IBM

IBM *@*server xSeries ServeRAID Technology

Reliability through RAID technology

Executive Summary:

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. Now, IBM *@server* xSeries and 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 the technology of grouping several physical drives in a computer into an *array* that you can define as one or more logical drives. Within this technology, data is stored redundantly in various ways to enhance integrity and availability. And because RAID controllers 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 controllers to maximize throughput, plus tune a controller's performance to meet the needs of specific applications. 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 physical drive.

When multiple physical drives are grouped 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 those of nonarrayed drives. This increased speed makes the system better able to meet the *throughput* (the amount of data processed 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 increase in throughput, but also a decrease in response time compared to a nonarrayed drive. The combination of parallel transfers and simultaneous responses to multiple requests allows disk arrays to provide a high level of performance in network environments.

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.

.

Exploring RAID Technology

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.

Stripe-unit size

With RAID technology, data is *striped* across an array of physical drives. This data-distribution scheme complements the way the operating system requests data. The collection of stripe units, from the first drive of the array to the last drive of the array, is called a *stripe.* The granularity at which data is stored on one drive of the array before subsequent data is stored on the next drive of the array is called the *stripe-unit size.*

You can control the stripe-unit size and maximize the performance of your ServeRAID controller by setting a stripe-unit size to a value that is close to the size of the system I/O requests. You can set the stripe-unit size to 8KB, 16KB, 32KB, or 64KB. For example, performance in transaction-based environments, which typically involve large blocks of data, might be optimal when the stripe-unit size is set to 32KB or 64KB. However, performance in file and print environments, which typically involve multiple small blocks of data, might be optimal when the stripe-unit size is set to 8KB or 16KB.

Supported RAID levels

Disk arrays are used to improve performance and reliability. The amount of improvement depends on the applications running on the server and the RAID levels that you assign to the logical drives.

The ServeRAID controllers support RAID level-0, 1, 1E, 5, 5E, 00, 10, 1E0 and 50.

Blueprint for Data Protection

In a non-RAID environment, users with single-drive computers have no means of protecting their data or ways to create redundant data online. Data protection and redundancy might not seem important to many people, or they simply might not have thought of the need—until their hard disk drive fails and all data is lost, with no means of recovery. At the very least, 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 solution to help save data in the event of drive failure, provide redundancy and enable recovery while the system is running. That solution is RAID technology, which has been developed over the last two decades.

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.

RAID can be controlled by specialized software or by a RAID controller 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 level -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 equals the capacity of the physical drives times the number of physical drives. RAID 0 requires a minimum of one drive and, depending upon the level of firmware and the stripe-unit size, supports a maximum of 8 or 16 drives.

RAID 0 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, because multiple drives are operating simultaneously.

RAID 0 offers substantial speed enhancement, but it provides no data redundancy or fault tolerance. So if one disk fails, all data is lost and all disks must be reformatted. RAID 0 provides the largest storage capacity of the RAID levels that are offered, because no room is taken for redundant data or data-parity storage.

The following illustration shows an example of a RAID 0 logical drive.
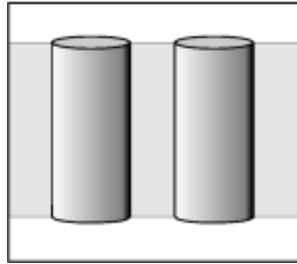
Start with two physical drives.
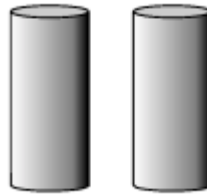
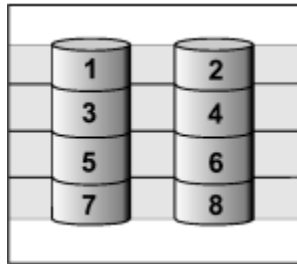Create an array using the two physical drives.
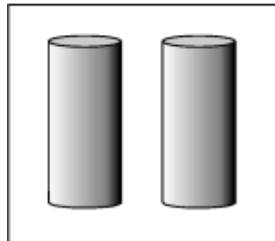
Then, create a logical drive within that array.

The data is striped across the drives, creating blocks.

Notice that the data is striped across all the drives in the array, but no redundant data is stored.

RAID 0 is primarily used 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.

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.

The advantages of RAID 0 are:

- Provides a large logical disk drive through drive spanning

- Provides performance acceleration through data striping.

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

RAID level-1

RAID 1 provides 100% data redundancy and requires only two physical drives. RAID 1 employs the concept of data mirroring. Mirroring creates a single logical disk drive from two physical disk drives. With RAID 1, the first half of a stripe is the original data; the second half of a stripe is a *mirror* (a copy) of the data. The written to the other drive in the RAID 1 array. 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 one of the physical disk drives fails, 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 being concentrated on a single drive. Write performance is superior to RAID 5 but less than RAID 0, because two write operations 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. In general, RAID 1 provides the highest level of performance, but with less storage efficiency.

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.
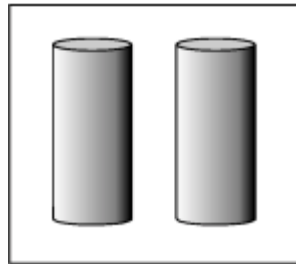
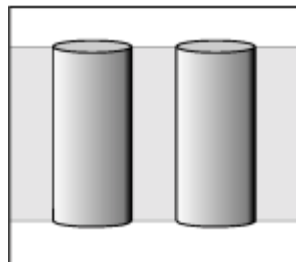The following illustration shows an example of a RAID level-1 logical drive.

Start with two physical drives.
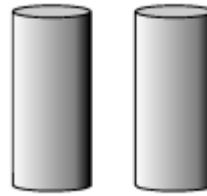
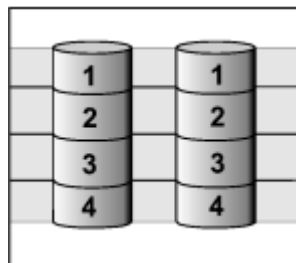Create an array using the two physical drives.

Then, create a logical drive within that array.

The data is striped across the drives, creating blocks.

Notice that the data on the drive on the right is a copy of the drive on the left.

The advantages of RAID 1 are:

- Redundancy through mirrored copy of data

- Read performance superior to RAID 0 and RAID 5

- Write performance superior to RAID 5

- Critical-mode performance superior to RAID 5

The disadvantages of RAID 1 are:

- Write performance less than RAID 0

- Higher capacity overhead than RAID 5

- Requires even number of physical disks

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

RAID level-1 Enhanced

If you have three or more physical drives in the array, RAID 1E is the logical choice.

RAID 1E combines mirroring with data striping. This RAID level stripes data and copies of the data across all of the drives in the array. The first set of stripes are the data, and the second set of stripes are mirrors of the first data stripe contained within the next logical drive. As with the standard RAID level-1, the data is mirrored, and the capacity of the logical drive is 50% of the array capacity of the grouping of physical drives in the array.

RAID 1E requires a minimum of three drives and, depending upon the level of firmware and the stripe-unit size, supports a maximum of 8 or 16 drives.
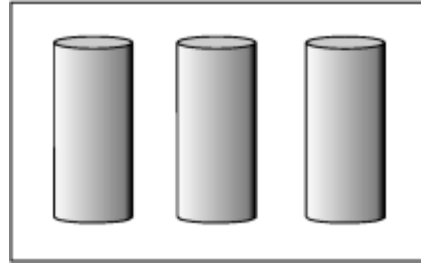
The following illustration is an example of a RAID level-1E logical drive.
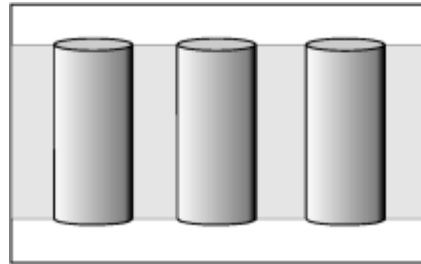
Start with three physical drives.
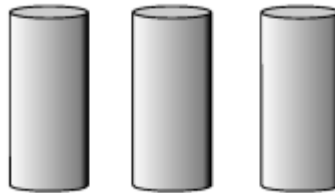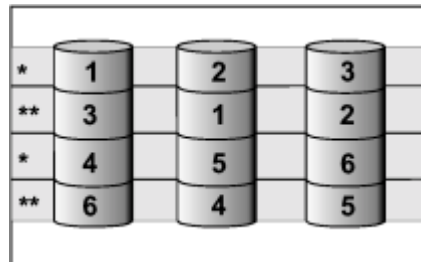
Create an array using the physical drives.

Then, create a logical drive within that array.

The data is striped across the drives, creating blocks.

Notice that the stripe labeled * is the data stripe and the stripe labeled ** is the copy of the preceding data stripe. Also notice that each block on the mirror stripe is shifted one drive.

*	1	2	3
**	3	1	2
*	4	5	6
**	6	4	5

With RAID level-1E, if one of the physical drives fails, the ServeRAID controller switches read and write requests to the remaining functional drives in the RAID level-1E array. RAID 1E shares the characteristics of RAID 1 but additionally allows more than two drives, including odd numbers of drives.

The advantages of RAID 1E are:

- Shares the characteristics of RAID 1,

- Allows more than two drives, including the ability to add an odd number of drives

The disadvantage of RAID 1E is a 50% storage efficiency.

RAID level-5

RAID 5 employs data striping and block interleaving in a technique designed to provide fault-tolerant data storage that does not require duplicate disk drives (as does RAID 1 mirroring). RAID 5 spreads both the data and the parity information across the disks one block at a time. This enables maximum read performance when accessing large files and improves 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 drives and, depending upon the level of firmware and the stripe-unit size, supports a maximum of 8 or 16 drives.

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, RAID 5 is best used when capacity efficiency is most important, because performance will be lower in critical mode.

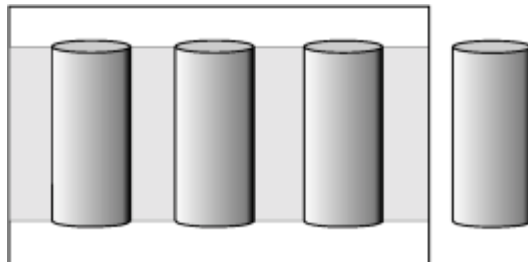The following illustration is an example of a RAID level-5 logical drive.

Start with four physical drives.
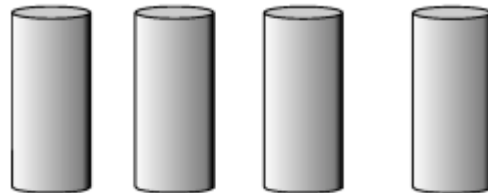
Create an array using three of the physical drives, leaving the fourth as a hot-spare drive.
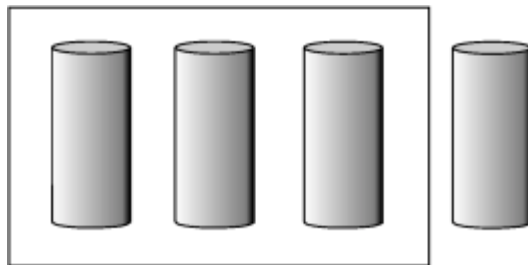
Then, create a logical drive within that array.

The data is striped across the drives, creating blocks. Notice that the storage of the data parity (denoted by ˚) also is striped, and it shifts from drive to drive.

A parity block (˚) contains a representation of the data from the other blocks in the same stripe. For example, the parity block in the first stripe contains data representation of blocks 1 and 2.
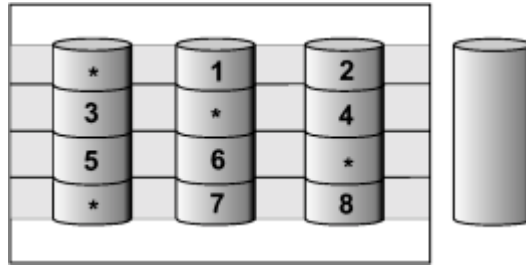
If a physical drive in the array fails, the ServeRAID controller switches read and write requests to the remaining functional drive in the RAID level-5 array, which is a hot-spare drive.
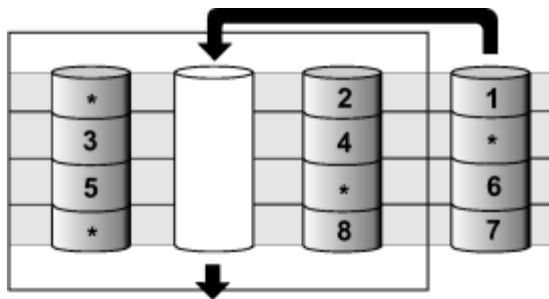
The advantages of RAID 5 are:

- 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:

- Lower write performance than RAID 0 and RAID 1

- Read performance equal to RAID 0 but less than RAID 1

- Least efficient critical-mode performance

RAID level-5 Enhanced

RAID 5 Enhanced (RAID 5E) complements 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. You can think of RAID 5E as "RAID level-5 with a built-in spare drive." This technique provides a performance improvement for smaller arrays (three to five drives) with typical data transfer sizes (8KB of data).

Reading from and writing to four disk drives is more efficient than three disk drives and an idle hot spare and therefore improves performance. Additionally, the spare drive is actually part of the RAID 5E array, as shown in the following example. With such a configuration, you cannot share the spare drive with other arrays. If you want a spare drive for any other array, you must have another spare drive for those arrays.

Like RAID 5, this RAID level stripes data and parity across all of the drives in the array. When an array is assigned RAID level-5E, the capacity of the logical drive is reduced by the capacity of two physical drives in the array (that is, one for parity and one for the spare).

RAID 5E offers both data protection and increased throughput in addition to the built-in spare drive, and it is firmware specific. It requires a minimum of four drives and, depending upon the level of firmware and the stripe-unit size, supports a maximum of 8 or 16 drives.
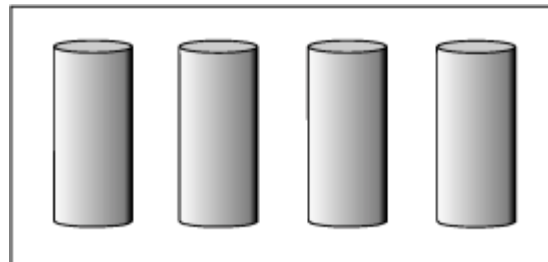
Note:

For RAID 5E, you can have only one logical drive in an array. When using RAID 5E, you can have a maximum of seven logical drives on the controller.

The following illustration is an example of a RAID level-5E logical drive.
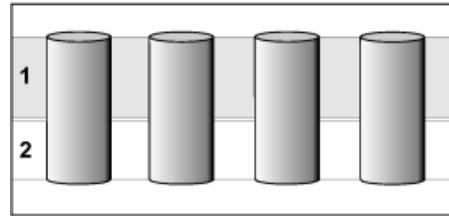
Start with four physical drives.
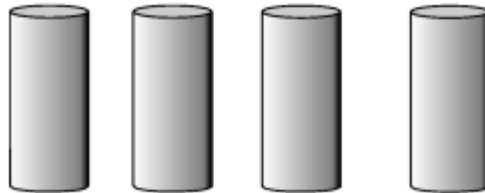
Create an array using all four physical drives.

Then, create a logical drive (labeled as 1) within the array.

Notice that the distributed spare drive is the free space (labeled as 2) shown below the logical drive.

The data is striped across the drives, creating blocks in the logical drive.

The storage of the data parity (denoted by *) is striped, and it shifts from drive to drive as it does in RAID level-5.

Notice that the spare drive is **not** striped.
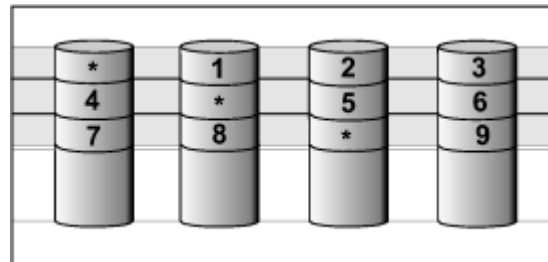
If a physical drive fails in the array, the data from the failed drive is compressed into the distributed spare drive. The logical drive remains RAID level-5E.

When you replace the failed drive, the data for the logical drive decompresses and returns to the original striping scheme.
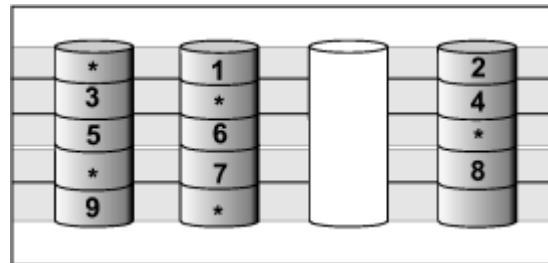
The advantages of RAID 5E are:

- Improved throughput by actively using the "spare" heads

- Active use of hot-spare drive allows constant monitoring of the "health" of all drives

RAID level-x0

RAID x0 refers to RAID levels-00, 10, 1E0, and 50. RAID level-x0 includes more physical drives in an array—the benefits are larger logical drives, increased performance and increased reliability. RAID levels-0, 1, 1E, 5 and 5E cannot use more than 16 physical drives in an array. However, RAID levels-00, 10, 1E0 and 50 include more physical drives by managing an array of arrays, or a *spanned array*. The operating system uses the spanned array logical drive the same as a regular array logical drive.

Note:

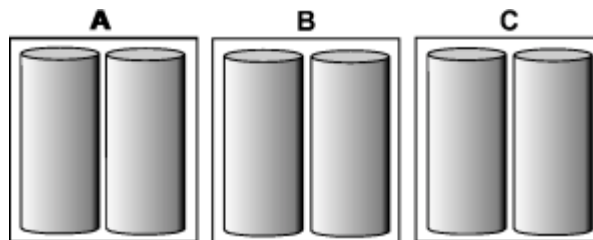RAID level-x0 is available only on the IBM ServeRAID-4 Ultra160 SCSI controller.

RAID x0 requires a minimum of two drives and supports a maximum of 60 drives.

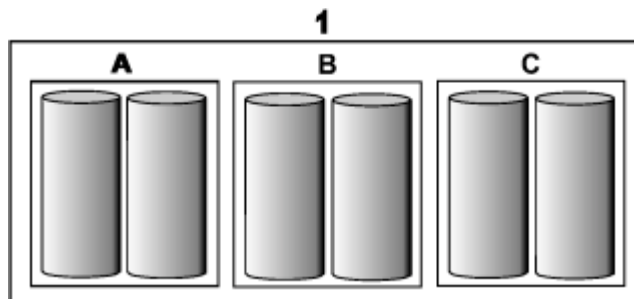The following illustration is an example of a RAID level-10 logical drive.

Start with six physical drives.

Create three arrays (labeled A, B and C), each array using two physical drives.

Then, create a *spanned* array (labeled as **1**) that spans the three arrays.

A sublogical drive is created within *each* array (A, B and C). Then the data is striped across the physical drives in the array, creating blocks.

Notice that in each array, the data on the right side of the drive is a copy of the drive on the left. This is because the sublogical drives (A, B and C) are RAID level-1 in a RAID level-10 implementation (see the following table).
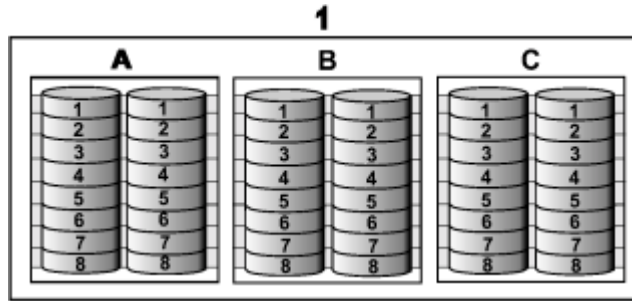
Then, create a logical drive within the spanned array (**1**). The data is striped across this logical drive, creating blocks. Notice that none of these blocks is redundant. This is because the logical drive (**1**) is RAID level-0 in a RAID level-x0 implementation (see the following table).
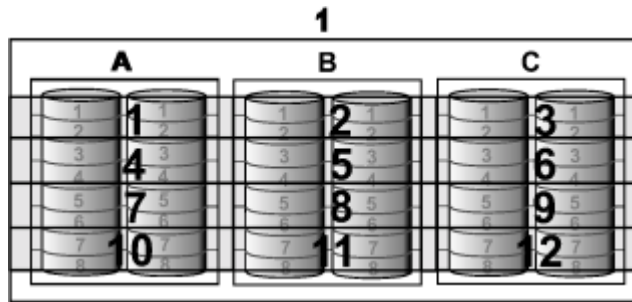
RAID level	Sublogical drive	Spanned array logical drive
00	RAID level-0	RAID level-0
10	RAID level-1	RAID level-0
1E0	RAID level-1E	RAID level-0
50	RAID level-5	RAID level-0

With RAID level-10, 1E0 and 50, if one of the physical drives fails in a sublogical drive, the ServeRAID controller switches read and write requests to the remaining functional drives in the sublogical drive. With RAID level-00, a physical drive failure within the sublogical drive results in loss of data.

Selecting a RAID level and performance tuning

When selecting a RAID level for your system, consider the following:

RAID level	Data redundancy	Hard disk drive capacity utilization	Read performance	Write performance	Built-in spare drive
RAID 0	No	100%	Superior	Superior	No
RAID 1	Yes	50%	Very high	Very high	No
RAID 1E	Yes	50%	Very high	Very high	No
RAID 5	Yes	67% to 94%	Superior	High	No
RAID 5E	Yes	50% to 88%	Superior	High	Yes
RAID 00	No	100%	Superior	Superior	No
RAID 10	Yes	50%	Very high	Very high	No
RAID 1E0	Yes	50%	Very high	Very high	No
RAID 50	Yes	67% to 94%	Superior	High	No

Drive-state descriptions

This section provides descriptions of the physical and logical drive states. ServeRAID publications frequently refer to these states during instructions and procedures.

Physical drive-state descriptions

The following table provides descriptions of the valid physical drive states.

Drive state	Definition
Defunct	A physical drive in the online, hot-spare, or rebuild state has become defunct. It does not respond to commands, which means that the ServeRAID controller cannot communicate properly with the drive. If a physical drive has become defunct, refer to "Rebuilding a defunct drive" in the *IBM ServeRAID-4 Ultra160 SCSI Controller User's Reference* on the *IBM ServeRAID Support* CD for more information.
Hot spare	A hot-spare drive is a physical drive that is defined for automatic use when a similar drive fails.
Online	The drive is online. It is functioning properly and is part of an array.

Drive state	Definition
Rebuilding	The drive is being rebuilt. For more information on rebuilding a drive, refer to "Rebuilding a defunct drive" in the *IBM ServeRAID-4 Ultra160 SCSI Controller User's Reference* on the *IBM ServeRAID Support* CD for more information.
Ready	The ServeRAID controller recognizes a ready drive as being available for definition.
Standby hot spare	A standby hot spare is a hot-spare drive that the ServeRAID controller has spun down. If an online drive becomes defunct and no suitable hot-spare drive is available, a standby hot-spare drive of the appropriate size automatically spins up and enters the rebuild state.

Logical drive-state descriptions

The following table provides descriptions of the valid logical drive states.

Drive state	Definition
Blocked	During a rebuild operation, the ServeRAID controller sets the state of any RAID level-0 logical drives associated with a failed array to the blocked state; then, it reconstructs the data that was stored in RAID level-1, RAID level-1E, RAID level-5 and RAID level-5E logical drives. After the rebuild operation finishes, you can unblock the RAID level-0 logical drives and access them once again. However, the logical drive might contain damaged data. You must either re-create, reinstall, or restore the data from the most recent backup disk or tape to the RAID level-0 logical drive.
Critical migrating	A logical drive in the critical state that is undergoing a logical drive migration (LDM).
Critical system	The ServeRAID controller uses this reserved state during a logical drive migration (LDM) when the logical drive is in the critical state.
Critical	A RAID level-1, RAID level-1E, RAID level-5 or RAID level-5E logical drive that contains a defunct physical drive is in the critical state. A critical logical drive is accessible, despite a physical drive failure. **Attention:** If the state of the logical drive is critical, refer to "Rebuilding a defunct drive" in the *IBM ServeRAID-4 Ultra160 SCSI Controller User's Reference* on the *IBM ServeRAID Support* CD for more information.
Migrating	The logical drive is undergoing a logical drive migration; that is, a change in RAID levels, a change in logical drive size, an increase in free space, or a RAID level-5E compression or decompression.

Drive state	Definition
Offline	The logical drive is offline and not accessible. This state occurs when one of the following is true. • One or more physical drives in a RAID level-0 logical drive are defunct. • Two or more physical drives in a RAID level-1, RAID level-1E, or RAID level-5 logical drive are defunct. • Three or more drives in a RAID level-5E logical drive are defunct. If any of these is true, refer to "Rebuilding a defunct drive" in the *IBM ServeRAID-4 Ultra160 SCSI Controller User's Reference* on the *IBM ServeRAID Support* CD for more information.
Okay	The logical drive is okay. It is in a good, functional state.
System	The ServeRAID controller uses this reserved state during logical drive migration (LDM).

.

IBM ServeRAID Technology

RAID technology can be implemented with all IBM @server xSeries and Netfinity servers. If your xSeries or 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, embedded processors to perform these RAID functions.

What's more, IBM RAID technology is tested to perform with the following operating systems:

- Microsoft® Windows NT® and Microsoft® Windows 2000

- Novell NetWare

- SCO UnixWare

- SCO OpenServer

- IBM OS/2® Warp

- Linux®

.

IBM ServeRAID-4 Ultra160 SCSI Family of Controllers

Adding an IBM ServeRAID-4 Ultra160 SCSI Controller to your IBM xSeries or Netfinity system can help you gain power, performance and convenience—and it can boost RAID disk performance to enterprise level databases or application networks. Improved levels of system throughput are possible with these fourth-generation RAID controllers that provide twice the bandwidth of Ultra2 SCSI and an advanced 64-bit performance PCI interface. ServeRAID controllers deliver higher performance and offer a choice of power and performance to match your data protection requirements as follows:

- ServeRAID-4Lx is an industry-leading, entry-level controller featuring 66MHz/64-bit PCI bus performance, one Ultra160 SCSI channel and 32MB of ECC cache.

- ServeRAID-4Mx is a midrange controller featuring 66MHz/64-bit PCI bus performance, two Ultra160 SCSI channels and 64MB of battery-backed ECC cache.

- ServeRAID-4H is a high-performance controller with extra high-reliability features, four Ultra160 SCSI channels and 128MB ECC, removable, battery-backed cache.

Standard features of the ServeRAID controllers are:

Feature	ServeRAID-4H	ServeRAID-4Mx	ServeRAID-4Lx
Arrays (max.)	8	8	8
Battery-backup cache	Yes	Yes	No
Cache memory	128MB	64MB	32MB
Hard disk drives (max.)	60	30	15
Logical drives (max.)	8	8	8
Microprocessor	266MHz	100MHz	100MHz
SCSI channels	4	2	1
SCSI transfer speed (max.)	160MBps	160MBps	160MBps

Conclusion

.

RAID is the technology of grouping several hard disk drives in a computer into an array that can be defined 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 multiple physical hard disk drives are grouped 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 those of 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.

IBM continues to be a pioneer in new storage and RAID technologies to improve and expand its role in data storage and protection. IBM will 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.

Additional Information

Visit **ibm.com**/pc/ww/eserver/xseries/scsi_raid.html for more information on IBM ServeRAID-4 Ultra160 SCSI controllers.

Visit **ibm.com**/eserver/xseries for more information on IBM xSeries servers. From the xSeries home page, select **Library** for a list of the types of documentation available.

.