



IBM Netfinity RAID Technology

Reliability through RAID technology

Not long ago, business-critical computing on industry-standard platforms was unheard of. Proprietary systems were simply the only systems capable of providing the power and reliability required for deploying applications critical to a company's operations. IBM's Netfinity® server technology allows companies to use advanced, industry-standard technology to build a reliable foundation for business-critical computing. One technology that has made this possible is Redundant Array of Independent Disks (RAID) technology.

RAID is a blueprint for combining two or more disks to create an array of disks. Through hardware or software functionality, multiple physical disks are treated as one logical disk. Data is stored redundantly in various ways to enhance integrity and availability. And because RAID adapters can significantly improve data transfer rates, this technology is extremely effective when implementing demanding, transaction-oriented applications. High-performance interface technology and powerful, dedicated processors allow RAID adapters to maximize throughput, plus tune an adapter's performance to meet the needs of specific applications.

RAID subsystems are commonly used as the cost-effective foundation of a business-critical storage strategy. By employing the advanced fault tolerance of RAID technology, companies can effectively implement networked business systems that require large amounts of storage space for data and applications that must be available for their businesses to continue operating.

This paper explores the basic functionality of RAID implementations, discusses the numerous RAID options available and looks at how RAID technology is poised to take advantage of new storage technologies as they emerge. Finally, the paper provides an overview of RAID technology offered by IBM Netfinity servers.

Blueprint for Data Protection

In a non-RAID environment, users with single-drive computers have no means of protecting their data and no way to create redundant data. Data protection and redundancy might not seem important to many people, or they simply might not have thought of the need for them—until their hard drive fails and all data is lost, with no means of recovery. At the very least, such users should back up their data every day, perhaps saving it to a diskette. Another alternative is to purchase a tape backup unit and use that to save data. But there is a better way, one that can help save data in the event of drive failure, provide redundancy and enable recovery while the system is running. That way is RAID technology, which has been developed over the last 12 years.

RAID technology provides several different ways to use multiple disks to increase availability and performance. A number of RAID specifications (or *levels*) have been defined, each offering unique capabilities in the areas of throughput and fault tolerance. All levels (except RAID 0) provide fault tolerance: if a single disk in the array fails, access to *all* the data stored on the array is still available. The failed disk can be replaced while the array is in use. In addition, all levels (except RAID 0) provide fault tolerance through a hot-spare drive that automatically replaces a failed drive in order to maintain redundancy.

A RAID can be controlled by specialized software or by a RAID adapter that uses a dedicated array processor to offload RAID functions from the CPU. Hardware-based arrays usually provide better performance than software-based arrays.

RAID 0

RAID 0 begins with a concept called *drive spanning*. Drive spanning allows multiple physical disk drives to be logically concatenated into a single logical disk drive. The capacity of the logical drive created via spanning is the capacity of the physical drives times the number of physical drives.

RAID 0 then uses a technique called *data striping* to distribute data evenly across the physical drives in such a manner as to maximize I/O performance. Striping divides the logical drive into data blocks called *stripes*, which are then distributed over the physical disk drives. The layout is such that a sequential read of data on the logical drive results in parallel reads to each of the physical drives. This results in improved performance since multiple drives are operating simultaneously.

RAID 0 does not provide fault tolerance. If one disk fails, all data is lost and all disks must be reformatted.

There are two key advantages to RAID 0: It provides a large logical disk drive through drive spanning; and it provides performance acceleration through data striping.

The disadvantage of RAID 0 is that it provides no redundancy.

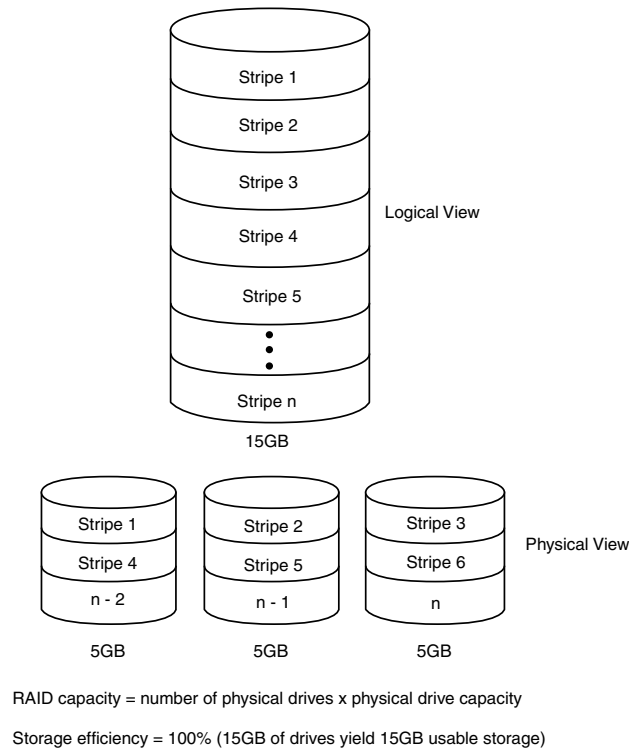


Figure 1. RAID 0: One file broken into multiple groups of sectors and striped across multiple disks

The major uses of RAID 0 are in situations where no redundancy is required or where redundancy can be provided through the use of transaction logs, which make it possible to recreate data from the last status recorded in the log.

As noted, however, with no built-in redundancy and no data protection, RAID 0 does not support a hot-spare drive; if one drive fails, all data is lost. For comparison purposes, RAID 0 is regarded as the baseline against which to measure the performance of the other RAID levels.

RAID 1

RAID 1 employs the concept of *data mirroring*. Mirroring creates a single logical disk drive from two physical disk drives. All data written to the logical drive is written to BOTH physical disk drives, thus creating a pair of drives containing exactly the same data. In the event of failure of one of the physical disk drives, the data will be available via the remaining disk drive. A hot-spare drive can then be used to reestablish the mirror relationship and the redundancy that results while the failing drive is being replaced.

The read performance of RAID 1 is superior to RAID 0 or RAID 5 because read operations are distributed over a pair of drives instead of concentrated on a single drive. Write performance is superior to RAID 5 but worse than RAID 0 since two writes must occur instead of one. The storage efficiency overhead for RAID 1 is 50% because two physical drives are required to get a single drive's capacity. This means that two 5GB drives are required to yield a single 5GB logical drive. This is much worse than RAID 5. In general RAID 1 provides the highest level of performance, but with the poorest storage efficiency.

Dependable protection for your business-critical data

RAID 1 employs an optional hot-spare disk, so that if a disk fails, data performance and redundancy can be recovered while the failed disk is being replaced. The hot-spare disk takes the place of the failed drive in the array. The RAID algorithms regenerate the lost data onto the new drive, thus repairing the array. During the transition period in which the regeneration is taking place, the array is said to be in *critical* or *degraded mode*, which results in slowed performance and the loss of redundancy.

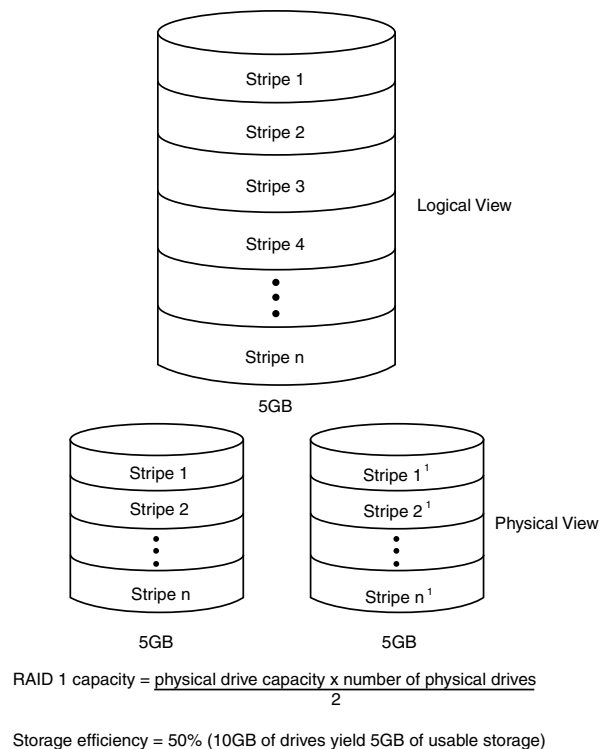


Figure 2. RAID 1: Mirroring

The advantages of RAID 1 are as follows:

- Redundancy through mirrored copy of data
- Read performance superior to RAID 0 and 5
- Write performance superior to RAID 5
- Critical-mode performance superior to RAID 5

The disadvantages of RAID 1 are as follows:

- Write performance worse than RAID 0
- Higher capacity overhead than RAID 5
- Even number of physical disks required

RAID 1 is most useful when performance is more important than capacity and when a configuration is limited to two drives.

RAID 1 Enhanced

RAID 1 Enhanced (also known as *RAID 1E*, *Hybrid RAID 1* or *RAID 6*) combines mirroring with data striping—data is striped across each disk in the array. The first set of stripes are the data stripes, and the second set of stripes are the mirror (copies) of the first data stripe, but shifted one drive.

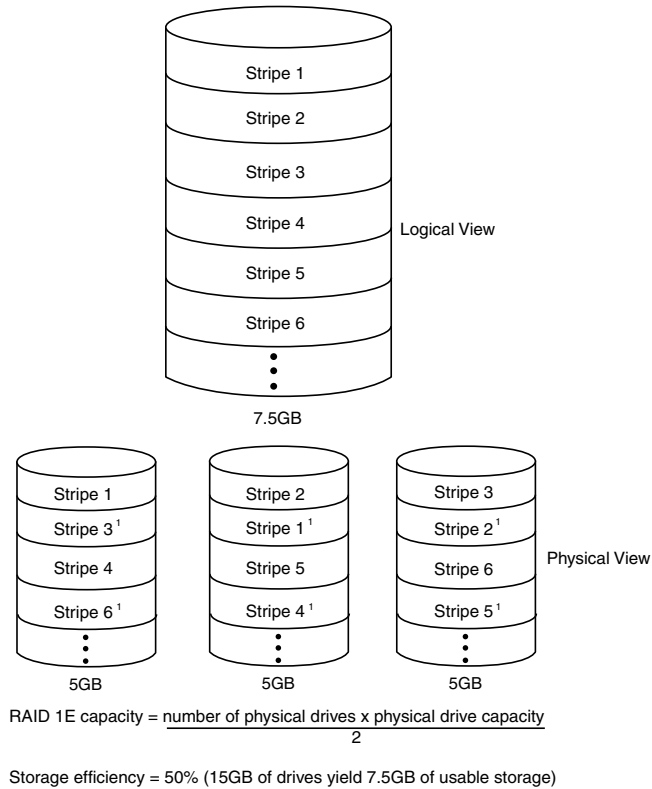


Figure 3. RAID 1E: Stripes and mirrors data across all disks

RAID 1E shares the characteristics of RAID 1 but additionally allows more than two drives, including odd numbers of drives.

RAID 10

RAID 10 combines mirroring with data striping. RAID 10 is supported by Fibre Channel storage subsystems and provides mirroring of two RAID 0s. RAID 10 is also called *RAID 1+0*. In the future, ServeRAID™ will support RAID 10.

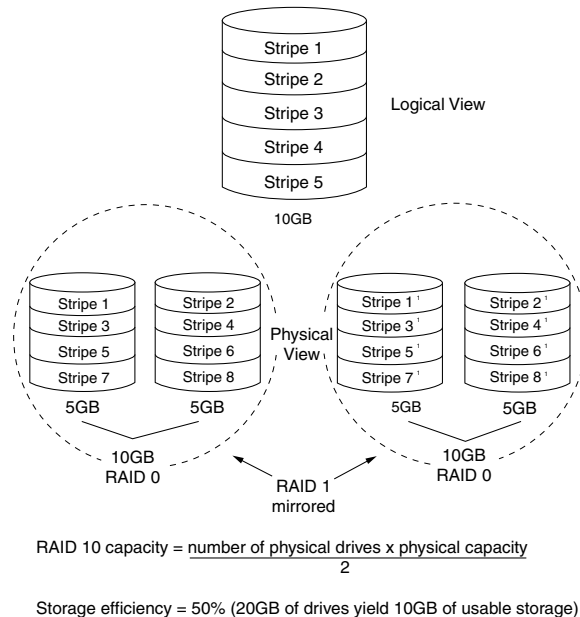


Figure 4. RAID 10: Data mirroring and striping

RAID 3 and Raid 4

RAID 3 stripes data, one group of bits or bytes at a time, across all the data drives. Parity information, used to reconstruct missing data, is stored on a dedicated drive. RAID 3 requires one parity disk and at least two data disks. RAID 3 also requires all drives to be rotationally synchronized.

Use of a parity disk instead of mirroring greatly reduces the amount of additional disk space required for redundancy. However, having a single parity drive results in a performance bottleneck during write operations. RAID 3 is recommended only in workloads that are mostly read oriented or where write performance is not essential.

RAID 3 can provide a performance enhancement in very large block transfers; RAID 3 is often employed when dealing with very large data blocks such as graphics or imaging files. The data protection RAID 3 provides is excellent: If any disk fails, the data can still be accessed by using the information from the other disks and the parity disk to reconstruct it.

However, because RAID 3 was developed for directly controlled, ESDI-like drives, it is not applicable to today's SCSI and Fibre Channel drives. It is therefore considered to be obsolete and is rarely used. Most implementations claiming RAID 3 support are actually simulating RAID 3 by using RAID 5.

Dependable protection for your business-critical data

RAID 4 is similar to RAID 3 but differs in the size of the stripe unit. RAID 4 uses larger stripes to improve the write performance of the array.

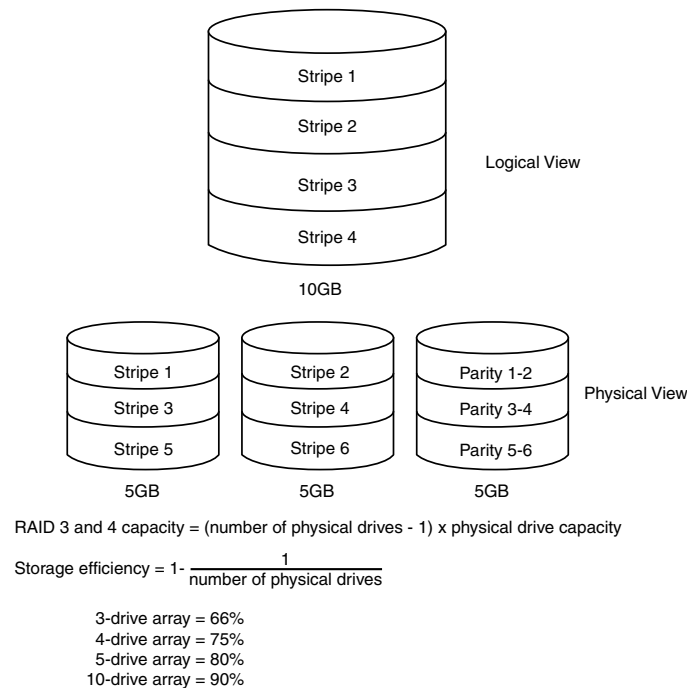


Figure 5. RAID 3 and RAID 4: Data striping with parity disk

As in RAID 3, use of a parity disk instead of mirroring greatly reduces the amount of additional disk space required for redundancy. However, having a single parity drive results in a performance bottleneck during write operations. RAID 4 is recommended only in workloads that are mostly read oriented or when write performance is not essential.

RAID 4 is an older RAID version that shares some characteristics of RAID 3 and, like RAID 3, is rarely used today.

RAID 5

RAID 5 employs data striping and block interleaving in a technique designed to provide fault-tolerant data storage but that does not require duplicate disk drives (as does RAID 1 mirroring). RAID 5 spreads both the data and parity information across the disks one block at a time for maximum read performance when accessing large files and to improve array performance in a transaction processing environment. Redundancy is provided via parity information, which is striped across the drives to remove the bottleneck of storing all of the parity data on one drive.

RAID 5 requires a minimum of three disks—the capacity equivalent of one drive per array is used for the parity data, regardless of the size of array. In performance, RAID 5 is superior to RAID 3 and 4, although the performance boost is, to some extent, limited to smaller block transfers, such as transfers the size of typical network or Internet files. RAID 5 provides data protection—if any disk fails, the data can still be accessed by using the information from the other disks along with the striped parity information.

Dependable protection for your business-critical data

With built-in parity redundancy and much lower overhead in the number of drives needed for redundancy, a three-drive RAID 5 has significantly more capacity than a three-drive RAID 1. However, it performs poorly in critical mode. With these characteristics, RAID 5 is best used when capacity efficiency is most important.

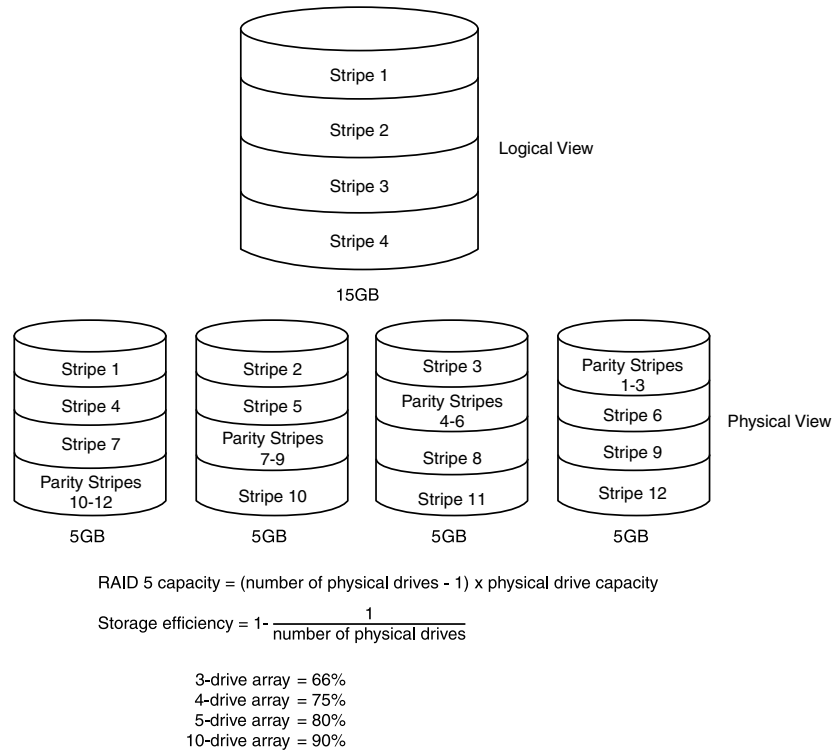


Figure 6. RAID 5: Parity information interleaved with data

The key advantages of RAID 5 are as follows:

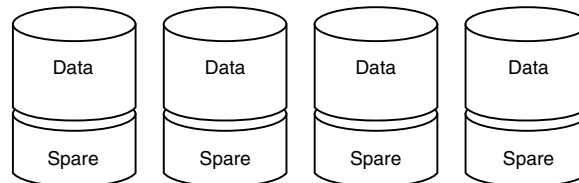
- Redundancy provided through dedicated parity; no action is necessary if a drive fails
- Least capacity overhead in number of drives
- Minimal additional drives necessary to implement redundancy

The disadvantages of RAID 5 are as follows:

- Much worse write performance than RAID 0 and worse write performance than RAID 1
- Read performance equal to RAID 0 but worse than RAID 1
- Poorest critical-mode performance

RAID 5 Enhanced

RAID 5E is an IBM enhancement of RAID 5. Traditional RAID 5 has active drives plus an inactive hot-spare drive. In a RAID 5E array, however, the hot-spare drive is incorporated as an active element in the array; data is laid out in such a way that the spare space is striped across all drives in the array. This technique provides a performance improvement for smaller arrays (three to five drives) with typical data transfer sizes (8kB of data).



- 15-20% throughput improvement by actively using the "spare" heads
- Active use of hot spare allows constant monitoring of the "health" of all drives
- X-architecture innovation from IBM Almaden Research

Figure 7. RAID 5E: Incorporating a hot-spare drive in the array

IBM RAID Technology

RAID technology can be implemented with all IBM Netfinity servers. If your Netfinity server is not RAID enabled, you can purchase IBM RAID hardware to convert your server to be RAID functional. Hardware-based arrays use a dedicated array processor to offload RAID functions from the CPU. The dedicated array processor is designed to provide better performance than software-based arrays. IBM RAID controllers use dedicated, imbedded processors to perform these RAID functions.

By supporting RAID levels 0, 1, 1E, 10, 5 and 5E, IBM gives you more freedom to select the best RAID configuration for your particular needs. What's more, IBM RAID technology is tested to perform with the following operating systems:

- Microsoft® Windows NT® Server 4.0 and higher
- Novell NetWare 3.12 and higher
- SCO UNIXWare Version 7
- OS/2® Warp
- OS/2 Warp Server
- OS/2 SMP
- SCO OpenServer 5.0

In the second half of 1999, IBM will assist customers in installing Linux on Netfinity servers. The specific versions of Linux will be those distributed by Red Hat, SuSE, Caldera and Turbo Linux. IBM intends to work with these Linux distributors to pave the way for co-marketing, development, training and support initiatives that will help customers deploy Linux. Delivery dates depend on the Linux providers.

IBM will also work with independent software vendors to optimize Linux applications on Netfinity servers, to provide customers with powerful and reliable enterprise and e-business solutions on the Linux platform.

IBM ServeRAID 3HB and 3L Ultra2 SCSI Adapters

Adding the IBM ServeRAID 3 Ultra2 SCSI Adapters to your IBM Netfinity server system can help you gain power, performance and convenience—and boost RAID disk performance to enterprise databases or application networks. Improved levels of system throughput are possible with these third-generation RAID adapters that provide twice the bandwidth of Wide Ultra SCSI and an advanced 64-bit performance PCI interface. The ServeRAID 3HB Ultra2 SCSI Adapter supports 32MB of read/write EDO cache standard, with an additional 32MB of battery-backed write cache to mirror the standard EDO cache, for superior data hit rates and improved system performance and data integrity. The IBM ServeRAID 3HB Ultra2 SCSI Adapter takes advantage of the latest Ultra2 SCSI technology, offering you three independent SCSI channels—each capable of supporting up to 15 devices, each at up to 80MBps data throughput—to handle up to 45 drives with a single adapter.

The IBM ServeRAID 3L Ultra2 SCSI Adapter supports 4MB of cache and a single Ultra2 SCSI channel capable of supporting up to 15 devices.

The IBM ServeRAID 3HB and 3L Ultra2 SCSI Adapters provide reliable data protection and remarkable configuration flexibility. The adapters support RAID levels 0, 1, 1E, 5 and 5E, allowing you to choose the combination of power and performance that matches your current needs. The ServeRAID 3HB and 3L Ultra SCSI Adapters both offer Logical Drive Migration (LDM) to let you make crucial changes without shutting down your system. They handle capacity changes to existing RAID arrays in background mode and allow new RAID arrays to be added while your server continues to operate.¹

ServeRAID-3HB adapters can be used in pairs to support Windows NT or Novell NetWare shared-disk clusters. Active PCI hot-add and adapter failover are also supported.

Because the IBM ServeRAID 3HB and 3L Ultra SCSI Adapters store critical RAID configuration information in three different places—in flash ROM, on the nonvolatile RAM on the adapters and in a reserved area on the attached disk—you achieve redundancy for the most crucial portion of your RAID subsystem data. Take the disks from one server and install them in another in any order. A one-step array initialization/synchronization feature reduces the time required to prepare a new RAID array to accept data. And background data scrubbing allows disk media errors to be corrected before they can cause a problem.

You can also tune your adapter's performance to meet the needs of your applications. By setting the stripe size at different levels (8, 16, 32 or 64kB), you control how much data is written to each drive in a given array. So while transaction-based processing, for example, would benefit from a

¹ Different operating systems use different methods for assigning drives and disk space. Some operating systems will require a reboot to access new disk space.

Dependable protection for your business-critical data

32 or 64kB stripe size, file-and-print transactions requiring multiple small transactions would most likely perform better with an 8kB stripe size.

The ServeRAID 3HB Ultra2 SCSI Adapter has two 0.8mm external connectors. These industry-standard, Very High Density Interface (VHDI) connectors allow two external cables to be attached directly to the adapter. In addition, a third channel can be routed out the back of the adapter with a third channel cable option without taking a valuable adapter slot. This design simplifies installation of external data storage enclosures for customers installing external RAID solutions.

In addition, a number of SCSI cable options up to 20m (62ft) in length are available that support external attachment of SCSI devices.

FLASHCOPY. If you are running Windows NT, you can use the FLASHCOPY command to create a quick backup copy of logical drive data. You can use the backup copy for tasks such as tape backup and multi-server rollout (for example, drive cloning). The FLASHCOPY command sets up a link between the source and target logical drives, and then creates a snapshot impression of the source data on the target logical drive. To create the snapshot impression, the size of the target logical drive must be equal to or larger than that of the source logical drive, and the partition size and type must be the same on the source and target logical drives. Once the snapshot impression is created, the target logical drive will contain a copy of the source logical drive data, and the source logical drive can be put back into use.

Logical Drive Migration. Until now, adding drives to an existing configuration meant shutting down your system, backing up your data to tape, adding and/or removing disks, defining a new configuration and, finally, restoring your data. It took a lot of time and, because it involved so many steps, it often meant that you took a lot of chances in the process. Thanks to a technology known as *Logical Drive Migration (LDM)*, the IBM ServeRAID 3HB and 3L Ultra2 SCSI Adapters let you make the changes you need—adding or subtracting drives, redefining arrays, even changing RAID levels—without shutting down the system. It handles capacity changes in background mode while the server continues to operate. You can easily migrate from RAID 0 to RAID 5 (see Figure 7).

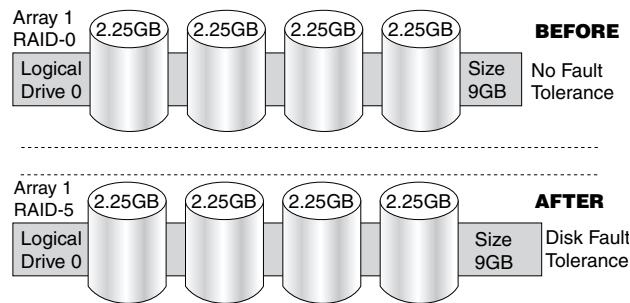


Figure 7. Migrate from RAID 0 to RAID 5

Add one disk to the system, and employ LDM to re-stripe data across all drives, resulting in a RAID 5 configuration that preserves usable storage space, but adds fault tolerance. Finally, LDM lets you expand the capacity of a RAID 5 array by adding up to three new disks at a time—optionally, the original logical drive size can be expanded or remain the same with spare

Dependable protection for your business-critical data

space for additional logical drives (see Figure 8). Drive volumes can also be spanned at operating system level.

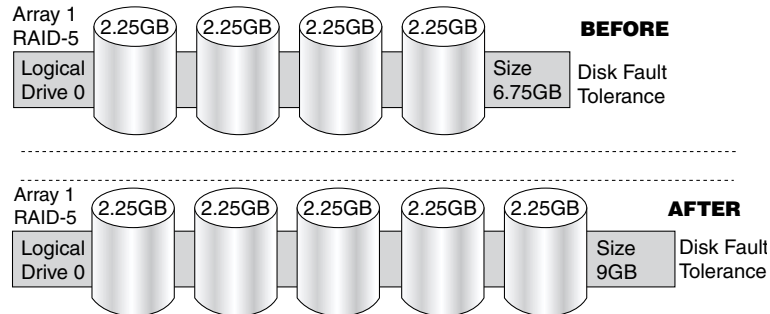


Figure 8. Expand existing RAID arrays

Fibre Channel RAID. Most server-class hard disk drives are attached to servers or RAID controllers via some version of SCSI, SSA or Fibre Channel interfaces. All disk drives use the same basic components such as magnetic disk platters, read/write heads, seek actuators, and most of the electronics needed to control access to data. These basic components are usually the parts that limit the performance of the disks regardless of which interface is used to connect the disks to the server or RAID controller. Rarely is the bandwidth of the disk interface a limiting factor in system-level performance.

The choice of which interface is best suited for a given application depends on a number of factors such as availability of supply, cost, scalability and performance. Of these three interfaces, SCSI disks are used most widely, are the simplest to design, cost the least to implement and are readily available from a number of suppliers. SSA and Fibre Channel disk drives are not as widely used, are generally available from only a single supplier (IBM for SSA and Seagate for Fibre Channel), are more complex in design, and generally have higher power requirements and higher costs than SCSI drives. Even if the disk drives were the same price, the cost of other components that support SSA or Fibre Channel disk drives such as hot-swap disk enclosures and RAID controllers generally are more expensive than their SCSI equivalents.

Even with these known limitations, there are configurations where SSA and Fibre Channel disk drives are preferable to SCSI disk drives. In cases where the number of disks that need to be attached to a system or RAID controller is very large, SSA and Fibre Channel disks can provide significant advantages. Up to 126 disks can be connected to an SSA or Fibre Channel interface, whereas a maximum of 15 disks can be connected to a SCSI interface. SCSI also has the limitation of a maximum of several meters of cabling used to connect the disks to the systems or RAID controllers compared to several thousand meters of optical fiber cabling allowed by SSA and Fibre Channel. SCSI cables are also very bulky, containing 68 wires, and are expensive compared with SSA and Fibre Channel cables, which usually contain 4 to 6 wires or 2 optical fibers. SSA and Fibre Channel disks support dual-interface circuits, which provide fault tolerance for some types of failures. However, SCSI RAID adapters or Fibre Channel-attached RAID subsystems that use SCSI disks attached to multiple SCSI buses can typically be configured to tolerate failures of a SCSI bus as well.

Because of its longer distance support, the capability to support large numbers of devices, and the ability to support switched fabric topologies, Fibre Channel has been selected overwhelmingly by the computer industry as the interface of choice for connecting multiple

servers to storage subsystems. The attachment of multiple storage devices to multiple servers is typically called a Storage Area Network. The use of Fibre Channel for this purpose should be distinguished from the use of Fibre Channel to attach disk drives to a server or RAID controller. For example, the current Netfinity Fibre Channel RAID subsystem uses 6 independent Ultra2 SCSI buses for attaching up to 60 Ultra or Ultra2 SCSI disk drives to a server or cluster of servers. This allows the reuse of existing investments in SCSI disks and enclosures while still getting the benefits of Fibre Channel technology. By using RAID-1 or RAID-5 disk arrays that span across two or more of these SCSI buses, a SCSI bus failure can be eliminated as a single point of failure. This is sometimes referred to as *orthogonal RAID* technology.

For attaching disks drives, the majority of customers' needs today are easily met with SCSI disk technology and with the consistent improvements that have been made to SCSI technology such as Ultra, Ultra2 and most recently the announcement of Ultra3 SCSI; SCSI disks will likely remain the most popular type of disks installed in servers in the foreseeable future. Netfinity servers support a wide variety of disk and RAID adapters and controllers that can use the same SCSI disks from the low end to the high end of the Netfinity product line. This strategy protects the investment in SCSI disks and allows scalability to many TB of capacity in standalone or clustered server applications. IBM intends to continue to support SCSI-disk-based solutions into the foreseeable future, as well as offer newer generations of SSA and Fibre Channel disks in situations where SCSI may become a limiting factor in the total enterprise storage solution.

For information about Netfinity Fibre Channel RAID products, please refer to the "IBM Netfinity Fibre Channel Directions" white paper at www.ibm.com/netfinity.

Conclusion

IBM continues to pioneer new storage and RAID technologies to improve and expand its role in data storage and protection. We intend to provide RAID solutions that take advantage of the latest SCSI technology. In addition, IBM offers Fibre Channel-attached storage subsystems to further improve external data rates and support N-way clustered storage through serial interfacing.

RAID is the technology of grouping several hard disk drives in a computer into an array that you can define as one or more logical drives. Each logical drive appears to the operating system as a single drive. This grouping technique greatly enhances logical-drive capacity and performance beyond the physical limitations of a single hard disk drive.

When you group multiple physical hard disk drives into a logical drive, the ServeRAID controller can transfer data in parallel from the multiple drives in the array. This parallel transfer yields data-transfer rates that are many times higher than with non-arrayed drives. This increased speed makes the system better able to meet the throughput (the amount of work in a given amount of time) or productivity needs of the multiple-user network environment.

The ability to respond to multiple data requests provides not only an impressive increase in throughput, but also a decrease in response time. The combination of parallel transfers and simultaneous responses to multiple requests allows disk arrays to provide a high level of performance in network environments.

Additional Information

For more information on IBM Netfinity direction, products and services, refer to the following white papers, available from our Web site at www.pc.ibm.com/netfinity.

Management

Integrating IBM Netfinity Manager with Microsoft Systems Management Server
Integrating IBM Netfinity Manager with Intel LANDesk Server Manager
IBM Netfinity Manager 5.2
IBM Netfinity Manager Plus for Tivoli Enterprise Overview
IBM Netfinity Advanced Systems Management
IBM Netfinity Advanced Systems Management for Servers
IBM ServerGuide for Netfinity and PC Server Systems

Other Topics

Capacity Planning for Netfinity on Windows Terminal Server
Enterprise Storage Solutions
Fibre Channel Solutions for Enterprise Storage
IBM Chipkill Memory
IBM Netfinity X-architecture
IBM ClusterProven Program on Netfinity
IBM Netfinity Predictive Failure Analysis
IBM Netfinity Cluster Directions
IBM Netfinity Web Server Accelerator
Lotus Domino Clusters Overview
Lotus Domino Clusters Installation Primer
Implementing Microsoft IIS on Netfinity 5500 M10
IBM Netfinity Availability Extensions for Microsoft Cluster Server
IBM Netfinity ESCON Adapter
IBM Netfinity Hot-Plug Solutions
IBM Netfinity Storage Management Solutions Using Tape Subsystems
IBM Netfinity Storage Area Networks
IBM Netfinity 8-Way SMP Directions
IBM Netfinity Server Ultra2 SCSI Directions
IBM Netfinity Server Quality
IBM Netfinity 5000 Server
IBM Netfinity 5500 Server Family
IBM Netfinity 7000 M10 Server
IBM Netfinity 8500R Overview
Achieving Remote Access Using Microsoft Virtual Private Networking
At Your Service...Differentiation beyond technology

Dependable protection for your business-critical data



© International Business Machines Corporation 1999 IBM
Personal Computer Company
Department LO6A
3039 Cornwallis Road Research Triangle Park
NC 27709
Printed in the United States of America

7-99

All rights reserved

For terms and conditions or copies of IBM's limited warranty, call 1 800 772-2227 in the U.S. Limited warranty includes International Warranty Service in those countries where this product is sold by IBM or IBM Business Partners (registration required).

References in this publication to IBM products or services do not imply that IBM intends to make them available in all countries in which IBM operates. IBM reserves the right to change specifications or other product information without notice.

IBM Netfinity systems are assembled in the U.S., Great Britain, Japan, Australia and Brazil and are comprised of U.S. and non-U.S. components.

Are you Year 2000 ready? Visit www.ibm.com/pc/year2000 or call 1 800 426-3395 (and request document number 10020 from our faxback database) for the latest information.

IBM, Netfinity, OS/2, ServeRAID and S/390 are trademarks of International Business Machines Corporation in the United States and/or other countries.

Microsoft, Windows, Windows NT and the Windows logo are trademarks or registered trademarks of Microsoft Corporation.

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