

# Understanding RAID

*If you run any mission-critical servers, it's essential to have some form of fault-tolerance. RAID is one of the more useful solutions.*

*By David Stott*

The term "disk array" has always been used to represent a variety of multiple hard disk drive configurations. In recent years the term has taken on a new meaning - RAID, describing a configuration consisting of a number of inexpensive disk drives, a parity drive, and a specialised controller. The Redundant Array of Inexpensive Disks (RAID) is a significant advancement in array technology.

## History

The term RAID applies to an architecture that safeguards data - if a disk fails, data is reconstructed. Data is "striped" across several disks. An extra disk is used to store parity information, which is used to reconstruct data. This architecture ensures that users can always access the data they need at any time.

One side-effect of using RAID, of course, is that the MTBF figures for a RAID subsystem are statistically worse than if only a single drive is involved. If you have a RAID system consisting of, say, four drives and one controller, each with an MTBF of five years, one component of the subsystem will fail, on average, every 12 months. However, against this is the fact that the data held on the RAID subsystem will be safe and it only takes a couple of minutes to replace the faulty drive and for the subsystem to start rebuilding the set.

RAID technology was originally developed in 1987 at the University of California at Berkeley. The original intention of the development team was to reduce the cost of mass storage by combining small, cheap drives to replace larger, expensive disks. The Berkeley engineers also wanted to provide a level of protection by including redundant information to ensure that

a disk failure would not cause the loss of access to data.

Protection against loss of data due to disk failure has been achieved through RAID technology. Unfortunately, RAID arrays no longer reduce the cost of data storage. In fact, the cost of a RAID array is usually higher than standard disk drives. However, RAID technology does provide greater performance, data integrity and data availability than standard disk storage. While I/O has always lagged behind CPU performance, the disparity is greater today. The appropriate RAID solution can significantly help to close that gap.

There are six different levels of RAID and each one is designed to provide greater resilience than the previous level.

## Level 0

RAID 0 is an independent array without parity redundancy that accesses data across all drives in the array in a block format. To accomplish this, the first data block is read from, or written to, the first disk in the array. The second block, meanwhile, is read from/written to the second disk and so on. RAID 0 only addresses improved data throughput, disk capacity and disk performance. RAID level 0 was not defined by the Berkeley engineers but has become a commonly used term.

RAID level 0 refers to striping data across multiple disks without any redundant information. Striping can be used to enhance performance in either a request-rate-intensive or transfer-rate-intensive environment.

No fault-tolerance is supported at this level. If a disk subsystem in the array fails, the entire system fails. The same thing would happen if all data was on a single drive. The primary advantage is higher disk access rates

than a single drive. Disk access increases with the number of drives, to the limit of the SCSI channel.

## Level 1

RAID 1 provides fault-tolerance differently than RAID 3 or 5. In RAID 1, any time data is written to a disk, an exact duplicate write is also made to a second (or "mirror") disk automatically and transparently to the system, application and user. The mirrored disk thus is an exact duplicate of the data disk.

The interface to the drives can be through a single controller, which produces the performance of a single drive for reads and writes. Incorporating two controllers (duplexing) can reduce the single-point-of-failure risk. Duplexing can improve I/O data rate by allowing a zig-zag read or by writing to both drives simultaneously. When mirroring with a single controller, data is written first to the data drive and then to the mirrored drive. This slows down write operations.

Mirrored disks have been used by most fault-tolerant, transaction-processing systems. They are an attempt to improve the reliability of the disk storage device rather than improve transfer rates. The MTBF of a mirrored disk subsystem greatly exceeds the expected life of a system with a single set of disk drives utilising conventional methods of implementation.

Unlike other levels, data is recoverable if a drive fails and may be recoverable if both drives fail (although this will require the services of a specialist). The biggest disadvantage is that only half of the total disk capacity is available for storage. Capacity can only be expanded in multiples of two drives.

Of the RAID levels, level 1 provides the highest data availability since two complete copies of all information are

maintained. In addition, read performance may be enhanced if the array controller allows simultaneous reads from both members of a mirrored pair. During writes, there will be a minor performance penalty when compared with writing to a single disk. Higher availability will be achieved if both disks in a mirror pair are on separate I/O buses.

### Level 2

RAID 2 stripes data to a group of disks using a byte stripe. A hamming code symbol for each data stripe is stored on the check disk. This code can correct as well as detect data errors and permits data recovery without complete duplication of data. This RAID level is also sometimes referred to as RAID 0+1. It combines the benefits of both striping and RAID 1. RAID 0+1 can be tuned for either a request-rate-intensive or transfer-rate-intensive environment.

RAID 2 arrays sector-stripe data across groups of drives, with some drives relegated to storing Error Checking and Correction (ECC) information within each sector. However, since most disk drives today embed ECC information within each sector as standard, RAID 2 offers no significant advantages over RAID 3 architecture.

### Level 3

RAID 3 is a striped parallel array where data is distributed by bit or byte. One drive in the array provides data protection by storing a parity check byte for each data stripe. As with RAID 0, disks are accessed simultaneously but the parity check drive is introduced for fault-tolerance.

Data is read/written across the drives one bit at a time and the parity bit is calculated and either compared with the parity drive in a read operation or written to the parity drive in a write operation. Thus with each byte written, a unique parity check is calculated to maintain the data integrity. This will allow the system and disk array to continue to have 100% operational functionality even when there is a failed drive subsystem in the array.

In the event of a failed drive, data

can continue to be written to and read from the other data drives. The parity bit allows the "missing" data for the failed drive to be reconstructed. The failed drive can be replaced while the system remains online ("hot-swapped") and the data is then reconstructed by the array controller in any of three modes. Clearly, if another drive fails while the reconstruction is taking place, data loss will result.

RAID 3 has the advantage over lower RAID levels in that the ratio of check disk capacity to data disk capacity decreases as the number of data drives increases. It has parallel data paths and therefore offers high transfer rate performance for applications that transfer large files. Array capacity can be expanded in single drive or group increments.

With RAID 3, data chunks are much smaller than the average I/O size and the disk spindles are synchronised to enhance throughput in transfer-rate-intensive environments.

RAID 3 is well suited for CAD/CAM or imaging type applications. Since parity is used, a RAID 3 stripe set can withstand a single disk failure without losing data or access to data.

### Level 4

In RAID 4, parity is interleaved at the sector or transfer level. As with RAID 3, a single drive is used to store redundant data using a parity check byte for each data stripe. Parallel data paths and sector or block level distribution across the data drives allows for independent drive operations and multiple I/O operations to execute in parallel.

RAID 4 is identical to RAID 3 except that large stripes are used, so that records can be read from any individual drive in the array apart from the parity drive. This allows read operations to be overlapped.

RAID 4 offers high read performance and good write performance. RAID 4 is a general solution, especially where the ratio of reads to writes is high. This makes RAID 4 a good choice for small block transfers, which are typical for transaction processing applications.

Write performance is slow because the parity drive has to be written for

each data write. Thus the parity drive becomes a performance bottleneck when multiple parity write I/Os are required. In this instance, RAID 5 is a better solution because parity information is spiralled across all available disk drives.

These days, RAID 4 systems are almost never implemented, as they offer no significant advantages over RAID 5.

### Level 5

RAID 5 combines the throughput of the block interleaved data striping of RAID 0 with the parity reconstruction mechanism of RAID 3 without requiring an extra parity drive. This level of fault-tolerance incorporates the parity checksum at the sector-level with the data and checksum striping across drives instead of to a dedicated parity drive.

This technique allows multiple concurrent read/write operations for improved data throughput while maintaining data integrity. A single drive in the array is accessed only when either data or parity information is being read from or written to that specific drive subsystem.

RAID 5's strength is handling large numbers of small files. It allows improved I/O transfer performance because the parity drive bottleneck of Level 4 is eliminated. While RAID 5 is more cost-effective because a separate parity drive is not used, write performance suffers because it requires an extra rotation of the disk. By adding cache memory to a RAID 5 array, write performance is improved. It is important that the cache be supported by a battery-backed power supply.

In graphic arts and imaging applications, the weakness of RAID 5 versus RAID 3 is the write penalty from the striped parity information. In RAID 3 there is no write penalty. RAID 5 is usually seen in applications with large numbers of small read/write calls. RAID 5 does offer higher capacity utilisation when the array has less than seven drives. With a full array, CPU utilisation is about equal between RAID 3 and 5.

In RAID level 5, both parity and data are striped across a set of disks. Data chunks are much larger than the average I/O size. Disks are able to sat-

# RAID

isfy requests independently which provides high read performance in a request-rate-intensive environment. Since parity information is used, a RAID 5 stripe can withstand a single

disk failure without losing data or access to data.

Unfortunately, the write performance of RAID 5 is poor. Each write requires four independent disk accesses

to be completed. First, old data and parity are read off separate disks. Next, the new parity is calculated. Finally, the new data and parity are written to separate disks. Many array vendors use write caching to compensate for the poor write performance of RAID 5.

## Pros And Cons Of RAID At A Glance

### RAID 0

#### Advantages:

- High performance.
- No cost penalty - all storage is available for use.

#### Disadvantages:

- Significantly reduced data availability.
- No fault-tolerance.

### RAID 1

#### Advantages:

- Excellent data availability.
- Higher read performance than a single disk.

#### Disadvantages:

- Expensive - 50% waste of space.
- Moderately slower write performance.

### RAID 2

#### Advantages:

- Excellent data availability.
- High performance.

#### Disadvantages:

- Expensive - requires twice the desired disk space.

### RAID 3

#### Advantages:

- Good data availability.
- High performance for transfer rate intensive applications.
- Cost effective - only one extra disk is required for parity.

#### Disadvantages:

- Can satisfy only one I/O request at a time.
- Poor small, random I/O performance.

### RAID 4

#### Advantages:

- Good data availability.
- High performance for read operations.
- Cost effective - only one extra disk is required for parity.

#### Disadvantages:

- Poor write performance.
- Poor small, random I/O performance.

### RAID 5

#### Advantages:

- Good data availability.
- High performance in request rate intensive applications.
- Cost effective - only one extra disk is required.

#### Disadvantages:

- Poor write performance.
- No performance gain in data transfer rate intensive applications.

## RAID Advisory Board

Formed in August of 1992, the RAID Advisory Board (RAB) has grown from eight to over 50 members. Membership is diverse and includes representatives from suppliers of enclosures, controllers, software and subsystems, PCs, mainframes and drives.

The Board has established three key programmes which support its goal of promoting the understanding and utilisation of RAID and related storage technologies:

- Education.
- Standardisation.
- Certification.

Seven committees have been assigned specific tasks in support of these programs:

- Functional Test.
- Performance Test.
- RAID-Ready Drive.
- Remote Array Monitoring.
- Host Interface.
- RAID Enclosure.
- Education.

The RAID Advisory Board definition of a disk array is "a collection of disks from one or more commonly accessible disk subsystems, combined with a body of Array Management Software. Array Management Software controls disk operation and presents the disks as one or more virtual disks to the host operating environments". Array Management Software may reside either in the host computer or in the disk subsystem. Both array types can be implemented as either internal or external drive systems.

## Non-standard Levels

Despite the efforts of the RAB to set overall standards for the design and implementation of RAID systems

there have been several instances where manufacturers have produced systems which are outside the scope of current definitions. Often such manufacturers had perfectly acceptable reasons for not adhering to the RAB standards, mainly to bring about performance improvements or even greater resilience. However, this situation means that there are several pseudo-RAID levels around which tend to cause confusion. Therefore a short summary is called for.

RAID levels recognised by the Berkeley papers and the RAB are:

**Level 1:** Disk mirroring.

**Level 2:** Redundancy through hamming code.

**Level 3:** Striped array plus parity.

**Level 4:** Independent striped array plus parity.

**Level 5:** Independent striped array with distributed parity.

**Level 6:** Level 5 with double parity.

Level 0: (disk striping) is recognised by the Berkeley papers but not by the RAB.

Combinations of existing RAID levels not recognised in the original Berkeley papers or by the RAB are:

**Level 10:** Levels 1 and 0.

**Level 7:** Independent striped array plus two parity drives.

**Level 53:** Levels 0 and 3.

## EDAP

Companies such as American Megatrends Inc have produced their own flavour of RAID systems with products such as MegaRAID and FlexRAID which provide enhanced facilities and capabilities over and above the RAB definitions. As a result of this confusion the RAB has decided to introduce a completely new classification scheme which will emphasise the practical functionality of systems rather than their underlying technology.

The RAB has identified certain criteria for establishing the degree to which a disk system, array controller and disk enclosure provide Extended Data Availability and Protection (EDAP) capabilities. For disk systems, the EDAP Criteria are mapped to three major classifications:

- 1 The term "Failure Resistant" defines the lowest level of EDAP capability, which for disk systems basically means the ability to protect against data loss and loss of access to data due to a disk failure.
- 2 The term "Failure Tolerant" defines the next major level of EDAP capability, which for disk systems basically means the ability to protect against data loss and loss of access to data due to the failure of any component within the storage system, not just a disk.
- 3 The term "Disaster Tolerant" defines the highest level of EDAP capability, which for disk systems basically means the ability to protect against data loss and loss of access to data due to the failure of one entire zone of a disk system physically separated into two or more zones.

To distinguish which set of EDAP Classification Criteria apply, the classification terms, as applicable, are followed by the term "disk system", "array controller" or "disk enclosure". Complete classification descriptions for disk systems therefore always consist of four words. The following are the three major classifications for disk systems:

- Failure Resistant Disk System (FRDS).
- Failure Tolerant Disk System (FTDS).
- Disaster Tolerant Disk System (DTDS).

To provide further classification differentiation, plus-signs are used. For example, the classification "Disaster Tolerant Disk System +" (DTDS+) differs from "Disaster Tolerant Disk System" (DTDS) by the fact that the zones must be separated by at least 10 Km for the former and 1 Km for the latter.

Each product category has its own Classification Report. These contain the vendor assertions for each of the applicable EDAP Classification Criteria and the comments of the RAB Examiner who has inspected the report. When completed by the vendor and successfully inspected by the RAB Examiner, the Classification Report is submitted with a RAB Gold Logo Li-

cense to the RAB where, after approval, it becomes the source document for the information displayed on that portion of the RAB Web site's CTIC assigned to the vendor. Classification Reports are also available to the public.

## Conclusion

Over the last seven years or so, RAID has established itself as an appropriate technology for increasing the robustness and reliability of a wide range of computer systems. As corporates make more and more use of mission-critical applications they will continue to depend on RAID solutions to provide fault-tolerance and failure resistance.

The new RAID classification will help non-technical users appreciate better the benefits of RAID technology from a business perspective. It will also help to banish the confusion that has arisen regarding the various current RAID levels. RAID in general has a bright and exciting future which the RAB is keen to promote.

However, in this fast-moving world of technology, one final point is worth mentioning. If you have a RAID subsystem and you need to replace a faulty drive, it's almost certain that you will need to replace it with a drive of precisely the same type as the remaining drives. If your RAID system is, say, three years old, it's highly likely that such drives will be unavailable. Therefore, if you're setting up a RAID system, it makes sense to buy a few spare drives and to test them regularly.

*This article replaces "How To Choose RAID", PCNA 52, file F0908, which you can safely discard from your files if you wish.*

**PCNA**

## The Author

David Stott (flingmo@cix.co.uk) is a freelance writer and journalist.