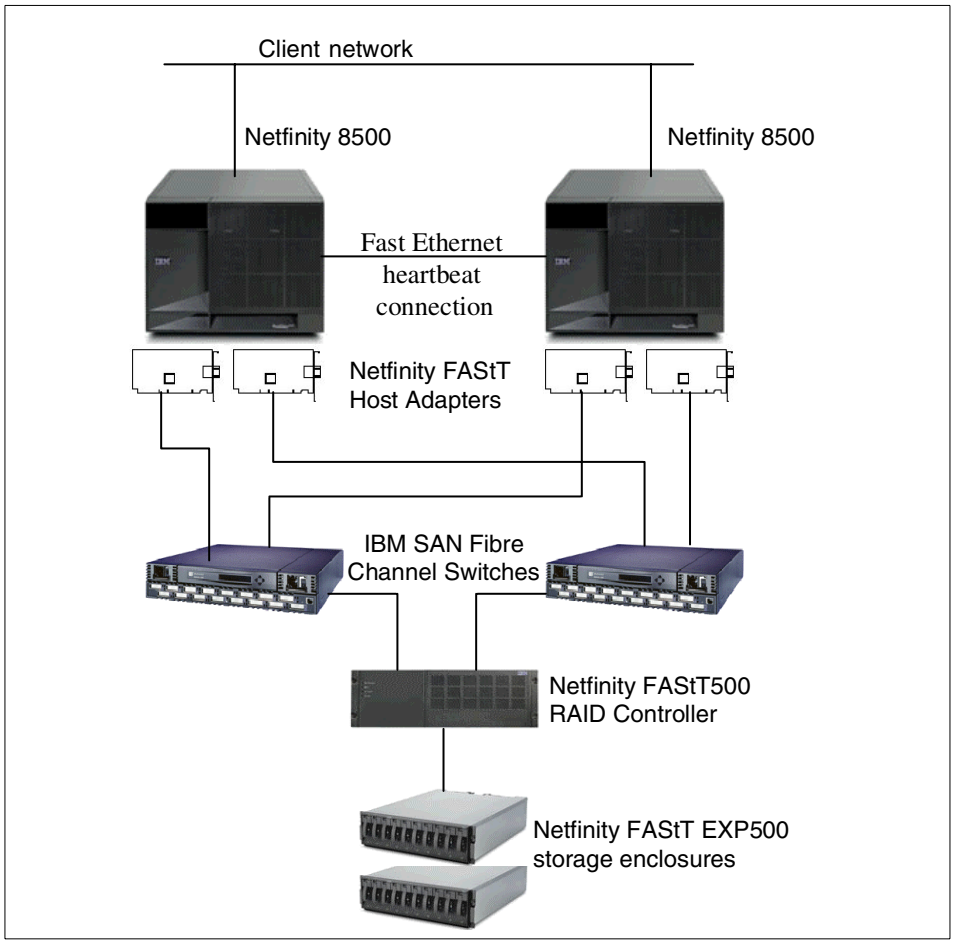

Figure 73. Two-node MSCS cluster with a fully redundant Fibre Channel disk subsystem

Clustering in Microsoft Windows 2000

Clustering is substantially enhanced in Windows 2000 Advanced Server and Windows 2000 Datacenter Server. Apart from improved availability and manageability, these new versions have increased scalability by supporting SMP servers that support a maximum of eight processors for Advanced Server and 32 processors in Datacenter Server. Increased memory capacity is also supported: up to 8 GB of RAM in Advanced Server and 64 GB in Datacenter Server.

The Windows 2000 Advanced Server supports a two-node cluster, as in Windows NT 4 Enterprise Edition. Windows 2000 Datacenter Server supports a four-node cluster.

5.2.3.3 Oracle Parallel Server

xSeries Fibre Channel is certified with Oracle Parallel Server (OPS) using the xSeries Cluster Enabler for OPS software (see 4.2, “Oracle Parallel Server and high availability” on page 122). This solution was certified for Oracle Parallel Server in two-, four-, six- and eight-node cluster configurations on Netfinity 7000M10. New certified configurations include up to two Netfinity 8500R servers connected to Fibre Channel storage.

5.2.4 Serial storage architecture

Serial storage architecture (SSA) technology provides a high-performance serial interface designed to connect I/O devices to their host adapters. It is a two-way signal connection (transmit and receive), providing full-duplex communication between host and devices on a nonarbitrated loop.

SSA provides an alternative to Fibre Channel for high-capacity data storage with the capability of separating disks from the servers by large distances. There is a perception that SSA is an IBM proprietary interface. While it is true that the first SSA interface was developed by IBM, in 1991 this technology was made available as one of the serial storage interface options for the SCSI-3 standard.

The current mainstream architecture for transmitting data to high-bandwidth, high-capacity devices is SCSI. The most recent iteration of the SCSI specification extends its bandwidth up to 160 MBps over a 68-wire parallel connection. It is recognized, however, that further increases in bandwidth are becoming increasingly limited. SSA is one approach to circumvent the practical limitations of the SCSI bus.

Devices in an SSA network are called nodes. From a communication point of view, a node can be either an initiator or a target. An initiator issues commands, and targets respond with data and status. Typically, the server-based SSA adapter is an initiator node and the SSA disk drives that it commands are target nodes. Each SSA node is given a unique address at the time of its manufacture, which allows initiators in a loop to determine what other SSA nodes are attached to the loop.

SSA allows more than one initiator to be present in a loop. In this case, commands and data from multiple initiators can be directed to the same or different targets and intermixed freely. An architectural limit of 128 nodes is imposed by SSA, a limit which is often further restricted by the implementation of the host adapter.

SSA has the flexibility to be implemented using several different topologies. Those in common use include string, loop, and switch topologies. The characteristics of each of these are as follows:

- String topology

  A string is a simple linear network of two or more nodes.The node at each end of the string can be a single-port node, while the others must be dual ported. A string configuration is also created when a loop is broken by disconnecting or breaking one of the cables.

  When a break occurs in a string, communication with devices beyond the break is lost, making this solution highly unsuitable for a cluster installation.

- Loop topology

  The most popular form of SSA installation uses a loop topology. A loop contains only dual-ported nodes with a maximum of 127 nodes in one loop. If a break occurs in the loop (for example, if a cable is disconnected), each device on the loop adjusts its routing methods, under direction from the master initiator, so that frames are automatically rerouted to avoid this break. This allows devices to be removed from or added to the loop while the subsystem continues to operate without interruption.

  A loop topology allows you to achieve the maximum bandwidth for your adapter and hence this is the preferred topology for clustered implementations. Today's IBM SSA adapters for Intel-based servers provide simultaneous full-duplex 40 MBps connections over two ports. That means 40 MBps reads and 40 MBps writes at the same time on each port adding up to 80 MBps data transfer per port transferred over a single cable.

In a loop topology, two ports are used, which means communication is done over two ports using two separate cables. This results in a total maximum transfer rate of 160 MBps per independent loop, controlled by only one serial interface chip (SIC) on the adapter. IBM's SSA adapters for Intel-based servers have four SICs for two independent loops of 160 MBps each.

- Switch Topology

  The switch topology is the most complex topology and theoretically supports up to two million nodes. It is the most fault-tolerant configuration and allows a virtually unlimited network configuration.

> **Note**
>
> Switch configurations are not currently implemented by IBM in the Netfinity server environment but are available from other vendors.

Two techniques of the SSA architecture provide a significant performance enhancement in SSA networks:

- Cut-through routing

  Cut-through routing (or worm-hole routing) means that a node may forward a frame character by character as it is received; it does not have to wait for confirmation that the frame passed its cyclic redundancy check (CRC).

- Spatial reuse

  One of the characteristics that distinguishes a loop (where each link between nodes is a separate connection) and a bus (where each node connects to the same piece of wire) is that a loop allows the possibility of spatial reuse. This is the technique whereby links that are not involved in a particular data transaction are available for use in another transaction.

  This means that a loop with two initiators (as in a cluster) can support transactions between each initiator and its associated device simultaneously. A SCSI bus, as a counterexample, can handle transactions from only one initiator at a time.

5.2.4.1 IBM Advanced SerialRAID/X Adapter

The IBM Advanced SerialRAID/X Adapter and IBM serial disk systems offer flexibility, scalability, and cost-effective data protection. The Advanced SerialRAID/X Adapter uses IBM serial technology to provide flexible storage solutions for both clustered servers and single-server configurations in the Windows and NetWare environments.

The Advanced SerialRAID/X Adapter enables users to connect up to 96 serial disk drives in two loops over a server's PCI bus. These serial loops use full-duplex, point-to-point communication to achieve a bandwidth of up to 160 MBps. This bi-directional serial loop architecture can improve availability, since single cable or disk failures do not prevent access to data.

The Advanced SerialRAID/X Adapter enables clustered Intel-based servers to share IBM serial storage disk systems. In configurations with two adapters in a loop, both adapters can read or write to all disk arrays. The distributed lock mechanism on the adapters enables real data sharing while enhancing data integrity. Failover protection is supported for Microsoft Windows NT 4.0 with Microsoft Cluster Server or Novell Cluster Services for Novell NetWare 5.x. If one server fails, a secondary server takes on the applications of the other. At the time of publication, there was no support for Windows 2000.

For even higher availability, the two-way RAID 0+1 function enables local or remote data mirroring — up to 10 km by using optical extenders. This enables true remote site failover for large RAID 0+1 arrays of up to 291 GB (with 36.4 GB disks).

The Advanced SerialRAID/X Adapter supports non-RAID, RAID 5, 1, 0, and 0+1 and has the flexibility to create large volume sizes. Up to 96 disk drives can be configured with up to 546 GB of capacity for a RAID 5 array. This enables a total RAID 5 storage capacity of more than 3.2 TB per adapter.

Full-stride writes have been implemented to increase RAID 5 performance by writing the entire width of the array. Up to 16,000 I/Os-per-second performance is available with 70:30 read/write workloads in clustered non-RAID environments. Up to 9,000 I/Os per second and 160 MBps have been measured in clustered RAID-5 configurations with array reads.

In addition, 32 MB Fast Write Cache is standard on the adapter for dual- or single-host configurations to provide significant write response time performance benefits. The data on the cache card is protected by battery backup against power loss.

The Advanced SerialRAID/X Adapter includes the SSA Remote System Manager (RSM), which enables the use of local or remote disk system management utilities, such as array creation, and hot spare or system attachment. The RSM can be integrated into IBM Netfinity Manager 5.2 for remote control and monitoring on Windows NT servers.

The Advanced SerialRAID/X Adapter is supported on selected Netfinity server models. Operating system support includes Microsoft Windows NT 4.0

Server with Microsoft Cluster Server, Novell NetWare 4.2 with Novell High Availability Services (HAS), and Novell NetWare 5.0 with Novell Cluster Services. The Advanced SerialRAID/X Adapter and the IBM 7133 Advanced Models are included on the Microsoft Hardware Compatibility List and are Novell NetWare and Microsoft Cluster Certified with selected servers. For supported server models and OS please visit:

`http://www.hursley.ibm.com/~ssa/pcserver/`

Ultra SCSI is the standard offering in the industry today and IBM provides reliable drives, enclosures, and RAID controllers for stand-alone or two-node clustered servers at a very competitive price point.

For clustered servers, or higher performance requirements, Ultra160 (LVD) SCSI may be more appropriate as it provides greater bandwidth (up to 160 MBps) and longer cabling lengths (up to 12 m) which can be beneficial in physical planning for cluster systems that use shared external disk subsystems.

Netfinity's Fibre Channel RAID subsystem is the preferred technology for clustered systems and high availability storage configurations. It provides multisystem attachment through a FC-AL (Fibre Channel Arbitrated Loop) network configuration.

A detailed comparison of these three disk technologies can be found in the redbook *Netfinity Server Disk Subsystems*, SG24-2098.

5.3 Storage area networks

As the number of nodes used in Intel server clusters moves beyond two, the technical demands placed on a common storage subsystem become more complex and challenging. In addition, large installations with several clustered systems and a heterogeneous server environment generally require separate storage solutions for each platform, which soon turns out to be an expensive investment.

A solution that consolidates storage attached to multiple hosts into one manageable infrastructure addresses both of these issues, promising performance, flexibility, and cost effectiveness. Storage area network (SAN) is the name that has been coined for such solutions. By the year 2002, IBM estimates that 70% of all medium- and large-sized customers will implement SANs to manage and share the volumes of data created as they transform themselves into e-businesses.

A concise definition of a storage area network might be: A dedicated, high-speed network of directly connected storage elements, designed to move large amounts of data between systems and host-independent storage devices.

So, what is new and different about a SAN in comparison with other storage solutions? It is the idea of providing access to data by way of a high-speed network using, for example, SSA or Fibre Channel technologies, which are designed for reliable, accurate, and high-performance server-to-server, server-to-storage, and storage-to-storage communication. Figure 74 shows a conceptual diagram of a SAN:

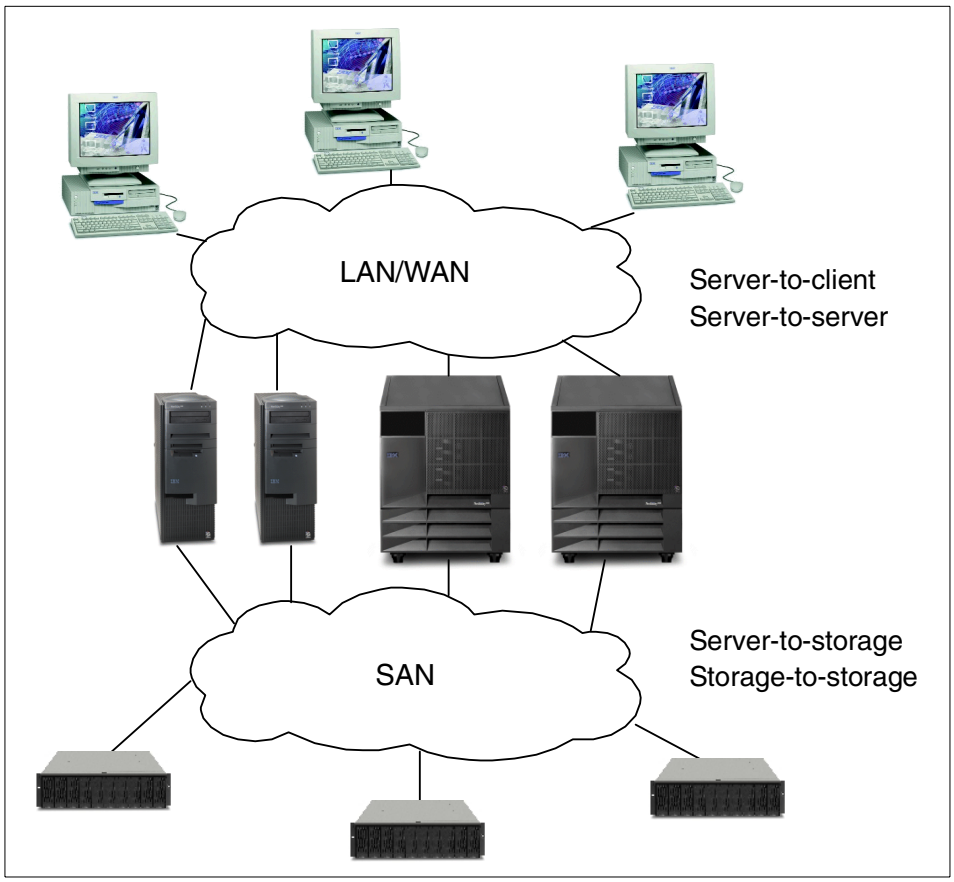

Figure 74. SAN principles

A SAN is not a specific technology but rather a business solution attempting to resolve the sometimes conflicting issues of cost, connectivity, distance, and performance by providing a better IT infrastructure.

5.3.1 SAN components

IBM has recently announced a number of elements that allow sophisticated SANs to be constructed:

- **IBM SAN Fibre Channel Switches**

  The IBM SAN Fibre Channel Switch provides the connectivity necessary to make enterprise information accessible and available across multiple host systems, from various storage systems or devices forming a manageable storage area network (SAN). Utilizing the ANSI standard Fibre Channel interconnect technology, the switch provides 100 MBps transfer bandwidth on each Fibre Channel connection port. The IBM Fibre Channel Switch is offered in eight-port (Model S08) and 16-port (Model S16) versions as shown in Figure 75. Both models come with four short-wave GBICs as standard.

Figure 75. The IBM Fibre Channel Switches 2109, Models S08 (top) and S16 (bottom)

  The IBM SAN Fibre Channel Switch provides:

  – High performance with low latency. The switch's non-blocking architecture provides multiple simultaneous connections, each capable of up to 100 MBps, with maximum latency of two microseconds.

  – Ports designed to support F, FL, and E-port modes of operation.

  – Internal power-on self test and a Web browser interface usable from any Java-enabled browser on the Inter/intranet provide configuration monitoring and diagnostics.

– Cascading switches enable flexibility in configurations to accommodate many different needs, and assure high availability, and optimum performance.

– StorWatch Specialist management software.

– Automatic fabric discovery.

– The switch can be configured with either short-wave or long-wave Gigabit Interface Converters (GBICs), for distances up to 10 kilometers between connections.

- **IBM SAN Data Gateway Router**

  The IBM SAN Data Gateway Router is used to connect Ultra SCSI and SCSI tape libraries to a Fibre Channel network. The device is a protocol converter that provides a single Fibre Channel port and two Ultra SCSI ports which may be either single-ended or differential. It supports Fibre Channel attachment of IBM xSeries and other Intel-based servers running Microsoft's Windows NT to IBM Magstar 3590, Magstar MP 3570, Magstar MP 3575, and DLT Tape Libraries.

  Specific details on server models, adapters, operating system releases, storage product attachments, and configuration options are available at this Web site:

  `http://www.ibm.com/storage/SANGateway`

  Whitepapers discussing SAN-related topics including IBM’s strategy for SANs can be found at:

  `http://www.storage.ibm.com/ibmsan/whitepaper.htm`

- **Vicom Fibre Channel SLIC Router**

  The Vicom SLIC Router FC-SL is one of Vicom's connectivity products that enables Fibre Channel-based servers to connect to IBM's SSA-based disk subsystems.

  The Router attaches with a single short-wave or long-wave GBIC and a Fibre Channel cable. Optionally, one additional GBIC may be ordered to extend the Fibre Channel loop to other equipment.

  For more information, refer to the Vicom Web page at:
  `http://www.vicom.com`

- **McDATA Enterprise Fibre Channel Director**

  McDATA Enterprise Fibre Channel Director is a high-speed switched fabric of centrally managed, heterogeneous servers and storage systems.

  – 32-port, any-to-any dynamic switching

– Fabric expansion ports (E_Ports)—add additional ED-5000s and/or E_Ports seamlessly
– Provides connections from 2 m to 20 km (100 km with repeaters)
– 100 MB/second full-duplex performance per port
– Hardware redundancy of major critical components
– Internal data-path protection
– Industry-standard Class 2 and 3 Fibre Channel technology

For more information, see `http://www.mcdata.com/efcdirector`

- **IBM Fibre Channel RAID Storage Server**

  The Fibre Channel RAID Storage Server is a high-performance high-availability storage solution for servers running Microsoft Windows NT, Novell NetWare, AIX, HP-UX, or Sun Solaris operating systems. This entry-level to mid-range storage solution can provide data storage for business-critical applications that require availability and performance.

  – The industry-standard Fibre Channel host attachment with configuration flexibility for distances of up to 10 kilometers.

  – Dual active controllers with a wide range of RAID protection (0, 1, 3, 5, and 0+1) and built-in redundancy.

  – Redundant RAID controllers, power units, and cooling fans that are hot pluggable, for continuous operation during maintenance.

  More information about the Fibre Channel RAID Storage Server can be found on the Web at: `http://www.ibm.com/storage/fcss`

5.4 Cluster interconnects and access to LAN

A cluster interconnect serves various purposes for different cluster implementations but a common function is the heartbeat signal that keeps track of which servers are up and which have failed.

Other types of information that are sent over the cluster interconnect include:

- Resource state information

  The cluster software informs cluster nodes of the status of cluster groups and resources and on which nodes the resources currently reside.

- Cluster commands

  Cluster software on one node can issue commands to the other nodes. For example, when moving a resource, the cluster software can tell its current owner to take it offline and then tell the new owner to bring it online.

- Application commands

  A cluster-aware application might use the interconnect to communicate among copies of the application running on multiple nodes. This is also called function shipping.

- Application data

  A cluster-aware application might use the interconnect to transfer data between servers, also referred to as I/O shipping.

Typically, there is only one cluster interconnect, shared by all nodes, but for some implementations redundant cluster interconnects might be implemented. There is also the possibility of utilizing the public LAN for heartbeat signals, but due to possible congestion and uncertain response times this is not generally recommended. It is common, however, to configure the public network as an alternate cluster interconnect to minimize the impact of a primary interconnect failure.

If the cluster interconnect fails, a “split brain” situation (referred to as cluster partitioning) can occur in which all nodes are active but think the others are down since it does not receive a correct heartbeat signal from them. This is a situation that MSCS tries to avoid by use of the quorum disk. If cluster communication is lost over all interconnects, then all nodes with attachment to the quorum disk (both nodes in a two-node cluster) compete to get control of the quorum disk. This is done by a challenge-response protocol exploiting the SCSI device reservation feature.

The winning node will finally hold a SCSI reservation for the quorum disk and then make decisions about the cluster and resource status. All nodes without an operational cluster connection to the winning node (in a partitioned two-node cluster, the other node) suspend cluster services and stop access to common resources. Because SSA and FC are SCSI-3 protocols, this feature is also implemented by SSA and FC devices.

In a Netfinity Availability Extensions for MSCS configuration, the chance of split brain occurring is slight so a quorum device is not generally used.

A cluster configuration can use virtually any network technology as its interconnect, with 100Base-T Ethernet being perhaps the most popular at present. Cluster interconnect performance can potentially affect cluster performance under certain conditions:

- The cluster consists of a large number of nodes.
- The cluster is running thousands of cluster groups and/or resources.
- The cluster uses the interconnect to load balance applications.

- The cluster is running a scalable, cluster-aware application that uses the interconnect to transfer high volumes of traffic.

In the next few sections we examine a number of possible cluster interconnects for Netfinity clusters. 100Base-TX Ethernet is the most common cluster interconnect. There are other faster interconnect solutions on the market, such as 100/16/4 token-ring, ATM, FDDI, Gigabit Ethernet and Giganet's VI (Virtual Interconnect). Faster interconnect is preferred in clustering solutions with full disk mirroring over interconnect, such as StandBy Servers.

Fast access and a redundant path to the LAN is very important when we speak about clustered systems. The topology for LAN and networking equipment should be planned with attention to LAN segments and users groups on the LAN.

The following teaming options can be used to increase throughput and fault tolerance when running on Windows 2000/NT 4.0 or NetWare 4.1x, 5.x or later.

- Adapter Fault Tolerance (AFT)

  Provides automatic redundancy for your adapter. If the primary adapter fails, the secondary adapter takes over. Adapter Fault Tolerance supports from two to four adapters per team. Adapter Fault Tolerance (AFT) provides the safety of an additional backup link between the server and buffered repeater or switch. If you have a buffered repeater, switch port, cable, or adapter failure, you can maintain uninterrupted network performance through an adapter team. AFT is implemented with a primary adapter and one or more backup, or secondary, adapters. During normal operation, the secondary adapters have their transmit function disabled. If the link to the primary adapter fails, the link to the secondary takes over automatically.

- Adapter Load Balancing (ALB)

  Allows balancing the transmission data flow among two to four adapters. ALB also includes the AFT option. ALB can be used with any 100BASE-TX switch. Adaptive Load Balancing (ALB) is a simple and effective way to balance the transmission load of your server among two to four Netfinity 10/100 Ethernet adapters. Using ALB you can group your Netfinity 10/100 Ethernet adapters into teams. The ALB software continuously analyzes transmit loading on each adapter and balances the rate across the adapters as needed. Adapter teams configured for ALB also provide the benefits of AFT. Received data is not load balanced.

> **Note**
>
> For maximum benefit, ALB should not be used under NetBEUI and some IPX environments.

To use ALB, your adapters must be configured to your server as a team and be linked to the same network.

- Cisco Fast EtherChannel (FEC)

  Fast EtherChannel (FEC) is a performance technology developed by Cisco Systems, Inc. to increase throughput between switches. It creates a team of two or four adapters to increase transmission and reception throughput. FEC also includes the AFT option. FEC can only be used with a switch that has FEC capability.

  Unlike ALB, FEC can be configured to increase both transmission and reception channels between your server and switch. FEC works only with FEC-enabled Cisco switches such as the Catalyst 5000 series. With FEC, as you add adapters to your server, you can group them in teams to provide up to 800 Mbps at full duplex, with a maximum of four Netfinity 10/100 Ethernet Adapters. The FEC software continuously analyzes loading on each adapter and balances network traffic across adapters as needed. Adapter teams configured for FEC also provide the benefits of AFT.

5.4.1 Ethernet

Ethernet is the most common cluster interconnect in use today, partly because Ethernet is a well-established standard, making the components (NICs, hubs, and so on) comparatively cheap for good performance. As the number of nodes in a cluster increases, the suitability of Ethernet as a communication medium decreases. Thanks to its collision-detection protocol, multiple nodes using a significant amount of the available bandwidth will run into inefficiencies that cause increased latency in data transfer. For this reason, there is a move toward switched fabrics for the interconnect.

5.4.1.1 Netfinity 10/100 Ethernet PCI Adapter 2

The Netfinity 10/100 Ethernet PCI Adapter 2 provides IEEE 802.3-compliant 100BASE-TX and 10BASE-T Ethernet connectivity for servers over an unshielded twisted pair link through a single RJ-45 connector.

- 32-bit PCI 2.1 bus mastering architecture
- Half-duplex and full-duplex operation at both 10 Mbps and 100 Mbps
- auto-negotiation of speed and duplex mode

- PCI HotPlug
- Adapter fault tolerance
- Adaptive load balancing (ALB)

The new Netfinity adapter offers system manageability features that were previously only available on desktop adapters, including Wake on LAN, a PXE 2.0-compliant remote boot EPROM, a Tivoli Management Agent, and SNMP/DMI instrumentation.

5.4.1.2 10/100 EtherLink Server Adapter by 3Com

The 10/100 EtherLink Server Adapter by 3Com provides optimized device drivers and advanced networking features for use on server products. This adapter delivers a high-performance, full-duplex connection to an Ethernet (10 Mbps) or Fast Ethernet (100 Mbps) network from an xSeries/Netfinity server or other Intel-based system. The adapter also offers the following features and benefits:

- Resilient Server Links (Dissimilar Failover Capability) - Failover capability offers continuous network connectivity for the server. Includes support for two different third-party adapters.
- Self-Healing Drivers - Continuous monitoring of network connection and take recovery action as necessary without rebooting the system.
- Bidirectional Load Balancing - Improves network performance by balancing or distributing network traffic among each of the server adapters aggregated into a group. Provides transmit load balancing for third-party adapters.
- HotPlug support to minimize maintenance impact by allowing replacement, removal, and addition of NICs to servers without powering off or restarting the server.
- Performance Monitoring - Enables comprehensive low-cost network management in switched and high-speed networks using RMON-1/RMON-2 for application response time and system-level monitoring.
- Enhanced Remote Wake Up with Keepalive - Enables the remote turning on of PCs to allow lower-cost after-hours updates and inventories and prevents sleeping PCs from becoming inaccessible to remote management applications.
- Multiple VLANs - Supports up to 64 VLANs allowing server/client traffic to avoid unnecessary router bottlenecks (using IEEE 802.1Q).
- Flow Control 802.3x - Delivers higher performance due to more efficient data transfers.

5.4.2 Gigabit Ethernet

IBM has two versions of Netfinity Gigabit Ethernet adapters. Netfinity Gigabit Ethernet SX Adapter is designed for fiber optic wiring and Netfinity Gigabit Ethernet Adapter is for Category 5 twisted pair cabling.

5.4.2.1 The Netfinity Gigabit Ethernet SX Adapter

The Netfinity Gigabit Ethernet SX Adapter provides a 1000Base-SX connection for servers over a multimode fiber optic link when attached to its duplex SC connector. Designed to operate in servers with 32-bit or 64-bit PCI bus slots, it provides the maximum data throughput available in an Ethernet adapter. Operating at 1000 Mbps, with efficient utilization of the host CPU, it is compliant with the IEEE 802.3z Gigabit Ethernet standard for easy integration with existing Ethernet and Fast Ethernet networks. The adapter also supports the IEEE packet prioritization and VLAN tagging capabilities.

The adapter offers fault tolerance when paired with a redundant adapter in the same machine for excellent connectivity. When used in a hot plug capable server, it supports online serviceability, reducing downtime and maintenance costs. By using a redundant adapter configured in backup mode, the adapter's fault tolerance protects the server's network link from failure of the cable, connector, or adapter. If the primary link fails, data traffic is automatically transferred to the redundant link, which is transparent to both users and applications.

5.4.2.2 The Netfinity Gigabit Ethernet Adapter

The IBM Netfinity Gigabit Ethernet Adapter offers a high-performance network connection for IBM server customers who need Gigabit Ethernet throughput over their existing Category 5 twisted pair cabling.

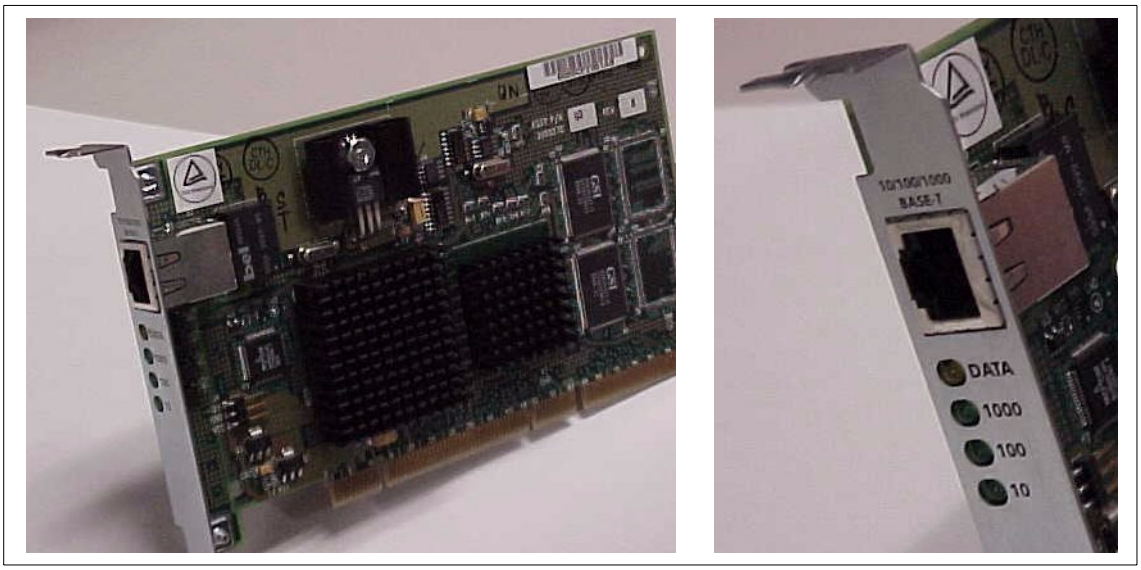

Figure 76. IBM Netfinity Gigabit Ethernet Adapter

The Netfinity adapter provides a simple migration path to Gigabit Ethernet with its support for 10 Mbps, 100 Mbps, and 1000 Mbps over Category 5 twisted pair cabling. No cabling upgrade is required in most cases to move from Fast Ethernet to Gigabit. The adapter's auto-negotiation capability enables it to detect the speed and duplex mode of the hub or switch at the other end of the link and configure the adapter automatically. The adapter can be used at 10 Mbps or 100 Mbps at first and will be ready for a seamless transition to Gigabit Ethernet when the server needs the additional bandwidth.

The Netfinity Gigabit Ethernet Adapter provides:

- Gigabit performance with Alteon-based chip technology and device drivers
- Up to 100 meters for 1000BASE-T
- Support for PCI 2.2-compliant, 64-bit, 66 MHz buses in the server
- Optional Jumbo Frames support
- On-board processing of checksums and PCI bus interrupt coalescence
- Adapter fault tolerance by using redundant pairs
- Active PCI hot plug support when installed in an Active PCI-capable Netfinity system
- Adapter Load Balancing for outgoing server traffic across up to four identical adapters

Server bandwidth can scale up to 8000 Mbps in a full-duplex switched environment. These adapters also provide link protection through adapter fault tolerance.

The adapter complies with IEEE 802.3ab and 802.3u LAN standards. It has received the Microsoft Server Design Guide 2.0 (SDG 2.0) logo and has drivers for:

- Microsoft Windows 2000/NT 4.0
- Novell NetWare, Versions 4.x and 5.x
- SCO UnixWare 7.1
- Linux

5.4.3 Giganet solution

For many cluster configurations, existing industry-standard technologies such as Ethernet and ATM switching provide ample bandwidth for internode communications and messaging in small and medium-sized clusters. However, as the clusters scale to 32 or more nodes per cluster, optimized interconnect technologies will become increasingly more important, providing internode communication with extremely high bandwidth and very low latency.

Giganet develops and markets solutions that enable the creation of high-performance, Intel-based, server farm networks. The Giganet cLAN product suite includes host adapters and switches that offer plug-and-play capabilities to scale computing resources incrementally. cLAN server-to-server communication allows applications to bypass the operating system, resulting in very high throughput rates with minimal latency delays and low CPU utilization.

Information about DB2 clusters and cLAN interconnects is in 4.3.1, “DB2 scalability” on page 134

Giganet’s products are in the ServerProven Compatibility Program. See:

`http://www.pc.ibm.com/us/compat/serverproven/giganet.shtml`

Note

The IBM RS/6000 SP switch will be replaced by the Giganet cLAN Cluster Switches and Host Adapters.

5.5 Uninterruptible power supplies

How can servers be protected against power failures? Usually, redundant power supplies and redundant power cords connected to different power lines are used for basic protection.

The failure of one of these components does not affect the operation of the server. To protect the system against a complete power loss a UPS is required. If software such as PowerChute Plus is installed on the system and a communication link is set up between the UPS and the server, PowerChute Plus can stop the applications and shut down the operating system.

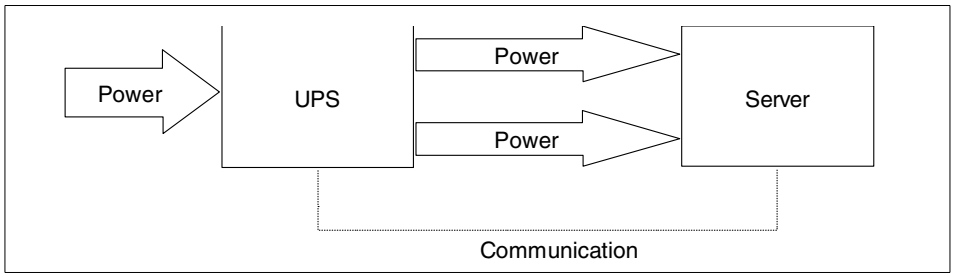

Figure 77. Figure 1. UPS with one server

For larger solutions such as an MSCS configuration, a single UPS may not provide sufficient protection.The runtime capacity of one UPS is too little in most cases for two servers with shared storage. Additionally, a single UPS is a single point of failure. Thus you need two or more UPSs. However, in an MSCS environment with one or two UPSs you have certain problems and a solution with more than two UPSs is not available. These problems are:

- **Application handling**

  In an MSCS environment you cannot simply stop an application. Furthermore, the cluster application must be set to offline with the cluster administration tool. If the application is stopped by normal procedures (such as PowerChute Plus “Application Shutdown” or an application-specific stop procedure) then the application would be considered as failed and restarted by the cluster resource monitor.

- **Server status**

  If a power loss occurs on one server, communication about the status of the other server is required. If only the local server is affected by the power loss, then it makes sense to move all cluster resources from this node to the surviving one. Otherwise, if no other node is available or if the other node will also shut down, then the resources must be set to offline.

- **Server power cabling**

  Some servers have an N+1 redundancy in power supplies. If you connect a server with three power supplies and three power cords to two power circuits, you have at least two power supplies on the same circuit. If this circuit fails, with the remaining one power supply the server will not work. In the case of a server with two power cords, you could try to connect them to different UPS units, but then you have the communication problem in the next item.

- **UPS monitoring**

  PowerChute Plus can monitor only one UPS at a time, independent on the type of communication link to this UPS (serial line or network). If you attach more than one UPS to a server, then the server cannot receive signals from all UPS units. Thus you cannot attach more than one UPS to a server.

- **Storage protection**

  You must guarantee that the shared storage is the last component in a cluster to fail, because the cluster service stops immediately when loosing access to shared storage is lost.

- **Storage power cabling**

  Shared storage typically consists of more than one component (RAID controller and multiple drive enclosures). The wrong order of failures of these components may destroy your RAID arrays. For example, if a drive enclosure with more than one drive of a RAID-5 array fails before the RAID controller fails, the RAID controller would mark the whole array as dead.

A team of highly skilled technical specialists from APC, CSG and IBM developed ready-to-use solutions for clustered servers. As a sample installation they used MSCS running on two Netfinity 8500R with fully redundant FC array common disk subsystem. For detailed information, read the redpaper *Implementing UPS Configurations with Microsoft Cluster Server* available from `http://www.redbooks.ibm.com`

5.6 Backup and restore

It is rarely necessary to justify the need to perform regular and comprehensive backups of data on computer systems to experienced IT staff. If data is lost through hardware failure, software failure (including viruses, worms, and other attacks), user error or malicious intent, you can use a good backup to recover the majority of your data.

While clustering helps to protect your computing systems against downtime, it cannot guarantee that your data will be available. As we have seen, today's Intel-based clustering solutions do not provide complete protection against failure. If one server has developed problems and a failover has occurred, the cluster has done its job. Unlikely as it may be, should a second problem now occur, you have the real risk of losing data. So backups are important for clustered systems even if you ignore the possible problems caused by users.

User data and program files for applications held in directories and files in the operating system's file system need to be protected. However, to effect a full recovery, your backup needs to include data the operating system uses to maintain and manage operation of the whole system, such as resource control and cluster-specific information.

To adequately protect your systems, you should ensure that the following provisions are considered in your plan:

- The use of redundant hardware
- The use of partitioning and replication
- A tested backup-and-restore system
- A disaster recovery plan

When creating a data protection plan for your network, it is important to remember that system information may not be as readily accessible as user data and application files (for example, access to the Windows Registry is strictly controlled). You need to ensure that the backup solution you select can correctly and accurately back up your system and restore it.

You should perform a restoration test at regular intervals to ensure that your backups are sound. It is better to discover your backup process has a problem that prevents restoration during a test instead of when you really need it to work.

For more information about tape backup products and backups in Netfinity clustered environments, see these redbooks:

- *Netfinity Tape Solutions*, SG24-5218
- *Using TSM in a Clustered Windows NT Environment*, SG24-5742

The latter redbook covers Tivoli Storage Manager in an MSCS environment. Also, refer to 6.2.2, "Tape backups in a clustered environment" on page 225.

5.6.1 Backup strategies

Backups are commonly classified into three methods:

- Full

  A complete copy of all data is made. When large amounts of data have to be backed up, this can be very time consuming. Even if your servers are not used overnight, there may be insufficient time to complete a full backup. In extreme cases, or when no downtime is available, special techniques, such as ServeRAID's FlashCopy, may be required to take backups. Full backups simplify recovery because all of your data is stored in a single backup session.

- Incremental

  Data that has changed since the last backup is copied to the backup medium. Incremental backups are efficient and often quick as the amount of data that changes is usually a small percentage of the total data.

  Recovery can be time consuming. To get to the most recent copy of your data, you have to recover the last full backup and then every incremental backup that followed it.

- Differential

  Data that has changed since the last *full* backup is copied. This can be a good compromise between the other methods. As time progresses, more and more data changes and thus has to be copied. However, it is likely to remain significantly less than the total amount of data.

  Recovery also benefits because you only need to retrieve the last full backup and the most recent differential backup to restore your most recent data.

Several approaches to backup have been developed. One common method is to maintain three backups on a rotating basis at, for example, weekly intervals. In this way, you have access to a backup that is very recent, one that is a week older and one a week older than that (sometimes called a son, father, grandfather backup system, for obvious reasons). Assuming your business works a regular five-day week, your backup regime might look something like this:

1. At the weekend, change to the set of backup media containing the oldest backup.
2. On Saturday, take a full backup of your data.
3. Each evening, Monday to Friday, take an incremental backup.

Tape and tape drive vendors can usually offer advice on backup strategies. One that needs 10 tapes to provide backups going back three months can be found at:

`http://www.hp.com/tape/papers/strategy.html`

5.6.2 Data protection guidelines

Careful planning can minimize the amount of time you have to spend managing backups. Ensure that you have a sound data protection plan as part of your systems management strategy.

The following guidelines are suggested when designing a data protection plan:

- Use replication as the first level of protection if possible (Windows NT, Novell, Notes).
- In multiple server networks, maintain at least three replicas of important system data stored on servers across the network.
- Choose a backup product that is certified for your operating system.
- Periodically check with your vendors to ensure you have the latest version of the backup program, device drivers, and so on.
- Stay current with the newest operating system patches.
- Establish a backup strategy or routine and make sure it is executed. The frequency at which you back up data may depend on how much and how often your data changes. Make sure you include regular tests so that you can restore your system from the backup data.
- Keep good records of where backups and replicas are located.
- Always take a full backup before making major modifications to your installation (such as installing service packs).
- Verify the completeness of each backup-and-restore session. After each backup-and-restore session, check the error and log files to make sure the process completed successfully and did not skip crucial portions of your data.
- After reinstalling the operating system from the original media, remember to reapply OS patches and recopy updated drivers, software, and utilities before proceeding with a restore. Check whether your backup software will perform a full recovery from tape. This will save you from having to reload your operating system.

Chapter 6. Maintenance

To be successful in a clustered environment, businesses must be open 24 hours a day, 7 days a week. Those that operate e-businesses should expect to have customers accessing their Web site from multiple time zones, so availability must indeed be 24x7. The technology that innovative businesses are demanding to achieve this is clustering. Clustering allows the linking together of two or more servers to provide the reliability and availability that are essential in round-the-clock operations, all in a very cost-effective manner.

Understanding how to maintain you cluster can be vital to your success in keeping your business-critical applications available for your users.

This chapter provides some basic insight of maintenance and implementation for your Netfinity Cluster solution. It also provides some available offerings by IBM to help maintain your cluster.

6.1 Registering your customer profile

Since clustering is a highly complex technology, IBM offers a variety of programs to help you configure, maintain, and update your Netfinity cluster solution.

- **IBM Netfinity Solution Assurance program**

  Customer environments are growing more complicated and the business-critical applications customers are running require an IT subject-matter expert to carefully check over any proposed solution. The IBM Netfinity Solution Assurance program offers exactly that: your requirements are carefully examined and then IBM helps determine a solution that will meet your needs. Once the solution is determined, the Solution Assurance process will do the following:

  – Verify that the IT components of the solution are compatible
  – Provide a review from an IBM Advanced Technical Support expert
  – Develop action items to reduce high-risk factors
  – Facilitate the transition from solution design to solution implementation

- **Cluster registration**

  You can register their specific configuration data with the IBM HelpCenter. IBM will help verify hardware components prior to installation, and proactively stock parts and deploy personnel to improve service and

response times. Having your configuration on file aids the HelpCenter address problems more quickly.

- **Registering your profile**

  Profiling is the term used where you can register your systems on the IBM support page, `http://www.pc.ibm.com/support`, then customize how the support information appears on your browser and in your e-mail in-box.

  By doing this, IBM can help you perform routine maintenance health checks and keep you up to date with any changes that have occurred in your Netfinity solution. You can profile your systems and components by going to the following Web site:

  `http://www.pc.ibm.com/support`

  Once you have completed your profile, you will be able to:

  – Diagnose and summit problems using the IBM Online Assistant.

  – Participate in IBM's discussion forum.

  – Receive e-mail notifications of technical updates related to your profiled products.

  – Instantly access information related to your profiled products from your own personalized page.

6.2 Maintaining your cluster

Every company should implement a well-organized maintenance plan or change log, to keep up with the changes to their hardware and software. Not only will this aid them with troubleshooting problems, but it will help prevent unexpected server downtimes.

If you are running business-critical applications that requires 24-hour uptime, there are many considerations you should take in account. These include:

- Maintenance logs
- Tape backups
- Preventative maintenance

6.2.1 Maintenance log

A maintenance log is basically a change management log. It records your hardware configuration, all BIOS and firmware settings, and software that is installed on your systems. You should use this log to record any updates or changes to your hardware and software. The record should include date, time, who made the change, and what exactly was changed.

This log should be a handwritten document as well as a published Web document for your company's administrators.

A weekly hardware log review is advisable to help examine and possibly prevent any server issues. Choosing the logs that you will monitor will depend on your platform (for example, ServeRAID, SSA or Fibre Channel). Each hardware platform has its own set of logs for each type of RAID controller. These logs can detect and inform you if there are any problems with the RAID controllers, adapters, and the hard drives.

The Advanced Systems Management processors, when used in conjunction with Netfinity Manager or Netfinity Directory, provides an event log that monitors the local node's hardware and system functions. Reviewing this log can be useful by providing predictive failure analysis regarding your system's hardware.

Tools at the operating system level are available as well. For example, when implementing a Microsoft cluster and your cluster is heavily used, a weekly review of the Performance log can aid you in seeing any trends happening with the hardware in your cluster. Depending on your hardware platform, these trends can give you insight in helping you make any necessary adjustments.

If you have made any changes to your system and experienced an unexpected failure, the maintenance log can be useful in aiding you in back-tracking any changes that have been made to your cluster. This way you can analyze the problem and integrate any necessary changes to get your cluster up and running.

When registering your cluster with the IBM HelpCenter as discussed in 6.1, "Registering your customer profile" on page 223, you can ask the IBM HelpCenter for a cluster template that can assist you in implementing a maintenance log for your company.

All of the Netfinity servers come with a Netfinity Server Library. This library contains each Server's records and a specifications chapter that can serve as a hardware template for your system configurations.

6.2.2 Tape backups in a clustered environment

Loss of data could be costly for your company. That is why it is important to integrate a tape backup strategy plan that will ensure data integrity.

As your data grows, your backup window time frame must stay the same, placing more and more pressure on the backup process. Understanding how

to implement a tape backup solution can be advantageous to your company, especially in a highly complex clustered environment.

For example, in an MSCS solution, if you plan to back up your servers locally, then you should have a tape drive on each node in the cluster with a backup application that is cluster aware. This will enable you to back up the local server and the virtual server.

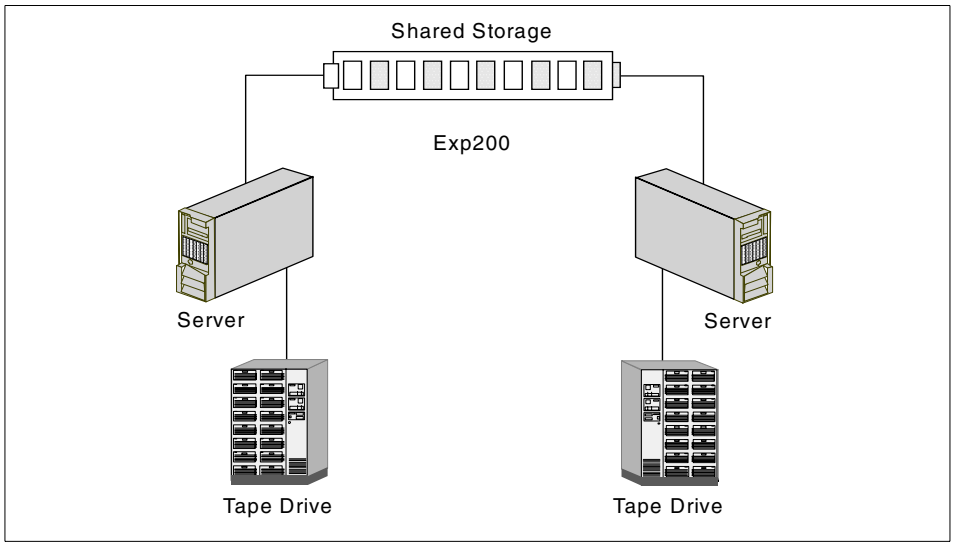

Figure 78. Tape drive configuration in a cluster

Most tape backup applications are not able to determine which node the virtual server is actually located on. If a fail over occurs on one of the nodes and the backup application is not cluster aware, the backup solution will get confused and could cause inconsistent and corrupted backups.

Note: Tape drives are not able to fail over in a Microsoft Cluster Server solution.

Perform your tape backup over the network. This means that you have a separate server on the network with a tape backup application that is capable of backing up the servers over a network.

Since the virtual server is always available, you do not need a cluster-aware tape backup application for your cluster solution.

Legato offers a product call SnapShot Server that creates an image of your live volumes at designated intervals and either hold them or sends them to

your backup solution. This utility gives you the ability to have a central point of control to backup open files and databases without affecting the users. See the Legato Web site for more information:

`http://www.legato.com/products`

For more specific information about tape backup procedures in clustered environments, review:

- 5.6, “Backup and restore” on page 219
- The redbook, *Netfinity Tape Solutions*, SG24-5218
- The Tivoli Storage Manager redbook, *Using TSM in a Clustered Windows NT Environment*, SG24-5742

6.2.3 Preventive maintenance

You should plan at least four hours of downtime a month for maintenance issues. This does not mean you have to take four hours every month, but it will help you plan for any unexpected or necessary changes or updates. The following items should be performed monthly:

- Run `CHKDSK` on all volumes or drives.
- Reboot the servers to clear the system memory.
- Do a trial restoration of a backup to a test system to ensure that the data actually gets restored and is not corrupted.

The `CLUSTER` command can be beneficial to you when implementing a disaster recovery plan for your clustered environment. In the event of a multiple node failure, you will spend hours trying to recreate all of your cluster resources through the graphical interface to Microsoft’s Cluster Management Console.

The other way of creating cluster resources is through the command-line. Put the `CLUSTER` commands used to create the resources into a batch file, then test the batch file to ensure that it works and there are no problems with execution.

Figure 79 shows the syntax of the `CLUSTER` commands used to create groups and resources:

```
CLUSTER /LIST[:domain-name]
CLUSTER [[/CLUSTER:]cluster-name] <options>

<options> =
  /PROP[ERTIES] [<prop-list>]
  /PRIV[PROPERTIES] [<prop-list>]
  /PROP[ERTIES][:propname[,propname ...] /USEDEFAULT]
  /PRIV[PROPERTIES][:propname[,propname ...] /USEDEFAULT]
  /REN[AME]:cluster-name
  /VER[SION]
  /QUORUM[RESOURCE][:resource-name] [/PATH:path] [/MAXLOGSIZE:max-size-kbytes]
  /SETFAIL[UREACTIONS][:node-name[,node-name ...]]
  /REG[ADMIN]EXT:admin-extension-dll[,admin-extension-dll ...]
  /UNREG[ADMIN]EXT:admin-extension-dll[,admin-extension-dll ...]
  NODE [node-name] node-command
  GROUP [group-name] group-command
  RES[OURCE] [resource-name] resource-command
  {RESOURCETYPE|RESTYPE} [resourcetype-name] resourcetype-command
  NET[WORK] [network-name] network-command
  NETINT[ERFACE] [interface-name] interface-command

<prop-list> =
  name=value[,value ...][:<format>] [name=value[,value ...][:<format>] ...]

<format> =
  BINARY|DWORD|STR[ING]|EXPANDSTR[ING]|MULTISTR[ING]|SECURITY|ULARGE
```

Figure 79. Cluster command line syntax

Figure 80 is an example of a batch file that has been created to make a cluster group named ClusterGroup and an IP Address resource.

```
cluster %1 group ClusterGroup /create
cluster %1 resource _IBM_ClusterGroup /create /group:ClusterGroup /type:"_IBM_HIDDEN"
cluster %1 resource ClusterIPAddress /create /group:ClusterGroup /type:"IP Address"
cluster %1 resource ClusterIPAddress /adddep:_IBM_ClusterGroup
cluster %1 resource CLusterIPAddress /priv /address=203.3.1.150
cluster %1 resource ClusterIPAddress /priv /subnetmask=255.255.255.0
cluster %1 res      ClusterIPAddress /priv /network=public
cluster %1 res      ClusterIPAddress /priv /EnableNetbios=1
```

Figure 80. IP address batch file example

Figure 81 is another batch file example of creating a Network Name resource.The name of the resource is ITSOCLUSTER.

```
cluster %1 resource ClusterName /create /group:ClusterGroup /type:"Network Name"
cluster %1 resource ClusterName /adddep:ClusterIPAddress
cluster %1 resource ClusterName /PRIV name=ITSOCLUSTER
cluster %1 res      ClusterName /priv
cluster %1 res      ClusterIPAddress /priv
echo "1"
```

Figure 81. Network Name batch file example

The next procedure would be gathering of all the individual batch files and creating one centrally managed batch file. This is illustrated in Figure 82:

```
cluster %1 group ClusterGroup /create
cluster %1 resource _IBM_ClusterGroup /create /group:ClusterGroup /type:"_IBM_HIDDEN"
cluster %1 resource ClusterIPAddress /create /group:ClusterGroup /type:"IP Address"
cluster %1 resource ClusterIPAddress /adddep:_IBM_ClusterGroup
cluster %1 resource CLusterIPAddress /priv /address=203.3.1.150
cluster %1 resource ClusterIPAddress /priv /subnetmask=255.255.255.0
cluster %1 res      ClusterIPAddress /priv /network=public
cluster %1 res      ClusterIPAddress /priv /EnableNetbios=1
cluster %1 resource ClusterName /create /group:ClusterGroup /type:"Network Name"
cluster %1 resource ClusterName /adddep:ClusterIPAddress
cluster %1 resource ClusterName /PRIV name=ITSOCLUSTER
cluster %1 res      ClusterName /priv
cluster %1 res      ClusterIPAddress /priv
echo "1"
cluster %1 group RSTestGroup /create
cluster %1 resource _IBM_RSTestGroup /create /group:RSTestGroup /type:"_IBM_HIDDEN"
cluster %1 resource RSFileShare /create /group:RSTestGroup /type:"File Share"
cluster %1 resource RSFileShare /adddep:_IBM_RSTestGroup
cluster %1 resource RSFileShare /priv /sharename=rsfs
cluster %1 resource RSFileShare /priv /path=c:\rsfs
cluster %1 res      RSFileShare /priv /MaxUsers=-1
cluster %1 resource RSGenApp /create /group:RSTestGroup /type:"Generic Application"
cluster %1 resource RSGenApp /adddep:_IBM_RSTestGroup
cluster %1 resource RSGenApp /priv /commandline=calc.exe
cluster %1 resource RSGenApp /priv /InteractWithDesktop=1
cluster %1 resource RSGenApp /priv /currentdirectory=c:\winnt.0\system32
cluster %1 res      RSGenApp /priv
cluster %1 res      RSFileShare /priv
echo "2"
cluster %1 group RSTestGroup /prop /AutoFailbackType=1
cluster %1 group ClusterGroup /prop /AutoFailbackType=1
cluster %1 group RSTestGroup  /setowners:%2
cluster %1 group ClusterGroup /setowners:%2
```

Figure 82. Group of batch files combined into one large file

If your cluster suffers any kind of disaster, you can execute one large batch file that will execute the small batch files and, in return, will recreate all of your resources the way you originally set them up. Therefore, these precautionary procedures will greatly reduce your downtime and speed up the recovery process.

6.3 Creating a test cluster

Before implementing any changes or updates to your production LAN, employing a test cluster on a test LAN can remarkably reduce your downtime and at the same time eliminate many costly mistakes.

The following are reasons why a test cluster can be an invaluable resource to your production LAN:

- **Change management**

  Change management is defined as implementing one change at a time to your cluster solution. Use this management technique to introduce changes to your BIOS, firmware, system and operating system. You will be able to see how the changes affect your Netfinity cluster. This is also a perfect way to test out any service pack. This way, if there are any adverse effects, it will not affect your clients and the production LAN.

- **Software analysis**

  You can use your test cluster to perform analysis for your corporate or third-party application software to see how it reacts with your cluster solution.

- **Testing a backup**

  A test cluster can ensure that your backup solution is functioning properly. By doing a full restore of data, you can also verify that your data is not corrupt or inconsistent. This will also give you practice in deploying your disaster recovery plan. You can also experiment with different backup applications to see if they meets your business needs.

- **Parts availability**

  Normally it is good practice to have a 10% overstock of vital parts on hand. This is so you can replace a failed option at the time of failure. Otherwise, you would have to wait on a service technician to come out and service the option and the technician might not have the option in stock. This means that there will be a considerable amount of downtime that you can not afford. With a test cluster, you already have working parts available at your disposal, so you can replace any failing part whenever needed.

6.4 Applying updates and replacing components

When implementing clustering environments, it is important that you do *not* set a preferred owner for either the cluster group or the quorum resource.

If you need to shut down either server in a clustered environment, you *must* manually transfer all resources to the node that is to remain online *before* invoking a shutdown. If you are installing software on a machine in a production clustered environment, be aware that the installation process may automatically shut down or reboot the machine. Ensure that all resources are transferred away from the machine on which software is being installed. If the possibility exists that the installation will restart the machine, you need to stop Cluster Service on the server that is to be shut down.

Note: It is important that you follow the correct steps when you want to shut down a cluster. Likewise, you should also follow the correct sequence when starting up the cluster again. Failure to do so may result in unexpected failovers, split-brain configuration, or defunct drives. See the publications *IBM Netfinity ServeRAID Cluster Solution* or *Netfinity FAStT Storage Manager for Windows NT Installation and Support* for information about the correct procedure. These are available from `http://www.pc.ibm.com/support`

6.4.1 Installing service packs

Before you install any service pack, you must take the following precautionary steps:

- Close all open applications and disable all virus scanners to make sure that the installation program can replace system files that are in use. Contact your application vendor for instructions before performing an upgrade.
- Update the Emergency Repair Disk.
- Perform a full backup of the cluster, including the registry files.
- Check your computer's available memory and disk space against prerequisites for the service pack.

6.4.1.1 Windows 2000 Advanced Server

At the time of publication, Service Pack 1 (SP1) is the latest update for Windows 2000. Microsoft currently states that SP1 is not a required update. Our analysis of the list of fixes shows that there are no updates in SP1 that apply to Cluster Service nor Network Load Balancing.

If you do apply the update, simply follow the Windows NT 4.0 Enterprise Edition procedure for installing a Service Pack in a clustered environment, as described in the following sections.

Note: Unlike Windows NT 4.0 Enterprise Edition, Windows 2000 Advanced Server will perform a rolling upgrade automatically. This means that once you install the service pack, it will then automatically pause the second node to install the service pack. Once the service pack upgrade is complete on the first node, that node is paused. At this point, the second node is unpaused and it begins the service pack upgrade.

6.4.1.2 Windows NT 4.0 Enterprise Edition

You can eliminate the downtime of your cluster services and reduce your administrative tasks by performing a rolling upgrade of the operating system. In a rolling upgrade, you upgrade the operating system on each node one at a time. This allows the other node to be available to handle client requests.

A rolling upgrade consists of the following steps:

1. Start — Each node is running.
2. Node 1 upgrade — Node 1 is paused, and Node 2 handles all cluster resource groups while you upgrade the operating system of Node 1 to the latest service pack.
3. Restart Node 1 — Node 1 rejoins the cluster.
4. Node 2 upgrade — Node 2 is paused, and Node 1 handles all cluster resource groups while you upgrade Node 2 to the latest service pack.
5. Restart Node 2 - Node 2 rejoins the cluster.

Note: If you are not sure that your application can be upgraded, contact your application vendor to see if there are any instructions for any service pack updates.

6.4.2 ServeRAID updates and recovery

IBM periodically makes updated versions of the ServeRAID firmware, device drivers, and utility programs available from the IBM Support Web page:

`http://www.pc.ibm.com/support`

When you update any RAID software, you must update *all* RAID software at the same time to ensure compatibility with all of the levels of the software. This means the following components:

- BIOS/firmware

- Device drivers
- ServeRAID Manager
- IBM ServeRAID Cluster Solution software

You should ensure the ServeRAID controllers have the latest IBM ServeRAID software installed. However, when using Windows Cluster Service (MSCS), be sure that you use a version of the ServeRAID software that has been certified by Microsoft for Windows NT and Windows 2000 compatibility.

6.4.2.1 Failed adapter replacement

There is the possibility that one of the servers, one of the controllers, or one of the hard disk drives in a cluster might fail. This section presents procedures you can use to recover from these situations.

If you have to replace your ServeRAID controller, you must reconfigure the controller after you have installed it in your server. It is important to note that this procedure requires specific configuration settings for the ServeRAID controller.

If the controller is not functional, you will need a record of these settings.This is where the maintenance log, discussed in 6.2.1, “Maintenance log” on page 224, can be very helpful.

You must have the following information to configure your new ServeRAID controller:

- SCSI initiator IDs
- Controller and partner name
- Stripe-unit size
- Unattended mode setting

There are some guidelines to follow when replacing your failed controller:

- Note which SCSI cables are connected to the SCSI channel connectors on the controllers. This is so you can replace the cables back to the correct SCSI channels.
- Note which PCI slot the controller is installed in.
- Do not reconnect the SCSI channel cables to the controller until you have ensured that you have the same level of ServeRAID BIOS and firmware on both controllers in the cluster.

6.4.2.2 Recovering from a power failure during failover

If a power failure occurs during a failover, it is possible that the two controllers in the active-passive pair might be in a state where some logical drives are

configured on one controller and some logical drives are configured on the other controller. It is also possible that there might be one logical drive that does not show up on either controller.

To recover from this problem, run the `IPSSEND` command-line utility. `IPSSEND MERGE` needs to be run once for every merge group ID that you configured in the pair on the controller that you want to become active. Then, run `IPSSEND UNMERGE` once for every merge group ID that you configured in the pair on the passive controller. Then, restart Windows to pair the controllers again.

6.4.2.3 Other notables

Here are some other pointers for ServeRAID clustering:

- After a failover occurs in an MSCS server, the red drive lights in the IBM Netfinity EXP15 enclosure might turn on for physical drives that are online. There is no hardware problem—the lights turn off within an hour.The ServeRAID is unaware of the drive state change, so it has to wait for the EXP15's ESM board to update the ServeRAID controllers.
- Do not install Microsoft Cluster Service during the installation of Windows 2000. You will install the Cluster Service later.
- You must have both servers turned on during the installation of Cluster Service if your servers are connected by a cross-over cable. Windows 2000 will mark your network card as failed and you will not be able to select it as the private network interface during setup.
- Windows 2000 requires an NTFS partition for the operating system and the shared cluster disks.

6.4.3 Fibre Channel updates

The IBM Netfinity Fibre Channel Storage Manager V7.0 is a Java-based tool for managing Netfinity Fibre Channel storage products. The new interface has a similar look-and-feel to ServeRAID management for an easier transition from one to the other.

You cannot upgrade the Netfinity Fibre Channel controller while running Windows 2000 if you are not at least at Version 7.01. If you plan to migrate or install Windows 2000 Advanced Server, you must upgrade all the firmware (appware, bootware, and NVSRAM) on the controller while running Windows NT 4.0 Enterprise Edition.

For expedience, IBM recommends Version 04.00.01.00 or higher for the controllers. If you installed Windows NT4.0 EE just to update the controllers,

IBM recommends formatting the drives and doing a clean install of Windows 2000 Advanced Server.

If you are running Novell, you must have a Windows NT 4.0 or Windows 2000 client on the network to administrator and perform updates on the Fibre Channel solution. This client must also have an Ethernet connection. Consult the *Networked SYMplicity Storage Manager Installation and User's Handbook, Second Edition* for more information.

Support for more than 8 LUNs

To support more than eight logical unit numbers (LUNs), you must install Windows NT Service Pack 5 or greater.There are additional registry parameters that must be set in the IBM Fibre Channel Host Bus Adapter, and you must disable the miniport's extended LUN mapping. Refer to the host adapter's readme file for the latest information on these settings. This is not needed in Windows 2000.

Note: Disabling the miniport's extended LUN mapping does not apply to the IBM Netfinity QLA2200 Host Adapter because this is the adapter's default setting. This adapter also does a true fabric login.

6.4.3.1 Scripts

The following script files are necessary for IBM Netfinity Fibre Channel Storage Manager V7 or higher. These files change NVSRAM settings in the user-configurable area of NVSRAM based on special criteria. This area is used to make changes that might be required by a user to satisfy different operating environments, such as clustering or Novell NetWare.

Table 28. Fibre Channel script files

Script	Purpose
SoftResetOn.scr	Enables soft reset setting to IBM Netfinity FC RAID Controllers for Microsoft clustering environments (required for clustering)
SoftResetOff.scr	Disables soft reset setting to IBM Netfinity FC RAID Controllers for Microsoft Clustering environments
ClusteringOff.scr	Disables IBM Netfinity FC RAID Controllers for Microsoft Clustering
ClusteringOn.scr	Sets up IBM Netfinity FC RAID Controllers for Microsoft Clustering
ResetOn.scr	Enables propagated reset for Microsoft Clustering environments to IBM Netfinity RAID Controllers (required for clustering)

Script	Purpose
WinNT4.scr	Enables IBM Netfinity FC RAID Controllers for Windows NT
Windows2000.scr	Enables IBM Netfinity FC RAID Controllers for Windows 2000
Novell.scr	Enables IBM Netfinity FC RAID Controllers for Novell NetWare

You can execute the script files by clicking **Enterprise Management Tools > Netfinity Fibre Channel Storage Manager Script Tool > Editor**. You can then load the appropriate script file and select **Run**. Users may also modify individual NVSRAM bytes for Novell or UNIX. See the Fibre Channel readme files for installing NVSRAM settings.

6.5 IBM offerings

IBM offers a number of facilities to help you get the most out of your cluster configuration:

- **TechConnect Certification**

  TechConnect is the IBM Netfinity training, certification and information resource. Members receive quarterly CD-ROM mailings containing over 10,000 pages of news and technical information. TechConnect also sends a separate Drivers and Fixes CD, giving convenient access to the latest updates.

  `http://www.pc.ibm.com/techconnect/index.html`

- **Discussion forums**

  IBM-moderated discussion forums provide a way for owners of Netfinity servers to get up-to-date and accurate help. Once registered, click the discussion forum link to get interactive support from experienced IBM technicians and other IBM customers. (Enter the URL all on one line):

  `http://www.pc.ibm.com/support?lang=en_US&page=hottips&brand=IBM+PC+Server&doctype=&subtype=Cat`

- **Online Assistant**

  Use the Online Assistant, an IBM diagnostic tool that the HelpCenter technicians use to diagnose customer problems, to help solve your problems. Use this tool to answer commonly asked questions. All your answers will be saved and transferred to a technician if the online Assistant does not solve your problem.

  `http://www.pc.ibm.com/support?lang=en_US&page=help&subpage=online_assistant&brand=IBM+PC+Server&doctype=&subtype=Cat`

Note: You must register and create a profile before accessing this site. Also, direct support with a technician is available only for customers whose systems are still in warranty.

- **Training for customers**

  IBM PC Institute offers highly qualified instructor-led lab classes that use the latest IBM technology. To register call 1-800-235-4746. To obtain more information about scheduling a class in your area, send requests to the following address: pci@us.ibm.com.

  http://www.pc.ibm.com/training/pci_schedule_ww_techconnect.html

- **IBM SmoothStart Services**

  IBM SmoothStart Services are installation services designed to help accelerate the productive use of your IBM solution. IBM takes care of all the complex details of cluster server installation and configuration so your staff doesn't have to. Upon completion of the installation, IBM will transfer basic operational skills to your staff.

  http://www-1.ibm.com/services/its/us/spsspcclstr.html

- **99.9% Availability Guarantee Program**

  IBM Netfinity servers continue to improve your hardware availability through the Netfinity 99.9% Availability Guarantee Program. The program has been enhanced for more flexible server hardware configurations and larger clusters.

  Services included with this solution to protect your investment with 99.9 percent availability include:

  – Pre-sales Solution Assurance
  – Installation services
  – Setup for Remote Connect and remote monitoring
  – Warranty upgrades and maintenance options
  – Project manager
  – Weekly review of system logs

  Each server has Fibre Channel-connected external storage enclosures, UPS power protection, monitor and keyboard. A supported client workstation must be available for Remote Connect V1.2 using Netfinity Manager V5.2 or greater. Operating system: Microsoft Windows NT Server Enterprise Edition 4.0 with Service Pack 4 or greater and Resource Kit Tools.

  There is no current support model for Windows 2000 and Netfinity Director. IBM plans to support this model in the future. For more

information about the 99.9% Availability Guarantee Program, visit the following Web site:

`http://www.pc.ibm.com/ww/netfinity/999guarantee.html`

6.6 Summary

As you can see, planning, implementing, and maintaining all have important roles when deploying a Netfinity clustering solution. It would be beneficial to your company to consult with an IBM solutions provider because IBM offers a wide variety of programs to help assist you in accomplishing all of these tasks.

Continual analysis is necessary to ensure that there are no unscheduled downtimes. It is of the utmost importance to have the available knowledge of all hardware, its location, version numbers, and configuration for every server in your cluster. The more information the installation team has about the configuration, the better they can plan for rollouts and maintenance.

Chapter 7. Putting the pieces together

Implementing a cluster should not be undertaken lightly. A cluster will probably require additional investment in hardware and software. Your administrative staff will certainly have to learn how to install and manage the new environment. We also strongly recommend that a new cluster implementation be thoroughly tested before being put into production, so that the way your applications behave during failover, failback, and when other cluster events occur is completely understood. Your users may need education to know what to do when a server fails. All of this costs time and money.

In this chapter, we give you some general considerations about aspects of cluster implementation, followed by some more specific details about individual clustering solutions. The closing sections discuss site preparation and disaster recovery.

One area we specifically do not cover in this chapter is network planning. More often than not, a cluster is introduced into an existing network and has to accommodate the topology already in place.

7.1 Deciding to cluster

If you are considering a clustered system in your business, you need to make a number of decisions about the goals of the project, that is the reasons for taking this big step. You might want to review Chapter 2, “Why clustering, why xSeries and Netfinity?” on page 5 to refresh your memory about clustering benefits.

Here are our suggested steps for implementing a cluster:

1. Review your suite of applications and network resources. Which applications and other resources do you wish to make more available?
2. What level of availability do you require? How many 9s do you need (see Table 1 on page 9)?
3. Is a cluster necessary, or can this level of availability be achieved another way? Redundant components such as power supplies, fans, network adapters, disk controllers and RAID arrays, and early failure warning systems such as the predictive failure analysis feature built into IBM disk drives all help to improve system uptime.

4. If you do not need a cluster to provide availability, do you need other clustering benefits such as load balancing, scalability, or online systems maintenance?

Having decided that clustering offers the characteristics you are looking for, you now need to select a specific clustering product. Answers to these questions will help:

5. Which operating system does your application require?
6. What clustering solutions are available for this platform?
7. Which of these cluster solutions support your application?

To answer these questions, check the compatibility listings that many hardware and software vendors display on their Web sites. IBM Netfinity Servers compatibility information is maintained at these URLs:

`http://www.pc.ibm.com/ww/netfinity/clustering/clusterproven.html`
`http://www.pc.ibm.com/ww/netfinity/serverproven/index.html`

Now that the basic specification of the selected cluster solution is understood, the practical implications of your choice have to be analyzed:

8. What hardware (servers, disk subsystems, network adapters, and so on) supports the selected operating system, clustering solution, and application?
9. Do you know your applications well enough to specify the required hardware to implement them in a clustered environment? How much memory, disk space, and spare capacity (to handle failover) do you need?
10. Do you have or can you afford the hardware and software to implement the cluster?
11. Do you understand the physical requirements of the cluster (power, cooling, security, floor or rack space, and access for service, for example)?
12. Have you considered how factors external to the cluster might affect availability? Network components that are unprotected can bring operations to a halt.
13. Have you thought about backup-and-restore or disaster-recovery procedures?
14. Do you have the necessary skills available to implement a cluster, or at least the time to develop, hire, or outsource them?

Assuming you have answered all of the above questions appropriately, there are now just a few steps remaining:

15. Have you developed an installation and test plan to make sure the cluster performs as expected? Remember to include tests to observe the behavior of clients during cluster events.
16. Does your plan include education for administrators and users?
17. Is one of the deliverables of the plan a document that fully explains the design of the cluster, defines the cluster resources and their interdependencies, and spells out the operational procedures for the cluster?

Note

This last point is important, particularly for clusters implemented specifically for high availability alone. In this case, a cluster could operate without incident for an extended period of time; modern hardware is, after all, very reliable. You need to be sure that an administrator will not have to trust to memory to recall just how the cluster is set up and operates. Even worse, it could be that no one from the original implementation team still works in the department when that first failure occurs. Good documentation is crucial. An example of this is discussed in 7.2.1, “Maintenance log” on page 230.

7.2 Planning for specific cluster products

We now examine particular products that were discussed in earlier chapters, summarizing the protection offered by each product, and highlighting any specific planning implications.

7.2.1 Microsoft Cluster Server

MSCS in itself does not protect your data from all types of problems. MSCS helps to ensure the availability of data to your users but does not protect the data itself. As we discussed in 3.1, “Microsoft Cluster Server” on page 19, the disk subsystem can be a single point of failure. You should plan to use MSCS with other high-availability products and techniques such as redundant hardware (ECC memory, fans, power supplies, and adapter cards), RAID arrays, backup schemes, UPS devices, and disaster recovery strategies.

Failures that MSCS and Windows NT can address by failover include:

- Server connections

- Server hardware such as CPU or memory
- Operating system or application failure
- Network hub failure by using redundant network connections
- Serial I/O operations can be failed over using Serial I/O Multiport Boards and Expandable Subsystems from IBM.

Failures that MSCS cannot address include:

- Power. Use redundant power supplies in conjunction with UPS solutions.
- Disk. Use RAID to protect against drive failures.
- Data loss. Implement a backup solution.
- Network cards. Specify fault-tolerant network adapter cards.
- Routers. Configure redundant links.
- Dial-up. Attach multiple modems and telephone lines.
- Major disaster. Ensure you have a disaster recovery plan.

7.2.1.1 Clustering applications with MSCS

Most applications can be used in an MSCS environment. The primary restrictions are that a clustered application must be able to place its data on any disk in the server, specifically on the common disk subsystem and communicate using TCP/IP. It is worth examining these restrictions in more detail to understand how they affect applications.

- Data

  The application should store sufficient information on the common disk subsystem to enable a full recovery. By this, we mean that the application should be able to find a consistent status after failover. Some data loss is still likely, but the application must recover to a consistent state. If, instead, the application stores status information in the registry, for example, then it may be unrecoverable because the replication of registry keys may not take place quickly enough and thus be incomplete at the moment of the crash.

  If the application cannot fully recover from the information on the common disk, operator intervention is necessary, and a cold standby server with external disks would provide the same function without the complexity of a cluster.

- TCP/IP

  The application must be able to run in an MSCS virtual server environment. That means it has to communicate using TCP/IP. But it means also that the application has to accept names and addresses from the virtual server. For example, if the application queries the `HKEY_LOCAL_MACHINE\SYSTEM\COMPUTERNAME` registry key instead of calling `gethostbyname()`, it will retrieve the name of the physical node. If this name

is then used within application objects, which may be stored on common disks, the application will not start correctly after a failover.

There must be some minimal support to set up the application in a clustered environment. This may be a cluster-aware installation procedure or just a hint in the readme file. But without any help, you will need to do some experimentation to determine which registry keys need to be replicated and which must not. This is not a trivial task.

7.2.1.2 IBM ClusterProven

Applications can either be aware of the cluster, and make application-specific resources available to MSCS for direct manipulation, or they can be *generic* applications. Cluster-aware applications respond intelligently to *LooksAlive* and *IsAlive* polling and can be managed using MSCS administration tools or the IBM Cluster Systems Management (ICSM) software. Check with the application vendor to assess the level of support provided for operation with MSCS.

IBM ClusterProven program provides two levels of certification, ClusterProven and Advanced ClusterProven, which give a good indication of exactly how cluster-aware an application with certification is. ClusterProven certification requires basic switchover and graceful recovery capability. In other words, it meets the requirements to operate correctly in a clustered environment as discussed in the preceding paragraphs.

An application that meets the Advanced ClusterProven criteria implies the solution also delivers one or more of these benefits to the customer:

- Proactive notification prior to failure (graceful shutdown)
- Application monitoring (intelligent LooksAlive/IsAlive polling)
- Recovery of transactions upon failure (which may require work from client applications)
- Start/Stop scripts (applies when running Legato Co-StandbyServer for Windows NT)
- Dynamic workload balancing
- Cluster serviceability and diagnosis
- Reduction of costly downtime for planned upgrades (rolling upgrade)

7.2.1.3 Failover options

As described in 3.1.11, “Failover” on page 36 and 3.1.12, “Failback” on page 39, if an application fails, rather than the operating system or the server itself, MSCS can be configured to restart the application on the same server a specific number of times before transferring it to the other server. You can

specify how many times this failover will occur before MSCS gives up and leaves the application in a nonoperating state.

When configuring MSCS, you can also specify which server or node is the preferred node for a particular application. If that node fails, the application is restarted on the second node. If the preferred node comes back online, MSCS can be configured to automatically return the application back to its preferred node. You can specify whether this failback occurs immediately or during specified hours in the day.

7.2.1.4 Cluster configurations

As stated in Chapter 2 of *Microsoft Cluster Server Administration Guide*, MSCS can be set in one of a number of configurations depending on your availability requirements. Here are some examples:

- Active-active

  This is typically how an MSCS configuration is set up. Applications run in production on both servers in the cluster. Should either server fail, the applications on that server are restarted on the surviving server. The surviving server may run in a degraded state due to the extra workload placed upon it.

- Active-hot spare

  In this cluster configuration, the second node does not normally perform any useful work, but should the first server fail, all applications will restart on the second node with no performance degradation. The downside of this configuration is the additional hardware costs to implement the hot spare.

- Mixed failover

  While you may have applications that are suitable candidates for MSCS, you may also have applications (such as those that are based on IPX and cannot use TCP/IP) that cannot be clustered. Note that applications that use NetBIOS over TCP/IP work well with MSCS.

  In this situation, you can configure MSCS to allow failover of some applications but not others. If a hardware failure occurs, those IPX applications will not fail over, but when the server comes back online, they can be automatically restarted.

7.2.1.5 Determining groups

After deciding what form your cluster will take, you need to group your applications and resources together. Review Chapter 2 of *Microsoft Cluster Server Administration Guide*.

The key steps in determining how your groups should be organized are:

1. Determine how many IP addresses and network names are needed.
2. Design a naming schema.
3. Decide how many disks (which may fail over independently) must be seen by Windows NT.
4. Determine how to implement this number of disks using the subsystem you have chosen. The following table summarizes the situation for the three subsystems we have already discussed:

Table 29. Disk subsystem to Windows NT disk correspondence

Disk subsystem	Subsystem mapping to Windows NT disks
SCSI (ServeRAID)	One array -> one logical drive -> one Windows NT disk
SSA	One array -> one system resource-> one Windows NT disk
Fibre Channel	One LUN -> one Windows NT disk

   Note that for Fibre Channel subsystems, LUNs of the same array may fail over independently, but for safety and performance reasons some files should be distributed to different arrays, for example, separate database logs, archived logs and tablespaces.

5. Do not forget hot spares and room for later expansion.

These are the factors you should consider when setting up your groups:

- All the resources that have dependencies defined between them must be located in the same group.
- Resources can belong to only one group.
- Whole resource groups fail over, not individual resources. When a resource fails, the entire group fails over to the other node.
- It can be helpful to draw a dependency tree depicting all the resources and how they interact with each other.
- Refer to 3.1.1, “Resources” on page 21 for details about resources, resource types, and dependencies between them.

6. If you are using multiple storage units (EXP enclosures), then you should consider making your arrays orthogonal. Orthogonal means that the array is configured such that only one drive is housed in each storage unit. This way, if you have a storage unit failure, your LUN will only go into a critical mode as opposed to going completely offline.

7.2.2 Legato Co-StandbyServer

This product is an alternative to MSCS. The main difference between the two is that MSCS uses a common disk subsystem while Co-StandbyServer is

implemented by using disk mirroring technology. In addition, Co-StandbyServer supports only two virtual servers (failover groups) whereas MSCS can create as many as you wish.

Failures that Co-StandbyServer can address include:

- Server connections
- Server hardware such as CPU or memory
- Application failure
- Network connections (using redundant network connections)
- Network printer
- Data (because data is mirrored)
- Serial I/O operations can be failed over using serial I/O multiport boards and expandable subsystems

Other positive features of Co-StandbyServer:

- The distance between the two servers can be much greater than with MSCS. WAN distances can separate the servers if a high-speed, low-latency link is available. This makes implementing disaster recovery processes much easier.
- Both active/active and active/passive configurations are supported.
- Co-StandbyServer uses native Windows NT APIs to cluster other resources such as IP addresses, shares, and NetBIOS computer names.
- Support for all storage technologies is provided, including RAID subsystems.
- Co-StandbyServer uses TCP/IP protocols and industry-standard network cards for mirroring traffic between clustered servers.
- Hardware certification is not required.
- Matching hardware and software configurations are not required.

Possible drawbacks of Co-StandbyServer:

- Mirroring can be expensive when a large amount of storage is needed.
- You will probably want to implement RAID subsystems in each server to protect against disk failure while one of the servers is down.

7.2.2.1 Hardware

Co-StandbyServer requires two servers running Windows NT Server 4.0 with Service Pack 3 or later. The server hardware need not be identical; 30 MB of hard disk space is needed on the system drive for the Co-StandbyServer files. Hardware specifications for each server include:

- Three physical disk drives (or three logical drives configured in a disk array) for an active/active environment, or two drives per server for an active/passive environment.

- Two network interface cards. One is used for the client network, the other is for the recommended dedicated 100 Mbps or faster mirror link. The dedicated link should only have the TCP/IP protocol bound to it.

7.2.2.2 Organizing cluster resources

To organize cluster resources, they are assigned to a failover group, which is the smallest unit that can fail over.

In the Co-StandbyServer Management Console, server resources are displayed in the Resource window under their associated server. After a resource is clustered, it is displayed under its associated failover group in the Cluster window, and under each server in the Resource windows. In the Resource windows, the cluster resource is displayed as active on one server and inactive on the other. As a failover group moves from one host to the other, these states reverse to show where the cluster resource is currently active.

If the Automatic Failover option is enabled, the group on a failing host will move to the surviving host and still be available. If the Automatic Failover option is not enabled and a host fails (which moves its failover group to the Inactive Groups folder), or if you manually inactivate the group by moving it to the inactive folder, the resource becomes unavailable. These attributes are important to understand so that as you cluster resources you consider what failover properties are currently assigned and what additions must be made to the failover group.

It is important to note that you have less flexibility in defining groups than with MSCS because Co-StandbyServer supports only two of them.

7.2.2.3 Failover options

Co-StandbyServer's bidirectional failover enables either node to take over the functions and identity of the other. This includes IP addresses, shares, print functions, server names, and applications. Resources are organized into two failover groups. Each failover group has its own user-configured failover properties that control actions such as:

- Automatic failover
- Alerts
- Command file execution

7.2.2.4 Failover properties

You can set up failover properties to influence events before and after failover. There are six properties.

- **Name** — Failover group name.

- **Active On** — Current host server.
- **Automatic Failover** — When enabled (checked) and both the network and Legato link connections indicate that a server has failed, Co-StandbyServer activates affected failover groups (both could be running on a single system) on the surviving server.
- **Delay** — The amount of time Co-StandbyServer waits after a server failure is detected before initiating failover. The delay can be from 0 to 4800 seconds.
- **Alerts button** — Used to access and configure the command files that are executed when a server fails.
- **Files button** — Used to access and configure the command files that are executed when a Failover Group moves between hosts.

Alerts are only executed when a failure is detected. Command files are executed even when a failover occurs manually.

7.2.2.5 Application support

Legato provides an application support script for many of the most commonly used applications, such as Microsoft SQL Server, Microsoft IIS Server, Microsoft Exchange Server, Microsoft DHCP Server, Lotus Notes, Oracle, Sybase and Informix.

A current list of the available application support scripts can be found at:

`http://www.legato.com/products/availability/costandbyserver/nt/ntco_tools.cfm`

7.2.3 Novell StandbyServer

Novell StandbyServer for NetWare is a hardware-independent, high-availability solution that protects the primary server with a standby server. Data is mirrored between the primary and standby servers to create a fully redundant system protecting users against both hardware and software failures.

Failures that StandbyServer can address include:

- Server hardware such as CPU or memory
- Network connections (using redundant network connections)
- Loss of a hard disk drive

Other positive features of StandbyServer:

- The standby server can perform useful work in a Utility Server configuration.
- The distance between the two servers can be large. WAN distances can separate the servers if a high-speed, low-latency link is available. This makes implementing disaster recovery processes much easier.
- Industry-standard interconnects (compliant with ODI and IP) are supported.
- Server hardware does not have to be identical.
- Installation and maintenance are simple.
- NetWare 5 is supported.

Possible drawbacks of StandbyServer:

- Mirroring increases the overall cost of storage.
- There is no support for application failure.

If you operate a Novell network, StandbyServer is an excellent solution to protect your servers and data. More information on these products can be found in *Novell NetWare 5.0 Integration Guide*, SG24-5847.

7.2.4 Application clustering

A lot of information regarding planning for applications that provide their own clustering has already been given in Chapter 4, “Application clustering” on page 95. We recommend these additional resources and the respective product documentation:

- For a more detailed discussion of implementing Lotus Domino on Netfinity servers: *Netfinity and Domino R5.0 Integration Guide*, SG24-5313.
- For information about DB2 and its clustering capabilities: *The Universal Guide to DB2 for Windows NT*, ISBN 0-13-099723-4
- For information about implementing Oracle Parallel Server on Netfinity hardware see *Oracle Parallel Server and Windows 2000 Advanced Server on IBM Netfinity*, SG24-5449.

7.3 Site preparation

When you plan a Netfinity cluster installation, there are important considerations beyond the hardware and software specifications. We have already mentioned documentation of cluster operations, backup-and-restore processes, and disaster recovery planning, for example. It is also essential

that you take into account any requirements for the physical installation of the machines.

The five main elements to consider in site preparation are:

- Physical space
- Electrical power supply
- Cooling and fire protection
- Security
- Cable management

Information to help you determine the exact requirements of your installation are provided in the products specifications.

7.3.1 Space requirements

Obviously, the space requirements for your cluster will depend on the exact hardware configuration. Each Netfinity server has its dimensions specified in the product literature and the announcement information. The external dimensions of each unit are given, along with any additional clearance requirements for cooling and cabling.

Many customers choose to implement clusters as rack-based systems. For Fibre Channel-based clusters, rack hardware must be used as this is the only configuration in which FC components are supported. IBM provides a software tool that ensures you specify all the necessary components to complete rack assembly. This program and other references to assist you in configuring your systems can be found at:

`http://www.pc.ibm.com/us/products/server/download.html`

For large installations, particularly rack-based systems installed on raised flooring, it is possible that the weight of the systems needs to be assessed to ensure a safety hazard is not created. Individual server weights are given in their specifications, and to determine the weight of a rack configuration use the IBM Netfinity Rack Configurator.

IBM also specifies the noise emissions of all Netfinity servers. You may need to assess overall noise emission to ensure no ergonomic or legal limits for your work environment are exceeded.

Note

All of IBM's Netfinity Server family are designed to meet stringent IBM and international standards for safety, noise output, and electrical emissions.

7.3.2 Electrical power supply

The electrical power supplied to your server systems must be sound. A competent electrician should ensure that the wiring, switches, and other electrical equipment are installed correctly and capable of carrying the load you intend to place upon them. As discussed in 5.5, “Uninterruptible power supplies” on page 218, it is of benefit to clusters (and high availability in general) to supply power through two independent mains circuits, so this should be considered.

Using a UPS to protect your servers from power problems is a good idea. In Table 30 we show the run-time estimates for an APC Smart-UPS 3000RMB under several different loads:

Table 30. Run-time estimates with APC SU-3000RMIB

Total configuration load in Watts	Time in minutes for a SU-3000RMIB 30 RIxxx
200	104
400	52
500	38
600	31
700	26
800	22
900	18
1100	14
1300	10
1500	8

Do not forget to calculate the load for all components taking power through the UPS. It is easy to forget items such as monitors. Of course, your cluster will function perfectly without monitors so it is not so important to protect them. Supplying them directly from the mains circuit will extend the run time for the servers and other devices on the UPS.

7.3.3 Cooling and fire protection

Netfinity servers come fitted with cooling fans to ensure that the internal components of the system do not exceed their rated operating temperatures. Cooling fans can only do their job, however, if the air outside the box is below

the maximum allowed for the system. The operating range for Netfinity servers is 10 to 35 degrees Celsius, at a relative humidity between 8 and 80 percent.

Netfinity servers have either standard or optional facilities to monitor important internal temperatures and fan status. When internal temperatures rise due to high external temperatures or fan failure, the Netfinity Manager software will alert administrative staff as preset limits are exceeded. In extreme circumstances, it is possible for some systems to shut themselves down to prevent damage.

7.3.4 Security

There are no specific security considerations for a cluster. The same common-sense measures should be applied as for normal server systems.

As with any important server, physical security is important to prevent unauthorized tampering. At the most basic level, use the keylock and system passwords on the servers themselves to prevent access, and implement virus protection to prevent damage from program or macro viruses and worms.

For more control, consider these stronger security measures:

- Operating your servers in keyboardless and displayless mode
- Tying systems down with optional U-bolts
- Securing cables with optional cable covers

All of these measures help to minimize inadvertent or malicious interruptions to your servers' operation.

Note

If you plan to restart your server remotely or restart servers automatically after a power problem, do not set up a power-on password. Use keyboardless and displayless mode instead.

The best protection for your most critical servers is to locate them in a locked, air-conditioned computer room.

7.3.5 Cable management

When planning the installation of your Netfinity Enterprise racks, you need to schematically plan out your cable management for your servers and its

options. Poorly managed cabling can lead to problems with your servers and this will adversely affect your cluster.

All the Netfinity rack mounted servers come with rack-mounting kits.These kits provide the necessary parts to help you manage the cabling. If you are using multiple Netfinity racks, you might have to purchase additional power cables to meet the distance requirements.

The following are some cabling considerations:

- When cabling the server, make sure that the server can be fully extended out of the rack with the cables intact.
- When looping cables, make sure that they are no smaller than four inches in diameter.
- When using Fibre Channel cables, make sure that there are no 90 degree angles. The cable could splinter and this will cause a decrease in performance.
- Ensure that your are using the proper power cables because it might be necessary to use a longer power cable to connect a device to a UPS that is housed in another rack.
- You should label your cables for easy deployment and replacement.
- Ensure that there is no cabling on the ground or any loose cabling throughout the rack. Fibre channel cables are easily damaged if stepped on.

By implementing proper cabling procedures, you can save many hours of downtime.

7.4 Disaster recovery

Disaster recovery is the process of recovering data after a major accident has completely destroyed your server installation. A burst water pipe that causes irreparable damage to a system, or an explosion that destroys your computer room are, no doubt, unlikely events, but if such an accident occurred, it could cause a major loss of business.

Part of your normal backup routines should require safe storage of the backup media in a secure off-site location. This off-site media is available for system recovery if the on-site data is lost or damaged in a disaster. Procuring a replacement system and restoring data from backups can, however, be a time-consuming process. Clustering over a large geographic area can minimize any downtime due to disaster.

The first step in developing a disaster recovery plan is to define your requirements in terms of where your critical data resides and how often it should be backed up (see 5.6, “Backup and restore” on page 219 for more on this).

You can choose to implement a centralized backup methodology, where one system backs up all critical data on your network, or a decentralized methodology where each system has its own backup hardware. A centralized system is good from a management perspective because fewer trained staff and less backup hardware are needed. The disadvantages are that more data will have to travel over the network and more expensive backup equipment may be required to store the additional data in the time available. Designing and implementing the right hardware and software solution is the key to a successful backup and disaster recover strategy.

You need to account for growth in data volume, so we recommend that you buy a larger, faster backup system than you think you need. Remember to test the restore function at regular intervals in your plan.

7.4.1 Clustering for data protection

Backup and archiving are traditional and popular methods of data protection. Clustering offers an alternative way to protect your data against disaster.

7.4.1.1 Clusters using a common disk subsystem

Clusters with a common disk subsystem have a distance limitation that varies depending on the nature of the disk attachment technology. These technologies are described in 5.2, “Disk subsystems” on page 184. As we saw, the maximum distance we can achieve between a server and a disk subsystem is 11 km, using Fibre Channel (see Figure 83). This is obviously much better than the 24 m or so achievable with SCSI, and will probably be sufficient to protect against a local fire or an explosion. If you are unfortunate enough to be affected by an earthquake or flood, however, it may not be enough to guarantee that at least one server will remain functional.

You also have the problem that if disaster strikes the disk subsystem, you are still severely affected. Normal backups are therefore indispensable.

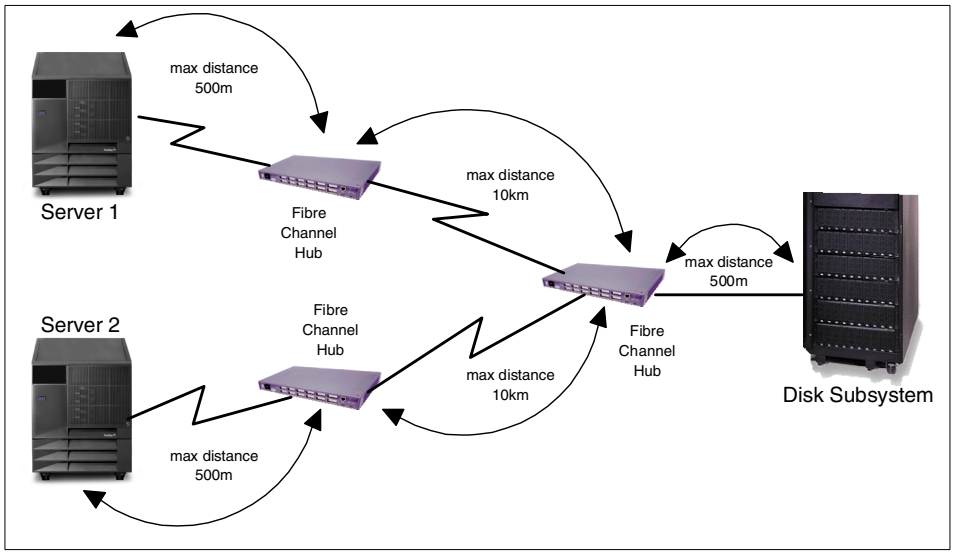

Figure 83. Maximum distance using Fibre Channel

7.4.1.2 Clusters using mirroring

Clusters that maintain copies of your data offer the best protection for your data. If you have two or more replicas of your information in different places and a server is destroyed, you have at least one copy remaining.

A clustered environment such as that shown in Figure 84 is very robust and will survive major calamity. The price of this level of availability is a higher amount of network traffic and more complicated management processes.

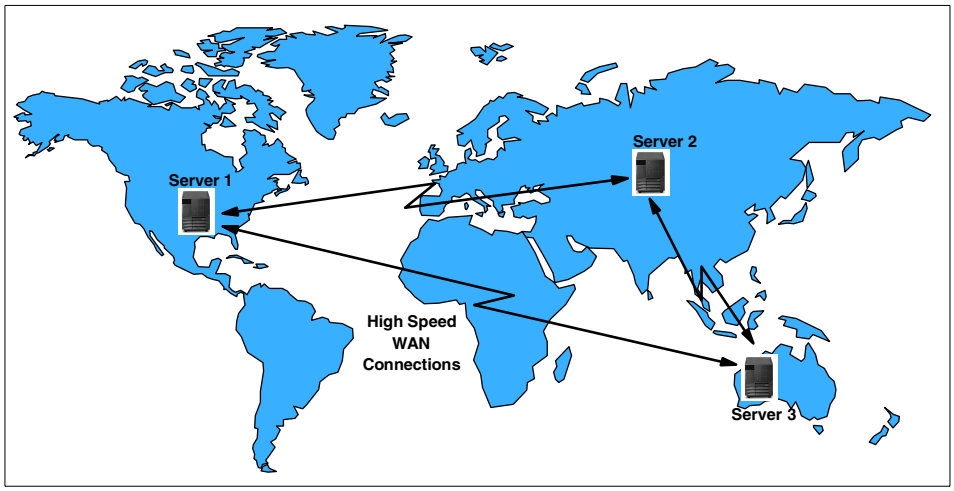

Figure 84. Shared nothing clustering

7.4.1.3 Clustering over long distances

When discussing long-distance clusters, there are several special issues to note:

- There is no way to configure a completely symmetric two-data center cluster without risking data corruption in the case of partitioning. Either you break symmetry and make one data center “more important” than the other (thus this data center can survive alone and, if this data center is damaged, operator intervention brings the other up and running), or you need a third center that contains the quorum resource.

  Without taking these precautions, you have the risk that a fault could result in applications running in both centers. Duplicate IP addresses may not be detected in the case of a complete backbone split, and different users will regard each center as the valid surviving server. Data will then be updated in both centers, making recovery difficult. Precautions against partitioning are critical in planning long-distance clusters.

- Network planning becomes much more important. In addition to the WAN problems, modem and ISDN connections also need to be covered by rerouting dial-in numbers.

- Regardless of advances during recent years, it is difficult to run a Windows NT lights-out data center a long distance away from your location. This requires the best tools available, additional redundancy (to avoid problems due to breaks in the network), and well-trained and experienced staff.

- Failover for the disaster case needs to be tested, and staff members need to be properly trained and drills performed at regular intervals. Testing over large distances is more expensive than a conventional cluster failover and the failback may require significant effort.
- You must be prepared to live with one data center for a long time. In the case of a real disaster, it may take weeks or months to recover a lost data center. During this period, you need to be sure your remaining data center has all the necessary processes and systems in place to survive all but another major disaster. These include backup procedures, implementation of redundant subsystems and all of the other high-availability techniques, which may even include local clusters.

This incomplete list shows that preparing for a major disaster is not a simple exercise. It needs a lot of foresight and planning to consider every possibility, perform a risk assessment for each one and decide on the appropriate steps to minimize the impact should the worst case situation arise.

7.4.2 Backup and recovery examples

Backup and, in particular, recovery are complex tasks that vary depending on your application, your operating system, your hardware, and your processes.

A backup, in concept, is easy to produce. You just need the right hardware and software tools and everything is ready to save your data. It is important to have a well-defined process to ensure your data is available for restoration when you need it. You can read more about this in 5.6, “Backup and restore” on page 219.

Recovery is more complex because you may be restoring to different hardware because of an upgrade or a disaster. You may also only want to recover selected data. So you have more variables to contend with when performing data recovery. Each time you recover data it could be different from the last.

Backup and recovery processes for a cluster are slightly different compared to those for stand-alone systems. Similar strategies may be used, but some details and restrictions have to be considered.

Note

While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Customers attempting to adapt these techniques to their own environments do so at their own risk. It is important to test your backup and recovery processes to ensure the integrity of your data.

7.4.2.1 Microsoft Cluster Server

To back up data stored on the common disk subsystem, you could install backup hardware on each cluster member and have each node back up its own data. A more efficient solution is to nominate a server outside the cluster as a backup server and schedule regular backups of the cluster data to that machine.

Here are some suggestions for backing up and restoring the individual cluster nodes:

Backup

1. Stop the clustering service on both servers.
2. Stop the shared disk service.
3. Disable the cluster disk device driver in Settings/Control Panel/ Devices (for example, the IBM ServeRAID cluster driver).
4. Make a full backup of the Windows Registry, Access Control List, and files as needed.
5. Make a full backup of \WINNT\CLUSTER directory.
6. Restart the cluster disk device driver.
7. Restart cluster services again.

Restore

1. Install Windows NT, and restore the system from your backup.
2. Disable cluster disk device driver.
3. Restore \WINNT\CLUSTER directory.
4. Enable cluster disk device driver.
5. Start cluster services again.

Note that Microsoft provides a utility with Windows NT 4.0 Enterprise Edition specifically for backing up cluster configuration data, called Cluster

Configuration Backup (ClusConB). More information can be found in the MSCS release notes (\MSCS\Readme.doc) found on the second product CD.

Note: there is no Cluster Configuration Backup utility in Windows 2000. Refer to Microsoft's backup and restore procedures on this subject covered in Knowledge Base article Q248998 “How to properly restore cluster information”.

A complication arises because Windows NT uses disk signatures to identify individual physical disks, including those in the common disk subsystem. Disk signatures are stored in the registry, which means you cannot restore a Windows NT backup onto another computer. So, if a failed cluster node is replaced with a new node, you must reinstall Windows NT Server Enterprise Edition on the new computer. Assuming the other cluster node is still working, you can use Cluster Administrator to evict the replaced node, then install MSCS on the new node, and join the existing cluster, and, finally, restore applications and data.

Some additional hints and tips for restoring Windows NT after replacing hardware can be found in the following Microsoft Knowledge Base articles:

Q112019Changing Primary Disk System After Installation
Q130928Restoring a Backup of Windows NT to Another Computer
Q139822How to Restore a Backup to Computer with Different Hardware
Q139820Moving or Removing Disks & Fault Tolerant Drive Configurations
Q113976Using Emergency Repair Disk With Fault Tolerant Partitions
Q172951How to Recover from a Corrupted Quorum Log
Q172944How to Change Quorum Disk Designation

See C.5, “Microsoft Knowledge Base articles” on page 276 for a complete list of articles relevant to Microsoft clustering.

7.4.2.2 Legato Co-StandbyServer for Windows NT

When a server failure causes an automatic failover of a failover group, the cluster is considered to be in a failed-over state because there is only one host server. This condition should be repaired as soon as possible to return the system to its original high-availability state. The steps for recovery depend on what caused the server to fail.

If a server hosting a failover group fails and the automatic failover option is armed, the failover group will activate on the surviving server. If the automatic failover option is not armed, you can move the failover group manually to the surviving server to activate the resources and make them available to users.

In most cases when the server is repaired and returned to the cluster, it automatically synchronizes the cluster volumes and is available for hosting the failover group(s). In other cases, such as when the system disk, a mirrored disk, or a network card fails, you should follow the procedures below.

System disk or the complete system failed

1. Uninstall Co-StandbyServer from the surviving machine.
2. Replace the failed components in the failed server and if you replaced the entire server, make sure that the new server meets the system requirements for Co-StandbyServer.
3. Reinstall Co-StandbyServer to the two servers.
4. Reconfigure the cluster.

Mirrored disk failed

1. Move the failover group to the server containing the surviving disk.
2. Remove cluster volumes associated with the failed disk.
3. Shut down the server with the failed disk.
4. Replace the failed disk.
5. Start the server containing the new disk.
6. Run Disk Administrator to assign a disk signature to the new disk.
7. Recreate the cluster volumes.
8. Move the Failover Group back to the repaired server.

Network card failed

When you suspect that a network card has failed, you can usually confirm its failure by checking the Windows NT event log. If you replace the card with hardware that uses the same driver, it is a simple process to become fully operational again:

1. Move the failover group to the healthy server.
2. Remove Cluster IP Addresses from the failed card.
3. Remove the network adapter driver using the Network applet in Control Panel.
4. Shut down the server with the failed network card.
5. Replace the failed network card.
6. Start the server.

If the network card is a different make from the original or if it uses a different driver, use the Network applet in Control Panel to install a new network adapter driver. Before continuing the wizard, stop the Legato Service:

1. From the Windows NT Control Panel, run the Service Control Manager.
2. Select the Legato Service and click **Stop**.
3. Configure the driver and network properties for the new network card. (Do not restart the server yet.)
4. From the Windows NT Control Panel, run the Service Control Manager.
5. Select the Legato Service and click **Start**.
6. Restart the server.
7. Re-create the Cluster IP Addresses.
8. Return the failover group to its former host.

7.4.2.3 Novell StandbyServer

When a primary server failure and the subsequent standby failover occur, the failed server should be repaired or replaced. After the failed server has been repaired, the following steps need to be taken to restore StandbyServer operation. Note that these steps will leave the cluster with what was the standby machine as the primary server and the replaced server will assume the standby role.

1. After servicing and powering up the replaced primary server you must reconnect all network adapters, disk devices, and dedicated link.
2. If a disk device that contained the NetWare SYS volume and unmirrored boot data failed, you must reinstall using standard DOS and NetWare installation tools. If you have a backup use that.
3. Start the failed server as the standby server.
4. If a disk device failed and was replaced, use NetWare's configuration utility program to create a new partition and remirror data from the new primary server to the new standby machine. If the failure was caused by some other component, not a disk device, then remirroring will automatically occur.
5. Check the mirror status of the disk devices. Once remirroring is completed, high availability will be restored.

If you want to revert the roles of the servers to their original state:

6. Down both servers.
7. Start the original standby machine as a standby server.

8. Start the original primary server as a NetWare server.
9. Check the mirror status of the disk devices and the StandbyServer user interface to ensure proper functionality.

7.4.2.4 Lotus Domino

You need to reinstall Lotus Domino only if a disk device fails and you have no backup. Domino can be backed up using standard backup procedures. With a good backup, a straightforward restore will get you up and running again. Do not forget notes.ini; you must make sure you restore this important file. The location of notes.ini file depends on your server's operating system.

All clustered databases are available on the other Domino servers in the Domino cluster. You could create new replicas and start replication to reinstate them. We recommend, however, that you back up database files and restore them from the backup. If you have a lot of data in the clustered databases, you will create a lot of traffic if they have to be restored through replication.

7.4.2.5 OPS and DB2 databases

Oracle Parallel Server and DB2 UDB have their own backup-and-restore systems. These application-level tools access backup hardware, such as a tape drive, through the operating system.

Database backup tools are not designed to back up the complete system, only their own data and databases. Other data and operating system files have to be backed up using standard techniques. You can find more about operating system backup earlier in this chapter and 5.6, "Backup and restore" on page 219.

Database backup and recovery key features

- Improves database administrator productivity by handling the database backup and recovery operations.
- Allows backing up the entire database, or a subset of the database, in one operation.
- Minimizes the possibility of operator error and detects database corruption.
- Minimizes space needed for backup creation by supporting incremental backups.
- Minimizes time needed for backup-and-restore operations by performing automatic parallel execution of backups and restores.
- Allows backups when the database is open or closed.

- Provides greater flexibility in recovery operations by allowing restore and recovery of a data file using a mixture of incremental backups and application of archived redo logs.
- Allows a database to be recovered to a point in time.
- Optionally can perform full block-checking when performing a backup, and can add a checksum to a backup.
- Can back up disk-archived logs to tape, and can restore tape-archived logs back to disk.

7.4.2.6 Oracle Parallel Server

The major components of the OPS Backup and Recovery subsystem include:

- Recovery Manager
- Oracle8 Server
- Recovery Catalog
- Graphical management software (optional) with Oracle Enterprise Manager or a third-party product
- Third-party media management software (required for tape)

Automated process of backup and recovery

The Oracle8 backup-and-restore feature is fully managed by the server. This greatly minimizes the likelihood of database administrator error in the backup or recovery steps, and frees the administrator from a substantial administrative task. The Recovery Manager utility manages the processes of creating backups and restoring or recovering from them. All of this work is done inside Oracle8.

Restoring the database, or part of it, for recovery, is very straightforward in Oracle8 because Recovery Manager finds the appropriate backups and restores them as needed. Recovery Manager automatically restores any archive logs needed for recovery as well.

Management of the backup process

Recovery Manager is a component of the backup and recovery system that manages creating backups and restoring or recovering from them. Recovery Manager maintains a recovery catalog that holds information about the backup files and archived redo log files, thereby freeing the administrator from having to track all the backup copies and archive logs. It uses the recovery catalog to:

- Automate both restore and recovery operations
- Perform automatic parallel execution of backups and restores
- Generate a printable report of all backup and recovery activity

Comprehensive recovery catalog
Reports can be run against the recovery catalog to determine which files need backing up, and which backups are no longer required.

The recovery catalog contains:

- Information about data file and archive-log backup sets.
- Information about data file copies.
- Information about archived redo logs and backup copies of them.
- Information about the tablespaces and data files of the target database.
- Named user-created sequences of commands called stored scripts.

7.4.2.7 DB2 UDB

DB2 UDB provides the Backup, Restore and Roll forward utilities to enable database recoverability. You have two ways for backup in DB2 UDB. You can use the Control Center Backup Database function or SmartGuide.

The first tool is the simplest of these. You can choose the media type and the directories or tape for backup. With the ability to select more than one path you can back up the database to more than one location. A backup image that is split over multiple locations will use the same time stamp.

Remote shared drives

You cannot use a shared drive from a remote machine to store your database backup image and you cannot use a remote shared drive to restore a database image to the local machine.

The second tool is the SmartGuide. The SmartGuide offers control of a complex administrative task in a user-friendly way. The Backup Database SmartGuide helps the database administrator define a database backup plan. The Database Backup SmartGuide has five pages:

- **Database** — On the Database page, you select the database you want to back up.
- **Availability** — On this page, you enter information about how your database is used. Based on this information, the Backup Database SmartGuide will determine what type of backup you should perform and how often your database should be backed up.
- **Protection** — On the Protection page, you select which level of protection you want for your database, fast or complete database recovery.

- **Rate of Change** — On this page, you can indicate how much your database is modified per day. The Rate of Change information determines how often the database should be backed up.
- **Recommendations** — The Recommendations page gives a summary of the Database Backup SmartGuide recommendations. You can review these and override them if necessary.

The Restore Database SmartGuide leads you through a number of steps to recover your database. It has three pages:

- **Restore Status** — You can select the database you want to restore.
- **Backup History** — You select the database backup image you want to use for restore operation.
- **Roll Forward** — Here you can apply the log files to redo all the changes made in the database since database backup was taken.

7.5 Summary

In this chapter, we pulled together some of the information distributed throughout the rest of the book. Clustering adds a new layer of complexity to server operations and has to be approached in a structured way to ensure a smooth implementation. The necessity for careful evaluation, planning, installation, testing, and documentation cannot be overstressed.

Clusters are primarily implemented on Intel-based systems today to provide high availability. It would be a pity for the potential benefits of clustering to be lost because of an implementation error. Careful consideration of the matters we have presented will go a long way toward ensuring you get the return on your investment that you have the right to expect.

Appendix A. Creating an Ethernet crossover cable

For high-speed interconnects between two cluster nodes, 100 Mbps Ethernet makes a good choice. The nodes can be connected without the need for a hub by using a crossover cable, which is simple to make if you have a crimping tool on hand.

Only four wires in the 8-pin RJ-45 connector are used to carry Ethernet signals. Both 10Base-T and 100Base-T use an identical connector so the same crossover cable will work for both. A crossover cable works much like an RS232 null modem cable, connecting one device's Transmit to the other device's Receive pins. Figure 85 shows an RJ-45 connector, seen from below, and describes how to connect the wires:

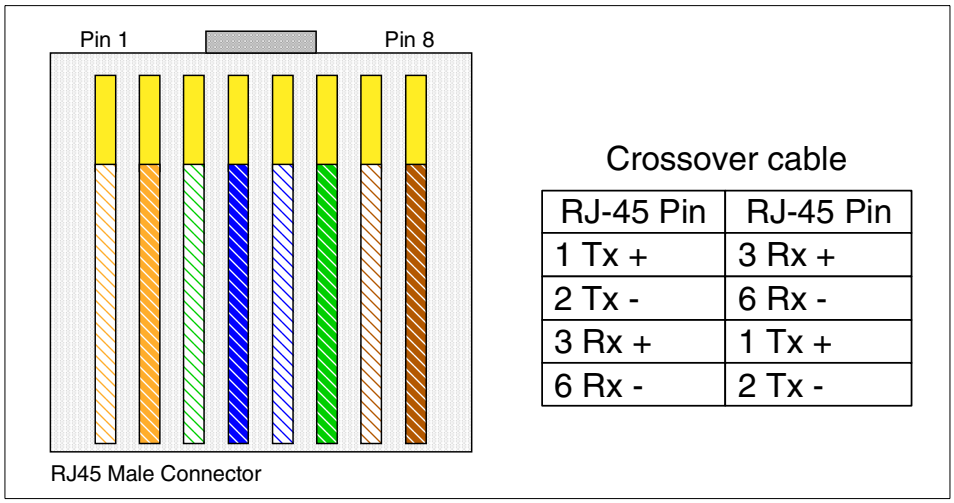

Figure 85. Ethernet 10/100 Base-T crossover cable wiring

The EIA/TIA 568A/568B and AT&T 258A standards define the wiring and allow for two different wire color schemes as shown in Table 31.

Table 31. CAT5 cabling color schemes

Pin	Signal	EIA/TIA 568A	EIA/TIA 568B AT&T 258A	10Base-T 100Base-T
1	Tx +	White/Green	White/Orange	X
2	Tx -	Green/White or Green	Orange/White or Orange	X
3	Rx +	White/Orange	White/Green	X

Pin	Signal	EIA/TIA 568A	EIA/TIA 568B AT&T 258A	10Base-T 100Base-T
4	N/A	Blue/White or Blue	Blue/White or Blue	Not used
5	N/A	White/Blue	White/Blue	Not used
6	Rx -	Orange/White or Orange	Green/White or Green	X
7	N/A	White/Brown	White/Brown	Not used
8	N/A	Brown/White or Brown	Brown/White or Brown	Not used

Note: Even though pins 4, 5, 7, and 8 are not used, it is mandatory that they be present in the finished cable.

Appendix B. Special notices

This publication is intended to help customers and business partners in determining the most appropriate clustering solutions for implementation in a variety of environments. The information in this publication is not intended as the specification of any programming interfaces that are provided by the software products discussed in the book. See the PUBLICATIONS section of the relevant IBM Programming Announcement for more information about what publications are considered to be product documentation.

References in this publication to IBM products, programs or services do not imply that IBM intends to make these available in all countries in which IBM operates. Any reference to an IBM product, program, or service is not intended to state or imply that only IBM's product, program, or service may be used. Any functionally equivalent program that does not infringe any of IBM's intellectual property rights may be used instead of the IBM product, program or service.

Information in this book was developed in conjunction with use of the equipment specified, and is limited in application to those specific hardware and software products and levels.

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.

Licensees of this program who wish to have information about it for the purpose of enabling: (i) the exchange of information between independently created programs and other programs (including this one) and (ii) the mutual use of the information which has been exchanged, should contact IBM Corporation, Dept. 600A, Mail Drop 1329, Somers, NY 10589 USA.

Such information may be available, subject to appropriate terms and conditions, including in some cases, payment of a fee.

The information contained in this document has not been submitted to any formal IBM test and is distributed AS IS. The use of this information or the implementation of any of these techniques is a customer responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. While each item may have been reviewed by IBM for accuracy in a specific situation, there is no guarantee that the same or similar results will be obtained elsewhere. Customers

attempting to adapt these techniques to their own environments do so at their own risk.

Any pointers in this publication to external Web sites are provided for convenience only and do not in any manner serve as an endorsement of these Web sites.

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

AS/400
Chipkill
ClusterProven
DB2
DB2 Universal Database
Domino
ESCON
EtherJet
e(logo)@
HelpCenter
IBM ®
Lotus
Lotus Notes
Magstar
Netfinity
Netfinity Manager
Notes
Nways
OS/2
OS/390
OS/400
OnForever
Parallel Sysplex
Predictive Failure Analysis
Redbooks
Redbooks Logo 
RS/6000
S/390
ServeRAID
ServerProven
SmoothStart
StorWatch
System/390
TechConnect
Tivoli
Wake on LAN

The following terms are trademarks of other companies:

Tivoli, Manage. Anything. Anywhere.,The Power To Manage., Anything. Anywhere.,TME, NetView, Cross-Site, Tivoli Ready, Tivoli Certified, Planet Tivoli, and Tivoli Enterprise are trademarks or registered trademarks of Tivoli Systems Inc., an IBM company, in the United States, other countries, or both. In Denmark, Tivoli is a trademark licensed from Kjøbenhavns Sommer - Tivoli A/S.

C-bus is a trademark of Corollary, Inc. in the United States and/or other countries.

Java and all Java-based trademarks and logos are trademarks or registered trademarks of Sun Microsystems, Inc. in the United States and/or other countries.

Microsoft, Windows, Windows NT, and the Windows logo are trademarks of

Microsoft Corporation in the United States and/or other countries.

PC Direct is a trademark of Ziff Communications Company in the United States and/or other countries and is used by IBM Corporation under license.

ActionMedia, LANDesk, MMX, Pentium and ProShare are trademarks of Intel Corporation in the United States and/or other countries.

UNIX is a registered trademark in the United States and other countries licensed exclusively through The Open Group.

SET, SET Secure Electronic Transaction, and the SET Logo are trademarks owned by SET Secure Electronic Transaction LLC.

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

Appendix C. Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this redbook.

C.1 IBM Redbooks

For information on ordering these publications see “How to get IBM Redbooks” on page 279.

- *TCP/IP Tutorial and Technical Overview*, GG24-3376
- *Netfinity Server Disk Subsystems*, SG24-2098
- *Lotus Domino R5 Clustering with IBM @server xSeries and Netfinity Servers*, SG24-5141
- *Netfinity Tape Solutions*, SG24-5218
- *Netfinity and Domino R5.0 Integration Guide*, SG24-5313
- *Oracle Parallel Server and Windows 2000 Advanced Server on IBM Netfinity*, SG24-5449
- *Using TSM in a Clustered Windows NT Environment*, SG24-5742

C.2 IBM Redbooks collections

Redbooks are also available on the following CD-ROMs. Click the CD-ROMs button at **ibm.com**/redbooks for information about all the CD-ROMs offered, updates and formats.

CD-ROM Title	Collection Kit Number
IBM System/390 Redbooks Collection	SK2T-2177
IBM Networking Redbooks Collection	SK2T-6022
IBM Transaction Processing and Data Management Redbooks Collection	SK2T-8038
IBM Lotus Redbooks Collection	SK2T-8039
Tivoli Redbooks Collection	SK2T-8044
IBM AS/400 Redbooks Collection	SK2T-2849
IBM Netfinity Hardware and Software Redbooks Collection	SK2T-8046
IBM RS/6000 Redbooks Collection	SK2T-8043
IBM Application Development Redbooks Collection	SK2T-8037
IBM Enterprise Storage and Systems Management Solutions	SK3T-3694

C.3 Other resources

These publications are also relevant as further information sources:

- *The Universal Guide to DB2 for Windows NT*, SC09-2800
- *Microsoft Cluster Server Administration Guide* (shipped with product)
- *Novell High Availability Server for NetWare User Guide* (shipped with product)

C.4 Referenced Web sites

These Web sites are also relevant as further information sources:

- `http://linux-ha.org` — High-Availability Linux Project
- `http://people.redhat.com/kbarrett/HA` — Red Hat High Availability Server Project
- `http://www-1.ibm.com/services/its/us/spsspcclstr.html` — IBM SmoothStart Services
- `http://www-4.ibm.com/cgi-bin/db2www/data/db2/udb/winos2unix/support/v7pubs.d2w/en_admin` — IBM DB2 v7.x publications
- `http://www.beowulf.org` — Beowulf Project
- `http://www.citrix.com` — Citrix home page
- `http://www.globalfilesystem.org` — Linux Global File System (GFS)
- `http://www.hp.com/tape/papers/strategy.html` — Hewlett Packard white paper: *A Recommended Backup Strategy*
- `http://www.hursley.ibm.com/~ssa` — IBM SSA Web site
- `http://www.ibm.com/servers/clusters` — IBM ClusterProven
- `http://www.ibm.com/storage/SANGateway` — Product information on the IBM Storage Area Network Data Gateway and Router family
- `http://www.ibm.com/storage/fcss` — Product information on the IBM Fibre Channel RAID Storage Server
- `http://www.inter-mezzo.org` — InterMezzo distributed file system (DFS) for Linux
- `http://www.legato.com/products` — Product information for Legato products include Co-StandbyServer
- `http://www.legato.com/products/availability/costandbyserver/nt/ntco_tools.cfm` — Application Scripts for Legato Co-StandbyServer for NT

- http://www.linuxvirtualserver.org — Linux Virtual Server (LVS) Home Page
- http://www.marathontechnologies.com/productinfo/Endurance_6200.html — Product information on the Marathon Endurance 6200
- http://www.mcdata.com/efcdirector — McDATA ED-5000 Director information
- http://www.microsoft.com/TechNet/win2000/understa.asp — Understanding Active Directory Replication, Chapter 14 from *Creating Active Directory Infrastructures by Curt Simmons*, published by Prentice Hall, PTR
- http://www.microsoft.com/windows2000/guide/datacenter/overview/default.asp — Windows 2000 Datacenter Server product overview
- http://www.microsoft.com/windows2000/guide/server/features — Windows 2000 Server family
- http://www.mosix.org — MOSIX home page
- http://www.notesbench.org — NotesBench Consortium
- http://www.novell.com/documentation/lg/nw5/docui/index.html#../usfile/nss__enu/data/hcf8v0n5.html — Novell Storage Services architecture
- http://www.oracle.com/database/options/parallel/certification/oracle8i_ibm.html — OPS 8i certified solutions from IBM
- http://www.pc.ibm.com/support — IBM Personal Systems Group support
- http://www.pc.ibm.com/support?lang=en_US&page=help&subpage=online_assistant&brand=IBM+PC+Server&doctype=&subtype=Cat — Online Assistant from the IBM HelpCenter
- http://www.pc.ibm.com/techconnect/index.html — IBM TechConnect
- http://www.pc.ibm.com/training/pci_schedule_ww_techconnect.html — IBM PC Institute Schedule for TechConnect courses
- http://www.pc.ibm.com/us/compat/nos/matrix.shtml — @server Proven xSeries Operating System compatibility
- http://www.pc.ibm.com/us/compat/serverproven/giganet.shtml — Giganet in the ServerProven program
- http://www.pc.ibm.com/us/compat/serverproven/index.htm — @server Proven xSeries participants
- http://www.pc.ibm.com/us/eserver/xseries/xarchitecture.html — IBM X-architecture

- http://www.pc.ibm.com/us/netfinity/clusterproven.html — IBM ClusterProven
- http://www.pc.ibm.com/us/netfinity/parallel_server.html — Netfinity Cluster Enabler for Oracle Parallel Server
- http://www.pc.ibm.com/us/netfinity/tech_library.html — xSeries technical library
- http://www.pc.ibm.com/us/products/server/download.html — xSeries Configuration Tools
- http://www.pc.ibm.com/us/techlink/wtpapers — xSeries White Papers
- http://www.pc.ibm.com/ww/netfinity/999guarantee.html — 99.9% Availability Guarantee Program
- http://www.pc.ibm.com/ww/netfinity/clustering/matrix.html — Netfinity clustering compatibility matrix
- http://www.pc.ibm.com/ww/netfinity/clustering/mscs.html — IBM Netfinity Availability Extensions for Microsoft Cluster Service
- http://www.pc.ibm.com/ww/netfinity/systems_management/nfdir/serverext.html — Netfinity Director UM Server Extensions
- http://www.storage.ibm.com/ibmsan/whitepaper.htm — IBM SAN White Papers
- http://www.t11.org — Fibre Channel specifications

C.5 Microsoft Knowledge Base articles

The following articles are referenced in or are related to this book. They can be found by searching on the article number at:

http://support.microsoft.com/search

Cluster Service

Q169414 Cluster Service May Stop After Failover
Q171277 Microsoft Cluster Server Cluster Resource Failover Time
Q171792 Using Microsoft Cluster Server to Create a Virtual Server
Q174070 Registry Replication in Microsoft Cluster Server
Q197047 Failover/Failback Policies on Microsoft Cluster Server
Q224999 How to Use the Cluster TMP file to Replace a Damaged Clusdb File
Q228904 Print Spooler Support on Microsoft Windows 2000 Server Cluster
Q229733 During Cluster Failover Print Queue Monitoring May Stop

Q238137 Windows 2000 Support for Clustered Network Shares
Q244700 “Device Is Not Ready” Error Message When Creating a Server Cluster MSMQ Resource
Q245762 Recovering from a Lost or Corrupted Quorum Log
Q247392 MSCS/Cluster Node Is Not Attached to All Cluster Defined Networks
Q247709 Cluster.exe Create Command Resource Name Must Match Resource Type Name Not Display Name
Q248998 How to Properly Restore Cluster Information
Q249194 MSCS/Cluster Does Not Form with Error Messages 170 and 5086
Q254287 BUG: When Microsoft Message Queuing Is Configured for Workgroup on a Windows 2000 Cluster, the Resource Fails to Come Online
Q254360 XADM: Content Indexing, Free/Busy Publishing Do Not Work on a Clustered Server
Q254651 Cluster Network Role Changes Automatically
Q259243 How to Set the Startup Value for a Resource on a Clustered Server
Q266274 How to Troubleshoot Cluster Service Startup Issues

Load Balancing

Q197862 WLBS Cluster Is Unreachable from Outside Networks
Q232190 Description of Network Load Balancing Features
Q242242 Using the “WLBS QUERY” Command to Determine the State of an WLBS/NLB Cluster
Q247297 Network Load Balancing Connection to a Virtual IP Address Not Made Across a Switch
Q248346 L2TP Sessions Lost When Adding a Server to an NLB Cluster
Q254110 Network Load Balancing May Stop in Unicast Mode with Some FDDI Network Adapters
Q256910 IP Address Assignment for NLB with Multiple Network Adapters
Q264645 IP Address Conflict Switching Between Unicast and Multicast NLB Cluster Mode

Networking

Q168567 Clustering Information on IP Address Failover
Q176320 Impact of Network Adapter Failure in a Cluster
Q193890 Recommend WINS Configuration for Microsoft Cluster Server
Q220819 How to Configure DFS Root on a Windows 2000 Server Cluster
Q224508 How to Migrate a DFS Root Configuration to a Windows 2000 Cluster
Q230356 Changing the IP Address of Network Adapters in Cluster Server

Q232478 Cluster Server DFS Root Resource Does Not Pass Status Check
Q241828 Changing the IP Address of a Cluster Adapter May Result in Failover
Q242600 Network Failure Detection and Recovery in a Two-Node Server Cluster
Q244331 MAC Address Changes for Virtual Server During a Failover with Clustering
Q252764 Cluster Service Generates Additional DNS Traffic
Q257577 gethostbyname() Does Not Return Cluster Virtual IP Addresses Consistently
Q259593 WINS Resource Does Not Come Online After Cluster Failover

Administration

Q228480 Error 1722 When Starting Cluster Administrator
Q262705 Cluster Administrator Incorrectly Rejects Trusted Domain SIDs as Local SIDs

Hardware

Q198513 Clustering Cannot Determine If a Shared Disk Is Working Properly
Q224075 Disk Replacement for Windows 2000 Server Cluster
Q257503 Clusdisk Hides Disks Without Signatures
Q258750 Recommended Private “Heartbeat” Configuration on a Cluster Server

General Clustering

Q235529 MSCS Virtual Server Limitations in a Windows 2000 Domain Environment
Q236328 Netdiag Resource Kit Tool Is Not Cluster Aware
Q242430 Network Adapter Not Available During Cluster Configuration with a Crossover Cable
Q243195 Event ID 1034 for MSCS Shared Disk After Disk Replacement
Q247720 Changing Server Status on a Server Cluster Node Affects Security Permissions
Q248998 How to properly restore cluster information
Q250355 Antivirus Software May Cause Problems with Cluster Services
Q251434 Rolling Upgrade of Cluster Nodes from Beta to Retail Not Supported
Q256926 Implementing Home Folders on a Server Cluster
Q257309 XADM: Message Tracking for Two Exchange 2000 Server Virtual Machines on a Cluster Are Written to the Same Log

How to get IBM Redbooks

This section explains how both customers and IBM employees can find out about IBM Redbooks, redpieces, and CD-ROMs. A form for ordering books and CD-ROMs by fax or e-mail is also provided.

- **Redbooks Web Site** ibm.com/redbooks

  Search for, view, download, or order hardcopy/CD-ROM Redbooks from the Redbooks Web site. Also read redpieces and download additional materials (code samples or diskette/CD-ROM images) from this Redbooks site.

  Redpieces are Redbooks in progress; not all Redbooks become redpieces and sometimes just a few chapters will be published this way. The intent is to get the information out much quicker than the formal publishing process allows.

- **E-mail Orders**

  Send orders by e-mail including information from the IBM Redbooks fax order form to:

	e-mail address
In United States or Canada	pubscan@us.ibm.com
Outside North America	Contact information is in the “How to Order” section at this site: http://www.elink.ibmlink.ibm.com/pbl/pbl

- **Telephone Orders**

United States (toll free)	1-800-879-2755
Canada (toll free)	1-800-IBM-4YOU
Outside North America	Country coordinator phone number is in the “How to Order” section at this site: http://www.elink.ibmlink.ibm.com/pbl/pbl

- **Fax Orders**

United States (toll free)	1-800-445-9269
Canada	1-403-267-4455
Outside North America	Fax phone number is in the “How to Order” section at this site: http://www.elink.ibmlink.ibm.com/pbl/pbl

This information was current at the time of publication, but is continually subject to change. The latest information may be found at the Redbooks Web site.

IBM Intranet for Employees

IBM employees may register for information on workshops, residencies, and Redbooks by accessing the IBM Intranet Web site at http://w3.itso.ibm.com/ and clicking the ITSO Mailing List button. Look in the Materials repository for workshops, presentations, papers, and Web pages developed and written by the ITSO technical professionals; click the Additional Materials button. Employees may access MyNews at http://w3.ibm.com/ for redbook, residency, and workshop announcements.

IBM Redbooks fax order form

Please send me the following:

Title	Order Number	Quantity

First name	Last name	
Company		
Address		
City	Postal code	Country
Telephone number	Telefax number	VAT number

☐ Invoice to customer number _____

☐ Credit card number _____

Credit card expiration date	Card issued to	Signature

We accept American Express, Diners, Eurocard, Master Card, and Visa. Payment by credit card not available in all countries. Signature mandatory for credit card payment.

Abbreviations and acronyms

AFT	adapter fault tolerance
AGP	advanced graphics port
ALB	adapter load balancing
AMD	Advanced Micro Devices
ANSI	American National Standards Institute
APC	American Power Conversion
API	application programming interface
ARP	address resolution protocol
ATC	advanced transfer cache
BDC	backup domain controller
BIOS	basic input/output system
CD-ROM	compact disk read only memory
CE	compute element
CM	cluster manager
CPU	central processing unit
CRC	cyclic redundancy check
DBA	database administrator
DBCA	Database Configuration Assistant
DHCP	Dynamic Host Configuration Protocol
DIMM	direct inline memory module
DLL	dynamic link library
DLT	digital linear tape
DNS	domain name system
DOS	disk operating system
ECC	Error Checking and Correcting
EE	Enterprise Edition
EEE	Enterprise Extended Edition
EIA	Electronics Industries Association
EIDE	enhanced integrated drive electronics
EMEA	Europe/Middle East/Africa
EPROM	electrically programmable read only memory
ERP	enterprise resource planning
ESCON	enterprise systems connection
ESM	Enclosure Services Monitor
FC-AL	Fibre Channel Arbitrated Loop
FDDI	Fiber Distributed Data Interface
FEC	Fast EtherChannel
FOS	Fail Over Service
FSB	front-side bus
GBIC	GigaBit Interface Converter
GFS	Global File System
GUI	graphical user interface
HA	high availability
HACMP	High Availability Cluster Multi-Processing
HCL	hardware compatibility list
HDD	hard disk drive
HH	half high
HPC	high performance computing
HTML	HyperText Markup Language
HTTP	Hypertext Transfer Protocol
HTTPS	secure form of HTTP
I/O	input/output
I20	Intelligent Input/Output
IBM	International Business Machines
ICM	Internet Cluster Manager
ICSM	IBM Cluster Systems Management
ID	identifier
IDE	integrated drive electonics
IE	Internet Explorer
IEEE	Institute of Electrical and Electronics Engineers

IIOP	Internet Inter-ORB Protocol
IIS	Microsoft's Internet Information Server
IMAP	Internet Mail Access Protocol
IOP	I/O processor
IP	Internet Protocol
IPX	Internet Packet Exchange
ISA	industry standard architecture
ISDN	Integrated Services Digital Network
ITSO	International Technical Support Organization
LAN	local area network
LDAP	Lightweight Directory Access Protocol
LDM	logical drive migration
LED	light emitting diode
LUN	logical unit
LVDS	low-voltage differential SCSI
LVS	Linux Virtual Server
MAC	medium access control
MMC	Microsoft Management Console
MOSIX	Multicomputer OS for UNIX
MP	multiprocessor
MPI	message passing interface
MPP	massively parallel processing
MSCS	Microsoft Cluster Server
MSDTC	Microsoft Distributed Transaction Coordinator
MSMQ	Microsoft Message Queue Server
MTS	Multithreaded Server
NAE	Netfinity Availability Extensions for MSCS
NAT	Network Address Translation
NCF	NetWare command file
NCITS	National Committee for Information Technology Standards
NCS	NetWare Cluster Services
NDIS	network driver interface specification
NDS	Novell Directory Services
NFS	Network File System
NIC	network interface card
NLB	network load balancing
NLM	NetWare loadable module
NSS	NetWare Storage Services
NTFS	NT File System
NTLM	NT Lan Manager
NVSRAM	non-volatile static random access memory
ODI	open data-link interface
OFS	Oracle Fail Safe
OLTP	online transaction processing
OPS	Oracle Parallel Server
OS	operating system
OSD	Operating System Dependent
PCI	peripheral component interconnect
PDC	primary domain controller
PFA	predictive failure analysis
PHP	Hypertext Preprocessor
POST	power on self test
PVM	parallel virtual machine
PXE	Preboot eXecution Environment
RAID	redundant array of independent disks
RAM	random access memory
RDAC	Redundant Dual ActiveController
RISC	reduced instruction set computer
RPM	revolutions per minute
RSM	Remote System Manager
SAN	storage area network
SCO	Santa Cruz Operation, Inc.

SCSI	small computer system interface
SDG	Server Design Guide
SDK	software development kit
SDRAM	static dynamic random access memory
SFT	system fault tolerant
SGI	Silicon Graphics Inc.
SIC	serial interface chip
SL	slim line
SLIC	subscriber-line interface circuit
SMP	symmetric multiprocessing
SMS	Systems Management Server
SMTP	simple mail transfer protocol
SNA	system network architeture
SP	Scalable POWERparallel
SQL	structured query language
SSA	serial storage architecture
TAF	Transparent Application Failover
TCP/IP	Transmission Control Protocol/Internet Protocol
TIA	Telecommunications Industries Association
TSM	Tivoli Storage Manager
UDB	Universal Database
UM	Universal Managability
UPS	uninterruptible power supply
URL	Uniform Resource Locator
VI	Virtual Interconnect
VIPX	Virtual IPX
VLAN	Virtual Local Area Network
WAN	wide area network
WINS	Windows Internet Naming Service

Index

D

E

F

G

H

I

K

L

M

N

O

P

Q

R

S

T

U

V

W

X

IBM Redbooks review

Your feedback is valued by the Redbook authors. In particular we are interested in situations where a Redbook "made the difference" in a task or problem you encountered. Using one of the following methods, **please review the Redbook, addressing value, subject matter, structure, depth and quality as appropriate.**

- Use the online **Contact us** review redbook form found at **ibm.com**/redbooks
- Fax this form to: USA International Access Code + 1 845 432 8264
- Send your comments in an Internet note to redbook@us.ibm.com

Document Number **Redbook Title**	SG24-5845-01 IBM @server xSeries Clustering Planning Guide
Review	
What other subjects would you like to see IBM Redbooks address?	
Please rate your overall satisfaction:	O Very Good O Good O Average O Poor
Please identify yourself as belonging to one of the following groups:	O Customer O Business Partner O Solution Developer O IBM, Lotus or Tivoli Employee O None of the above
Your email address: The data you provide here may be used to provide you with information from IBM or our business partners about our products, services or activities.	O Please do not use the information collected here for future marketing or promotional contacts or other communications beyond the scope of this transaction.
Questions about IBM's privacy policy?	The following link explains how we protect your personal information. **ibm.com**/privacy/yourprivacy/

IBM @server xSeries Clustering Planning Guide

(0.5” spine)
0.475”<->0.875”
250 <-> 459 pages

IBM @server xSeries Clustering Planning Guide

Redbooks

Describes the xSeries and Netfinity clustering hardware

Covers Windows, Linux and NetWare clustering

Helps you plan a cluster implementation

This redbook will help you to implement practical clustered systems using the IBM @server xSeries and Netfinity families of Intel-based servers. Clustering as a technology has been used for many years but has only recently started to become popular on Intel-based servers.

This is due, at least in part, to the inclusion of clustering technology in Windows NT 4.0 Enterprise Edition and now Windows 2000 Advanced Server and Windows 2000 Datacenter Server operating systems. After discussing the reasons why clustering has become such an important topic, we move on to specific clustering technologies.

Clustering comes in many forms. It can appear natively within an operating system, or it can be offered as an add-on product by either the operating system vendor itself or by third-party vendors. A third variation is clustering that is built into specific applications. Generally, these have been complex products such as relational databases.

By describing the characteristics and benefits of each approach, the book will help you to make informed decisions about the clustering solutions most appropriate for your applications.

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks

SG24-5845-01 ISBN 0738419397