

EtherDrive[®] Storage

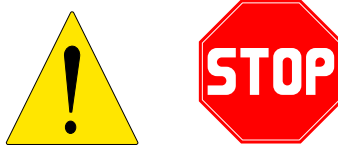


CLUSTERED MODULAR STORAGE

General Best Practices for Networked Storage

Coraid Staff

June 2008



IMPORTANT WARNING

**Disk redundancy (RAID) and high availability
are NO SUBSTITUTE FOR BACKUPS.**

**RAID does not preserve past versions of your files
and does not protect from disasters that
damage the entire disk server.**

NEVER BE CAUGHT WITHOUT BACKUPS!

Contents

Environment, Power, and Maintenance	5
The physical environment	5
Building load capacity.....	5
Cabling	5
Air conditioning	5
Electric power	6
Circuit load capacity	6
Lightning protection and surge suppression.....	6
Backup power.....	6
UPS best practices.....	7
Maintenance	8
Firmware updates	8
Spare parts	8
Documentation	8
Disk Performance.....	10
How disks work	10
Tracks and sectors.....	10
Latency and throughput	11
Sequential vs. random access	11
Using multiple disks	11
JBOD	11
RAID.....	12

RAID performance..... 12

Volume allocation 13

Tuning for specific application software 13

Caching 14

Network connections and components..... 15

Disk and RAID Reliability.....17

Disk reliability.....17

Effect of RAID on reliability 18

Networking improves fault tolerance..... 19

Type of disk doesn't matter 19

Environment, Power, and Maintenance

The physical environment

Building load capacity

An often overlooked element of system design is the weight that the product exerts on the rack and the supporting floor beneath it. Each disk shelf populated with drives can weigh more than 100 lbs. (45 kg). Make sure your installation will not exceed the load capacity of the floor (rated in pounds per square foot or kilograms per square meter). Neglecting this requirement can cause structural failure and loss of life.

Cabling

Gigabit Ethernet requires Cat 6 cables (not Cat 5 or 5e). Use high-quality cables and label each cable at both ends.

Air conditioning

All electronic equipment, including networked storage, must operate in an environment with controlled temperature and humidity. Even in mild climates (Seattle, San Francisco) where air conditioning is not normally needed in residences, computers require air conditioning because they give off heat of their own and because they must keep running without damage even on occasional hot, humid days.

Your HVAC contractor can assist with calculations to ensure adequate control over temperature and humidity. Ultimately, most of the electric power entering a computer system is dissipated as heat, so the electric power consumption provides a useful guide to the cooling requirements.

High-availability systems require redundant (backup) cooling systems.

In the event temperature and humidity cannot be maintained within the acceptable range, a controlled shutdown should be performed in order to protect equipment.

Electric power

Circuit load capacity

Do not view a 120-volt AC outlet as something that can be used indiscriminately. The outlets in your building are grouped into circuits, and each circuit is rated for a maximum of 15 to 20 amps. (Volts measure the pressure that makes electricity flow; amps measure the actual rate of flow, which depends on what is connected to the circuit.)

Rather than wait for circuit breakers to trip, map out the circuits in your building, label each outlet with its circuit number, and add up the total amperage required by the devices on each circuit. Amp ratings are generally printed on a label near where the power cord enters each piece of equipment. Many devices, especially disk drives, require more amps to start up than to continue running.

One shortcut for mapping circuits is to use a digital voltmeter. If two outlets show *exactly* the same voltage (to within 0.1 volt), they are probably on the same circuit. A circuit with an unusually low voltage is overloaded, but most overloads do not produce this symptom. A more sophisticated way to trace circuits involves a test instrument with a radio-frequency transmitter plugged into the outlet and a receiver that you can move around to trace the wires.

It goes without saying that UPSes and surge suppressors are no cure for overloaded circuits; they cannot output more energy than they take in.

Lightning protection and surge suppression

Install surge suppression devices at the utility ingress as a first line of defense against damage to your systems caused by voltage spikes from distant lightning and other sources. Your electrical contractor can recommend appropriate solutions.

NOTE: Surge protectors do not protect against direct lightning strikes, nor do equipment warranties cover lightning damage. Also, note that there is almost no way to tell whether a surge protector is in working order. Under normal conditions you can't tell whether the surges are being absorbed or are simply not present in the first place.

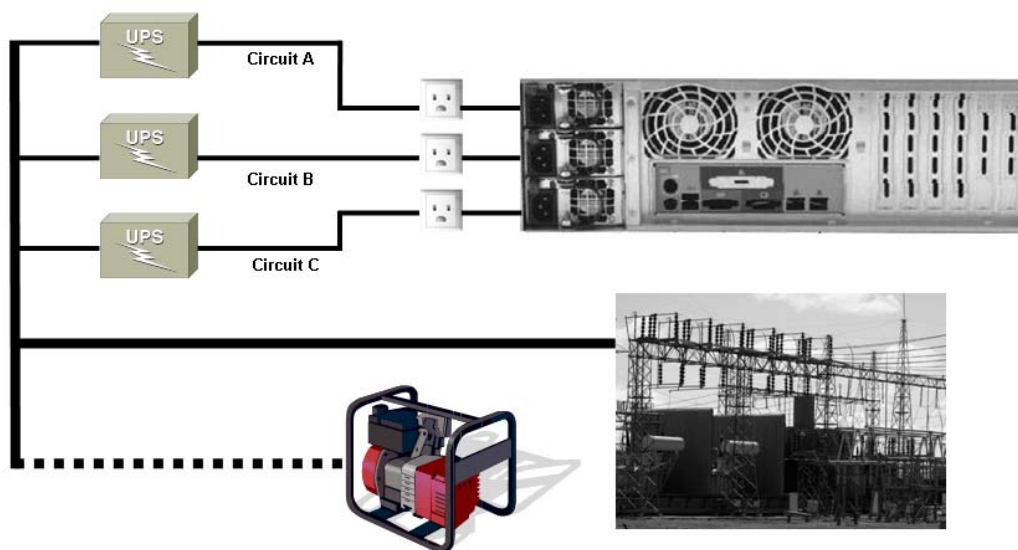
Backup power

Backup power, to protect against power failures, generally takes three forms:

- Multiple power supplies inside your equipment. If at all possible, these should be connected *to different circuits* of your building's wiring, so that if a circuit breaker trips, not all of them will be cut off. If connected to the same circuit, the

redundant power supplies only protect against their own failure, not that of the external power.

- Uninterruptible power supplies (UPSes) to protect against brief outages. A UPS contains batteries, a battery charger, an inverter to convert the battery power back to AC, and an automatic switch-over mechanism so that if the incoming AC power fails, AC will be generated from the batteries.
- Generators to supply power if the system must run during a utility power outage lasting more than a few minutes. *This is necessary if you require 24/7 availability.*



Redundant AC Power Configuration

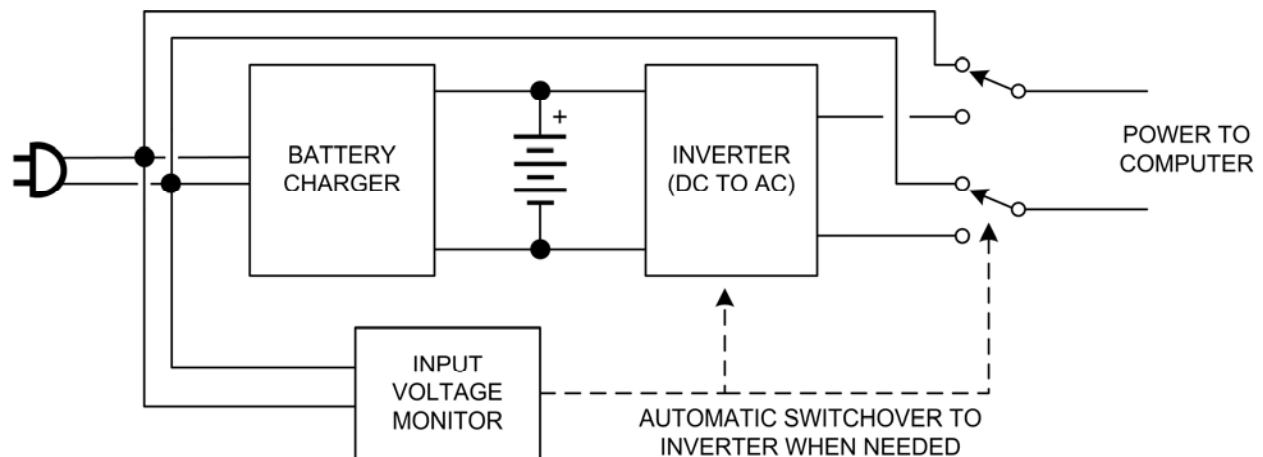
UPS best practices

As shown in the diagram on the next page, the purpose of a UPS is to power the computer from batteries for a few minutes during a brief power outage. Except when actuated by power loss, a UPS has no effect on the power to the computer, except that many UPSes also include some surge protection.

Be sure to size the UPS system to allow enough time for all critical systems to be powered down gracefully. Your UPS vendor can assist you with the relevant calculations. UPSes are generally rated in volt-amperes (VA, volts times amps) and are intended to run the system for 15 minutes or so (or appreciably longer with brand-new batteries).

The batteries in a UPS deteriorate with age. Other than by testing the run time of the UPS with an actual load, this deterioration is hard to detect, since the full-charge voltage does not change. Batteries that would run your computer for 15 minutes when new will probably not run it for 5 minutes when they are three years old. Replace the batteries regularly rather than waiting for failure.

Sudden battery failures also occur. A completely dead UPS usually needs nothing but new batteries, which need to charge for 12 hours after installation.



Inside a UPS

Maintenance

Firmware updates

Keep your storage appliances updated with the latest firmware. Do not wait for problems.

Spare parts

Maintain spare parts for high-failure-rate items such as hard drives, power supplies, and fans. If your application requires 24/7 uptime, keep spare parts on hand for all essential components.

Documentation

Do not keep your documentation in the technician's head. *Write everything down* so that a technician unfamiliar with your setup – or yourself, after two or three years of having your mind on other things – can work on the system efficiently.

Beware of employees who try to make themselves indispensable by failing to document systems, so that only they can fix them. Very soon, such a person forgets some details, and then *nobody* can maintain the equipment.

Keep documentation current. Out-of-date documentation can impede recovery efforts.

An important part of the documentation consists of labels affixed to hardware and cables. Keep these current, and above all, *do not let labels tell lies*. If a label gives false information, change it immediately.

Document the AC power system, including the amperage load on each circuit.

Document the connections of all system components to the network and to switch and router ports. Document the use of virtual networks (VLANs) and other logical structures.

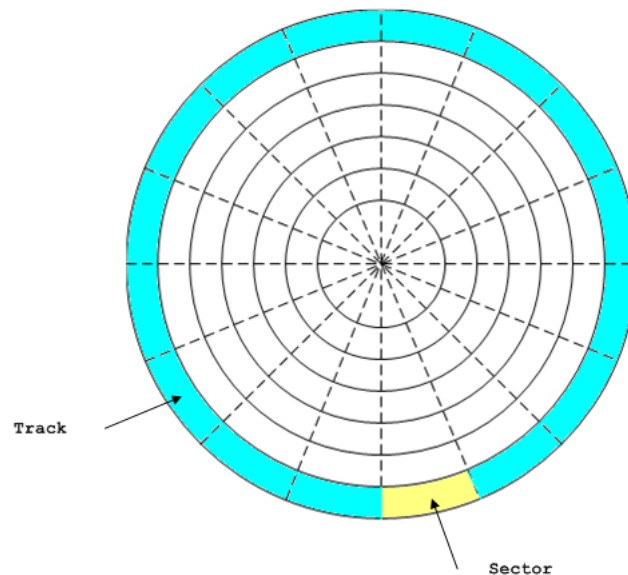
Read the documentation that comes with all your hardware. Then file it where any technician working on your system will be able to find it.

Disk Performance

How disks work

Tracks and sectors

Hard disks store data in small magnetic tracks on one or more disk platters coated with a thin magnetic film. Tracks are read and written by positioning a tiny electromagnet (the “head”) very near the disk platter directly over the track.



Disk Geometry

Latency and throughput

The time needed to retrieve data, after commanding the disk drive to read from the right location, is called **latency**. To retrieve data from the disk, the head must move to the right track, and then the disk must rotate until the right sector is under the head.

It follows that the latency of a disk drive depends on its rotational speed (RPM) and how quickly it can move the head to the right track (seek time).

The **throughput** of a disk drive is how fast it can transfer data to or from consecutive sectors. Throughput also benefits from higher RPM.

Sequential vs. random access

The trade-off between latency and throughput is likely to be different with networked storage or with large-scale servers than with a personal computer. In the days when a personal computer had one CPU running only one process at a time, the disk drive was often doing nothing but reading or writing one file in consecutive sectors. In that situation, throughput is high, and latency doesn't matter.

Nowadays, though, there is almost always concurrent demand for access to more than one location on the disk. This happens even with multitasking on a single computer (Windows 95 and up, MacOS, Linux, UNIX) and is greatly amplified on servers or storage appliances that serve multiple CPUs. The trend is for disk access to become more and more random (jumping around among non-adjacent sectors). Random access can be hundreds of times slower than sequential access.

Naturally, we can't allow two CPUs to write on the same file at the same time, or, worse, try to use the same disk sector for different purposes. In the absence of a cluster file system, hosts require exclusive access to a physical or logical disk volume. Consider, however, a single disk server partitioned into two logical disks for use by different CPUs. Each CPU thinks it has a disk all to itself, but in reality, if both CPUs are using the disk, the heads are likely to be moving back and forth constantly between two different regions.

Using multiple disks

JBOD

Multiple disks in a single disk server can alleviate the problem just described. The simplest configuration is JBOD ("just a bunch of disks") – each disk is served separately, either to different CPUs or to a single CPU as multiple drives. In that case, each disk works just as if it were installed in the computer that is using them, and they do not contend with each other.

RAID

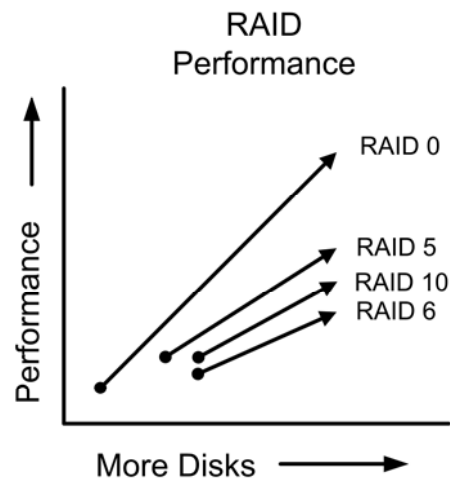
More commonly, multiple disks are arranged into a RAID (Redundant Array of Inexpensive Disks) to take advantage of **striping**, **redundancy**, or both.

The simplest configuration, RAID 0, “stripes” consecutive sectors across different disk drives so that each drive can be getting ready (moving to the appropriate sector) while the previous drive is reading or writing. This greatly improves performance because the whole system doesn’t have to wait when any particular drive is moving its heads.

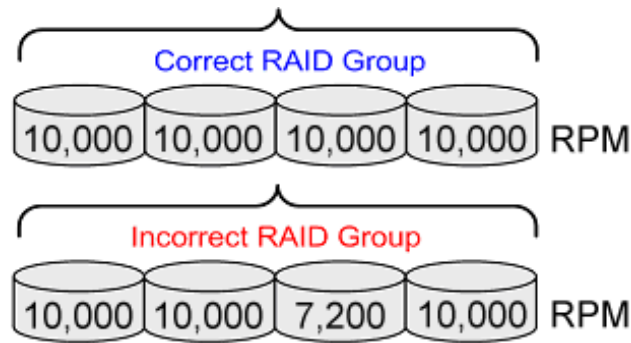
More advanced RAID systems (RAID 1, 5, 6, 10, etc.) also provide redundant (backup) disks so that the failure of a single disk drive does not cause loss of data. We will return to RAID configurations later in this booklet.

RAID performance

Any RAID system is faster than a single disk, and adding disks to a RAID group increases performance. The number of physical disks is sometimes called the **spindle count**.



Note that the disks in a RAID volume *should have the same rotational speed (RPM)* so that the sectors move in unison as expected by the striping algorithm. Performance deteriorates if they are mismatched:



Use Disks of Common Rotational Speeds in a RAID Volume

In general, to reduce latency in a RAID system:

- Use disks with higher rotational speeds.
- Match disk rotational speed within a RAID volume (LUN).
- Add more disks to the RAID volume.
- Allocate a separate RAID volume for each host/server to reduce random access.

Important note: RAID 0 is technically a misnomer, as it provides *no* redundancy. Do not use RAID 0 unless there is another layer of RAID underneath it or your data can be reproduced easily.

Volume allocation

Allocating each RAID volume to only one host computer will maximize performance. This eliminates unintentional random disk access, where two hosts are “fighting over” the position of the disk heads, and speeds performance. For example, partitioning a volume and allocating the partitions to separate hosts may yield lower performance than giving each host its own disk.

Tuning for specific application software

Different applications need different networked storage architectures. For example, streaming data applications, such as web servers, often read and deliver large files in their entirety. Highly transactional applications, on the other hand, perform frequent reads and writes of small files or portions of files; they access the disk in a more random way.

The more random the disk access pattern, the greater the performance benefit from many disks in parallel. This is achieved by striping the data across many disks using RAID 0, RAID 5, or RAID 10. In addition, some application software can be configured to balance the load across logical disk drives, which, in a networked storage application, need not be equivalent to physical drives. For example, an enterprise email service could load-balance by spreading user accounts across a storage pool.

Disk sharing (SAN) can be segmented to allocate portions of the storage pool to meet each application's requirements, while sharing a common connection medium. In some cases, file sharing (NAS) can provide higher performance than disk sharing. Because the file server is the only computer accessing the storage device, complete files can be accessed sequentially, without the risk of random disk access due to interleaved and overlapping I/O requests from cluster members.

It is important to think through the storage requirements of your actual application and choose the best way to map logical drives onto physical disks.

Caching

In general, data caching improves network storage performance. RAM caches are much faster than disk drives. The larger the cache, the greater the benefit to system performance.

There are many layers of cached data in a typical system: The host OS will manage its own memory cache, the file server will have a file and directory cache, the disk storage device will have a block cache, and the disk drive itself will have a cache on its controller to perform speed matching from the disk to the data interface.

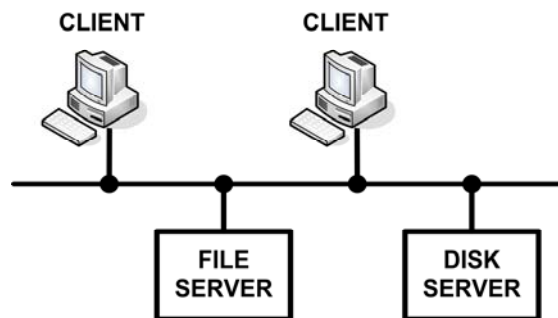
Caching write requests can increase performance, but can also create a risk to data integrity. If the caching system tells the host computer that data has been written to disk, but the data is actually still in the cache, then any failure of the cache (e.g., a momentary power outage) will lead to unrecoverable data loss. For this reason, most storage systems will not acknowledge the completion of a write until the write operation has been committed to disk. This significantly slows write performance when compared to write-to-cache performance.

Write performance can be optimized by distributing the write operation across many disks, such as a large RAID group. It can also be improved through immediate acknowledgement of write operations, even though the block has yet to be moved from the write cache and written to disk. This feature is justified if the storage controller implements a non-volatile cache (usually battery-backed RAM). As an alternative safeguard against data loss, storage devices may immediately acknowledge write

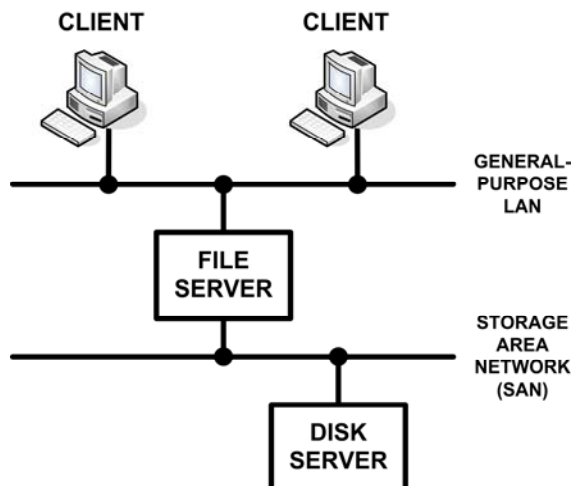
operations, but incorporate a very small buffer. The cached data is only at risk for short periods of time, typically a few seconds.

Network connections and components

A networked storage system is no faster than its network. While AoE or iSCSI disk sharing can take place over the same cables as other network traffic, it is usually preferable to give the disk-sharing system its own subnetwork (a storage area network or SAN, separate from the general-purpose LAN).



Single Network (OK for light usage)



Separate Networks (faster)

For best performance, each host should have two Network Interface Cards (NIC), one for each network.

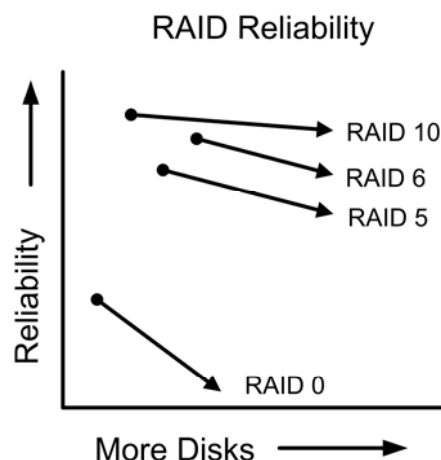
In an Ethernet SAN (AoE or iSCSI), NICs and switches should support 9-kilobyte jumbo Ethernet frames. Jumbo frames reduce the number of packets needed to transport data across the network by allowing an Ethernet packet to forward a much larger payload than the 802.3 standard of 1500 bytes. With jumbo frame support enabled, much more data is moved with each packet.

Disk and RAID Reliability

Disk reliability

Networked storage is expected to provide higher availability than direct attached storage. This section discusses some of the system design choices that can affect reliability and availability.

If you add disks to a multiple-disk system, *the reliability goes down* because there are more points of failure. For example, a RAID 0 (non-redundant) disk system with 3 disks is almost 3 times as likely to fail as one with one disk.¹ In general, adding disks to any RAID system, redundant or non-redundant, reduces its reliability:



¹ "Almost" because failures of separate disks are not mutually exclusive. To be precise, if P is the probability of one disk failing, then the probability of a failure of at least one of three is $3P - 3P^2 + P^3$. Without this kind of correction, the probability would sometimes seem to be more than 100%, which is impossible.

However, any redundant RAID system (RAID 1 and up) is much more reliable than a single disk. That's why we use redundant RAID.

Effect of RAID on reliability

With RAID, discrete disk drives are combined into a single logical block storage device (volume or LUN). The reliability of the RAID group depends on several factors.

RAID 0 is a stripe of data across N disks. It provides excellent performance, but it affords NO fault tolerance. RAID 0 is less reliable than a single disk, and the larger the number (N) disks in the RAID 0, the lower the reliability of the volume.

RAID 0 is a misleading name. It should be called AID (Array of Independent Disks), since there is no redundancy provided. A single disk failure renders the RAID 0 array non-functional.

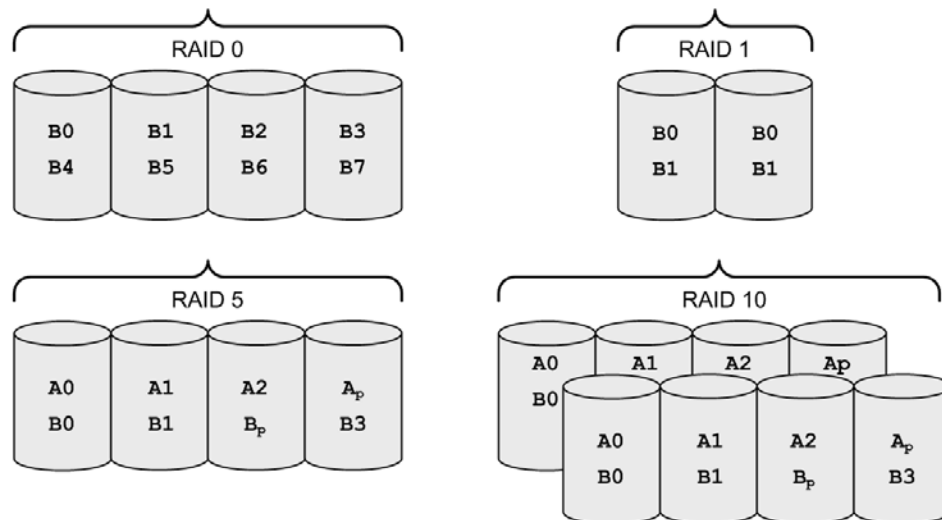
RAID 1 is a pair of mirrored disks. Two disks are paired and all data is copied to each disk. This is the most reliable configuration, but it has no performance advantage over a single disk (there is no striping).

RAID 5 is a set of three or more disks. RAID 5 provides $N-1$ disks of storage capacity with one unit of disk capacity used for parity storage. If any single drive fails, it can be reconstructed, and in addition, we get the performance benefits of $N-1$ -way striping. The parity sectors alternate between different physical drives so that a single drive failure cannot take all the parity bits with it.

The greater the number (N) of disks in the RAID 5 group, there is a higher the probability of a multi-disk failure (failure of two or more disks at the same time, causing data loss in spite of the redundancy).

RAID 6 is like RAID 5, but with two parity sectors so that even if two drives fail at the same time, they can be reconstructed. With N disks you get $N-2$ disks of storage capacity.

RAID 10 is a RAID 0 stripe on top of a number N of RAID 1 mirrored pairs. RAID 10 is commonly employed to increase speed or expand volume size in a more fault-tolerant manner than other RAID types. The risk of failure is very low, but, as with all multi-disk systems, it grows as disks are added.



Block Layout by RAID Level

Networking improves fault tolerance

A simple advantage of networked storage is that the disk drives are not captive inside the host computer that uses them. Thus, failure of a host can be circumvented by allowing another host to access the data stored on the network storage device.

By using a combination of file sharing (NAS) and disk sharing (SAN storage), it is possible to design very fault tolerant storage systems with no single point of failure.

Type of disk doesn't matter

There is no practical difference in reliability between different types of disk drives (PATA, SATA, SCSI) – internally, they use the same drive technology, though they differ in communication protocol and in cost. Special high-reliability enterprise-class disk drives may be justifiable on large RAID volumes where failures are considered intolerable.