DELL RAID PRIMER

DELL PERC RAID CONTROLLERS

Joe H. Trickey III

Dell Storage RAID Product Marketing

John Seward

Dell Storage RAID Engineering

http://www.dell.com/content/topics/topic.aspx/global/products/pvaul/topics/en/us/raid_controller?c=us&cs=555&l=en&s=biz

TABLE OF CONTENTS

Table of Contents

Introduction

This document provides basic information about using redundant array of independent disks (RAID) technology. It is a high-level overview that defines RAID, the advantages and disadvantages of various RAID levels, and guidelines to observe when implementing RAID.

About RAID

RAID is a way of storing data on multiple independent physical disks for the purpose of enhanced performance and/or fault tolerance. The physical disks combine to make up what is called a virtual disk. This virtual disk appears to the host system as a single logical unit or drive. For example, if you have physical disk 1 and physical disk 2 forming a RAID virtual disk, those two disks appear to the host system as one disk.

***NOTE*:** Virtual Disks are sometimes called *containers* or *arrays*.

There are several different RAID types or levels, which determine how the data is placed in the virtual disk. Each RAID level has specific data protection and system performance characteristics. The following are commonly used RAID levels:

- **RAID 0** — Striping without parity, improved performance, additional storage, no fault tolerance
- **RAID 1** — mirroring without parity, fault tolerance for disk errors and single disk failures
- **RAID 5** — striping with distributed parity, improved performance, fault tolerance for disk errors and single disk failures
- **RAID 6** — striping with dual parity, fault tolerance for dual drive failures
- **RAID 10** — mirroring combined with striping, better performance, fault tolerance for disk errors and multiple drive failure (one drive failure per mirror set)
- **RAID 50** — combines multiple RAID 5 sets with striping, improved performance, fault disk errors and multiple drive failures (one drive failure per span)
- **RAID 60** — combines multiple RAID 6 sets with striping, improved performance, fault disk errors and multiple drive failures (two drive failures per span)

These RAID levels are discussed in more detail later in this document. You can manage RAID virtual disks with a RAID controller (hardware RAID) or with software (software RAID).

Advantages of RAID

Depending on how you implement RAID, the benefits include one or both of the following:

• **Faster performance** — In RAID 0, 10, 50, or 60 virtual disks, the host system can access simultaneously. This improves performance because each disk in an virtual disk has to handle of the request. For example, in a two-disk virtual disk, each disk needs to provide only its requested data.

• **Data protection** — In RAID 1, 10, 5, 6, 50, and 60 virtual disks, the data is backed up on disk (mirror). In the RAID 5, 50, 6, or 60 virtual disks, the data is parity protected on a single multiple disks. RAID 10, 50, and 60 also allow the host to access disks simultaneously.

Supported RAID Levels

Dell™ systems that use RAID controllers may support RAID 0, 1, 5, 6, 10, 50, and 60 depending upon the controller. The following is a brief explanation of these levels.

RAID Architecture Fundamentals

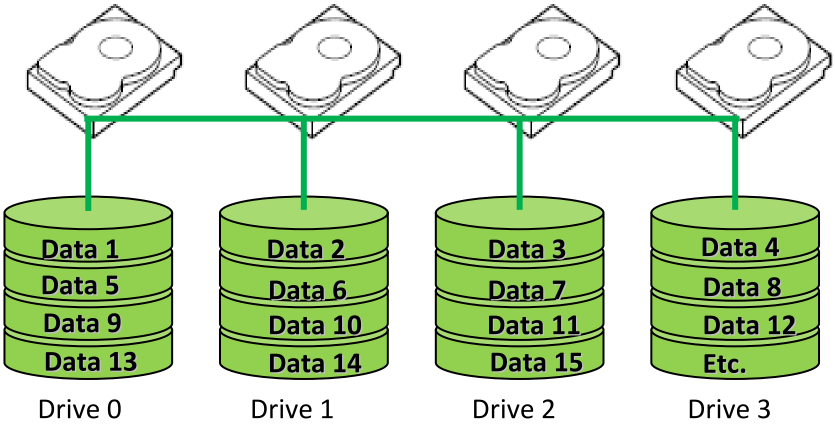
RAID 0 (Striped Virtual Disk without Fault Tolerance)

RAID 0, also known as striping, maps data across the physical drives to create a large virtual disk. The data is divided into consecutive segments or stripes that are written sequentially across the drives in the virtual disk. See Figure 1. Each stripe has a defined size or depth in blocks.

For example, a four-drive virtual disk may be configured with 16 stripes (four stripes of designated space per drive). Stripes A, B, C and D are located on corresponding hard drives 0, 1, 2, and 3. Stripe E, however, appears on a segment of drive 0 in a different location than stripe A; stripes F through H appear accordingly on drives 1, 2 and 3. The remaining eight stripes are allocated in the same even fashion across the drives.

RAID 0 provides improved performance because each drive in the virtual disk needs to handle only part of a read or write request. However, because none of the data is mirrored or backed up on parity drives, one drive failure makes the virtual disk inaccessible and the data is lost permanently.

Figure 1, Example of RAID 0



Advantages of RAID 0

- I/O performance is greatly improved by spreading the I/O load across many channels and drives (best performance is achieved when data is striped across multiple channels with only one drive per channel)
- No parity calculation overhead is involved
- Very simple design
- Easy to implement

Disadvantages of RAID 0

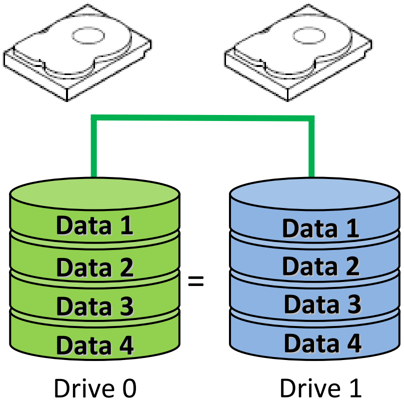
- Not a "True" RAID because the failure of just one drive will result in all data in a virtual disk being lost
- Should not be used for critical data

RAID 1 (Mirroring)

RAID 1 is achieved through what is called *disk mirroring*, and is done to ensure data reliability or a high degree of fault tolerance. In a RAID 1 configuration, the RAID management software instructs the subsystem's controller to store data redundantly across a number of the drives (mirrored set) in the virtual disk. See Figure 2.

In other words, if a disk fails, the mirrored drive takes over and functions as the primary drive.

Figure 2, Example of RAID 1 (Mirroring)



Advantages of RAID 1

- High performance up to twice the read transaction rate of single disks, and the same write transaction rate as single disks
- 100 percent redundancy of data means no rebuild of data is necessary in case of disk failure, just a copy to the replacement disk
- Typically supports hot-swap disks
- Simplest RAID storage subsystem design
- Fastest recovery of data after a drive failure, no data has to be “re-created” from parity codes on retrieval

Disadvantages of RAID 1

- Highest disk overhead of all RAID types (100 percent) results in inefficient use of drive capacity
- Limited capacity since the virtual disk can only include two disk drives

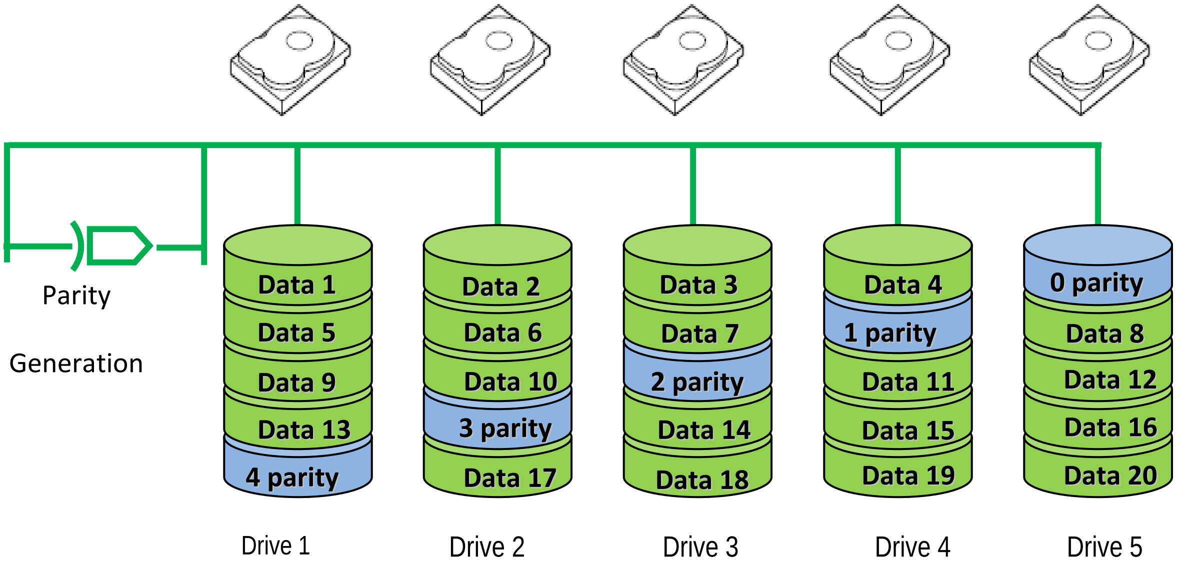
RAID 5 (Striping With Distributed Parity)

RAID 5 maps the data across the drives and stores parity information for each data stripe on different drives in the virtual disk. Data redundancy is maintained with a technique called *parity checking*. With this technique, the RAID controller writes information called parity bits on the disks. Parity data is distributed across disks in the RAID 5 virtual disk such that any 1 disk failure within the virtual disk allows data to be recreated from the remaining disks.

Parity is used to maintain data integrity and to rebuild lost data in case of drive failures. Parity bit data can be written on a single drive (this is RAID Level 3). But during periods of high write activity, the parity disk can become saturated with writes. This reduces the server’s write throughput. However, RAID

Level 5 reduces parity write bottlenecks by allowing all of the drives in the virtual disk to assume part of the parity responsibilities. This alleviates the single drive bottleneck, improving overall subsystem throughput. Figure 3 shows how the parity data is distributed among five hard drives. A RAID 5 virtual disk can preserve data if one drive fails. However, if two drives fail, the virtual disk will fail.

Figure 3, Example of RAID 5 (single virtual disk with 5 disks)



Advantages of RAID 5

- Most efficient use of drive capacity of all the redundant RAID configurations
- High read transaction rate
- Medium-to-high write transaction rate

Disadvantages of RAID 5

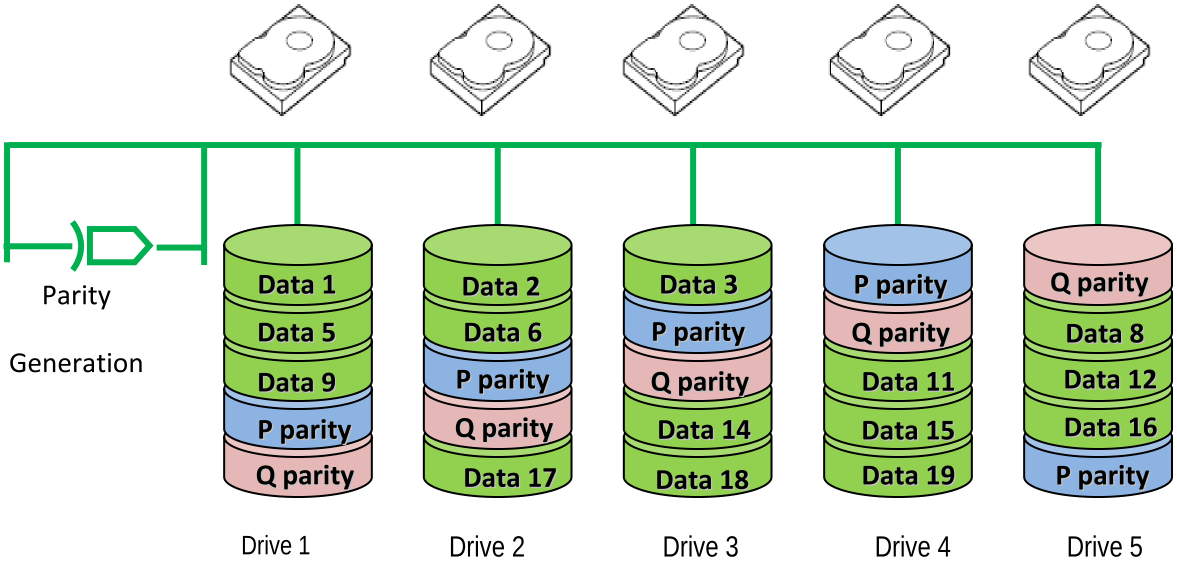
- Disk failure has a medium impact on throughput
- Most complex controller design
- Retrieval of parity information after a drive failure takes longer than with mirroring

RAID 6 (Striping With Dual Distributed Parity)

RAID 6 provides data redundancy by using data striping in combination with parity information. See Figure 4. Similar to RAID 5, the parity is distributed within each stripe. RAID 6, however, uses an additional physical disk to maintain parity, such that each stripe in the disk group maintains two disk blocks with parity information. The additional parity provides data protection in the event of two disk failures.

Figure 4 depicts the RAID 6 data layout. The second set of parity drives are denoted by Q. The P drives follow the RAID 5 parity scheme. The parity blocks on Q drives are computed using Galois Field mathematics. There is no performance hit on read operations. However, as a second independent parity data needs to be generated for each write operation, there is a performance hit during write. Due to dual data protection, a RAID 6 VD can survive the loss of two drives or the loss of a drive when one of its drives is being rebuilt.

Figure 4, Example of RAID 6 (single virtual disk with 5 disks)



Advantages of RAID 6

- Can survive the loss of two disks without losing data
- Data redundancy, high read rates, and good performance

Disadvantages of RAID 6

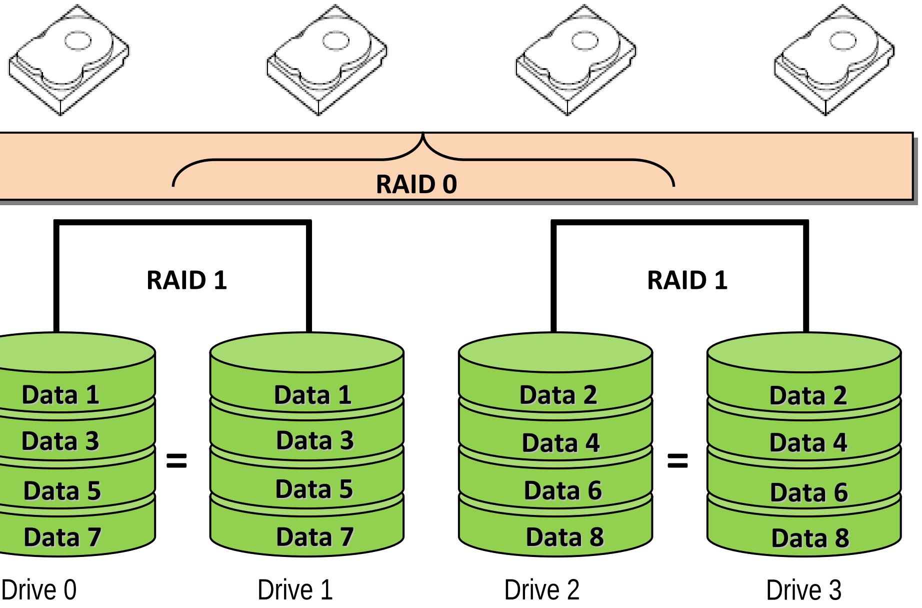
- Requires two sets of parity data for each write operation, resulting in significant decrease in write performance
- Additional costs because of the extra capacity required by using two parity blocks per stripe
- Retrieval of parity information after a drive failure takes longer than with mirroring

RAID 10 (Striping Over Mirrored Sets)

RAID 10 combines striping and mirroring to produce large virtual disks with high performance and fault tolerance. The performance gain comes from striping across mirror sets without the need for parity calculations. See Figure 5.

Although this delivers the highest performance, the drive storage overhead is 100 percent because the entire virtual disk is mirrored. This is an excellent solution for sites that require the highest level of performance and redundancy, as well as the fastest recovery of data after a drive failure.

Figure 5, Example of RAID 10 (1+0)



Advantages of RAID 10

- RAID 10 has the same redundancy as RAID level 1
- High I/O rates are achieved by striping RAID 1 segments

Disadvantages of RAID 10

- Most expensive RAID solution
- Requires 2n where n > 1 disks
- Very limited scalability at a very high inherent cost

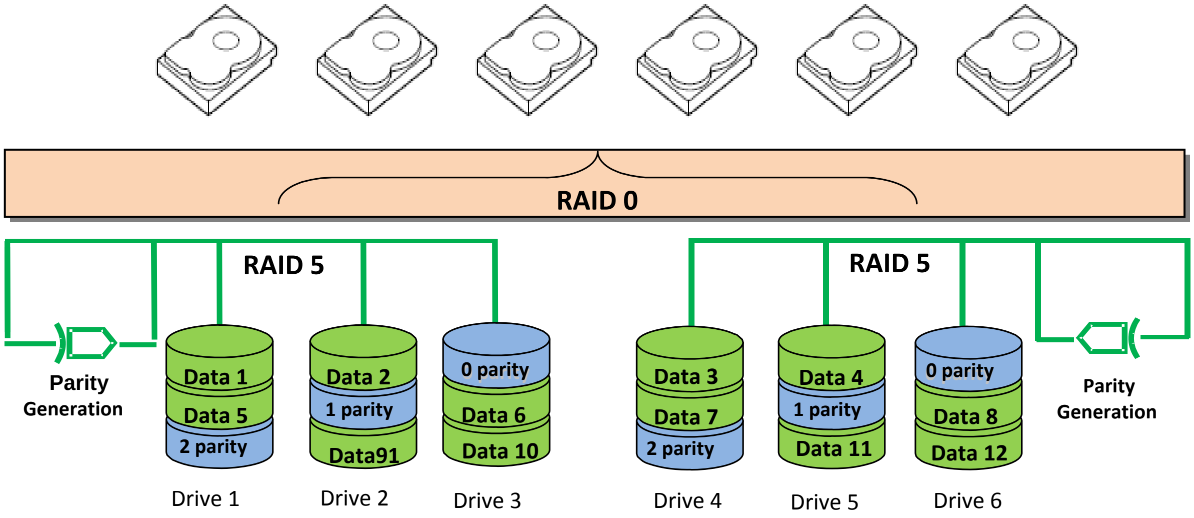
RAID 50 (Striping Across RAID 5)

RAID 50 is a variation of RAID 5 that maps data across two or more RAID 5 virtual disks. The RAID 5 subset must have at least three disks. Figure 6 indicates how the parity data is stored. RAID 50 strips data across each RAID 5 subset. RAID 50 provides a higher degree of fault tolerance since 1 drive per RAID 5 set may fail without data being lost.

A performance increase over RAID 5 may be realized depending on the configuration due to fewer disks reads per parity calculation.

For example if a comparison of a RAID 5 virtual disk with 6 disks were made to a RAID 50 virtual disk with two 3 disk RAID 5 virtual disks, the parity calculation on the RAID 10 virtual disk would require reading all 6 disks each time, where the parity calculation on the RAID 50 may require only reading 3. This may vary depending on several factors such as cache and data block sizes.

Figure 6, Example of RAID 50 (5+0)



Advantages of RAID 50

- Allows creation of largest RAID groups, up to 256 drives (theoretical)
- High read transaction rate
- Higher degree of fault tolerance due to parity calculation being done for each RAID 5 subset
- Potential for faster read transaction rates over large RAID 5 virtual disks
- Medium-to-high write transaction rate

Disadvantages of RAID 50

- Disk failure has a medium impact on throughput
- One of the more complex RAID implementations
- Less space efficient than RAID 5 since separate parity calculations are done for each RAID 5 subset
- Retrieval of parity information after a drive failure takes longer than using a mirrored solution

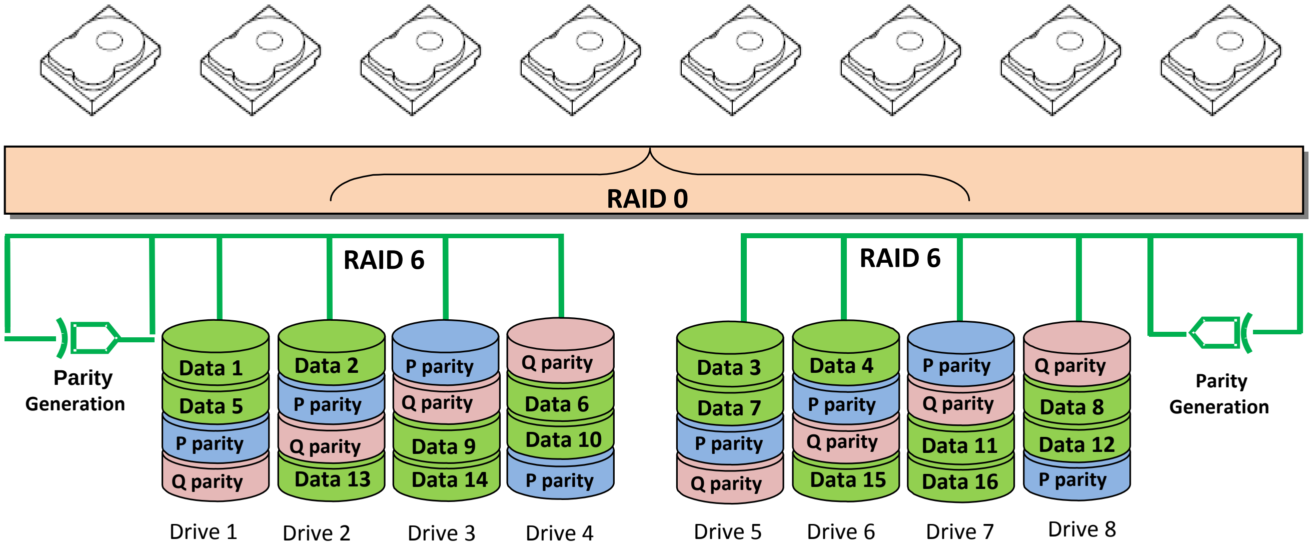
RAID 60 (Striping Across RAID 6)

RAID 60 is striping over more than one span of physical disks that are configured as a RAID 6. The RAID 6 subset must have at least four disks. For example, a RAID 6 disk group that is implemented with four physical disks and then continues on with a disk group of four more physical disks would be a RAID 60. See Figure 7.

RAID 60 strips data across each RAID 6 subset. RAID 60 provides a higher degree of fault tolerance since 2 drives per RAID 6 set may fail without data being lost. A performance increase over RAID 6 may be realized depending on the configuration due to fewer disks reads per parity calculation.

For example if a comparison of a RAID 6 virtual disk with 8 disks were made to a RAID 60 virtual disk with two 4 disk RAID 6 virtual disks, the parity calculation on the RAID 10 virtual disk would require reading all 6 disks each time, where the parity calculation on the RAID 60 may require only reading 4. This may vary depending on several factors such as cache and data block sizes.

Figure 7, Example of RAID 60 (6+0)



Advantages of RAID 60

- Allows creation of largest RAID groups, up to 256 drives (theoretical)
- High degree of fault tolerance due to 2 parity calculations being done for each RAID 6 subset
- Medium-to-high write transaction rate

Disadvantages of RAID 60

- One of the more complex RAID implementations
- Less space efficient than RAID 6 since separate parity calculations are done for each RAID 6 subset
- Retrieval of parity information after a drive failure takes longer than using a mirrored solution

Frequently Asked Questions

The following are common questions asked about RAID. Some answers depend on the capability of your RAID controller. See your RAID controller documentation for information about the features your controller supports.

Q: Do all drives in a RAID virtual disk have to be the same size?
A: All drives in a virtual disk do not have to be the same size. However, all drives in the virtual disk default to the smallest drive in the virtual disk. For example, if your virtual disk contains three 36-GB drives and one 18-GB drive, the maximum capacity of any drive in the virtual disk is 18 GB. On some controllers that support spanned RAID levels, it is possible to use different drive sizes in different spans and still use all the available disk space.

For example in a RAID50 config you can have one span with 4x300 GB drives and a second span with 4x500 GB drives and all the disk space will be available to the end user without any loss of redundancy. To reduce complexity and use drive space efficiently, use only unpartitioned drives of the same size when creating an virtual disk.

Q: What is Online Capacity Expansion (OCE)?
A: OCE refers to a method of adding storage space. You can add capacity to a virtual disk by adding a drive to the virtual disk. For example, if virtual disk 1 contains drives 1through 3, you can add an existing fourth drive to the virtual disk.

NOTE: OCE is not supported on non-spanned virtual disks with RAID 0, 1, 5, or 6 only.

Q: Can I hot swap a drive in a RAID configuration?
A: If your system supports hot-swappable drives (the ability to replace or insert a drive without powering down the system), you can replace a failed drive in a RAID virtual disk with a working drive that is the same size or larger than the other drives in the virtual disk. You can also insert spare drives to be configured into virtual disks or used as hot spares.

NOTICE: Never pull a drive from a virtual disk unless it is in a failed state.

Q: Can I upgrade controllers and use the same physical disks without data loss?
A: In most cases, you can upgrade controllers of the same family without losing data because configuration information is kept on the hard drive.

NOTICE: Data loss can occur if you move your disks to a new controller that is of a different family than the controller it is replacing unless specifically noted as supported by your controller documentation.

Q: Can I change the level of a RAID virtual disk?
A: Depending on the type of controller used, you may be able to migrate your virtual disk to a different RAID level. RAID Level Migration (RLM) is only supported on non-spanned RAID levels such as 0, 1, 5, and 6.

Q: How do hot spares work?
A: A hot spare is a drive that is on standby in case another drive fails. Depending on how the virtual disk is configured, the drive is either picked up automatically and the virtual disk is rebuilt, or you manually select the drive (or insert a new drive in the same slot as the failed drive) and rebuild the virtual disk.

Most Dell systems ship with the automatic rebuild feature enabled. How the hot spare works depends on how the virtual disk is configured. When a drive fails, the virtual disk rebuilds automatically using the hot spare. This is assuming that automatic rebuild is enabled (as it is by default on most Dell systems). If automatic rebuild is disabled, you must manually start the rebuild process. During a rebuild you may notice degraded performance on the drives.

Q: What is the difference between global and dedicated hot spares?
A: A dedicated hot spare is assigned to one or more virtual disks, whereas a global hot spare can be used for any redundant virtual disk that is on the same controller as the hot spare.

Q: What is hotspare enclosure affinity?
A: Enclosure affinity is used to set the preference for a hot spare to be used to rebuild a physical disk that resides in the same physical enclosure. This does not preclude the hot spare from being provisioned to a second enclosure if there are no other hot spares present. For example, if there are two enclosures and each enclosure has a hot spare with affinity set, then upon a drive failure the hot spare will be provisioned from the same enclosure as the failed drive.

NOTE: You can configure hotspare enclosure affinity only if you are using an external storage enclosure.

Q: How do I replace a failed drive?
A: If you replace a failed drive with a good drive, the rebuild is automatic. If you insert a drive of the same or larger capacity than the failed drive, auto-rebuild runs and reconstructs the data.

Q: What is the rebuild rate?
A: In RAID 1, 5, 6, 10, 50, and 60 virtual disks, you can rebuild a failed drive by re-creating the data that was stored on the drive before it failed. The rebuild rate is the priority given to the rebuild operation in relation to host I/O transactions. A rebuild rate of 100% means that rebuild I/O transactions are given the same priority as host transactions. Lowering the rebuild rate gives host I/O a greater priority.

Q: What is the best write cache policy to implement?
A: There are two types of write caching—write-back and write-through. In write-back caching, the RAID controller signals that a data transfer is complete when the controller cache has received all data in the transaction. In write-through caching, the RAID controller signals that a data transfer is complete when the disk subsystem has received all of the data. Write-back caching is typically faster for applications that perform mostly random IO, while writethrough caching provides more assurance that the data made it to the disk and does not require a battery backup for the cache.

If your RAID controller has a battery backup unit, the controller's cache retains data if there is a power loss, meaning that you can have the performance of write-back caching as well as data security.

NOTICE: It is recommended that you use a Battery Backup Unit (BBU) to avoid loss of data.

Q: What are stripe element size and width?
A: Disk striping, which enables data to be written across multiple hard drives, partitions each drive into stripe elements that can vary in size from 8 KB to 1 MB. The stripes are interleaved and the combined storage space consists of stripe elements from each drive. Stripe width is the number of stripe elements within a stripe. For example, a four-drive RAID 0 virtual disk has a stripe width of four. Disk striping enhances performance because multiple drives are accessed simultaneously, but it does not provide data redundancy.

NOTE: Disk striping is not supported on all storage controllers. For more information, see your storage controller documentation.

Q: Is disk spanning the same thing as RAID?
A: No. Disk spanning combines multiple drives and displays them in the operating system as one drive. For example, four 20-GB hard drives that are spanned appear as one 80-GB drive in the operating system. Disk spanning alone provides no data protection.

NOTE: Disk spanning does not use striping. It is a logical grouping to increase the capacity of the disk.

Q: How do consistency checks work?
A: In RAID, a consistency check verifies the redundant data in an virtual disk. For example, on a RAID 5 virtual disk, the consistency between the parity information and the actual data is checked. For a RAID 1 virtual disk, the consistency check will check to see if the data on both drives is the same. Some RAID controllers allow you to pause a consistency check and resume it later or to resume the consistency check after the system reboots.

Glossary

Array — A combination of two or more disks that appears to the system as one disk. Arrays are also referred to as containers or virtual disks.

Disk — A physical hard drive on which data is stored. Also referred to as a drive.

Hot spare — An extra, unused disk that is assigned as a backup disk that can take over when a primary disk fails without interrupting the system or requiring user intervention. A global hot spare can be used to replace any failed primary disk, whereas a dedicated hot spare replaces only a disk in a specific virtual disk to which it is assigned.

Parity — Redundant information that is associated with a block of information and used to rebuild a disk that has failed.

RAID — Storing data on two or more physical disks for the purpose of redundancy, improved performance, or both. You can implement RAID with a RAID controller (hardware RAID) or without a controller (software RAID). See "About RAID" on page 5 for more information.

SCSI — Acronym for small computer system interface, which is a type of interface between a system and devices such as hard drives, diskette drives, CD drives, printers, scanners, and other peripherals.

Volume — A logical or virtual entity that consists of portions of one or more disks. A volume may be formatted and may have a file system, a drive letter, or both.

SAS — Acronym for Serial Attached SCSI. SAS is a serial, point-to-point, enterprise-level device interface that leverages the proven Small Computer System Interface (SCSI) protocol set. The SAS interface provides improved performance, simplified cabling, smaller connectors, lower pin count, and lower power requirements when compared to parallel SCSI. PERC 6 controllers leverage a common electrical and physical connection
interface that is compatible with Serial ATA technology.

Striping — Spreading data over several disks on a bit, byte or cylinder level. The intention is to improve performance, through letting positioning and read/write operations overlap in time

Additional Resources

For more information regarding Dell's PERC & SAS iR RAID Controllers, please visit our Dell RAID Controller webpage at: http://www.dell.com/content/topics/topic.aspx/global/products/pvaul/topics/en/us/raid_controller?c=us&cs=555&l=en&s=biz

For information on Dell's Storage Enclosures (MD 1 Series), performance tuning tips, and best practices guide; please view our storage products catalog at:

http://www.dell.com/content/products/compare.aspx/sas?c=us&cs=555&l=en&s=biz