03 infra TI

RAID

Universidade Presbiteriana Mackenzie

MTBF; RAID Protection; Mirroring and Parity; RAID levels; write penalty

Por que RAID ?

Redundant Array Inexpensive Disks x Redudant Array Independent Disks

Performance limitation of disk drive

An individual drive has a certain life expectancy

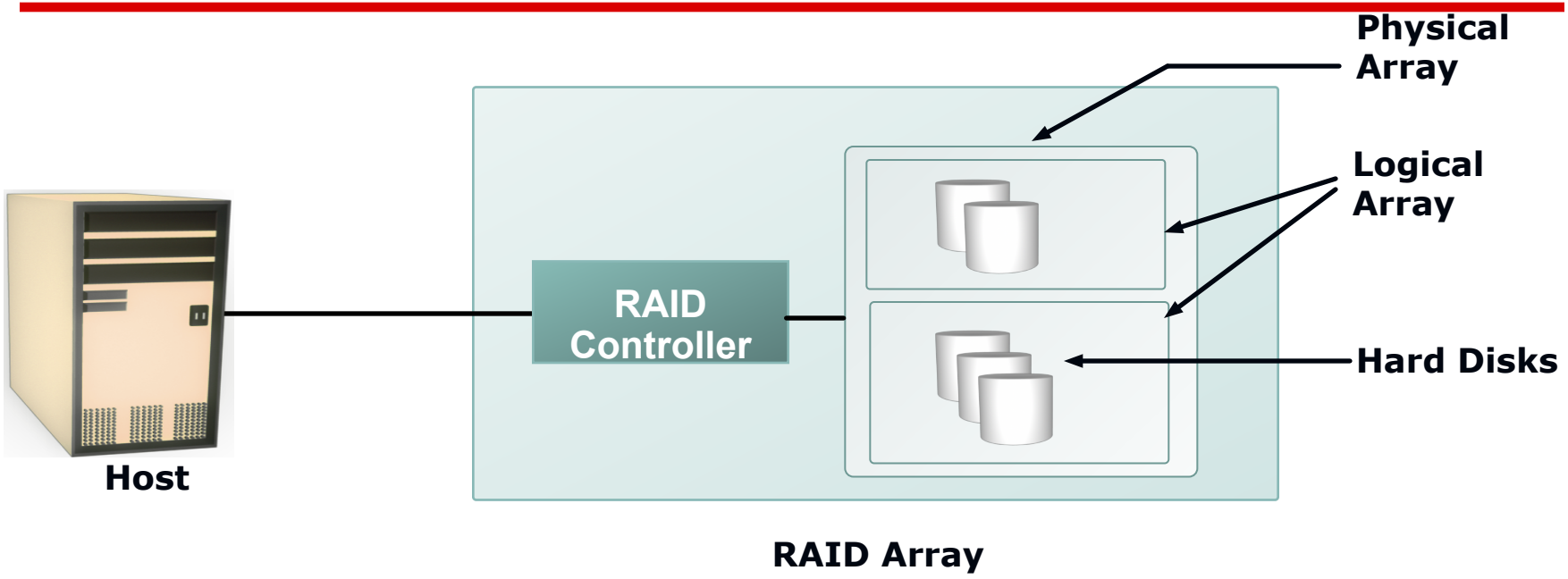
- Measured in MTBF (Mean Time Between Failure)
- The more the number of HDDs in a storage array, the larger the probability for disk failure.
  For example: If the MTBF of a drive is 750,000 hours, and there are 100 drives in the array, then the MTBF of the array becomes 750,000 / 100, or 7,500 hours

RAID was introduced to mitigate this problem

RAID provides:

- Increase capacity
- Higher availability
- Increased performance

Disk array components



RAID: SW vs. HW

Hardware (usually a specialized disk controller card) **Melhor escolha!**

- Controls all drives attached to it
- Array(s) appear to host operating system as a regular disk drive
- Provided with administrative software

Software **Unix, Oracle e outros sistemas**

- Runs as part of the operating system
- Performance is dependent on CPU workload
- Does not support all RAID levels

RAID levels

Table 3-1: Raid Levels

LEVELS	BRIEF DESCRIPTION
RAID 0	Striped array with no fault tolerance
RAID 1	Disk mirroring
RAID 3	Parallel access array with dedicated parity disk
RAID 4	Striped array with independent disks and a dedicated parity disk
RAID 5	Striped array with independent disks and distributed parity
RAID 6	Striped array with independent disks and dual distributed parity
Nested	Combinations of RAID levels. Example: RAID 1 + RAID 0

Disk Stripes

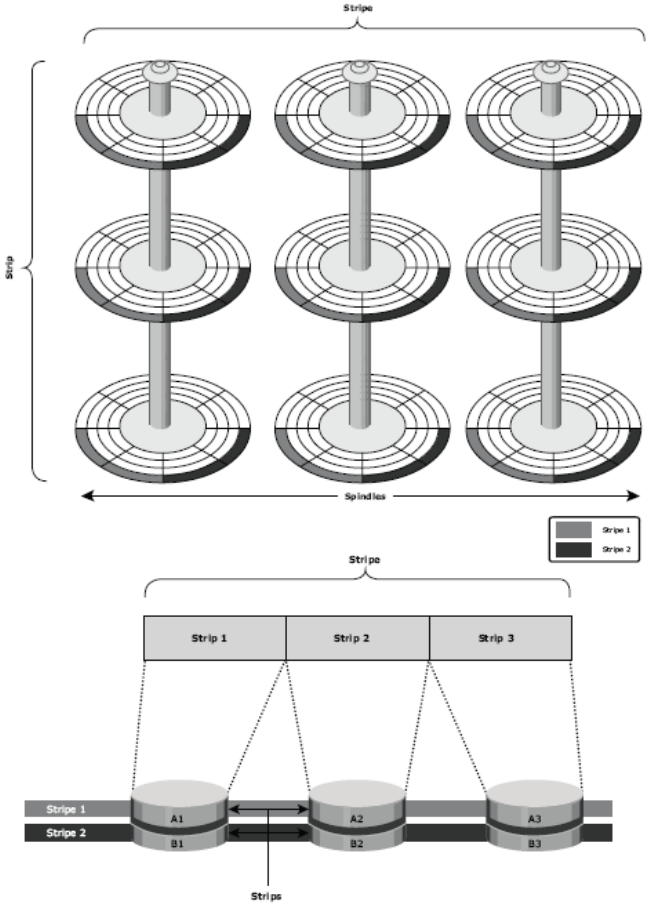


Figure 3-2: Striped RAID set

Mirroring & Parity

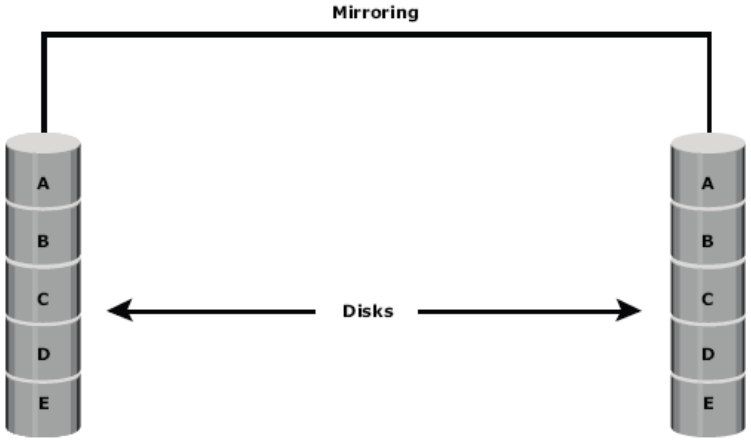


Figure 3-3: Mirrored disks in an array

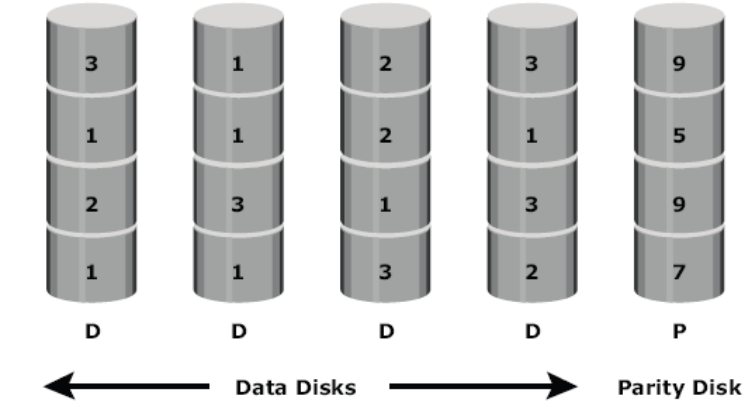


Figure 3-4: Parity RAID

RAID 0, RAID 1 and write penalty

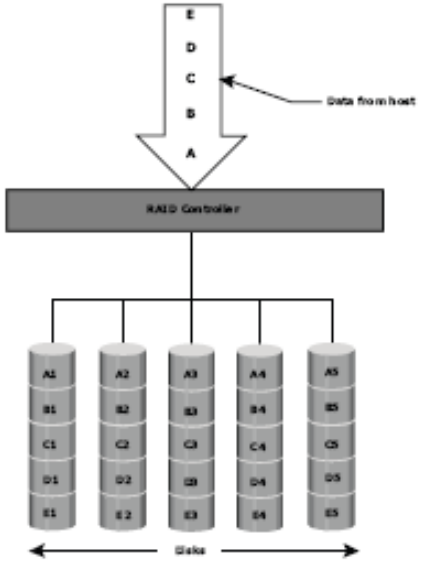


Figure 3-5: RAID 0

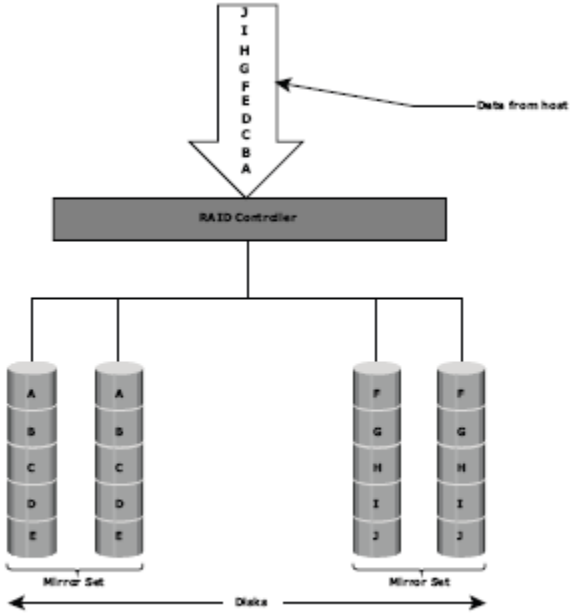


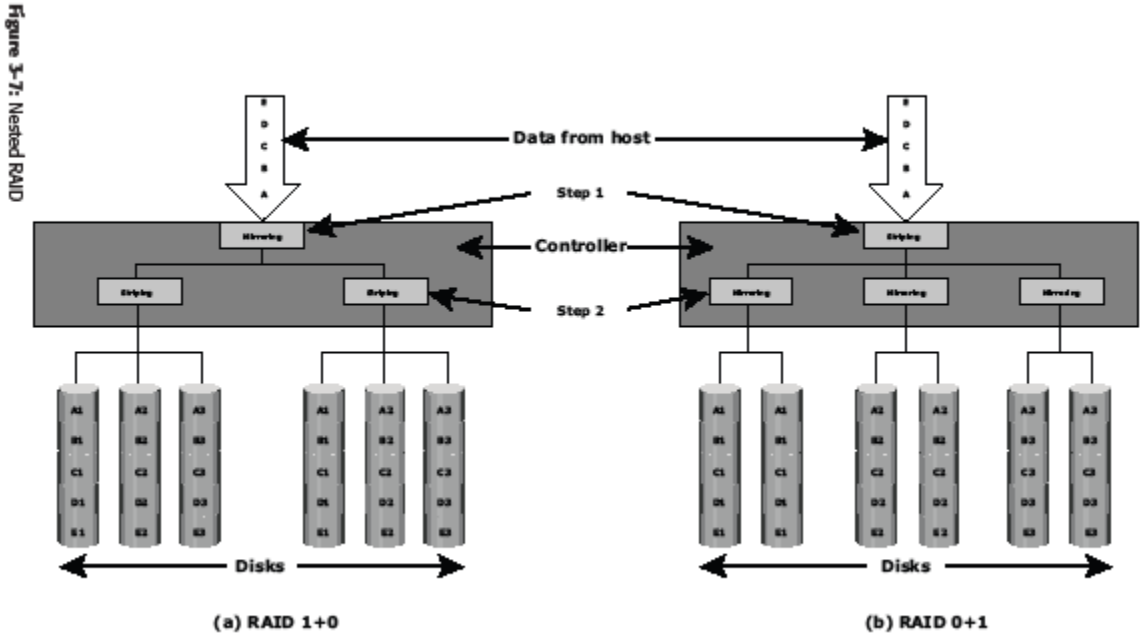
Figure 3-6: RAID 1

Write Penalty vs. Full Protection...

Nested RAID 1+0 0+1

RAID 1+0 – Striped Mirror

RAID 0+1 – Mirrored Stripe



RAID 3, 4

Stripes data for high performance and uses parity for improved fault tolerance. One drive is dedicated for parity information. If a drive files, data can be reconstructed using data in the parity drive.

For RAID 3, data read / write is done across the entire stripe.

Provide good bandwidth for large sequential data access such as video streaming.

For RAID 4, data read/write can be independently on single disk.

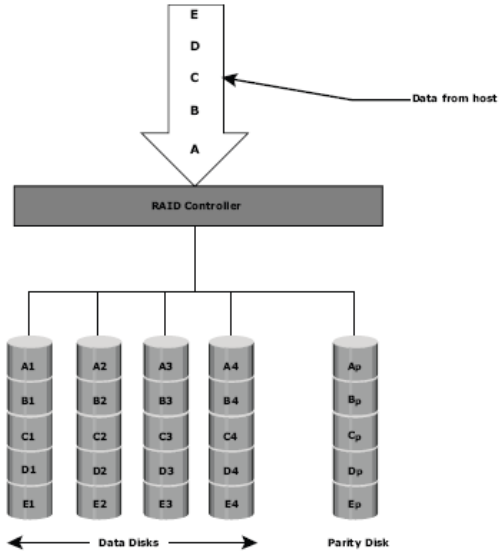


Figure 3-8: RAID 3

RAID 5, 6

RAID 5 is similar to RAID 4, except that the parity is distributed across all disks instead of stored on a dedicated disk.

This overcomes the write bottleneck on the parity disk.

It is largely used by Database systems

RAID 6 is similar to RAID 5, except that it includes a second parity element to allow survival in the event of two disk failures.

The probability for this to happen increases and the number of drives in the array increases.

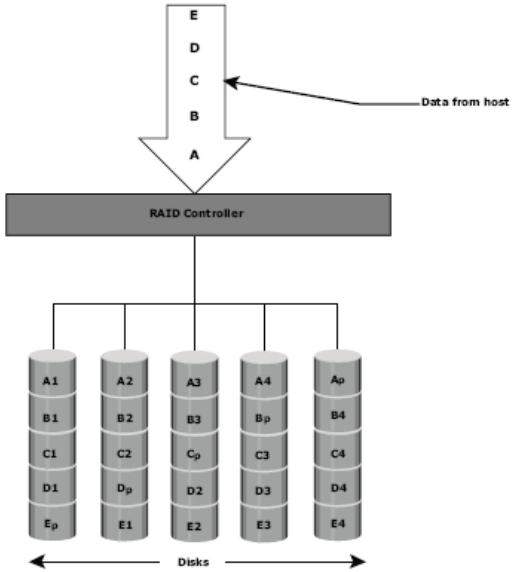


Figure 3-9: RAID 5

RAID Comparative

RAID	Min Disks	Storage Efficiency %	Cost	Read Performance	Write Performance
0	2	100	Low	Very good for both random and sequential read	Very good
1	2	50	High	Good Better than a single disk	Good Slower than a single disk, as every write must be committed to two disks
3	3	(n-1)*100/n where n= number of disks	Moderate	Good for random reads and very good for sequential reads	Poor to fair for small random writes Good for large, sequential writes
5	3	(n-1)*100/n where n= number of disks	Moderate	Very good for random reads Good for sequential reads	Fair for random write Slower due to parity overhead Fair to good for sequential writes
6	4	(n-2)*100/n where n= number of disks	Moderate but more than RAID 5	Very good for random reads Good for sequential reads	Good for small, random writes (has write penalty)
1+0 and 0+1	4	50	High	Very good	Good

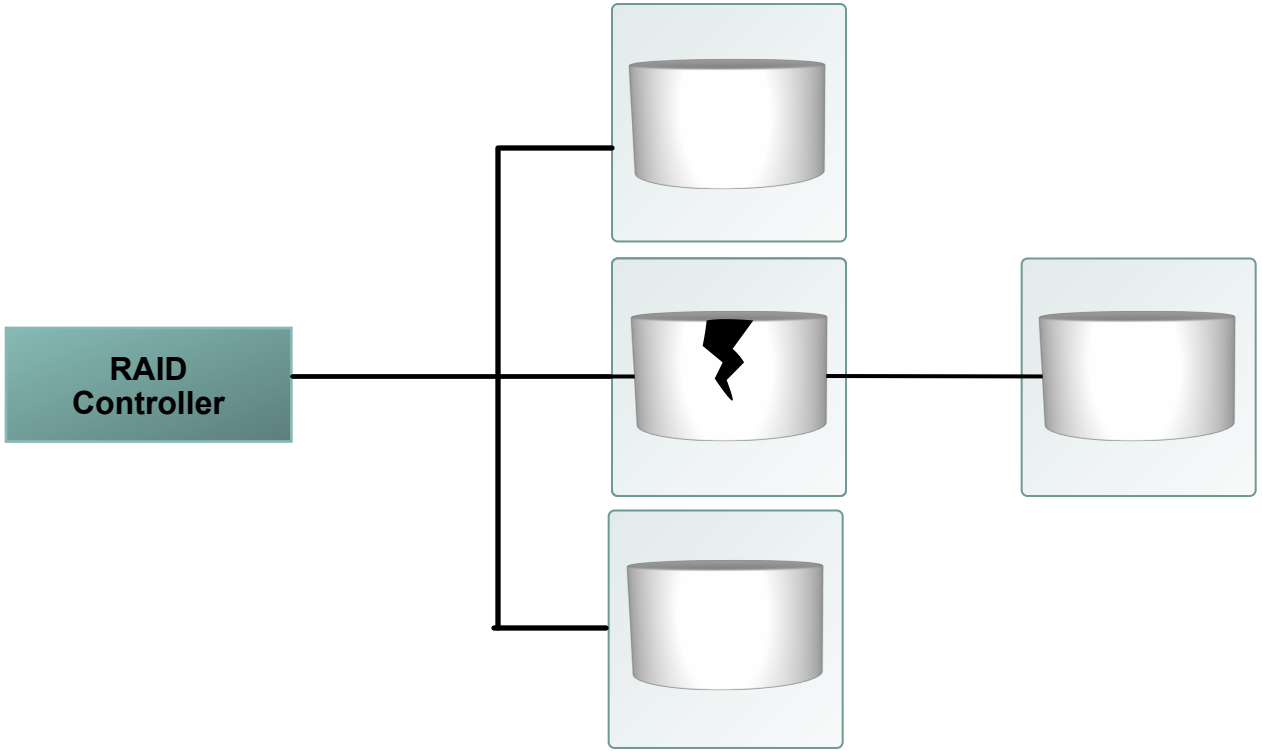
Compute penalty example

Consider an application that generates 5,200 IOPS, with 60 percent of them being reads.

The disk load in RAID 5 is calculated as follows:
RAID 5 disk load = 0.6 × 5,200 + 4 × (0.4 × 5,200) [because the write penalty for RAID 5 is 4]
= 3,120 + 4 × 2,080
= 3,120 + 8,320
= 11,440 IOPS

The disk load in RAID 1 is calculated as follows:
RAID 1 disk load = 0.6 × 5,200 + 2 × (0.4 × 5,200) [because every write manifests as two writes to the disks]
= 3,120 + 2 × 2,080
= 3,120 + 4,160
= 7,280 IOPS

Hot spare disks



Discussão e exercícios

Por que há uma penalidade de WRITE mas não de READ nos mecanismos de RAID?

Em geral as controladoras de disco local dos servidores implementam RAID 1 enquanto grandes sistemas de armazenamento em geral optam por RAID 5 ou suas variantes. Por que?

Compare os mecanismos de espelhamento e paridade.

Altere o exemplo de cálculo de write penalty na condição de que somente ¼ das operações são de gravação. Há penalty para o RAID 0?

Que tipo de gargalo RAID 3 apresenta quando comparado com o RAID 5?

Leitura recomendada

Capítulo 3

Information Storage and Management Storing, Managing, and Protecting Digital Information in Classic, Virtualized, and Cloud Environments

2nd Edition Edited by Somasundaram Gnanasundaram, Alok Shrivastava