



RAID:  
Awakening



*An Introduction To RAID*

[www.ami.com](http://www.ami.com)



# RAID Awakening

*An Introduction To RAID*

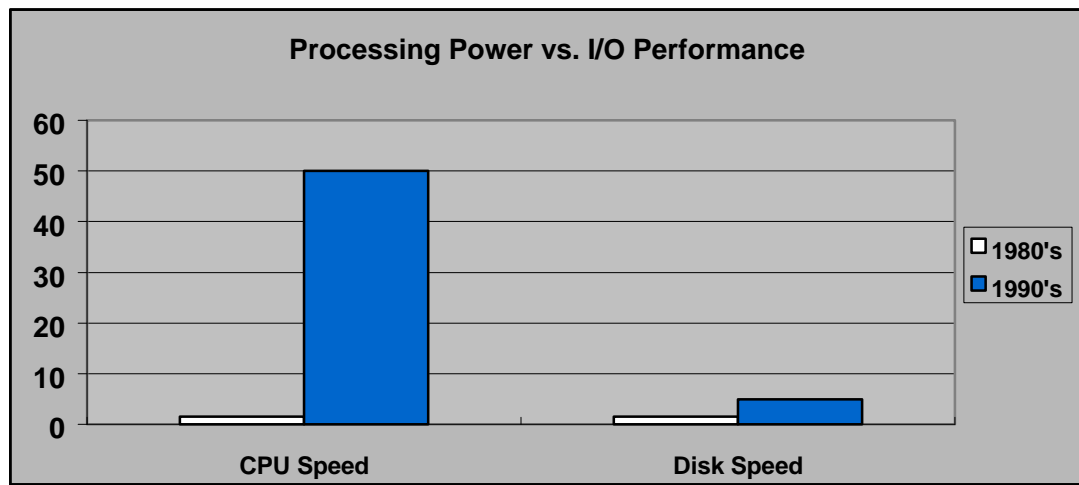
## Table of Contents

<b>Part I - Why Do We Need RAID ? .....</b>	<b>3</b>
<b>Part II - What is RAID ? .....</b>	<b>4</b>
Disk Spanning .....	4
Disk Striping .....	5
Disk Mirroring .....	5
Disk Striping with Redundancy.....	6
<b>Part III -RAID Levels.....</b>	<b>7</b>
RAID 10.....	8
RAID 50.....	8
<b>Part IV - Battery Backed Cache: Preventing the Data Loss .....</b>	<b>9</b>
WRITE-THROUGH CACHE .....	9
WRITE-BACK CACHE .....	9
THE SOLUTION: Battery Backed Cache .....	10

## Part I - Why Do We Need RAID ?

The term **RAID** which is an acronym for **R**edundant **A**rray of **I**ndependent **D**isks, originated at the University of California at Berkeley in the late 1980s. This technology has gained popularity in the last decade to meet two main challenges:

1. **To improve I/O performance at the same pace as computing performance** - although disk drive capacities have improved drastically, the actual performance has improved only 3-4 times in the last decade whereas the computing performance has improved over 50 times.
2. **To increase the reliability of storage subsystems** well in excess of the expected life time of the computer systems controlling them. The electromechanical components of a disk-subsystem operate more slowly, require more power and generate more noise and vibration than electronic devices, thereby reducing the reliability of data stored on disks.



Disk Array Technology has been developed to provide these features as compared to individual disks. Disk Array Technology can be broken down into more detailed models called RAID Levels. RAID Levels provide some or all of these desirable properties:

- It may improve the **I/O performance** by balancing the I/O load evenly across the disks. Striped arrays cause a stream of requests to be divided evenly across the disks in a set.
- It may improve **data reliability** by replicating data so that it is not destroyed if the disk on which it is stored fails. Mirrored arrays cause every block of data to be replicated on all disks of a set.
- It may simplify **storage management** by treating more storage capacity as a single manageable entity.

RAID provides a means of optimizing the use of a scarce resource (disk subsystem) by trading a little of a relatively abundant resource (computing power). Extra processing required by these RAID levels is minor and the benefits are huge.

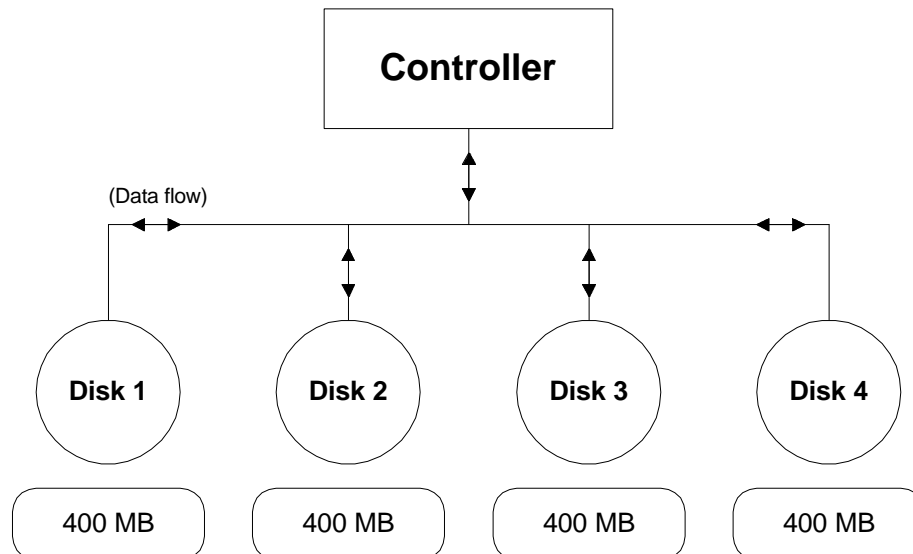
## Part II - What is RAID ?

RAID is an array of multiple small, independent disks which yields performance exceeding that of a Single Large Expensive Drive (SLED). This array appears as a single logical storage unit or drive to the Host computer. It can be made fault-tolerant by redundantly storing information in various ways. The redundant information enables regeneration of the data if one of the array's disks fails.

The original Berkeley paper outlined five disk array models called RAID Levels, each providing disk fault-tolerance and offering different trade-offs in features and performance. The basic concepts behind these levels are as follows:

### Disk Spanning

Disk spanning allows multiple disk drives to function like one big drive. This enables users to overcome limitations of their current disk space by combining existing resources or adding to current resources inexpensively.

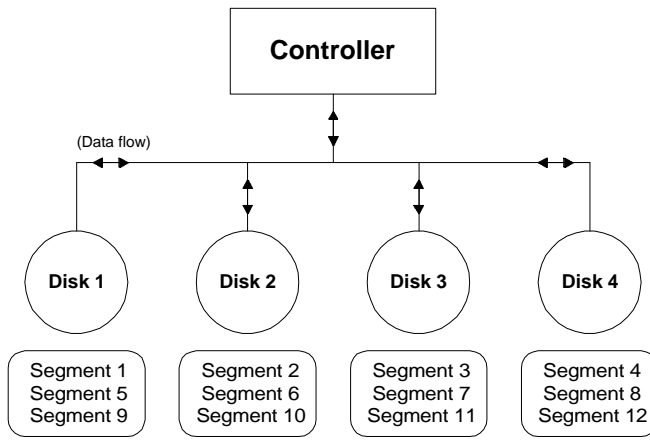


This figure shows an example of four 400Meg drives in a storage subsystem. Instead of seeing 4 drives as C, D, E and F, the user would only see one drive, C, comprising 1600 megs.

**Advantages** - Inexpensive, smaller, multiple disk drives can be added as needed  
**Disadvantages** - No performance improvement in terms of reliability or speed.

## Disk Striping

Disk striping writes data across multiple disks rather than onto one disk. It involves partitioning each drive's storage space into stripes which vary from one sector (512 bytes) to several megabytes. These stripes are interleaved round-robin, so that the combined storage space is composed alternately of stripes from each drive.



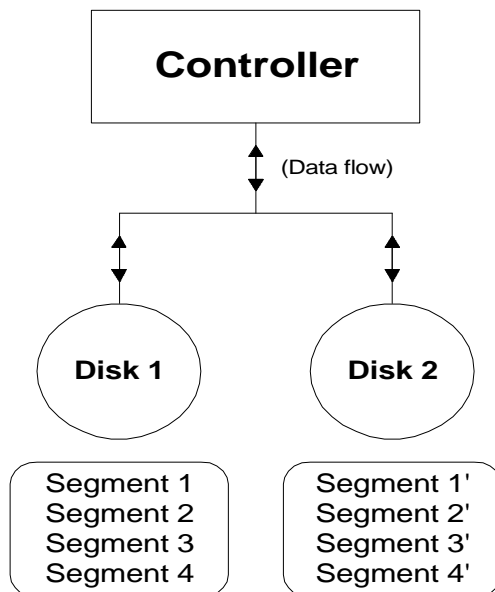
For example, segment 1 is written to drive 1, segment 2 is written to drive 2, segment 3 is written to drive 3, etc. When the system reaches the end of the drive list, it starts writing at the next available segment of drive 1.

**Advantages:** *Fast as it transfers data to multiple drives at once.*

**Disadvantages:** *No redundancy of the data*

## Disk Mirroring

The primary advantage of disk mirroring is 100% redundancy of data. The redundancy is achieved by simply duplicating all the data on one drive onto a second drive (or equivalent device). If one drive fails, the system can continue running since it can operate off the remaining good drive. Since the drive is mirrored, it doesn't matter which drive fails. Both contain the same data at all times and therefore either can act as the operational drive.



**Advantages:** *100% redundancy and better read performance.*

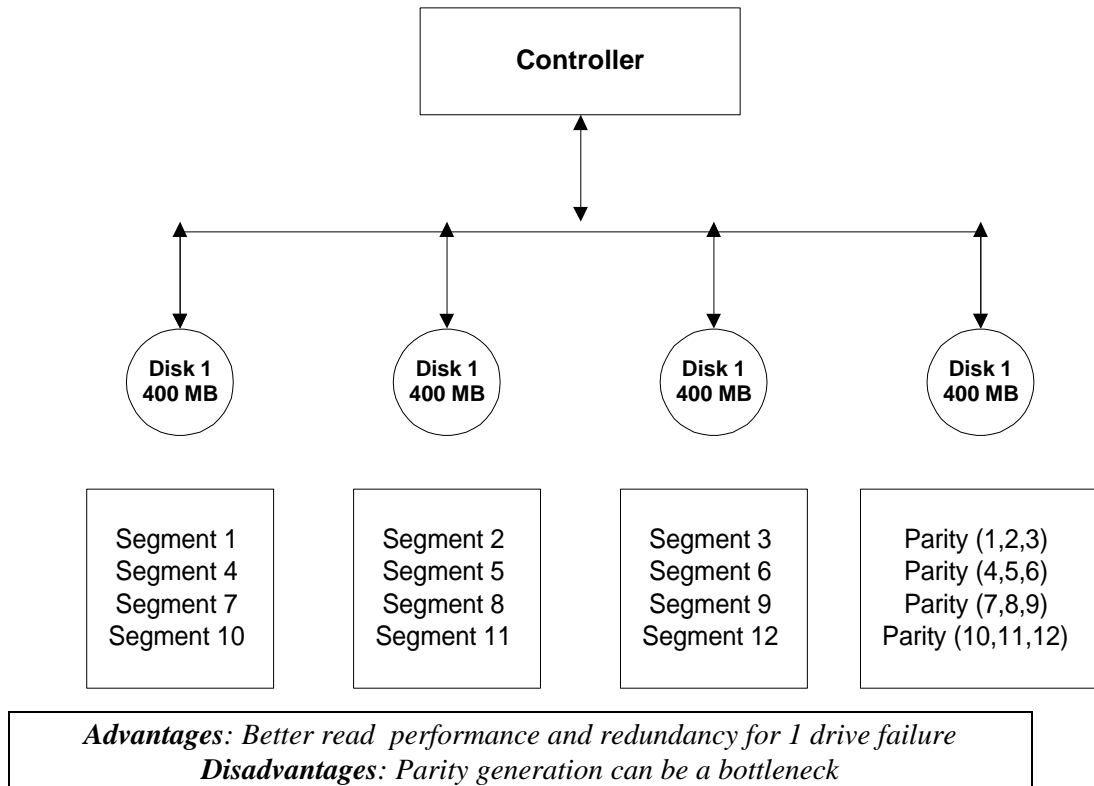
**Disadvantages:** *Expensive due to the duplication of each drive in the system*

## Disk Striping with Redundancy

Disk striping improves performance of a storage subsystem because we can write and read data to multiple disks rather than onto one disk. However, the disadvantage of disk striping is that the failure of any single drive can bring down the entire system.

Redundancy can be added to a striped disk system by using one of several parity schemes. An additional drive can be added to the array which contains only parity or parity can be distributed across all drives in the array.

For example, we can write the parity of segment 1,2,3 on to Disk 4 by XORing the data in segment 1,2 and 3. If drive 2 fails, we could XOR segments 1,3 and parity to get back missing segment.



## Part III -RAID Levels

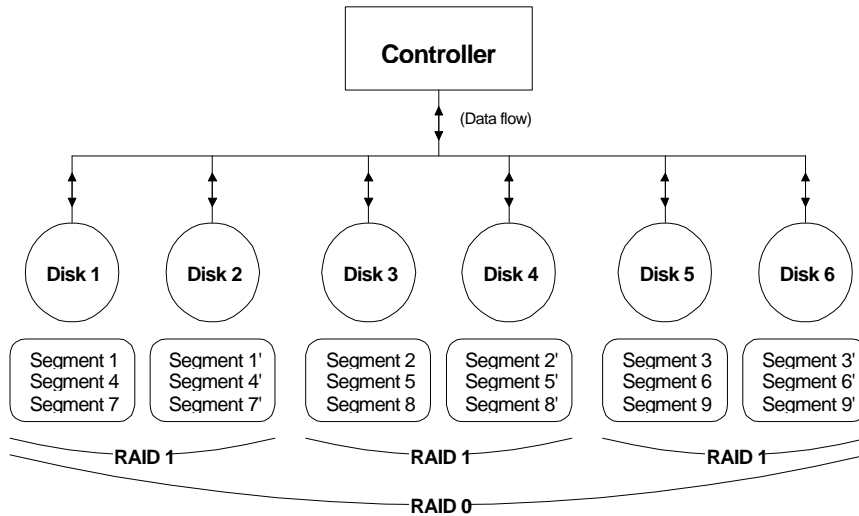
The original Berkeley paper outlined five disk array models called RAID Levels, each providing different levels of disk fault-tolerance and offering different trade-offs in features and performance. MegaRAID supports RAID Level 0,1,3 and 5 from the original models but it also supports level 10 and 50 which are defined by AMI.

<b>RAID Level</b>	<b>Description</b>	<b>Application</b>
0	Data is divided into blocks and distributed sequentially among disks (Pure Striping)	Data Collection from external sources at very high transfer rates, no Fault-tolerance.
1	Data written to one disk is duplicated onto another disk (Pure Mirroring)	Read-intensive, Fault-tolerant systems
3	Disk Striping with dedicated parity drive	Non-interactive applications that process large files sequentially
5	Disk Striping with distributed parity	Those with high read-request rates, with a low percentage of write requests, Transaction processing, office automation, on-line customer service etc.
10(AMIRAIID)	Mirroring of striped arrays, combination of levels 1 and 0.	For data storage whose value justifies the 100% redundancy of Mirrored arrays and needs enhanced I/O performance of striped arrays.
50(AMIRAIID)	Combination of RAID levels 5 and 0.	For data which must be kept on highly reliable storage and requires both high request rates and high data transfer performance at low cost.

Refer to the “Help” menu of MegaRAID Power console for detailed graphic description of the RAID levels. These levels are managed by the firmware on the MegaRAID controller and do not require any special components.

## RAID 10

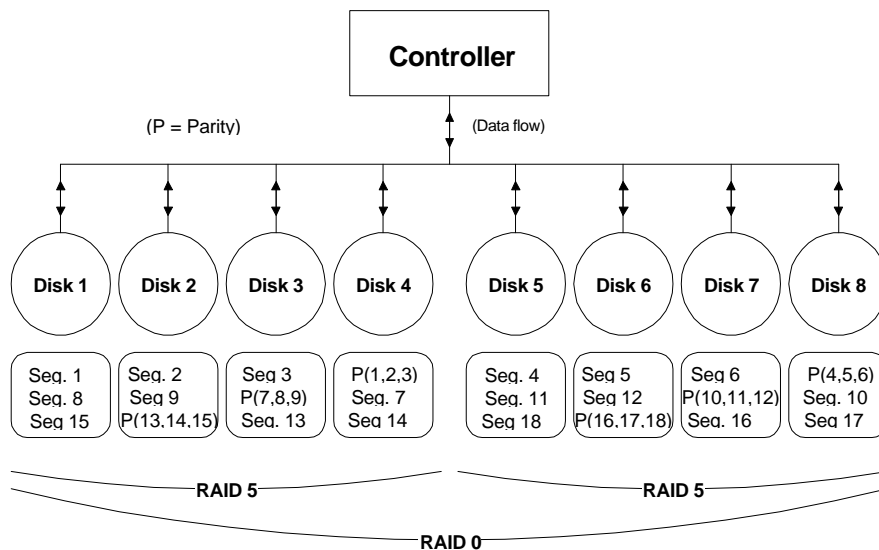
This RAID level is combination of levels 1 and 0 in a single array. The data is first striped across all the disks of RAID 0 and then it is duplicated using a similar striped array. This RAID level provides 100% data reliability through RAID 1 and enhanced I/O performance through striping of RAID 0 but at relatively high inherent hardware (disk and port).



The data reliability of a RAID 10 array is  $1/n$  time that of a 2-disk RAID level 1 array ( $n$  = no of stripes). The performance enhancement is achieved by involving more devices in a data transfer.

## RAID 50

This RAID level is combination of levels 5 and 0 in a single array to simultaneously provide RAID 5 like performance and reliability with striping-like I/O request rates at an inherent cost similar to that of RAID level 5. The simplest implementation of RAID 50 would use two RAID level 5 arrays and stripe data across them.



The data reliability of a RAID 50 array is equal to  $1/n$  times that of one of its member RAID 5 arrays ( $n$  = no. of stripes).

## Part IV - Battery Backed Cache: Preventing the Data Loss

When a file is written to a storage device such as a hard drive, the data in the file is transferred in blocks. A block of data is first sent to the SCSI controller, which after receiving the block sends an acknowledgment to the operating system (OS) that the data was received and stored on storage device. After receiving the acknowledgment, the OS considers the block to be safely stored in the storage device.

In a non-cache SCSI controller the data is sent immediately to the storage device by the controller before an acknowledgment is sent back to the OS; however, in a caching SCSI controller, an acknowledgment is sent to the OS after the data is stored in the cache. The controller does not wait for the data to be stored in the storage device before sending acknowledgment.

There are two methods used to handle data written to a storage device connected through a caching SCSI controller -

### WRITE-THROUGH CACHE

The first method is Write-through. In this method, the block of data is written to both the cache and the storage device once the data is received. Because the data is written to both the cache and the storage device, the data can be accessed fast from the cache if needed later, not waiting for the data to be retrieved from the slow storage device, and at the same time data is kept safe in the storage device. The negative side of this method is that the time to do a write is greater than the time to do a write to a non-cache storage device. The total write time is the time to write data to cache plus the time to write data to storage device.

### WRITE-BACK CACHE

The second method, called Write-Back method, does not have the write time limitation of the Write-Through method. The block of data is initially only written to the cache, not both the cache and the storage device. The write time for a Write-Back Cache is only the time to write the data to the cache, which is much less than the time to write to the storage device. Later, when cache space is needed, or when activity is low the SCSI controller writes the data to the slow storage device.

The limitation of this method is that the storage device for a period of time does not contain the new or updated block of data. If the data in the cache is lost, for example by the failure of the power supply, the data can not be recovered.

When the SCSI controller received the data and stored it in the cache, an acknowledgment was sent to the OS indicating that the data was received by the storage device, although the data was only stored in the cache. Once the data is placed in the cache, the SCSI controller is responsible for keeping the data valid and safe. The OS is only responsible for the data before acknowledgment is received.

## THE SOLUTION: Battery Backed Cache

To solve the problem of data loss associated with a Write-Back Cache, AMERICAN MEGATRENDS, Inc. has developed an optional Battery Backed Cache Module for MegaRAID PCI SCSI RAID controllers.

The module provides the cache of MegaRAID with uninterrupted power. For example, during a power supply failure, the module automatically switches the power source of cache from the regular +5 volt supply of the system to the regulated and conditioned +5 volts generated by a 5 cell nickel battery pack.

The module also places the MG9010, AMI's PCI-to-I960 Bridge, in "powerdown" mode. In this mode, the MG9010 only performs the refresh cycles, which are necessary for retention of data in the DRAM SIMM modules of the cache.

After power is restored to the controller, the module switches the power source of the cache back to +5 volt supply of the system. The MG9010 leaves powerdown mode, and all incomplete writes to storage devices are completed. No data contained in cache is lost. This provides a very inexpensive solution of preventing the data loss without any UPS attached to the system. It also protects the data in situations where UPS cannot for example power supply of the server

When there is no power failure, the 5-cell battery pack is being charged. The battery pack is charged at two rates: "Quick Charge" and "Trickle Charge". When the charge of the battery is low, after a prolonged power failure, the battery pack is charged at the "Quick Charge" rate. After the battery is fully charged, the "Trickle Charge" rate is used to maintain the charge of the battery pack.

*The Battery Back Cache Module is monitored continuously by the RAID controller for the charge in the battery and if for some reason the battery is not getting charged, the user is given a warning to replace the batteries.*

The majority of servers are installed with their own UPS which will protect the data in case of a main power failure, but there is no protection of data in case of an internal power supply failure. The battery-backup module fills this void and provides this protection, thus enhancing the overall fault-tolerance of the system.

### BATTERY BACK CACHE MODULE SPECIFICATION:

*Battery Pack Size: 5-cell NiCad Pack*

*Backup Time: 3 hours (Min) to 5 hours (Max)*

*Charge Time: 4 hours (Max) — (continuous charging at "Quick Charge" rate)*

*For More Information, Go To:*

[www.ami.com](http://www.ami.com)