

The RAID Advantage

Introduction

Electronic data processing evolved from virtually nothing 50 years ago to its virtual omnipresence in the industrialized societies of the world today. The technologies that have been harnessed to manipulate data converted to its lowest common denominators (zeros and ones) has made nothing short of a huge impact on the lives of people throughout the world. Digitized information, or data, is being used to enable everything from live conversations between continents via satellite to the advancement of scientific discoveries and research, to controlling the temperatures of different rooms in a home. The recently-emerged raft of on-line services provide not only the links to communicate with personal computers, but provide access to oceans of information to navigate, capture and use by anyone with a computer. Businesses like banks and credit card companies use massive computing systems to provide everyday conveniences like easier and faster access to money, in turn making it easier to bill or manage accounts. Even supermarkets and retail department stores are using powerful data-intensive information systems to do everything from managing inventories to monitoring consumer spending habits. The applications list goes on and on; everyone in virtually all walks of life is exposed in some manner or form to the impact of the ongoing revolution we call the "Information Age."

The engines behind this revolution, of course, are computers. Today's Pentium-class personal computers, RISC workstations, minicomputers, supercomputers and even (still!) mainframes provide the power that drives this infinite mass of data we rely upon to make everything from bank transactions to the purchase of groceries as easy as possible. The flow of data between computers, whether networked or linked via online services or the Internet, has become nothing less than a raging flood.

This astounding volume of data being transmitted between systems today has created an obvious need for data management. As a result, more and more servers -- whether they are PCs, UNIX workstations, minicomputers or supercomputers -- have assumed the role of information or data "traffic cops." The number of networked or connectable systems is increasing by leaps and bounds as well, thanks to the widespread adoption of the client/server computing model, the boom in home computer use and the rise of Internet access service providers.

Hard disc storage plays an important role in enabling improvements to networked systems, because the vast and growing ocean of data has to reside somewhere. It also has to be readily accessible, placing a demand upon storage system manufacturers to not only provide high-capacity products, but products that can access data as fast as possible and to as many people at the same time as possible. Such storage also has to be secure, placing an importance on reliability features that best ensure that data will never be lost or otherwise rendered inaccessible to network system users.

RAID: The solution to server gridlock and data integrity

The solution to providing access to many gigabytes of data to users fast and reliably has been to assemble a number of drives together in a "gang" or "array" of discs. These are known as RAID subsystems, which stands

for "redundant arrays of independent discs." Simple RAID subsystems are basically a clutch of up to five or six disc drives assembled in a cabinet that are all connected to a single controller board. The RAID controller orchestrates read and write activities in the same way a controller for a single disc drive does, and treats the array as if it were in fact a single or "virtual" drive. RAID management software that resides in the host system provides the means to manage data to be stored on the RAID subsystem.

RAID elements

Despite its multi-drive configuration, a RAID subsystem's disc drives remain "hidden" from users; the subsystem itself is the virtual drive, though it can be as large as 1,000 Gbytes. The phantom virtual drive is created at a lower level within the host operating system through the RAID management software. Not only does the software set up the system to address the RAID unit as if it were a single drive, it allows the subsystem to be configured in ways that best suit the general needs of the host system.

RAID subsystems can be optimized for performance, the highest capacity, fault tolerance or a combination of two or three of the above. Different so-called RAID levels have been defined and standardized in accordance with those general optimization parameters. There are six such standardized levels RAID, called RAID 0, 1, 2, 3, 4, or 5, depending upon performance, redundancy and other attributes required by the host system. The RAID software is what is used to configure the desired RAID level of features in an array described in more detail below.

The RAID controller board is the hardware element that serves as the backbone for the array of discs: it not only relays the input/output (I/O) commands to specific drives in the array, but provides the physical link to each of the independent drives so they may easily be removed or replaced. The controller also serves to monitor the "health" or integrity of each drive in the array to anticipate the need to move data should it be placed in jeopardy by faulty or failing disc drive (a feature known as "fault tolerance").

The array of RAID levels

The RAID 1 through 5 standards offer users and system administrators a host of configuration options. These options allow the arrays to be tailored to their application environments. Each of the various configurations listed below focus on maximizing the abilities of an array in one or more of the following areas: capacity, data availability, performance and fault tolerance.

RAID Level 0

An array configured to RAID Level 0 is an array optimized for performance, but at the expense of fault tolerance or "data integrity."

RAID Level 0 is achieved through a method known as "striping." The collection of drives (or "virtual drive") in a RAID Level 0 array has data laid down in such a way that it is organized in stripes across the multiple drives. A typical array can contain any number of stripes, usually in multiples of the number of drives present in the array. As an example, imagine a four-drive array configured with 12 stripes (four stripes of designated "space" per drive). Stripes 0, 1, 2 and 3 would be located on

corresponding hard drives 0, 1, 2 and 3. Stripe 4, however, appears on a segment of drive 0 in a different location than Stripe 0; stripes 5 through 7 appear accordingly on drives 1, 2 and 3. The remaining four stripes are allocated in the same even fashion across the same drives.

Practically any number of stripes can be created on a given RAID subsystem for any number of drives. Two hundred stripes on two disc drives is just as feasible as 50 stripes across 50 hard drives. Most RAID subsystems, however, tend to have between three and 10 stripes.

The reason RAID 0 is a performance-enhancing configuration is that striping enables the array to access data from multiple drives at the same time. In other words, since the data is spread out across a number of drives in the array, it can be accessed faster because it's not bottled up on a single drive. This is especially beneficial for retrieving very large files, since they can be spread out effectively across multiple drives and accessed as if it were the size of any of the fragments it is organized into on the data stripes.

The downside to RAID Level 0 configurations is that it sacrifices fault tolerance, raising the risk of data loss because no room is made available to store redundant data. If one of the drives in the RAID 0 fails for any reason, there is no way of retrieving the lost data as can be done in other RAID implementations described below.

RAID Level 1

The RAID Level 1 configuration employs what is known as "disc mirroring," and is done to ensure data reliability or a high degree of fault tolerance. RAID 1 also enhances read performance, but the improved performance and fault tolerance come at the expense of available capacity in the drives used.

In a RAID Level 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 array. In other words, the same data is copied and stored on different discs (or "mirrored") to ensure that, should a drive fail, the data is available somewhere else within the array. In fact, all but one of the drives in a mirrored set could fail and the data stored to the RAID 1 subsystem would remain intact. A RAID Level 1 configuration can consist of multiple mirrored sets, whereby each mirrored set can be a different capacity. Usually the drives making up a mirrored set are of the same capacity. If drives within a mirrored set are of different capacities, the capacity of a mirrored set within the RAID 1 subsystem is limited to the capacity of the smallest-capacity drive in the set, hence the sacrifice of available capacity across multiple drives.

The read performance gain can be realized if the redundant data is distributed evenly on all of the drives of a mirrored set within the subsystem. The number of read requests and total wait state times both drop significantly; inversely proportional to the number of hard drives in the RAID, in fact. To illustrate, suppose three read requests are made to the RAID Level 1 subsystem. The first request looks for data in the first block of the virtual drive; the second request goes to block 0, and the third seeks from block 2. The host-resident RAID management software can assign each read request to an individual drive. Each request is then sent to the various drives, and now -- rather than having to handle the flow of

each data stream one at a time -- the controller can send three data streams almost simultaneously, which in turn reduces system overhead.

RAID Level 2

RAID Level 2 is rarely used in commercial applications, but is another means of ensuring data is protected in the event drives in the subsystem incur problems or otherwise fail. This level builds fault tolerance around Hamming error correction code (ECC), which is often used in modems and solid-state memory devices as a means of maintaining data integrity. ECC tabulates the numerical values of data stored on specific blocks in the virtual drive using a special formula that yields what is known as a checksum. The checksum is then appended to the end of the data block for verification of data integrity when needed.

As data gets read back from the drive, ECC tabulations are again computed, and specific data block checksums are read and compared against the most recent tabulations. If the numbers match, the data is intact; if there is a discrepancy, the lost data can be recalculated using the first or earlier checksum as a reference point.

RAID Level 3

This RAID level is really an adaptation of RAID Level 0 that sacrifices some capacity, for the same number of drives, but achieves a high level of data integrity or fault tolerance. It takes advantage of RAID Level 0's data striping methods, except that data is striped across all but one of the drives in the array. This drive is used to store parity information that is used to maintain data integrity across all drives in the subsystem. The parity drive itself is divided up into stripes, and each parity drive stripe is used to store parity information for the corresponding data stripes dispersed throughout the array. This method achieves very high data transfer performance by reading from or writing to all of the drives in parallel or simultaneously but retains the means to reconstruct data if a given drive fails, maintaining data integrity for the system. RAID Level 3 is an excellent configuration for moving very large sequential files in a timely manner.

The stripes of parity information stored on the dedicated drive are calculated using the "Exclusive OR" function. Exclusive OR is a logical function between the two series that carries most of the same attributes as the conventional OR function. The difference occurs when the two bits in the function are both non-zero: in Exclusive OR, the result of the function is zero, wherein with conventional OR it would be one.

By using Exclusive OR with a series of data stripes in the RAID, any lost data can easily be recovered. Should a drive in the array fail, the missing information can be determined in a manner similar to solving for a single variable in an equation (for example, solving for x in the equation, $4 + x = 7$). Similarly, in an Exclusive OR operation, it would be an equation like $1 \bullet x = 1$. Thanks to Exclusive OR, there is always only one possible solution (in this case, 0), which provides a complete error recovery algorithm in a minimum amount of storage space.

RAID Level 4

This level of RAID is similar in concept to RAID Level 3, but emphasizes performance for different applications, e.g. Database TP versus large sequential files. Another difference between the two is that RAID Level 4

has a larger stripe depth, usually of two blocks, which allows the RAID management software to operate the discs much more independently than RAID Level 3 (which controls the discs in unison). This essentially replaces the high data throughput capability of RAID Level 3 with faster data access in read-intensive applications.

A shortcoming of RAID level 4 is rooted in an inherent bottleneck on the parity drive. As data gets written to the array, the parity encoding scheme tends to be more tedious in write activities than with other RAID topologies. This more or less relegates RAID Level 4 to read-intensive applications with little need for similar write performance. As a consequence, like its Level 3 cousin, it doesn't see much common use in commercial applications either.

RAID Level 5

This is the last of the most common RAID levels in use, and is probably the most frequently implemented. RAID Level 5 minimizes the write bottlenecks of RAID Level 4 by distributing parity stripes over a series of hard drives. In doing so it provides relief to the concentration of write activity on a single drive, which in turn enhances overall system performance.

The way RAID Level 5 reduces parity write bottlenecks is relatively simple. Instead of allowing any one drive in the array to assume the risk of a bottleneck, all of the drives in the array assume write activity responsibilities. The distribution frees up the concentration on a single drive, improving overall subsystem throughput.

RAID Level 5's parity encoding scheme is the same as Levels 3 and 4; it maintains the system's ability to recover any lost data should a single drive fail. This can happen as long as no parity stripe on an individual drive stores the information of a data stripe on the same drive. In other words, the parity information for any data stripe must always be located on a drive other than the one on which the data resides.

Other RAID levels

Other, less-common RAID levels have been developed as custom solutions by independent vendors (they are not established standards):

- RAID Level 6, which emphasizes ultra-high data integrity

- RAID Level 10 (also known as RAID Level 0 & 1), which focuses on high I/O performance and very high data integrity

- RAID Level 53, which combines RAID Level 0 and 3 for uniform read and write performance

Tailor-made RAID

Perhaps RAID technology's biggest advantage is the sheer number of possible adaptations available to users and systems designers. RAID offers the ability to customize a array subsystem to the requirements of its environment and the applications demanded of it. RAID's inherent variety of configuration options provides several ways in which to satisfy specific application requirements.

Customization, however, doesn't stop with a RAID level. Drive models, capacities and performance levels have to be factored in as well as what connectivity options that are available.

Interface Options

The newest parallel SCSI interface option is Ultra2 SCSI, an 80 Mbyte/sec interface standard. Ultra2 SCSI combines low-voltage differential (LVD) technology for extended cable length capabilities (up to 12 meters) as well as enhanced device support (up to 15 devices per controller card).

An emerging new serial interface standard known as Fibre Channel-Arbitrated Loop (FC-AL) is yet another interface option for RAID subsystems, and is the most powerful of them all. FC-AL is capable of up to 200 Mbyte/sec data throughputs (dual loop configurations) while allowing RAID subsystems or other connected peripherals to be placed as far as 10 kilometers from the host. It also enables easy connection of up to 126 disc drives on a single controller (compared to seven devices with conventional SCSI!). The potential impact of FC-AL alone will undoubtedly be enormous upon evolution of RAID subsystems. FC-AL can be operated in either single or dual loop configurations. The dual loop allows another level of redundancy by allowing two separate data paths for all attached devices.

SCA: Cleaning up the cable mess

Many of these interface options, including serial FC-AL and parallel Ultra2 SCSI, support the SCSI Single Connector Attachment (SCA) standard. SCA is an elegant means of eliminating the miles of wiring involved with connecting several drives via conventional backplane architectures. Before SCA, conventional connections involved two cables per drive: one for power and the other for data transmission. Arrays with more than a few drives would amass a lot of "spaghetti" at the rear of the rack, and especially large arrays would have an unwieldy mess of wire to connect the drives. SCA, however, allows for drives to be plugged directly into a backplane without cables. It not only rids subsystems of the mass of cabling previously required, but facilitates "hot plugging" (removal or insertion of a drive while the subsystem is on line) and improves the reliability of the system as a whole because of the substantially reduced number of connections.

ASA II: The only way to fly

Seagate has not only developed and shipped hard drives for RAID array applications (those include the award-winning Cheetah and Barracuda disc drives), it has innovated new technology to take advantage of the attributes of high-performance SCSI and FC-AL. Seagate's own Advanced SCSI Architecture II (ASA II) provides a way of maximizing a disc drive's performance for a given application. The ASA II chipset used in the drives mentioned above analyzes how the hard drive is being used (in this case, how it is being used in a RAID configuration) and optimizes the drive to provide maximum performance within the application. Essentially it is capable of minimizing disc command overhead through command queuing, event logging, disc sequencing and kernel management controls. Cache buffers on the disc are segmented to adapt to the application as well.

ASA II works in a similar fashion to the way RAID management software performs, but at the disc level (as opposed to the "virtual" disc level). Some systems, whether they are RAID or not, might require management of frequent reads for small amounts of data, while others require heavy

write management. Similarly, ASA II drives can be optimized for either multitasking or single-task environments. ASA II adapts the drives to meet the optimum performance requirement of the application for the disc within a system (in this case, a RAID subsystem).

Seagate's drives: the foundations to RAID

The award-winning Barracuda family of 7,200-rpm 3.5-inch disc drives provides solid value in cost-conscious systems where a balance of performance and cost-optimized designs are required. The Ultra2 SCSI and Fibre Channel Barracuda drives range from the low-profile (inch-high) 9- and 18-Gbyte Barracuda 18LP products to the 36- and 50-Gbyte, half-height Barracuda 36 and Barracuda 50. Barracuda drives have been used often in arrays with RAID levels 1, 2, 4 and 6.

The Cheetah family of 10,000-rpm 3.5-inch drives has also won awards for its blazing performance capabilities. This family includes the newly-announced Cheetah 36, a half-height drive that packs 36.4 Gbytes of formatted capacity into the 3.5-inch form factor. The Cheetah family also includes the 9.1- and 18.2-Gbyte Cheetah 18LP models. The Cheetah 36 and Cheetah 18LP are designed with 80 Mbyte/sec Ultra2 SCSI and 200 Mbyte/sec FC- AL interface options.

Seagate is a RAID partner, not a competitor

Seagate's technological and quality leadership in the RAID marketplace has made its drives a solid reputation among RAID subsystems manufacturers. Seagate is not a manufacturer of RAID arrays itself because the company does not compete with customers or potential customers at the subsystem level. Seagate's goal in the RAID market is to maintain its status as the premier supplier of quality, high-performance discs to makers of RAID subsystems. Seagate's Barracuda and Cheetah disc drive families are recognized in the industry as solid RAID foundations.

As the sea of digitized information continues to swell, technologies like Seagate's disc drives, RAID will play an ever more important role in capturing, managing and providing access for millions of users in the years to come.

By Tyson Heyn

June, 1995

Updated November, 1998

Copyright 1995, 1998 Seagate Technology, Inc. All rights reserved.