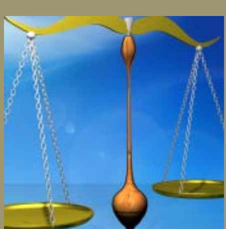
Managing the information that drives the enterprise

STORAGE

SEARCHSTORAGE.CO.UK

ESSENTIAL GUIDE TO

RAID Configuration



Evaluate the myriad options for protecting your data with RAID.

INSIDE

- Understanding key decisions in a RAID implementation
- Next generation of RAID: Innovation, transformation, paradigm shift
- RAID level comparison: Performance vs. protection

TechTarget

RAID: Basic (but not simple) data protection

DATA LIES AT the heart of the organisation, and data protection takes many forms—backup, archiving and the subject of this SearchStorage.co.UK *Essential Guide to RAID Configuration*.

The Redundant Array of Independent Disks was possibly the first-ever form of storage virtualisation. By taking many disks and aggregating them to present as virtual volumes, storage arrays were able to boost I/O performance and increase levels of data protection against drive failure.

In fact, by means of striping, mirroring and parity (all explained in this *Essential Guide*), the varying levels of RAID are able to multiply by many times the resilience and performance of a collection of single hard drives, adding up to much more than a sum of the parts.

Located at the heart of all data storage—the drive array—RAID is the first line of defence against mechanical disk failure, and for this reason knowledge of RAID is a fundamental for all storage professionals.

At the same time the many RAID levels provide numerous possible permutations of performance, data protection and cost, and the challenge for storage professionals is to know which is best-suited to their environment.

Located at the heart of all data storage—the drive array—RAID is the first line of defence against mechanical disk failure, and for this reason knowledge of RAID is a fundamental for all storage professionals.

The various RAID levels offer multiple choices when configuring storage arrays: striping data across many drives, dividing files by block or even at byte level; mirroring sets of drives; mirroring striped sets; striping mirrored sets; and using parity data—on a dedicated disk or disks or striped across many—to enable drive rebuilds in case of a disk failure.

The profusion of possible combinations of RAID characteristics, which all have different impacts on performance, resilience and cost, plus the fundamental importance of RAID to storage array management, is the reason we've compiled this Essential Guide for SearchStorage.co.UK readers. Our RAID content is perennially popular. Here you can learn the fundamentals of storage's most basic form of data protection in one handy guide. ⊙

Antony Adshead is the bureau chief for SearchStorage.co.UK.

RAID implementation choices

Learn about the key decisions that you'll confront when implementing RAID, including whether a software or hardware approach makes more sense for your organisation, as well as which level is best.

By ANTONY ADSHEAD

When implementing RAID, storage professionals face a lot of choices. In this article we examine key decisions that need to be made: whether to opt for hardware or software RAID; how to select RAID levels based on cost and performance; and, by way of a more detailed example, how to decide whether RAID 6 or RAID 10 is best suited to your needs.

SOFTWARE RAID VERSUS HARDWARE RAID: PROS AND CONS

RAID takes multiple disk drives and creates arrays that are resilient and highly available by mirroring and striping data across them. RAID can also incorporate the means to recover from disk failure using parity data.

The different ways that mirroring, striping and parity are used define the different RAID levels. Processing is required to carry out those actions, and that can take place on the host server's OS (software RAID) or in the storage array or controller (hardware RAID).

Software RAID. Disks attached to servers can be turned into RAID arrays using the built-in features of a number of operating systems. This is software RAID. All you need to do is connect the drives and configure the RAID level you want.

Software RAID does its processing using the server's resources, which adds to the processing load and could slow down RAID calculations and other operations carried out on that device. RAID 0 and RAID 1 place the lowest overhead on software RAID, but adding the parity calculations present in other RAID levels is likely to create a bigger impact on performance.

Numerous server OSes support RAID configuration, including those from Apple, Microsoft, various Linux flavours as well as OpenBSD, FreeBSD, NetBSD and Solaris Unix. Herein lies another difference: With software RAID you'll be restricted to the RAID levels your OS can support.

Software RAID is often specific to the OS being used, so it can't generally be used for partitions that are shared between operating systems.

Hardware RAID. With hardware RAID, because the processing work is done on a discrete controller card in the server or at the level of the storage subsystem, there's no added load to the server processor and buses. There will likely be more advanced features, such as drives being hot-swappable in case of failure. Hardware RAID is more expensive than software RAID but offers better performance and interoperability.

Whether software RAID or hardware RAID is the one for you depends on what you need to do and how much you want to pay. Hardware RAID will cost more, but it will also be free of software RAID's performance limitations.

WHICH RAID LEVEL HAS THE BEST PRICE-TO-PERFORMANCE COMPARISON?

RAID levels are defined by combinations of the following attributes: a) mirroring data between sets of drives; b) striping data across drives; c) striping mirror sets or mirroring striped sets; and d) using parity data to enable disk rebuilds. But a key aim of RAID is data protection, so if we take that goal into account alongside performance characteristics, what are the price-to-performance comparisons of the various RAID configurations?

RAID 0. In terms of raw price/performance, RAID 0 has the lowest cost. All disk space is used to store data; none is used for mirroring or parity data. Performance is good in terms of I/O, as data is striped across disks and there's no overhead created by parity calculations.

But RAID 0 offers the least data protection. If a disk fails, you'll have to accept the loss of that data or the drive being down while you restore it from other media.

RAID 1 and 10. RAID 1 offers excellent data protection by mirroring data between two identical sets of disks. Because of this duplication, you're immediately paying double for your useable capacity—if you have 4x 500 GB disks in a RAID 1 set, you'll only get 1 TB useable capacity, as half are used up for mirrored data. Also, RAID 1 doesn't stripe, so you'll lose out on performance compared with RAID 0.

By mirroring sets in a striped set of disks, RAID 10 gives the data protection of RAID 1 as well as the striping—and therefore performance—of RAID 0.

By mirroring sets in a striped set of disks, RAID 10 gives the data protection of RAID 1 as well as the striping—and therefore performance—of RAID 0. However, as with RAID 1, you pay for twice as much raw capacity than you can actually use.

RAID 5. RAID 5 stripes data and protects it by distributing parity data across all disks. Because no extra disks are occupied with mirroring, RAID 5 costs are immediately less than those for RAID 1 and its derivatives. Data protection is good because parity enables drive rebuilds, but performance takes a hit because of the processing overhead in calculating parity data.

Price-to-performance comparison results. While RAID 0 is the least costly, it cannot be used for databases unless you can withstand data loss or rebuild adequately from other backed-up media.

RAID 1 and 10 win on data protection but lose in terms of disk costs. RAID 10 offers the best performance and data protection, but at a cost.

RAID 5 offers the best tradeoff in terms of price and performance and includes data protection for database use.

RAID 6 VS RAID 10

When comparing RAID 6 vs RAID 10, many storage managers have a hard time deciding between the dual parity of RAID 6 and the mirrored data of RAID 10.

RAID 6 stripes data across disks and calculates dual distributed parity. Distributed parity provides fault tolerance against two drive failures. Dual parity means that while a failed disk is being rebuilt the array is still protected by the remaining parity data.

RAID 1+0 (RAID 10) is mirrored sets in a striped set. RAID 1+0 creates a striped set from subsets of mirrored drives. If disks fail, RAID 1+0 allows all the remaining disks to continue in use. The array can suffer multiple drive failures as long as no mirror set loses all of its drives.

To help you choose the RAID level that best meets the needs of your organisation, let's take a look at some of the advantages of RAID 6 vs RAID 10.

Because RAID 10 mirrors everything, an array requires double the disk capacity of the data to be stored.

RAID 6 gives more useable capacity the more disks you add. Because RAID 10 mirrors everything, an array requires double the disk capacity of the data to be stored. The remainder of the capacity constitutes the mirror. If a RAID 6 array comprises four disks, only 50% of that space is available as useable capacity, but the proportion of useable space increases as you add more drives. That means half the total capacity of a RAID 10 array will always be dedicated to protection, but with a RAID 6 array the useable capacity grows as the number of drives increases. For example, if you increased the number of disks in a RAID 6 array from four to eight, the space consumed by parity data would decrease from 50% to 25%.

RAID 6 requires more processing power. RAID 6 makes two parity calculations for each write operation, so it's slower to write than most other RAID levels.

RAID 6 can always protect against two simultaneous disk failures. Because RAID 6 doubles up its parity data, it can withstand two disks failing at the same time. Whether RAID 10 can handle two disk failures simultaneously depends on where they occur. If both the disks that fail are located in the same mirror, the other set can step in. You will lose all data if the same disks in both mirrors fail within the rebuild window (which should be relatively short, however).

RAID 10 rebuild times are faster. RAID 10 has among the fastest rebuild times possible because it only has to copy from the surviving mirror to rebuild a drive, which can take as little as 30 minutes for drives of approximately 1 TB. The key drawback of RAID 6 (vs RAID 10) is that the time it takes to rebuild the array after a disk failure is lengthy because of the parity calculations required, often up to 24 hours with even a medium-sized array.

RAID 10 doesn't need special hardware. Most controller hardware will support RAID 10 with good performance. Because RAID 6 doubles the parity calculations for every write, it requires specially designed controller hardware. ⊙

Antony Adshead is the bureau chief for SearchStorage.co.UK.

RAID: Basic Data Protection
Implementation Choices
Alternatives to RAID
Rating RAID Levels 0 to 6

RAID: Basic Data Protection

Implementation Choices

Alternatives to RAID

Rating RAID Levels 0 to 6

ALTERNATIVES TO RAID

The various forms of RAID have been around for a long time and have done a good job of protecting data. But high-capacity drives and new performance demands have spurred development of RAID alternatives.

By MARC STAIMER

REDUNDANT ARRAY OF INDEPENDENT DISKS (RAID) has been the standard for disk-based data protection since 1989 and is a proven and reliable method that's considered a basic data storage building block. Basic storage principles tend to change very slowly, and, despite its popularity and track record, change is coming to RAID.

To gain more insight into why alternatives to RAID might be appealing requires some understanding about RAID and the growing problems with the technology.

RAID SHORTCOMINGS IN THE 21ST CENTURY

The purpose of RAID is to protect data in the event a hard disk drive (HDD) fails. When that failure occurs, data from that failed HDD (or multiple HDDs) is re-created from parity data or copied from a mirror, depending on the type of RAID in use. Disk drives are electro-mechanical devices that have the highest probability of failure and the lowest mean time between failures (MTBF) of any storage system component.

It takes a lot of HDDs to keep up with the rapid growth of data that analyst firms like IDC, Gartner and Enterprise Strategy Group peg at somewhere between 50% and 62% per year. Statistically speaking, more hard disk drives mean more HDD failures. Disk drive manufacturers have continually increased HDD density, and today we have 2 TB SATA and will likely have 4 TB by the end of this year. Even high-performance SAS and Fibre Channel (FC) drive capacities are pushing 600 GB. RAID problems quickly become evident when a rebuild is required with such increasingly dense drives.

RAID: Basic Data Protection

Implementation Choices

Alternatives to RAID

Rating RAID Levels 0 to 6

Each RAID type has tradeoffs in write performance, read performance, level of data protection, speed of data rebuilds and the usable storage on each hard disk drive. For example, if guaranteeing data availability is the top priority, then some variation of mirroring or multiple mirrors (RAID 1, 10, triple mirror, etc.) will be required. Having full copies of the data on other HDDs or RAID sets simplifies protection and recovery of the data but at a severe and tangible cost because each mirror reduces usable storage by the same amount of the original data. In addition, system resources are required for every copy, which can impair I/O performance. Realistically, most organisations aren't this overprotective; most use RAID 5 and/or RAID 6.

When a HDD fails in a RAID 5 set, the system rebuilds the data on a spare drive that replaces the one that failed. The storage system then analyses every sector on every HDD in the RAID set to reconstruct the data. Such heavy utilisation of the other HDDs in the RAID set increases the likelihood of another HDD failure (usually a non-recoverable read error) by an order of magnitude, and this significantly increases the likelihood of data loss.

Ten or 20 years ago, when disk capacities were much lower, rebuilds were measured in minutes. But with disk capacities in the terabytes, rebuilds can take hours, days or even weeks.

Ten or 20 years ago, when disk capacities were much lower, rebuilds were measured in minutes. But with disk capacities in the terabytes, rebuilds can take hours, days or even weeks. If application users can't tolerate the system performance degradation that rebuilds cause, the rebuild is given a lower priority and rebuild times increase dramatically. Longer data reconstruction times typically equate to a significantly higher risk of data loss. Because of this, many storage shops are stepping up their use of RAID 6.

RAID 6 provides a second set of striped parity blocks that protect the data even if two HDDs fail or suffer a non-recoverable read error in the RAID set. The risk of data loss drops dramatically, but the extra stripe consumes additional usable capacity and system performance takes a bigger hit if two drives must be reconstructed simultaneously from the same RAID group. More disturbing is the increased risk of data loss if a third HDD fails or a non-recoverable read error occurs during the rebuild.

RAID: Basic Data Protection

Implementation Choices

Alternatives to RAID

Rating RAID Levels 0 to 6

There are other RAID issues such as "bit rot" (when HDDs acquire latent defects over time from background radiation, wear, dust, etc.) that can cause a data reconstruction to fail. Most storage systems include some type of background scrubbing that reads, verifies and corrects bit rot before it becomes non-recoverable, but scrubbing consumes system resources. And higher capacities mean more time is needed to scrub.

Another onerous RAID issue is that of documenting the chain of ownership when replacing a failed HDD. This comprises the documented trail (who, what, where, when) of the failed HDD from the time it was pulled to the time it was destroyed or reconditioned. It's a tedious, manually intensive task that's a bit less stringent if the HDD is encrypted.

Even more frustrating is that the vast majority of failed HDDs sent back to the factory for analysis or reconditioning (somewhere between 67% and 90%) are found to be good or no failure is found. Regrettably, the discovery happens after the system failed the HDD, the HDD was pulled, the data was reconstructed and the chain of ownership documented. That's a lot of operational pain for "no failure found."

Solid-state storage devices actually exacerbate the aforementioned RAID problems. Because solid-state drives (SSDs) can handle high-performance applications, they allow for storage systems with fewer high-performance HDDs and more high-density, low-performance hard disk drives. Tom Georgens, NetApp's CEO, recently noted, "Fast access data will come to be stored in flash with the rest in SATA drives." Lower cap-ex and op-ex for the system can end up translating into higher op-ex because of the increase in RAID problems.

> "Fast access data will come to be stored in flash with the rest in SATA drives."
>
> —Tom Georgens, CEO, NetApp

These RAID issues have inspired numerous vendors, academics and entrepreneurs to come up with alternatives to RAID. We can categorise those innovative alternatives into three groups: RAID + innovation, RAID + transformation and paradigm shift.

RAID + INNOVATION

Several vendors have addressed traditional RAID problems by taking an incremental approach to RAID that leverages its reliability while diminishing some of its shortcomings. IBM's EVENODD (implemented by EMC on Symmetrix DMX) and NetApp's RAID-DP (implemented on NetApp's FAS and V-series) have enhanced RAID 6 by reducing algorithm overhead while increasing performance.

NEC's RAID-TM, or triple mirror (implemented in its D-Series systems), aims to solve RAID 1 data loss risk if both the primary and mirror drive fail or if there's a non-recoverable read error. RAID-TM writes data simultaneously to three separate HDDs so if two HDDs fail or there are unrecoverable read errors in the same mirror, the app still has access to its data with no degradation in performance even while the drives are rebuilt. The advantage is performance; the disadvantage is far less usable capacity.

RAID-X is an IBM XIV Storage System innovation that uses a wide stripe to reduce RAID tradeoffs of performance and data loss risk.

RAID-X is an IBM XIV Storage System innovation that uses a wide stripe to reduce RAID tradeoffs of performance and data loss risk. It's basically a variation of RAID 10 that uses intelligent risk algorithms to randomly distribute block mirrors throughout the entire array. This approach allows XIV to reconstruct the data on very large, 2 TB HDDs in less than 30 minutes. As with all mirroring technology, the tradeoff is reduced usable capacity.

Hewlett-Packard's LeftHand Networks and Pivot3 provide similar variations of Network RAID for their x86-based clustered iSCSI storage. Network RAID leverages the concept of RAID but uses storage nodes as its lowest component level instead of disk drives. This allows it to distribute a logical volume's data blocks across the cluster with one to four data mirrors depending on the Network RAID level. Ongoing block-level, self-healing nodal health checks allow Network RAID to copy and repair the data to another node before a failure occurs. This decreases the probability of a hard disk drive fault or non-recoverable read error causing a performance-sapping rebuild; but like all mirroring technology, it reduces the amount of usable storage.

These are just some of the RAID + innovation technologies. Others are currently incubating, including proposals for RAID 7 (triple parity and more) or TSHOVER (triple parity).

RAID + TRANSFORMATION

There are also RAID alternatives that attempt to re-invent RAID. They typically use RAID and are layered on top of it in some way. The concept is to keep what's good about RAID and fix the rest.

RAID ENHANCEMENTS

			Max. HDD failure protection	Vendors
RAID + Innovation	RAID 6 EVENODD	An IBM innovation that uses only two additional redundant HDDs and consists of simple exclusive-OR computations. The advantage of EVENODD is that it only requires parity hardware, which is typically present in standard RAID 5 controllers. This reduces the number of exclusive-OR operations over the more common Reed-Solomon computations at approximately 50% (based on 15 drives). EVENODD has similar performance issues of RAID 6 when dual hard disk drive (HDD) rebuilds are occurring. One advantage of EVENODD is that it's based on XOR. A disadvantage is that it has a few hot spots in certain diagonal blocks that cause very poor short write performance.	2	EMC
	RAID-DP, or Row-diagonal Parity	Stores row parity across the HDDs in a RAID 4 group; the additional parity HDD stores diagonal parity across the HDD in a RDP group. The two RDP parity stripes provide data protection in the event of two HDD failures occurring within the RAID group. Performance is nearly equal to single-parity RAID 4 or RAID 5. Higher performance than standard RAID 6 but with similar performance issues when concurrently rebuilding two HDDs.	2	NetApp
	RAID-TM, or RAID Triple Mirror	RAID-TM delivers the high speed of RAID 1 while providing the high reliability and double HDD fault protection of RAID 6. RAID-TM writes data simultaneously to three separate HDDs. Even with two HDD faults or unrecoverable read errors in the same mirror, the application still has access to its data with no degradation in performance even while drives are rebuilt.	2	NEC
	RAID-X	RAID-X doesn't require a spare HDD, just spare capacity on existing HDDs in the storage system. The objects can be mirrored between any two types of HDDs (no need to match drive size or speed). Rebuild performance is extremely fast because data is mirrored. This is a variation of RAID 1 or RAID 10, but with the added protection of random distribution. A second drive failure can result in data loss that can only be mitigated with additional mirrorings. Useable storage can be restricted depending on the number of mirrors (minimally half).	2 plus, depending on mirror #	IBM (XIV)
	Network RAID	Lays out a logical volume's blocks across the cluster, providing reliability configurable on a per-volume basis to best meet the needs of each application's data. Depending on a logical volume's Network RAID level, one to four copies of each block are striped across the cluster. A volume's RAID level can be changed (auto-restriping) without data availability interruption. Also provides proactive block-level self-healing to decrease probability of a non-recoverable read error. However, each block copy reduces the amount of useable storage.	4	GridStor, Hewlett-Packard LeftHand, Pivot3
RAID + Transformation	Self-healing storage	Also know as heal-in-place storage. Uses series of automated repair sequences designed to eliminate or mitigate the majority of HDD failures and unnecessary RAID data rebuilds. Isolates HDD sectors it can't fix and rebuilds only the data lost on those sectors. More expensive upfront than traditional RAID, but with a much lower total cost of ownership. Heal-in-place storage requires a pool of unused HDDs for the fail-in-place capability.	2	Atrato, DataDirect Networks, Panasas, LSI, NEC, 3PAR and Xiotech
	BeyondRAID	BeyondRAID is essentially a virtualization engine on top of RAID that chooses the correct RAID algorithm based on the data protection required. It writes blocks that can actually alternate between data protection methodologies. If more storage capacity is required, additional HDDs can be inserted or small HDDs can be replaced with larger ones. Simple administration allows switching from single- to dual-disk redundancy with a single click. Protects against dual drive failures and adds transparent automatic data healing. It's data-aware, allows for mixed drive sizes, drive reordering, proportional rebuild time and self-management. Only available for small systems of up to eight drives.	2	Data Robotics
Paradigm Shift Alternatives to RAID	Erasure codes	Also known as a form of forward error correction (FEC), erasure coding adds additional information to a stored object that allows any data set to be completely resurrected from a subset of the total information. Multiple slices (storage objects) or subsets of a data set are distributed across multiple storage or server nodes. Additional information attached to a stored object equals greater resiliency of the data set, protecting against larger numbers of components (disk drives, storage objects or server nodes) that can be lost and still recover the complete data set. The additional information on each storage object also reduces the amount of useable storage. The biggest issue with erasure code-based storage is reduced write performance, especially small writes.	Depends on amount of additional object information	Cleversafe, EMC, NEC

Examples of transformation technologies include self-healing storage and BeyondRAID.

Self-healing storage. Xiotech's Intelligent Storage Elements (ISE) is a good example of self-healing storage. ISE tightly integrates RAID and HDDs and combines them into a single storage element.

Xiotech engineered ISE to resolve most RAID rebuild issues by eliminating 67% to 90% of the rebuilds. It starts by reducing HDD faults by proactively healing hard disk drives before a fault occurs using similar HDD reconditioning algorithms employed by the factory. It also uses advanced vibration controls and sealed systems called DataPacs to reduce the possibility of outside influences that cause HDD faults. When a fault does occur, it reacts by providing remedial component repair within the sealed DataPac using methods similar to those the original manufacturer uses. It analyses power cycles, recalibrates components, remanufactures the HDD and migrates data when required to other sectors or HDDs. If the fault persists, ISE isolates only the non-recoverable sectors and then initiates data reconstruction only for the faulty HDD sectors.

Xiotech engineered ISE to resolve most RAID rebuild issues by eliminating 67% to 90% of the rebuilds.

So, there are far fewer rebuilds, and when one is required there's much less to reconstruct. In addition, it's all automated so no manual intervention to pull failed drives is required. The result is equivalent to a factory-remanufactured HDD with only the components that are beyond repair taken out of service. The downside to this transformational technology is that it has higher up-front costs, although it lowers the total cost of ownership (Xiotech provides a five-year warranty).

Atrato's Velocity1000 (V1000) uses a self-healing technology called Fault Detection, Isolation Recovery (FDIR) in combination with Atrato's Virtualization Engine (AVE). FDIR watches component and system health and adds self-diagnostics and autonomic self-healing, but it doesn't attempt to remanufacture or recondition HDDs in place as Xiotech does.

Atrato puts 160 2.5-inch SATA drives in a 3U system called SAID (self-maintaining array of independent disks). The company uses its extensive SATA drive performance database of operational reliability testing (ORT) to monitor the installed drives' actual performance to detect SATA HDD deviations.

Atrato also deals with HDD faults by first attempting to repair the faulting HDD sectors (although not with manufacturer-level reconditioning, remanufacturing or component recalibration). If the fault or non-recoverable read error can't be repaired, the sector is isolated and only the affected data is reconstructed and remapped to virtual spare capacity. If a disk drive completely fails, it is reconstructed and remapped to the virtual spare capacity. Atrato reduces the number of rebuilds and rebuild times by reconstructing only affected data on virtual drives. Atrato backs its technology with a three-year warranty.

DataDirect Networks' DDN S2A technology heal-in-place approach to disk failure attempts several levels of HDD recovery before a hard disk drive is removed from service.

DataDirect Networks' DDN S2A technology heal-in-place approach to disk failure attempts several levels of HDD recovery before a hard disk drive is removed from service. It begins by keeping a journal of all writes to each HDD that show behavior aberrations and then attempts recovery operations. When recovery operations succeed, only a small portion of the HDD requires rebuilding using the journaled information so rebuild times are reduced and a service call may be avoided.

Panasas' ActiveScan technology continuously monitors HDDs and their contents to detect problems. ActiveScan monitors data objects, RAID parity, disk media and disk drive attributes. When a potential problem is detected, data is moved to spare blocks on the same disk. Future HDD failure is predicted through the use of HDD self-monitoring analysis and reporting technology (SMART), which permits action to be taken to protect data before a failure occurs. When a HDD failure is predicted, user-set policies pre-emptively migrate data to other HDDs, which eliminates or mitigates the need for reconstruction.

LSI and NEC both detect HDD sector errors while allowing operations to continue with the other drives in the RAID group. If an alternative sector can be assigned, the HDD is allowed to return to operation, avoiding a complete rebuild. Performance is maintained throughout the detection and repair process. This is a limited self-healing technology that reduces the number of rebuilds and helps maintain performance.

3PAR's InSpire Architecture is engineered to sustain high performance levels by leveraging advanced HDD error isolation to reduce the amount of data that requires reconstruction and by taking advantage of its massive parallelism to provide rapid rebuilds (typically fewer than 30 minutes). The system uses "chunklets" in their many-to-many drive relationships. That same massive parallelism allows 3PAR to isolate RAID sets across multiple drive chassis to minimize the risk of data loss if a chassis is lost.

BeyondRAID. Data Robotics' BeyondRAID sits on top of RAID and makes it completely transparent to the administrator. It transforms RAID from a deterministic offline process into one that's online and dynamic. Essentially self-managing, BeyondRAID chooses RAID sets based on the required data protection at any given point in time.

Data Robotics Inc.'s BeyondRAID sits on top of RAID and makes it completely transparent to the administrator.

But it's how BeyondRAID addresses RAID issues that truly makes it stand out. It protects against one or two HDD failures and has built-in automatic data self-healing (not storage self-healing). Data blocks are spread across all drives so data reconstruction is very fast. Because the system is "data aware," it allows for different drive sizes, drive re-ordering and proportional rebuild times. Because it tops out at eight SATA drives, it is most appealing for small and medium-sized businesses (SMBs), but it is a true fire-and-forget storage system.

RAID PARADIGM SHIFT: ERASURE CODES

Erasure codes are designed to separate data into unrecognisable chunks of information with additional information added to each chunk that allows any complete data set to be resurrected from some subset of the chunks. The chunks can be distributed to different storage locations within a data centre, city, region or anywhere in the world.

Erasure codes have built-in data security because each individual chunk contains insufficient information to reveal the original data set. A large enough subset of chunks from the different storage nodes is needed to fully retrieve the total data set, with the number of required chunks determined by the amount of additional information assigned to each chunk. More additional information means fewer chunks are required to retrieve the whole data set.

Erasure codes are resilient against natural disasters or technological failures because only a subset of the chunks is needed to reconstitute the original data. In actuality, with erasure codes there can be multiple simultaneous failures across a string of hosting devices, servers, storage elements, HDDs or networks, and the data will still be accessible in real time.

Also known as forward error correction (FEC), erasure coding storage is a completely different approach to RAID. Erasure codes eliminate all the RAID issues described here. It's a new approach and at this time only three vendors have erasure code-based products: Cleversafe's dsNet, EMC's Atmos and NEC's HYDRAstor.

Erasure codes appear to be better suited for large data sets than smaller ones. It's especially appropriate for cloud or distributed storage because it never has to replicate a data set and can distribute the data over multiple geographic locations.

RAID'S EVOLUTION

The issues with traditional RAID are well known and are escalating with higher disk capacities. The RAID alternatives described here address many of those problems, and more new approaches are on the way. Selecting the best fit for a particular environment requires research, testing, pilot programmes, patience and a willingness to take a risk with a non-traditional approach. ⊙

Marc Staimer is founder and senior analyst at Dragon Slayer Consulting in Beaverton, Ore., a consulting practice that focuses on strategic planning, product development and market development for technology products. You can contact him at marcstaimer@mac.com.

RAID: Basic Data Protection

Implementation Choices

Alternatives to RAID

Rating RAID Levels 0 to 6

Examining RAID levels: RAID 0 through RAID 6

Compare RAID levels and learn about each one's tradeoffs in performance and data protection.

By ARUN TANEJA and ANTONY ADSHEAD

BY COMBINING PHYSICAL DRIVES and presenting them as a single hard drive to the operating system, RAID technology allows storage pros to store the same data in different places on multiple disks. I/O operations can therefore be shared between drives, which means performance can increase and data storage protection can improve. For organisations considering a RAID deployment, there are multiple factors that need to be looked at, particularly the available levels of RAID technology and the specific needs of their data storage infrastructure.

Let's look at the functions of and differences between the various RAID levels.

RAID 0: Data striped across multiple drives

Technically speaking, RAID 0 is not a RAID level because it lacks redundancy, but it is customarily viewed as RAID. For example, if you have three disk drives, instead of writing everything to the first disk drive, the data is striped between them. So, the first chunk of the file is placed on disk one, the next chunk goes on disk two and the final chunk on disk three.

With such striping you have broken the file data down into pieces and placed them in the first, second and third arrays. So, instead of reading the data from one disk drive, you now read the data in parallel from all three disk drives and combine the data on the other end. Essentially, you end up leveraging the performance of three disk drives, and therefore access to that file is vastly improved.

So, RAID 0 offers a performance boost over a single drive because you have three disk drives being read in parallel and pumping data back at you. Therefore, with RAID 0, performance in terms of throughput of data is good, but its resilience is fragile. If one of those three disk drives dies, the whole file is ruined because you have a chunk missing.

RAID 1: Mirroring and performance improvements

Instead of striping, RAID 1 utilises mirroring, which is where you have two disk drives and whatever is written to disk one is simultaneously written to disk two. The idea is that if one of those two disks dies, you have another disk that is still working and therefore some resilience against complete loss of data should a drive fail.

You also achieve performance improvements with RAID 1. When everything is functioning correctly—i.e., both disk drives are spinning, behaving properly and you're reading data from both of them—your read performance will effectively double. Having said that, write speed is not improved by RAID 1 as writes take place to a single disk, unlike the simultaneous writes of data to many disks as in the striped RAID 0.

RAID 0+1 and RAID 1+0 (RAID 10)

It is quite common to combine the characteristics of RAID 0 and RAID 1 by creating two mirrors that are both striped (RAID 0+1) or two mirrored sets with data striped between them (RAID 10)

In RAID 0+1, the array will continue to operate if one or more drives in the same mirror set fail. But, but if drives on both mirror sets fail, all the data is lost.

If disks fail, RAID 1+0 performs better than RAID 0+1 because all the remaining disks continue in use. The array can suffer multiple drive failures as long as no mirror set loses all drives.

RAID 3: Byte-level parity

RAID 3 sees data striped across disks at the byte level (i.e., smaller than a block) with data protection taken care of by the addition of a parity disk. For example, if we were to use five disk drives in a RAID 3 group, four of those drives would be data disks and the fifth a dedicated parity disk. If any single drive fails, the RAID system will rebuild the lost data by referring to the other four disk drives that are still working.

Byte-level striping spreads a block across several disks and so often requires that all disks are read to extract a single block. This means RAID 3 isn't very good at random access and is best suited to large, sequential workloads, such as video files.

RAID 4: Block-level parity

RAID 4 is similar to RAID 3 in that a dedicated parity disk drive is always part of the RAID set. The difference is that striping takes place at the block level.

Because RAID 4 stripes at the block level, it is more suited to random access reads because blocks often reside on different disks that can be accessed simultaneously, as long as the array controller is capable of doing so.

Writes to RAID 4 can be subject to a heavy parity overhead as potentially data can be written to multiple disks simultaneously, but parity data for all those writes must go to the parity disk.

RAID 5: No dedicated parity

RAID 5 also uses block-level striping with single blocks of parity data spread across all drives and is probably the most popular RAID level. All five disk drives have a combination of data and parity.

If a single disk dies, you can re-create that information from the remaining drives. This means you can tolerate a single disk drive failure and performance is good because you are striping the data.

RAID 6: Tolerates failure of two disk drives

RAID 6 is similar to RAID 5 in terms of striping and parity across all drives in the set, with the major difference being that RAID 6 can tolerate two disk drives failing. That's because RAID 6 writes two blocks of parity data for each block of data written. ⊙

Arun Taneja is senior analyst and founder of Taneja Group. Antony Adshead is the bureau chief for SearchStorage.co.UK.