

Fault Tolerance and RAID

Any thing that can go wrong will go wrong at the worst possible time (*Murphy's Law*)

The question is not if your hard disk will break down but when? (*Corollary to Murphy's Law*)

What Is Fault Tolerance?

- Fault Tolerance is the ability of a system to continue functioning when a component on the computer fails.
- The term Fault Tolerance is typically used to describe disk subsystems, but it can also apply to other system components or the system as a whole.
- Fully fault tolerant computers use redundant disk controllers and un-interruptible power supplies as well as fault tolerant disk subsystems and clustered computers. Redundant controllers, etc.

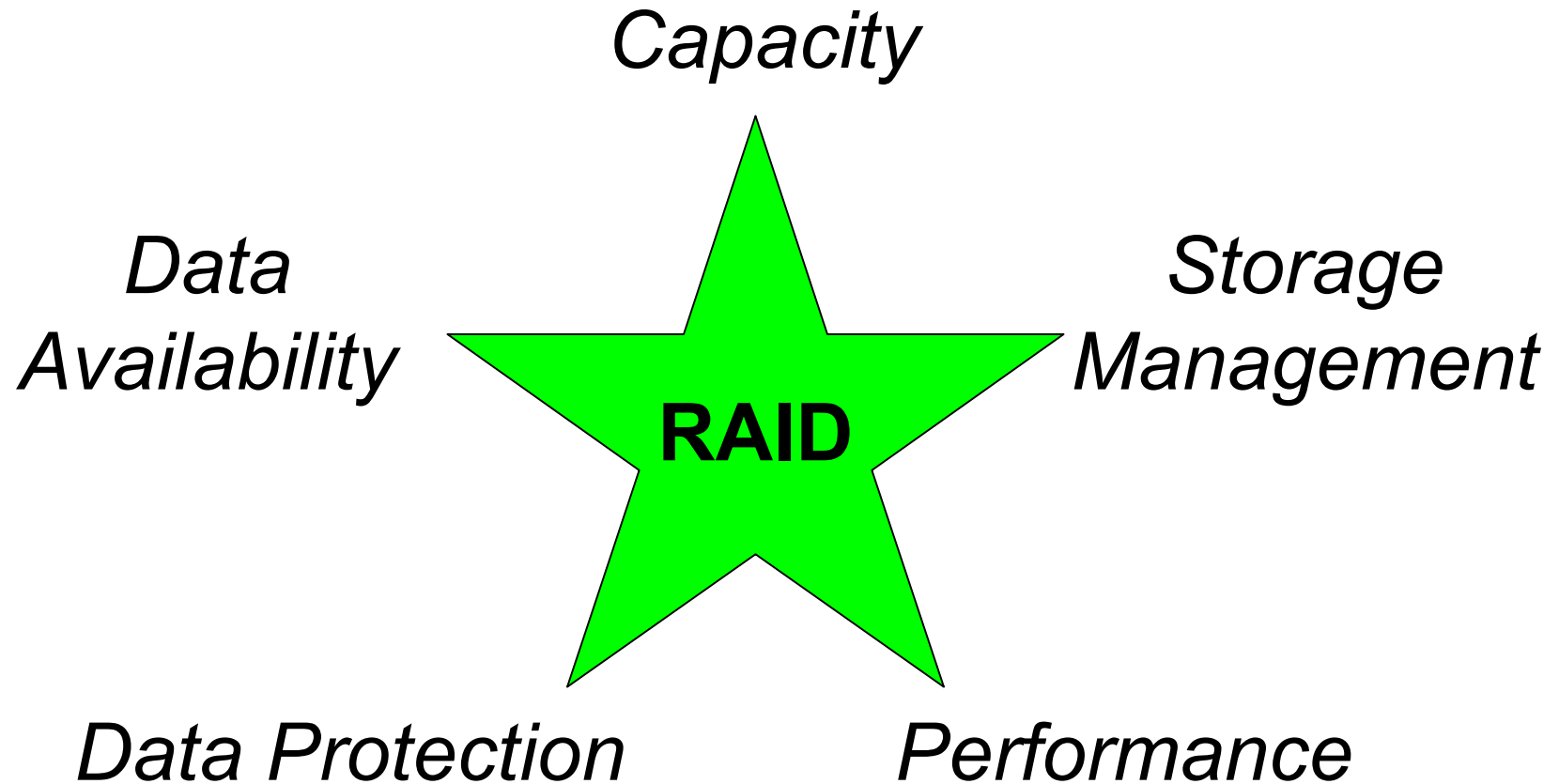
Where is Fault Tolerance?

- Redundant Power Supplies
- Redundant Cooling Fans
- Redundant Disk Subsystems
- Redundant NICs
- Backup Power Supply, UPS, Generators

Why Have Fault Tolerance?

- To ensure availability
- Availability is a Computer Security goal
- Availability:
 - Availability requires that computer systems assets are available to authorized parties.
 - *Availability*: A "requirement intended to assure that systems work promptly and service is not denied to authorized users." (*Computers at Risk*, p. 54.)

Redundant Array of Inexpensive Disks - RAID



RAID Conceived

- In 1987, Patterson, Gibson and Katz at the University of California Berkeley, published a paper entitled "A Case for Redundant Arrays of Inexpensive Disks (RAID)" .
- This paper described various types of disk arrays, referred to by the acronym RAID.
- The basic idea of RAID was to combine multiple small, inexpensive disk drives into an array of disk drives which yields performance exceeding that of a Single Large Expensive Drive (SLED).
- Additionally, this array of drives appears to the computer as a single logical storage unit or drive.

Why RAID?

- The Mean Time Between Failure (MTBF) of the array will be equal to the MTBF of an individual drive, divided by the number of drives in the array.
- Because of this, the MTBF of an array of drives would be too low for many application requirements.
- However, disk arrays can be made fault-tolerant by redundantly storing information in various ways.

Understanding RAID

- Five types of array architectures, RAID-1 through RAID-5, were defined by the Berkeley paper, each providing disk fault-tolerance and each offering different trade-offs 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.

Data Striping

- Fundamental to RAID is "striping", a method of concatenating multiple drives into one logical storage unit.
- Striping involves partitioning each drive's storage space into stripes which may be as small as one sector (512 bytes) or as large as several megabytes.
- These stripes are then interleaved round-robin, so that the combined space is composed alternately of stripes from each drive.

Data Striping

- In effect, the storage space of the drives is shuffled like a deck of cards. The type of application environment, I/O or data intensive, determines whether large or small stripes should be used.
- Most multi-user operating systems today, like NT, Unix and Netware, support overlapped 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.

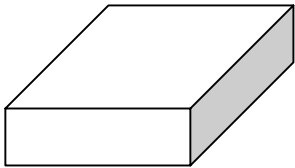
Data Striping

- In a multiple drive system without striping, the disk I/O load is never perfectly balanced. Some drives will contain data files which are frequently accessed and some drives will only rarely be accessed.
- In I/O intensive environments, performance is optimized by striping the drives in the array with stripes large enough so that each record potentially falls entirely within one stripe. This ensures that the data and I/O will be evenly distributed across the array, allowing each drive to work on a different I/O operation, and thus maximize the number of simultaneous I/O operations which can be performed by the array.

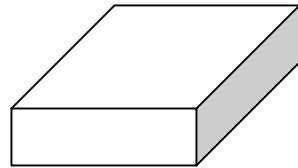
Data Striping

- In data intensive environments and single-user systems which access large records, small stripes (typically one 512-byte sector in length) can be used so that each record will span across all the drives in the array, each drive storing part of the data from the record. This causes long record accesses to be performed faster, since the data transfer occurs in parallel on multiple drives.
- Unfortunately, small stripes rule out multiple overlapped I/O operations, since each I/O will typically involve all drives. However, operating systems like DOS which does not allow overlapped disk I/O, will not be negatively impacted.
- Applications such as on-demand video/audio, medical imaging and data acquisition, which utilize long record accesses, will achieve optimum performance with small stripe arrays.

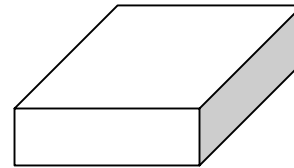
Data Striping



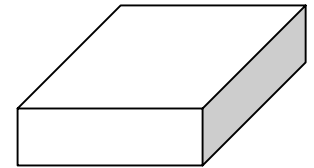
Disk 1



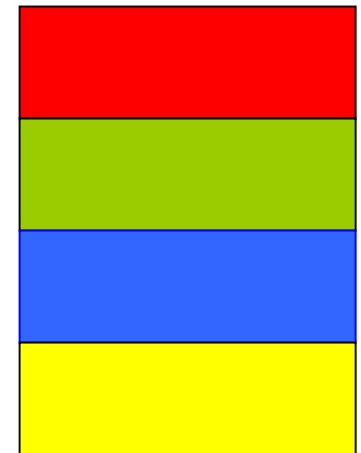
Disk 2



Disk 3



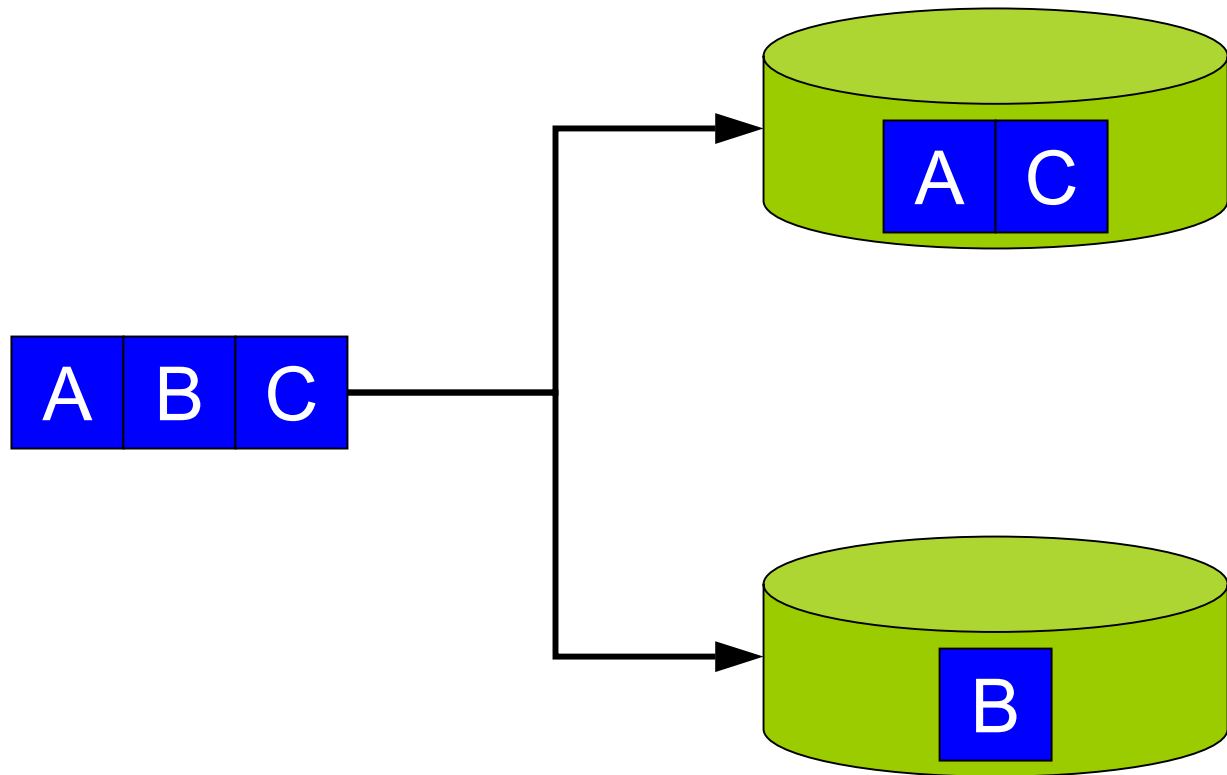
Disk 4



RAID-0

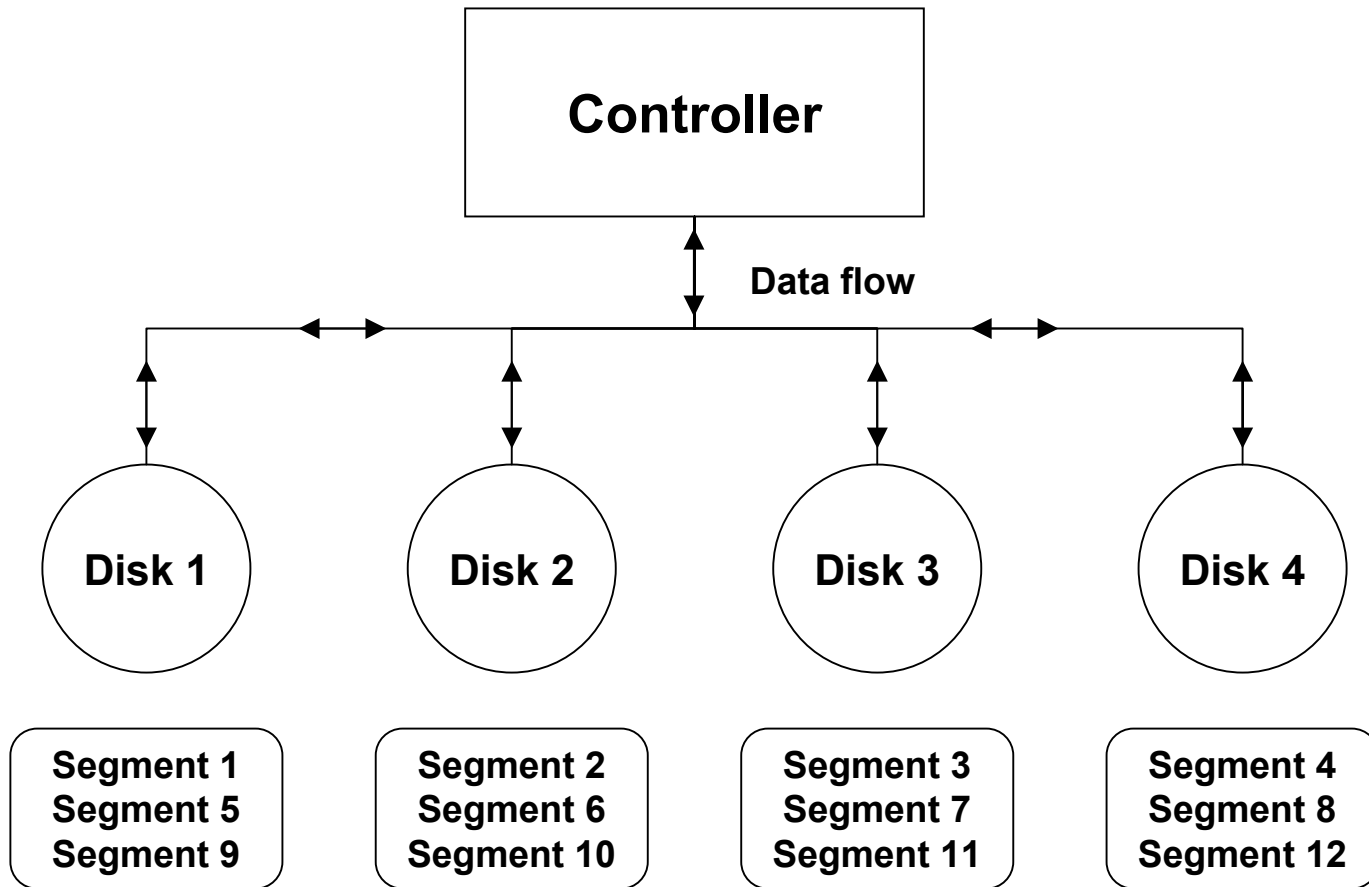
- RAID-0 is typically defined as a non-redundant group of striped disk drives without parity.
- RAID-0 arrays are usually configured with large stripes for I/O intensive applications, but may be sector-striped with synchronized spindle drives for single-user and data intensive environments which access long sequential records.
- Since RAID-0 does not provide redundancy, if one drive in the array crashes, the entire array crashes. However, RAID-0 arrays deliver the best performance and data storage efficiency of any array type.

RAID-0



Data is partitioned when it is stored

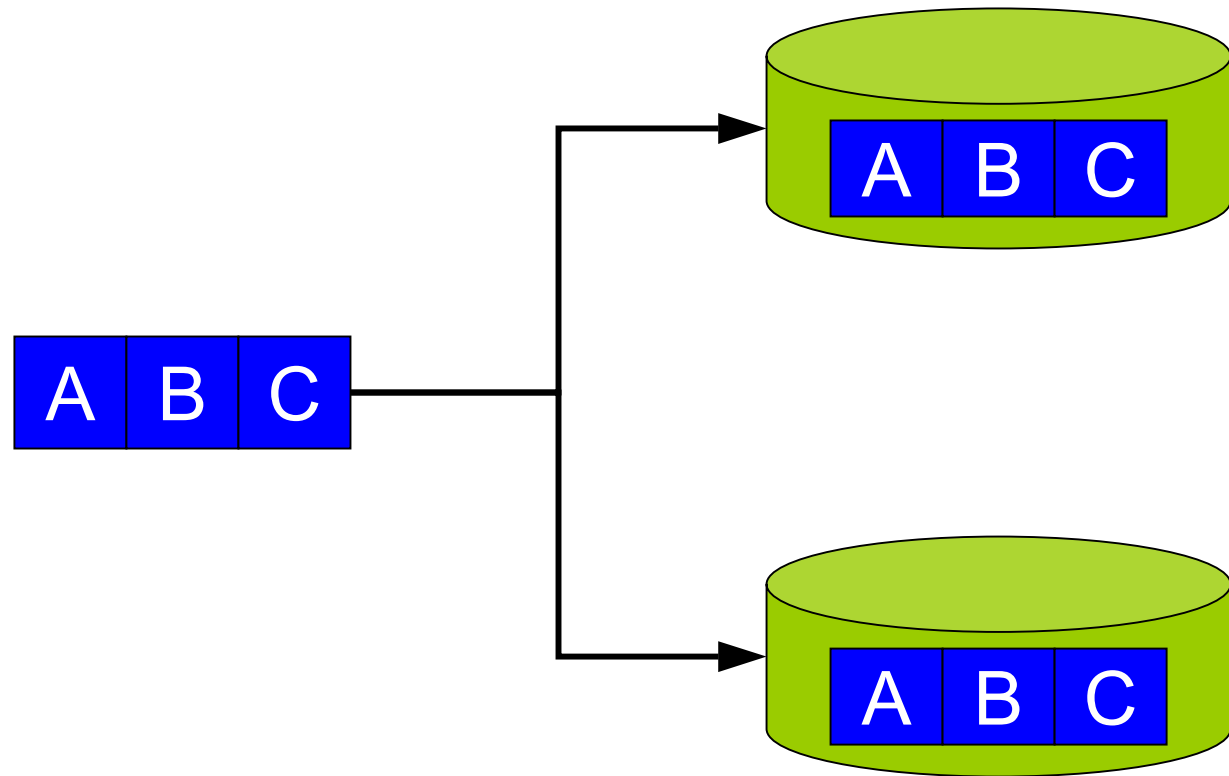
RAID-0



RAID-1

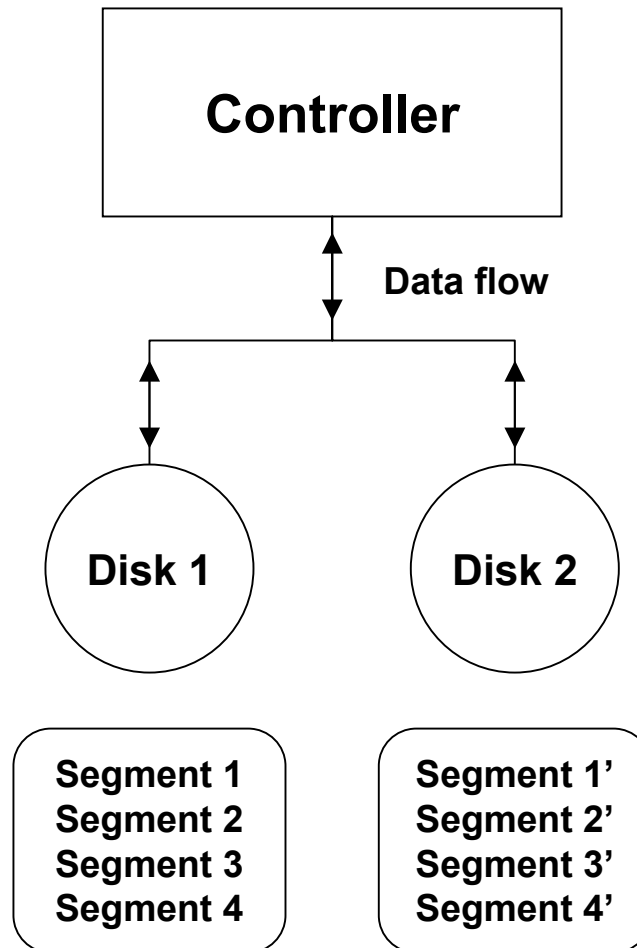
- RAID-1, better known as "disk mirroring", is simply a pair of disk drives which store duplicate data, but appears to the computer as a single drive.
- Striping is not used, although multiple RAID-1 arrays may be striped together to appear as a single larger array consisting of pairs of mirrored drives, typically referred to as "Dual-level array" or RAID 10.
- Writes must go to both drives in a mirrored pair so that the information on the drives is kept identical. Each individual drive, however, can perform simultaneous read operations.
- Mirroring thus doubles the read performance of an individual drive and leaves the write performance unchanged. RAID-1 delivers the best performance of any redundant array, especially in multi-user environments.

RAID-1



Identical data is stored on two separate disks

RAID-1



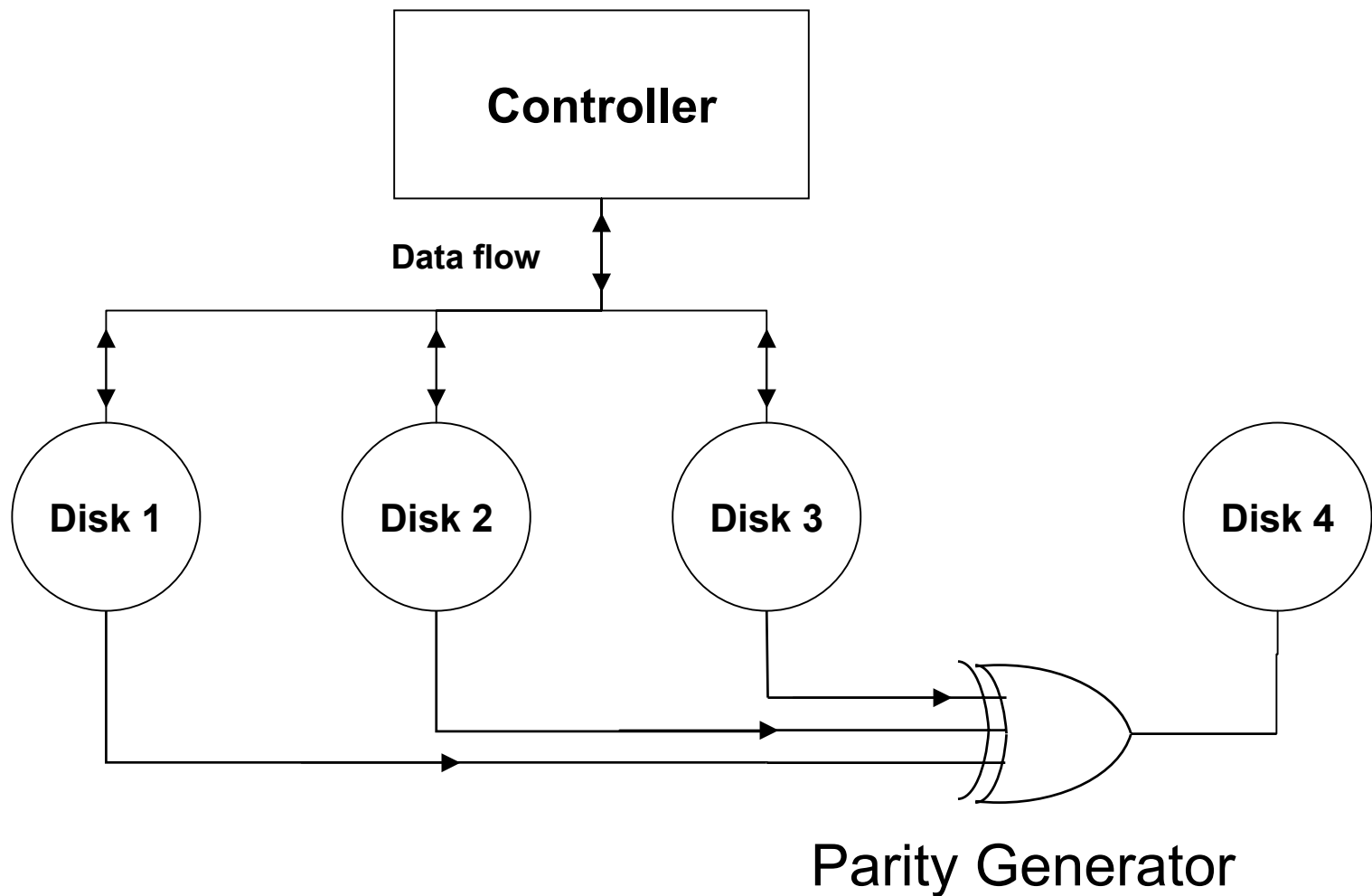
RAID-2

- RAID-2 arrays sector-stripe data across groups of drives, with some drives relegated to storing ECC information. Since most disk drives today embed ECC information within each sector, RAID-2 offers no significant advantages over RAID-3 architecture.

RAID-3

- 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 case of a hard 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, thereby optimizing data intensive environments. Since each I/O accesses all drives in the array, RAID-3 arrays cannot overlap I/O and thus deliver best performance in single-user, single-tasking environments with long records. Synchronized-spindle drives are required for optimum RAID-3 arrays in order to avoid performance degradation with short records.

RAID-3/4



RAID-4

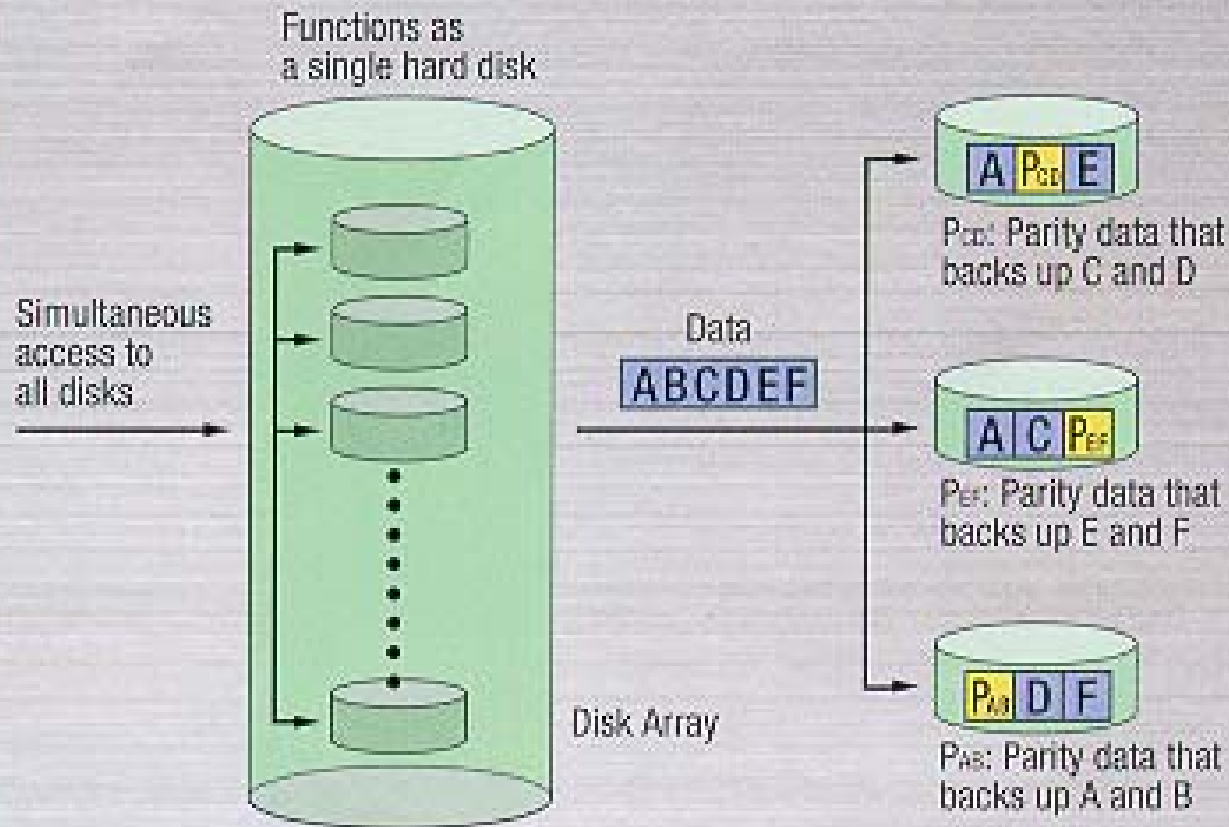
- RAID-4 is identical to RAID-3 except that large stripes are used, so that records can be read from any individual drive in the array (except the parity drive), allowing read operations to be overlapped. However, since all write operations must update the parity drive, they cannot be overlapped. This architecture offers no significant advantages over RAID-5.

RAID-5

- RAID-5, sometimes called a Rotating Parity Array, avoids the write bottleneck caused by the single dedicated parity drive of RAID-4. Like RAID-4, large stripes are used so that multiple I/O operations can be overlapped. However, unlike RAID-4, each drive takes turns storing parity information for a different series of stripes. 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 a single data drive, plus the parity drive for that record. Since, unlike RAID-4, different records store their parity on different drives, write operations can be overlapped.

RAID-5

RAID 5



Data is partitioned when it is stored;
each disk stores parity data that can be used to restore lost or damaged data

RAID-5

- RAID-5 offers improved storage efficiency over RAID-1 since parity information is stored, rather than a complete redundant copy of all data. The result is that any number of drives can be combined into a RAID-5 array, with the effective storage capacity of only one drive sacrificed to store the parity information. Therefore, RAID-5 arrays provide greater storage efficiency than RAID-1 arrays. However, this comes at the cost of a corresponding loss in performance.

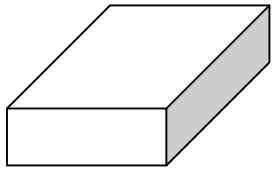
RAID-5

- When data is written to a RAID-5 array, the parity information must be updated. This is accomplished by finding out which data bits were changed by the write operation and then changing the corresponding parity bits. This is done by first reading the old data to be overwritten. This data is then XORed with the new data which is to be written. The result is a bit mask which has a one in the position of every bit which has changed. This bit mask is then XORed with the old parity information which is read from the parity drive. This results in the corresponding bits being changed in the parity information. The new updated parity is then written back to the parity drive. Therefore, for every application write request, a RAID-5 array must perform two reads, two writes and two XOR operations to complete the original write.

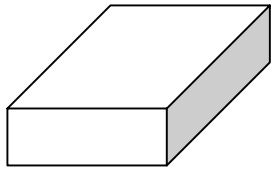
RAID-5

- The cost of storing parity, rather than redundant data, is the extra time taken during write operations to regenerate the parity information. This additional time results in a degradation of write performance for RAID-5 arrays over RAID-1 arrays by a factor of between 3:5 and 1:3. (i.e. RAID-5 writes are between $3/5$ and $1/3$ the speed of RAID-1 write operations.) Because of this, RAID-5 arrays should never be implemented in software and are not recommended for applications in which write performance is critically important.

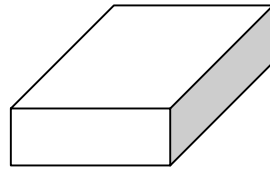
RAID-5



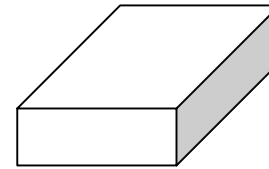
Disk 1



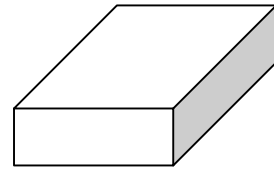
Disk 2



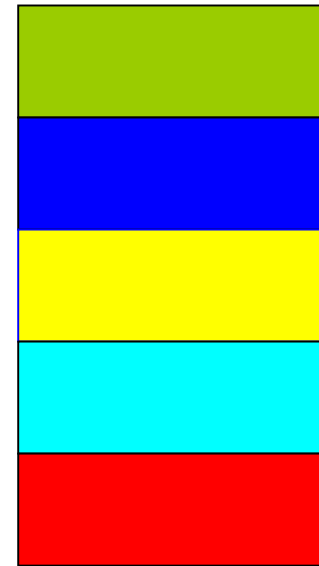
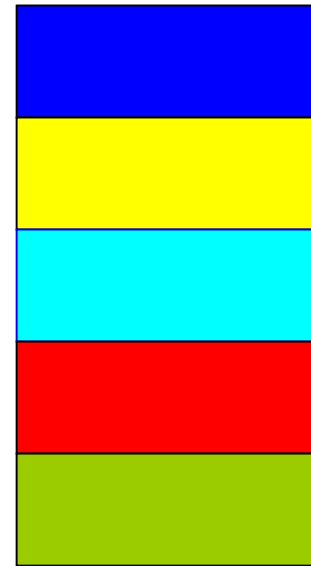
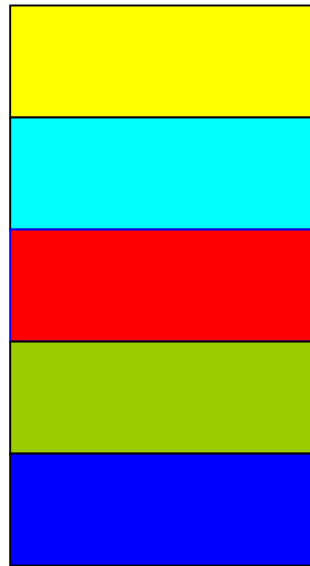
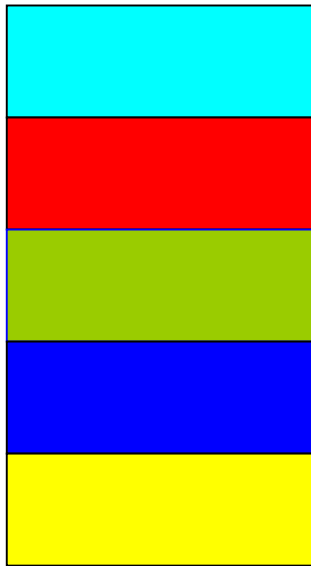
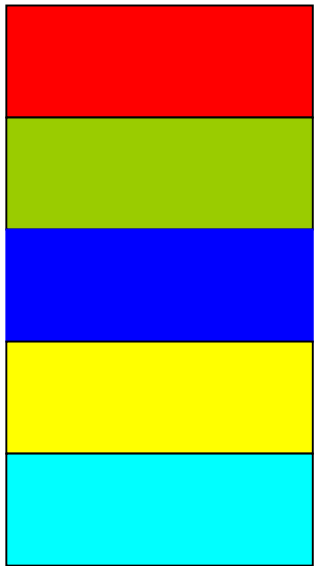
Disk 3



Disk 4



Disk 5



Parity information

Summary:

- RAID-0 is the fastest and most efficient array type but offers no fault-tolerance.
- RAID-1 is the array of choice for performance-critical, fault-tolerant environments. In addition, RAID-1 is the only choice for fault-tolerance if no more than two drives are desired.
- RAID-2 is seldom used today since ECC is embedded in almost all modern disk drives.
- RAID-3 can be used in data intensive or single-user environments which access long sequential records to speed up data transfer. However, RAID-3 does not allow multiple I/O operations to be overlapped and requires synchronized-spindle drives in order to avoid performance degradation with short records.
- RAID-4 offers no advantages over RAID-5 and does not support multiple simultaneous write operations.
- RAID-5 is the best choice in multi-user environments which are not write performance sensitive. However, at least three, and more typically five drives are required for RAID-5 arrays.

Software RAID

- Pure software RAID implements the various RAID levels in the kernel disk (block device) code. Pure-software RAID offers the cheapest possible solution: not only are expensive disk controller cards or hot-swap chassis not required, but software RAID works with cheaper IDE disks as well as SCSI disks. With today's fast CPU's, software RAID performance can hold its own against hardware RAID in all but the most heavily loaded systems. The Software RAID is becoming increasingly fast, feature-rich and reliable, making many of the lower-end hardware solutions uninteresting. Expensive, high-end hardware may still offer advantages in management, reliability, dual-hosting, hot-swap, etc. but are no longer required for low-end casual deployment.

RAID Disk Controllers

- Disk Controllers are adapter cards that plug into the ISA/EISA/PCI bus.
- Just like regular disk controller cards, a cable attaches them to the disk drives.
- Unlike regular disk controllers, the RAID controllers will implement RAID on the card itself, performing all necessary operations to provide various RAID levels.

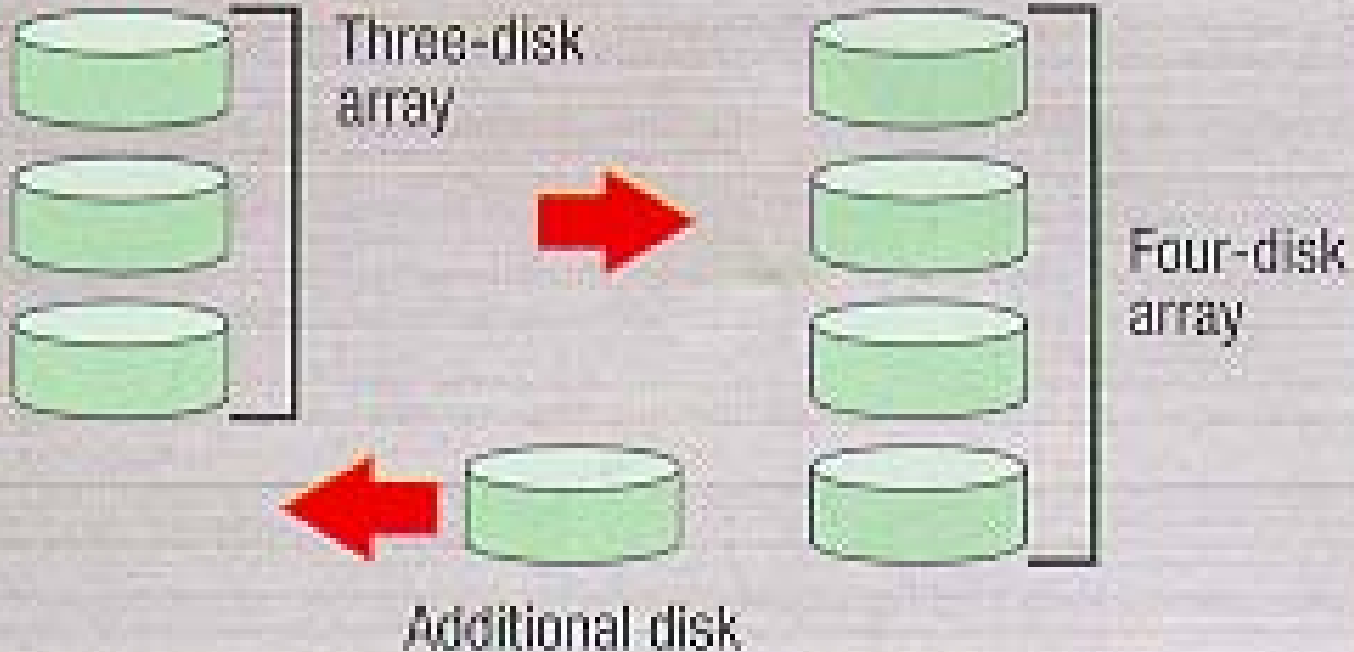
RAID Disk Controllers

- If the RAID disk controller has a modern, high-speed DSP/controller on board, and a sufficient amount of cache memory, it can outperform software RAID, especially on a heavily loaded system.
- However, using an old controller on a modern, fast 2-way or 4-way SMP machine may easily prove to be a performance bottle-neck as compared to a pure software-RAID solution.

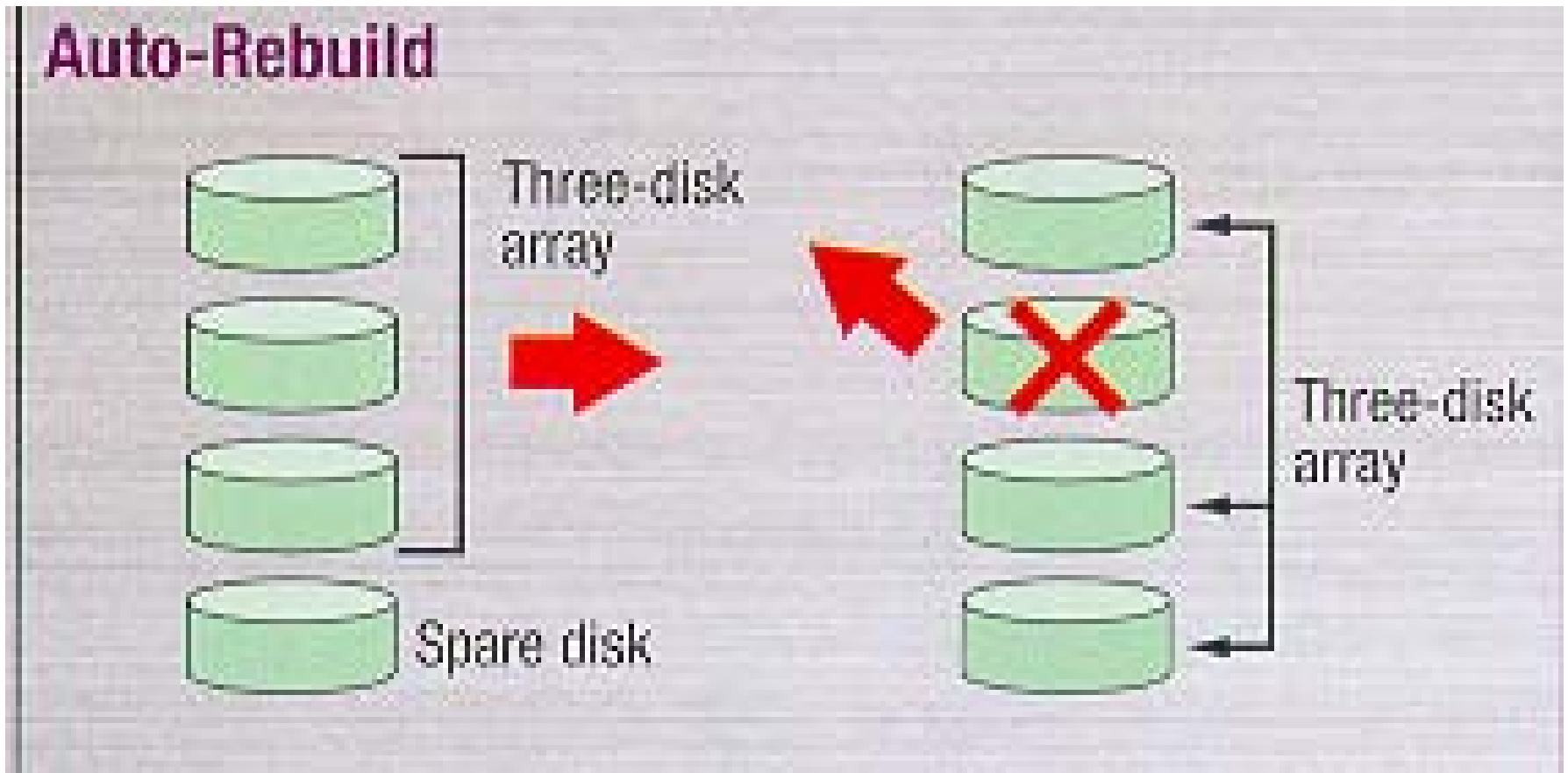
Hot Pluggability

Hot Plug Functionality

On-line Expansion



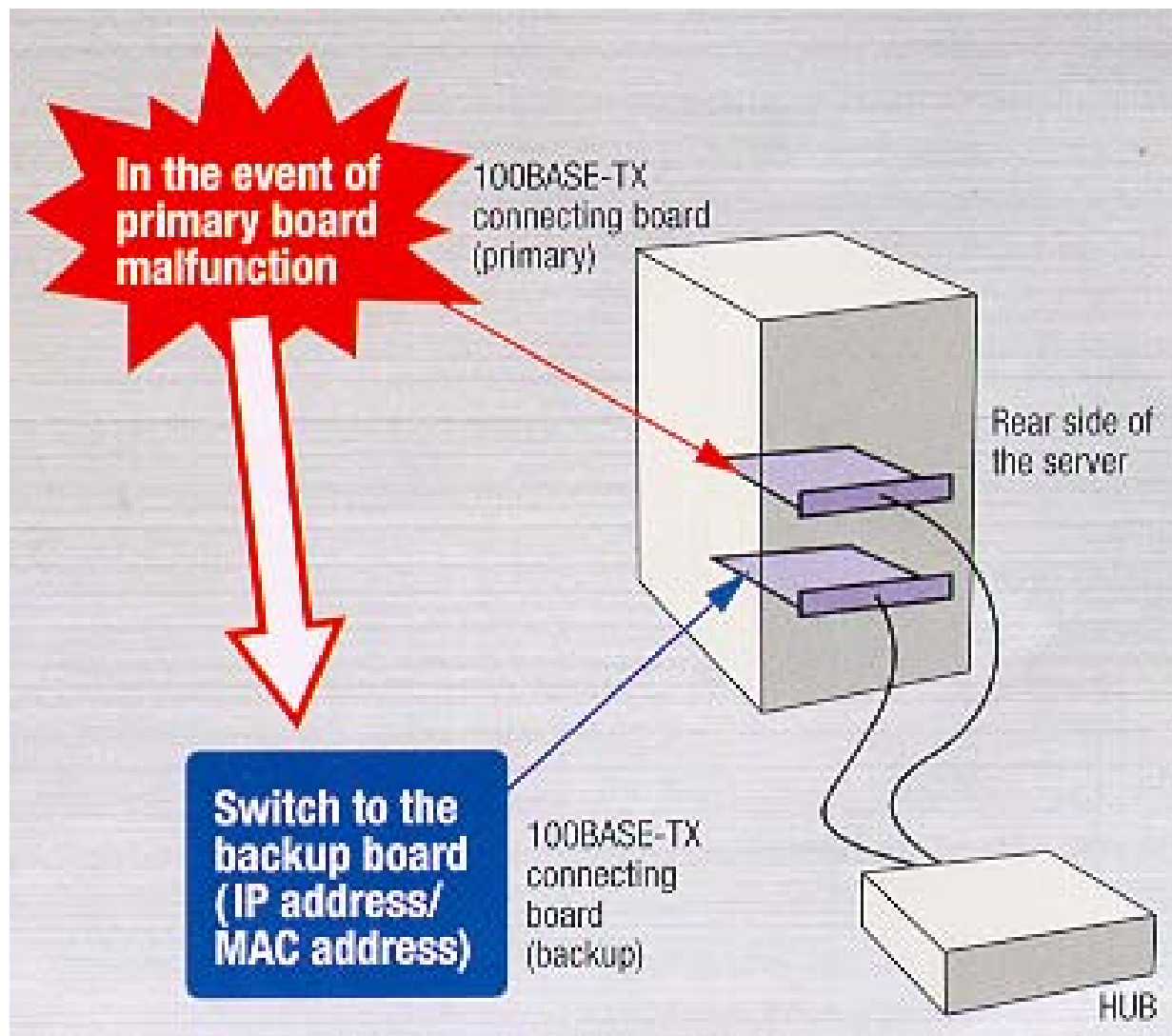
Hot Spare



Network Adapter Fault Tolerance

- Adapter Fault Tolerance provides link redundancy for two Network adapters. When configured, there becomes a primary and secondary server adapter.
- If the primary loses communication with the hub/switch, the secondary automatically takes over.
- The secondary adapter will take over for such reasons as cable connection problems, switch or hub port failure or adapter failure.

Network Card Fault Tolerance



Availability Features In a Server

- Hot-plug ready PCI slots
- Hot-plug hard drives allow replacement of failed drive without powering server down
- Redundant, hot-pluggable power supplies help remove the power supply as a single point of failure. Delivering consistent, reliable power. Three hot-pluggable redundant power supplies standard
- Individual power cords further increase the redundancy of the power supplies
- Redundant, hot-pluggable hard drive cooling fans and processor cooling fans make service replacement simple
- Redundant NIC solutions