The Difference Between PIO and DMA Modes

There are two modes in which data can be transferred between an ATA hard disk drive and the computers system bus, and they are PIO and DMA.

PIO - Programmed Input/Output mode is the slower of the two modes, having the capability of transferring data at a maximum burst rate of 16.7 MBytes per second. PIO mode is also very CPU intensive and has no built in error correction.

DMA or UDMA - Single and Multiword DMA transfers do not support CRC, and Single Word DMA is now considered obsolete and multiword DMA, the predecessor to Ultra DMA, was never widely implemented.

Ultra DMA, which is also referred to as Ultra ATA, incorporates a Cyclic Redundancy Check (CRC) for error detection and correction.

The following tables indicate the associated transfer or burst rates for different modes.

PIO Modes		Single Word DMA		Multiword Word DMA	
Mode	Burst Speed	Mode	Burst Speed	Mode	Burst Speed
Mode 0	3.33MB/s	Mode 0	2.08MB/s	Mode 0	4.17MB/s
Mode 1	5.22MB	Mode 1	4.17MB/s	Mode 1	13.3MB/s
Mode 2	8.33MB	Mode 2	8.33MB/s	Mode 2	16.7MB/s
Mode 3	11.1MB/s	n/a	n/a	n/a	n/a
Mode 4	16.7MB/s	n/a	n/a	n/a	n/a

Ultra DMA	
Mode	Burst Speed
Mode 0	16.7MB/s
Mode 1	25MB/s
Mode 2	33MB/s
Mode 3	44MB/s
Mode 4	66MB/s
Mode 5	100MB/s

ATA/UDMA Device Definitions

Hard drive technologies are expanding at a meteoric rate, like so many other computer technologies today. Just a short time ago, there were two terms to define hard drives, IDE and SCSI. Today, as new drive technologies develop and old ones expand, the terms that define them are added to an ever-growing list of definitions to remember.

Many times more than one term or definition is used to describe the same device type, and some terms have nothing more than the addition of a prefix or suffix to differentiate between technologies. In an effort to reduce some of the inherent confusion, we developed the following information to help you understand the differences between ATA, ATA-2, ATA-3, PIO, DMA, UDMA, etcetera.

ATA (AT Attachment -- AT refers to the IBM Personal Computer AT)
AT Attachment refers to a set of device interface standards or specifications that are used in the design of storage devices for most personal computers. This is often confused with the marketing terms "IDE" and "EIDE", which describe devices that conform to the ATA specification.

ATAPI (AT Attachment Packet Interface) is that part of the interface standard used to define removable media storage devices such as CD-ROM, DVD, Tape, ZIP, and JAZZ drives.

ATA (aka ATA-1 or IDE) -- ANSI document number.
ATA is the actual standard for what is commonly referred to as IDE. The ATA standard defines Programmed Input Output (PIO) modes 0, 1, and 2, and Direct Memory Access (DMA) mode 0. In 1999, at the recommendation of NCITS T13, ATA (ATA-1) was withdrawn as an ANSI standard.

ATA-2 (aka EIDE, FASTATA or FAST ATA) -- ANSI document number X3.279-1996. ATA-2 is the actual standard for what is known today as EIDE. ATA-2 introduces the higher speed data transfer modes of PIO Modes 3 and 4, in addition to Multiword DMA Modes 1 and 2. These modes allow the ATA interface to run at data transfer rates up to 16.6 MB/sec.

ATA-3 -- ANSI document number X3.298-1997.
While the ATA-3 specification has introduced some new features, such as S.M.A.R.T. and Security, it does not introduce any new PIO or DMA data transfer modes. Contrary to the beliefs of some, there is no such thing as PIO mode 5.

ATA/ATAPI-4 -- ANSI document number NCITS.317-1998.
ATA/ATAPI-4 specification adds and changes many things. Here is a brief list:

- It added new ATAPI command and reset protocols.
- It changed or modified many old ATA commands and features, such as the Format Track and Read/Write Long commands, making them obsolete.
- It added a data transfer protocol named Ultra DMA, added data integrity checking (via a CRC check) and created much higher data transfer rates (up to 33MB/second).
- It added a command overlapping and command queuing protocol for both ATA and ATAPI devices.
- It added many new, but minor, features for both ATA and ATAPI devices.

ATA/ATAPI-5
ATA/ATAPI-5 completed its final editing process at T13 and was published as an ANSI standard in the year 2000. ATA/ATAPI-5 deletes a few old commands, adds a few new commands, and changes the way a few commands operate. The single biggest change the occurred as the result of ATA/ATAPI-5 are the two new and faster Ultra DMA 66 data transfer modes.

ATA/ATAPI-6
In late 1999, ATA/ATAPI-6 became next logical, (and official) T13 project. The proposals that did not make it into ATA/ATAPI-5 were added in ATA/ATAPI-6 specifications. Those proposals included increasing the size of the LBA from 28 bits to 64 bits, increasing the Sector Count (transfer size in

sectors) from 8 bits to 16 bits, increasing the Ultra DMA timing mode (to ATA/100), and adding new commands for Audio/Visual (AV) applications. ATA/ATAPI-6 arrived in the marketplace in early in 2000.

Terms	**Original Standard**	**Max Bus Speed**
PIO Mode 0	ATA (ATA-1)	3.33 MB/sec
DMA Mode 0 or Multiword DMA Mode 0	ATA	4.16 MB/sec
PIO Mode 1	ATA	5.22 MB/sec
PIO Mode 2	ATA	8.33 MB/sec
PIO Mode 3	ATA-2	11.1 MB/sec
DMA Mode 1 or Multiword DMA Mode 1	ATA-2	13.3 MB/sec
PIO Mode 4	ATA-2	16.6 MB/sec
DMA Mode 2 or Multiword DMA Mode 2	ATA-2	16.6 MB/sec
Ultra DMA Mode 0	ATA/ATAPI-4	16.6 MB/sec
Ultra DMA Mode 1	ATA/ATAPI-4	25.0 MB/sec
Ultra DMA Mode 2, UDMA33 or ATA/33	ATA/ATAPI-4	33.3 MB/sec
Ultra DMA Mode 3	ATA/ATAPI-5*	44.4 MB/sec
Ultra DMA Mode 4, UDMA66 or ATA/66	ATA/ATAPI-5*	66.6 MB/sec
Ultra DMA Mode 5, UDMA100 or ATA/100	ATA/ATAPI-6**	100 MB/sec

* ATA/ATAPI-5 was not an officially approved standard as of 11/1/2000.
** ATA/ATAPI-6 wa not an officially approved standard as of 11/1/2000.

In PIO Mode, the systems microprocessor gets involved in transferring the data back and forth between the storage device and system memory through the input or output ports.

In DMA Mode, the task of transferring data is handled via a bus-mastering system controller called the DMA controller. The DMA controller is set up or programmed to handle the DMA transfer. This relieves the processor from the duty of handling data transfers, thereby allowing it do other tasks.

The maximum ATA cable length is 18 inches for both the 40- and 80-conductor cables.

The minimum ATA cable length is 10 inches for the 80-conductor cable. The 40-conductor cable has no minimum length specified.

The 40-conductor ATA cable is specified for use on devices that use up to and including Ultra DMA mode 2 or ATA/33 data transfer mode.

The 80-conductor ATA cable is optional for use on devices that use Ultra DMA mode 2 or ATA/33, but required for use on devices that use Ultra DMA mode 3 and higher, as defined in the ATA/ATAPI-5 specification.

SCSI vs. IDE

As the result of ever lowering costs, IDE RAID configurations are becoming increasingly popular. However, costs alone should not be your only consideration when comparing SCSI to IDE in RAID configurations with 4 or less hard disk drives. Although SCSI is often considerably more expensive on initial implementation, it offers many advantages that will save you money in the long term. Below you will find more information regarding additional advantages of SCSI over IDE in small RAID configurations.

SCSI RAID advantages:

- SCSI hard disk drives are optimized for 24 hours x 7 day operation.
- SCSI hard disk drives, in most cases, have better performance and are optimized to handle a large number of transactions in less time.
- SCSI offers better scalability and flexibility in RAID configuration.
- The SCSI bus offers more features to communicate with the devices. Bus bandwidth is not the only consideration!
- Most SCSI hard disk drives have a longer warranty period.

SCSI hard disk drives are optimized for 24 hours x 7 day operation.

The table below provides the typical mean time between failure (MTBF) values based upon anticipated operating environments for SCSI and IDE hard drives.

	SCSI	**IDE**
MTBF	1,000,000 hours	800,000 hours
Power on hours/month	732 (24 hour/day)	333 (11hour/day)
Access ratio	30%	20%

SCSI hard disk drives, in most cases, have better performance and are optimized to handle a large number of transactions in less time.

Faster data access saves time for your business. The table below provides typical data access times for SCSI and IDE high performance hard disk drives.

	SCSI	**IDE**
Seek time	3.4 ms (15,000 rpm)	8.2 ms (7200 rpm)
Sustained data rate	30 to 58 MB/s	20 to 40 MB/s
Cache	4 MB	2 MB

SCSI offers better scalability and flexibility for RAID systems.

One of the most significant advantages to SCSI is scalability. As an example, if you needed to expand your existing RAID system to increase the amount of storage space, up to 15 additional devices (+controller) can be added to a SCSI RAID system, while IDE hard drive controllers are limited to 4.

The SCSI bus offers more features to communicate with the devices. Bus bandwidth is not the only consideration!

- 160 MB per second bus speed for SCSI compared to 100 MB per second for IDE.
- Command tagged queuing - SCSI controllers and hard disk drives, in order to provide the best possible bus performance, realign both commands and data.
- Domain validation - SCSI controllers and hard disk drives exchange test sequences to insure that validity of the communication path before sending the data. This limits data failures commonly caused by cables, enclosures, etc. System integrators eagerly take advantage of this technology.
- Hot swapping drives and the use of spare drives can provide automatic RAID rebuild and minimize the risks when hard drive failures occur. This technology is only now becoming available to IDE controllers and drives, however it does not have the level of acceptance that SCSI hot swaps have.
- Packetization (Ultra 320 SCSI) - Command overhead is significantly reduced with this new technology.
- In contrast to IDE, SCSI supports a multi-controller operation. This offers an additional level of security because it minimizes the risks of a system crash if a controller fails.

What is RAID?
Some RAID Fundamentals

RAID stands for Redundant Array of Inexpensive (or sometimes "Independent") Disks. RAID is a method of combining several hard drives into one logical unit. It can offer fault tolerance and higher throughput levels than a single hard drive or group of independent hard drives.

The basic idea of RAID is to combine multiple inexpensive disk drives into an array of disk drives to obtain performance, capacity and reliability that exceeds that of a single large drive. The array of drives appears to the host computer as a single logical drive.

The Mean Time Between Failure (MTBF) of the array is equal to the MTBF of an individual drive, divided by the number of drives in the array. Because of this, the MTBF of a non-redundant array (RAID 0) is too low for mission-critical systems. However, disk arrays can be made fault-tolerant by redundantly storing information in various ways.

The Benefits of RAID

- Provides real-time data recovery with uninterrupted access when a hard drive fails.
- Increases system uptime and network availability.
- Protects against data loss.

- Multiple drives working in parallel increase system performance.

Types of RAID Architectures

There are five basic types of array architectures that were originally defined as RAID 1 through RAID 5. Each provides disk fault-tolerance but with different compromises in features and performance. In addition to these five redundant array architectures, it has become popular to refer to a non-redundant array of disk drives as a RAID 0 array.

RAID Levels

	RAID 0	RAID 1	RAID 0/1	RAID 5
Description	Data striping (no data protection)	Disk mirroring	RAID 0 and RAID 1 combined	Data striping with distributed parity
Minimum # of Drives	2	2	4	3
Benefit	Highest performance	Data protection through redundancy	Highest performance with data protection	Best balance of cost / performance / data protection

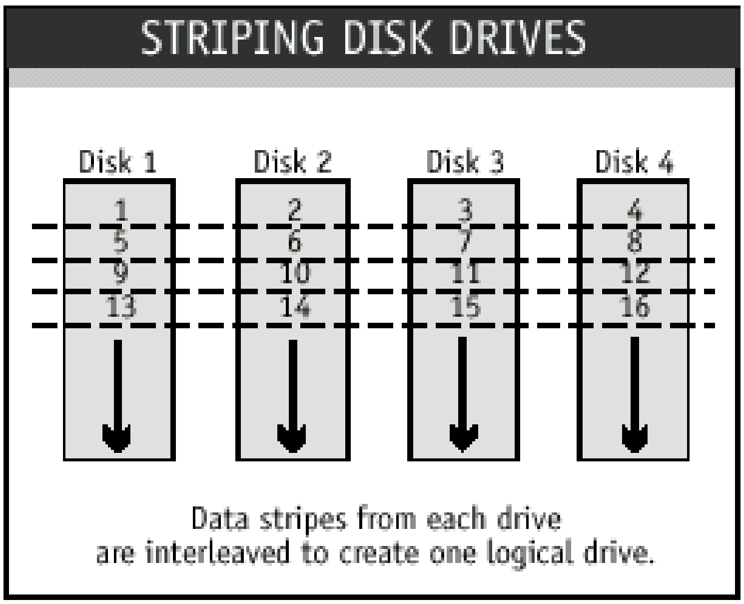
Raid Types

	Description	Advantages
Software-based RAID	Included in NOSs such as NetWare and Windows NT®. All RAID functions are handled by the host CPU which can severely tax its ability to perform other computations.	• Low price • Only requires a standard controller.
Hardware-based RAID	Processor-intensive RAID operations are off-loaded from the host CPU to enhance performance.	• Data protection and performance benefits of RAID • More robust fault-tolerant features and increased performance versus software-based RAID.
External Hardware RAID Card	Connects to the server via a standard controller. RAID functions are performed on a microprocessor located in the external RAID storage subsystem.	• OS independent • Build super high-capacity storage systems for high-end servers.

Disk Striping

Fundamental to RAID technology is striping. This is a method of combining multiple drives into one logical storage unit, or "stripe". Striping partitions the storage space of each drive into stripes, which can be as small as one sector (512 bytes) or as large as several megabytes. These stripes are then interleaved in a rotating sequence, so that the combined space is composed alternately of stripes from each drive. The specific type of operating environment determines whether large or small stripes should be used.

Most operating systems today support concurrent disk I/O operations across multiple drives. However, in order to maximize throughput for the disk subsystem, the I/O load must be balanced across all the drives so that each drive can be kept busy as much as possible. In a multiple drive system without striping, the disk I/O load is never perfectly balanced. Some drives will contain data files that are frequently accessed and some drives will rarely be accessed.

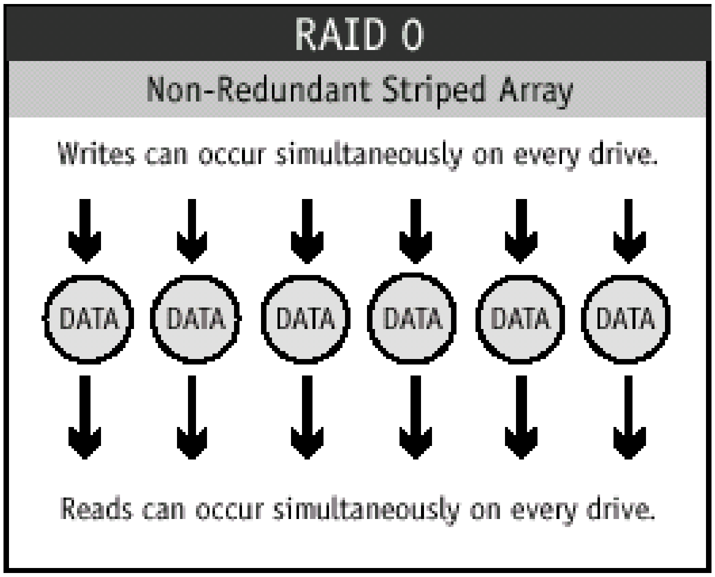


By striping the drives in the array with stripes large enough so that each record falls entirely within one stripe, most records can be evenly distributed across all drives. This keeps all drives in the array busy during heavy load situations. This situation allows all drives to work concurrently on different I/O operations, and thus maximize the number of simultaneous I/O operations that can be performed by the array.

RAID Level Definitions

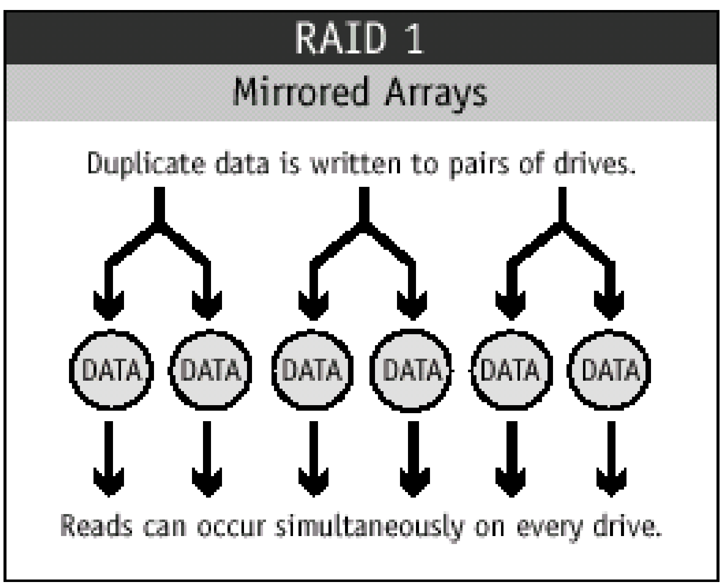
RAID 0 is typically defined as a group of striped disk drives without parity or data redundancy. RAID 0 arrays can be configured with large stripes for multi-user environments or small stripes for single-user systems that access long sequential records. RAID 0 arrays deliver the best data storage efficiency and performance of any array type. The disadvantage is that if one drive in a RAID 0 array fails, the entire array fails.

RAID Level 0 is not redundant, hence does not truly fit the "RAID" acronym. In Level 0, data is split across drives, resulting in higher data throughput. Since no redundant information is stored, performance is very good, but the failure of any disk in the array results in the loss of all data. This level is commonly referred to as striping.



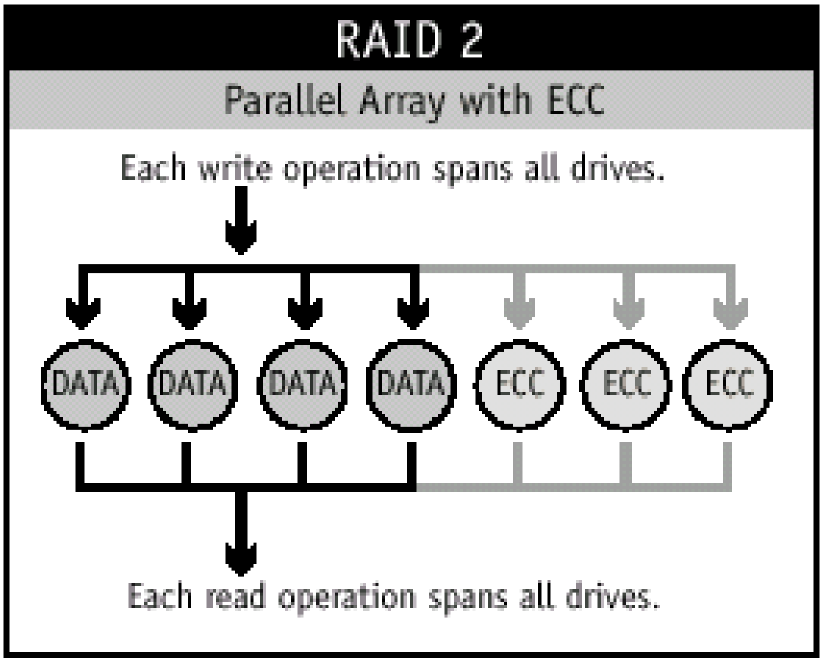
RAID 1, also known as disk mirroring, is simply a pair of disk drives that store duplicate data but appear to the computer as a single drive. Although striping is not used within a single mirrored drive pair, multiple RAID 1 arrays can be striped together to create a single large array consisting of pairs of mirrored drives. All writes must go to both drives of a mirrored pair so that the information on the drives is kept identical. However, each individual drive can perform simultaneous, independent read operations. While mirroring doubles the read performance of a single non-mirrored drive, the write performance is unchanged. RAID 1 delivers the best performance of any redundant array type, and in addition, there is less performance degradation during drive failure than in RAID 5 arrays.

Thus, RAID Level 1 is more commonly referred to as mirroring with 2 hard drives. It provides redundancy by way of the duplication of all data from one drive to another drive. Performance of a Level 1 array is slightly better than a single drive, and if either drive fails, no data is lost. This is a good entry-level redundant system, since only two drives are required. However, since one drive is used to store a duplicate of the data, the cost per megabyte is doubled. See also Disk Duplexing.



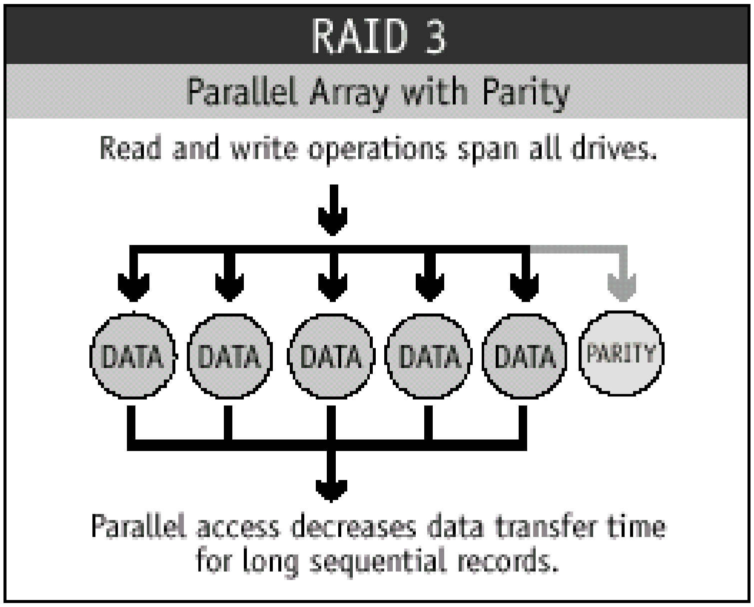
RAID 2 arrays sector-stripe data across groups of drives, with some drives assigned to store ECC information. Because all disk drives today embed ECC information within each sector, RAID 2 offers no significant advantages over other RAID architectures and is not supported by most popular RAID controller manufacturers such as Adaptec.

RAID Level 2 uses Hamming error correction coding, and is intended for use with drives which do not have built-in error detection. All SCSI drives support built-in error detection, therefore this level would be of little use when using SCSI drives.



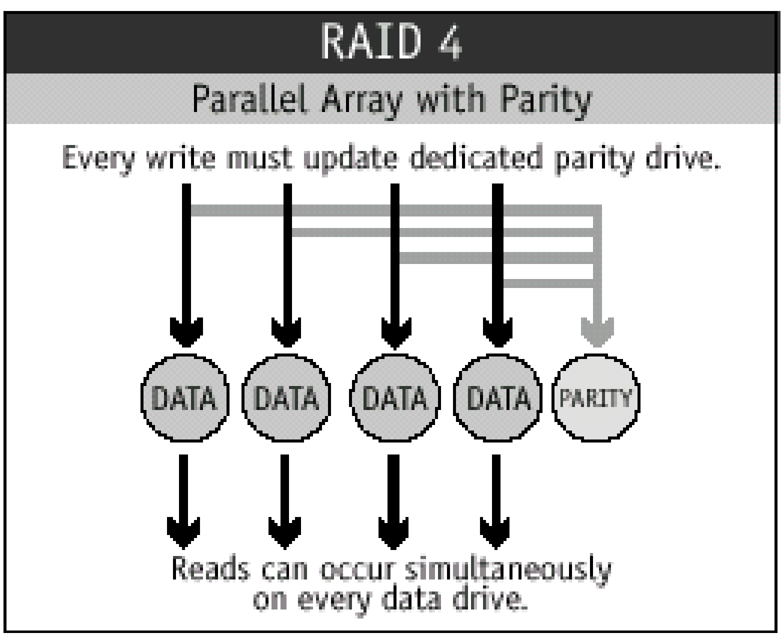
RAID 3, as with RAID 2, sector-stripes data across groups of drives, but one drive in the group is dedicated to storing parity information. RAID 3 relies on the embedded ECC in each sector for error detection. In the event of drive failure, data recovery is accomplished by calculating the exclusive OR (XOR) of the information recorded on the remaining drives. Records typically span all drives, which optimizes the disk transfer rate. Because each I/O request accesses every drive in the array, RAID 3 arrays can satisfy only one I/O request at a time. RAID 3 delivers the best performance for single-user, single-tasking environments with long records. Synchronized-spindle drives are required for RAID 3 arrays in order to avoid performance degradation with short records. Because RAID 5 arrays with small stripes can yield similar performance to RAID 3 arrays, RAID 3 is not supported by most popular RAID controller manufacturers such Adaptec.

RAID Level 3 stripes data at a byte level across several drives, with parity stored on one drive. It is otherwise similar to level 4. Byte-level striping requires hardware support for efficient use.



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 (except the parity drive). This allows read operations to be overlapped. However, since all write operations must update the parity drive, they cannot be overlapped. This architecture offers no significant advantages over other RAID levels and is not supported by most RAID controller manufacturers such as Adaptec.

RAID Level 4 stripes data at a block level across several drives, with parity stored on one drive. The parity information allows recovery from the failure of any single drive. The performance of a level 4 array is very good for reads (the same as level 0). Writes, however, require that parity data be updated each time. This slows small random writes, in particular, though large writes or sequential writes are fairly fast. Because only one drive in the array stores redundant data, the cost per megabyte of a level 4 array can be fairly low.



RAID 5, sometimes called a Rotating Parity Array, avoids the write bottleneck caused by the single dedicated parity drive of RAID 4. Under RAID 5 parity information is distributed across all the drives. Since there is no dedicated parity drive, all drives contain data and read operations can be overlapped on every drive in the array. Write operations will typically access one data drive and one parity drive. However, because different records store their parity on different drives, write operations can usually be overlapped.

This level is more commonly referred to as striping with distributed parity. RAID Level 5 is similar to level 4, but distributes parity among the drives. No single disk is devoted to parity. This can speed small writes in multiprocessing systems. Because parity data must be distributed on each drive during reads, the performance for reads tends to be considerably lower than a level 4 array. The cost per megabyte is the same as for level 4.

In summary:

- **RAID 0** is the fastest and most efficient array type but offers no fault-tolerance. RAID 0 requires a minimum of two drives. Any application which requires very high speed storage, but does not require redundancy, such as Photoshop temporary files as an example.

- **RAID 1** is the best choice for performance-critical, fault-tolerant environments. RAID 1 is the only choice for fault-tolerance if no more than two drives are used. Applications which require redundancy with fast random writes; entry-level systems where only two drives are available such as small file servers.

- **Level 0/1 or 10** (mirroring and striping) - Dual level raid, combines multiple mirrored drives (RAID 1) with data striping (RAID 0) into a single array. Provides highest performance with data protection.

- RAID 2 is seldom used today since ECC is embedded in all hard drives. RAID 2 is not supported by most RAID controller manufacturers.

- **RAID 3** can be used to speed up data transfer and provide fault-tolerance in single-user environments that access long sequential records. However, RAID 3 does not allow overlapping of multiple I/O operations and requires synchronized-spindle drives to avoid performance degradation with short records. Because RAID 5 with a small stripe size offers similar performance, RAID 3 is not supported by most RAID controller manufacturers.

- **RAID 4** offers no advantages over RAID 5 and does not support multiple simultaneous write operations. RAID 4 is not supported by most RAID controller manufacturers.

- **RAID 5** combines efficient, fault-tolerant data storage with good performance characteristics. However, write performance (and performance during drive failure) is slower than with RAID 1. Rebuild operations also require more time than with RAID 1 because parity information is also reconstructed. At least three drives are required for RAID 5 arrays. Similar to level 4, but may provide higher performance if most I/O is random and in small chunks such as would be experienced with database servers.

- **Level 0/5 or 50** - Dual level raid, combines multiple RAID 5 sets with data striping (RAID 0). Increased reliability and performance over standard RAID5 that can stand a multiple drive failure; one hard drive per RAID5 set.

Dual-Level Raid in Brief

We use Adaptec RAID Controllers, and here's why! In addition to the standard RAID levels, Adaptec RAID controllers can combine multiple hardware RAID arrays into a single array group or parity group. In a dual-level RAID configuration, the controller firmware stripes two or more hardware arrays into a single array.

NOTE: The arrays being combined must both use the same RAID level.

Dual-level RAID achieves a balance between the increased data availability inherent in RAID 1 and RAID 5 and the increased read performance inherent in disk striping (RAID 0). These arrays are sometimes referred to as RAID 0+1 or RAID 10 and RAID 0+5 or RAID 50.

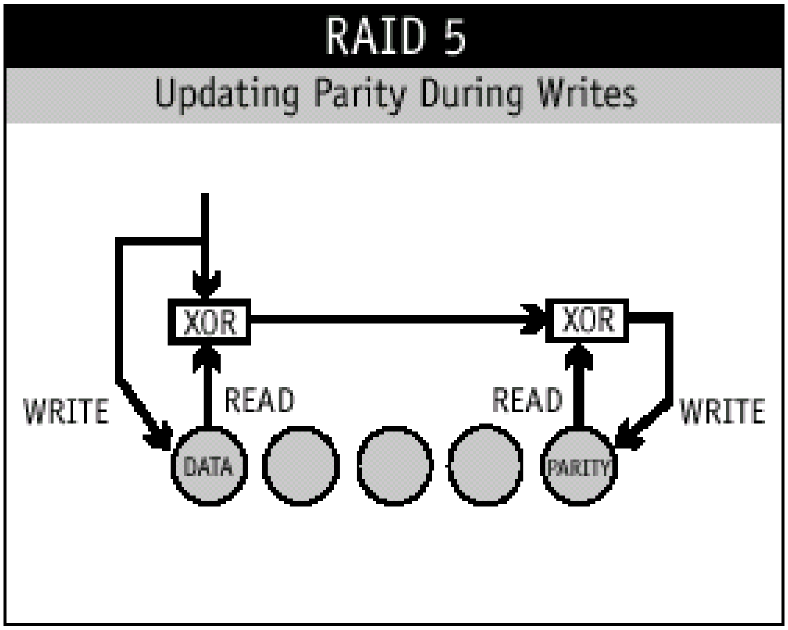
Creating Data Redundancy

RAID 5 offers improved storage efficiency over RAID 1 because only the parity information is stored, rather than a complete redundant copy of all data. The result is that three or more drives can be combined into a RAID 5 array, with the storage capacity of only one drive dedicated to store the parity information. Therefore, RAID 5 arrays provide greater storage efficiency than RAID 1 arrays. However, this efficiency must be balanced against a corresponding loss in performance.

The parity data for each stripe of a RAID 5 array is the XOR of all the data in that stripe, across all the drives in the array. When the data in a stripe is changed, the parity information is also updated. There are two ways to accomplish this: See also System Fault Tolerance.

1. The first method is based on accessing all of the data in the modified stripe and regenerating parity from that data. For a write that changes all the data in a stripe, parity can be generated without having to read from the disk, because the data for the entire stripe will be in the cache. This is known as full-stripe write. If only some of the data in a stripe is to change, the missing data (the data the host does not write) must be read from the disks to create the new parity. This is known as partial-stripe write. The efficiency of this method for a particular write operation depends on the number of drives in the RAID 5 array and what portion of the complete stripe is written.

2. The second method of updating parity is to determine which data bits were changed by the write operation and then change only the corresponding parity bits. This is done by first reading the old data which is to be overwritten. This data is then XOR'ed with the new data that is to be written. The result is a bit mask which has a 1 in the position of every bit which has changed. This bit mask is then XOR'ed with the old parity information from the array. This results in the corresponding bits being changed in the parity information. The new updated parity is then written back to the array. This results in two reads, two writes and two XOR operations. This is known as read-modify-write.



The cost of storing parity, rather than redundant data as in RAID 1, is the extra time required for the write operations to regenerate the parity information. This additional time results in slower write performance for RAID 5 arrays over RAID 1. Because Adaptec RAID controllers generate XOR in hardware, the negative effect of parity generation is primarily from the additional disk I/O required to read the missing information and write the new parity. Adaptec RAID controllers can generate parity using either the full- or partial-stripe write algorithm or the read-modify-write algorithm. The parity updated method chosen for any given write operation is determined by calculating the number of I/O operations needed for each type and choosing the one with the smallest result. To increase the number of full stripe writes, the cache is used to combine small write operations into larger blocks of data.

Technology Comparison

	UDMA	SCSI	Fibre Channel
Best Suited For	Low-cost entry level server with limited expandability	Low to high-end server when scalability is desired	Server-to-Server campus networks
Advantages	• Uses low-cost ATA drives	• Performance: up to 160 MB/s • Reliability • Connectivity to the largest variety of peripherals • Expandability	• Performance: up to 100 MB/s • Dual active loop data path capability • Infinitely scalable

Handling I/O Errors

Adaptec RAID controllers maintain two lists for each RAID 5 array: a Bad Parity List, and a Bad Data List. These lists contain the physical block number of any parity or data block that could not be successfully written during normal write, rebuild or dynamic array expansion operations. These lists alert the controller that the data or parity in these blocks is not valid. If the controller subsequently needs data from a listed block and cannot recreate the data from existing redundant data, it returns an error condition to the host.

Blocks are removed from the Bad Parity List or the Bad Data List if the controller successfully writes to them on a subsequent attempt.

Degraded Mode

When a drive fails in a RAID 0 array, the entire array fails. In a RAID 1 array, a failed drive reduces read performance by 50%, as data can only be read from the remaining drive. Write performance is increased slightly because only one drive is accessed. A RAID array operating with a failed drive is said to be in degraded mode.

RAID 5 arrays synthesize the requested data by reading and XOR'ing the corresponding data stripes from the remaining drives in the array. For RAID 5, the magnitude of the performance impact in degraded mode depends on the number of drives in the array. An array with a large number of drives will experience more performance degradation than an array with small number of drives.

Rebuilding a Failed Hard Drive

A failed drive can be replaced in a RAID 1 or RAID 5 array by physically removing the drive and replacing it or by a designated Hot Spare. Adaptec RAID controllers will rebuild the data for the failed drive onto the new drive or Hot Spare. This rebuild operation occurs online while normal host reads and writes are being processed by the array.

RAID 1 arrays are rebuilt relatively quickly, because the data is simply copied from the duplicate (mirrored) drive to the replacement drive. For RAID 5 arrays, the data for the replacement drive must be synthesized by reading and XOR'ing the corresponding stripes from the remaining drives in the array. RAID 5 arrays that contain a large number of drives will require more time for a rebuild than a small array.