Storage Technologies - 2

COMP 25212 - Lecture 9

Antoniu Pop

antoniu.pop@manchester.ac.uk

28 February 2018

Previous lecture

- Characterisation of storage technologies
  - Write once/many/not-too-many and Read many times
- Performance model
  - Seek, Search and Transfer time
  - Latency and Bandwidth
- Limitations
  - Mechanical constraints - latency and reliability issues
- RAID (Redundant Array of Independent/Inexpensive Disks)

Hard Disk Performance (recap)

Seek time Time for the **head** to reach the target **track**.

Search time Time for the target **sector** to arrive under the **head**. Also called *rotational latency*.

Transfer rate Amount of data that can be read / written per unit of time. Dependent on access patterns.

Aka. “sustained transfer rate” in contrast to “interface transfer rate”

Disk access time = seek time + search time + transfer time

Note: all values are average as they depend on many factors.

Learning Objectives - Storage 2

- Motivate RAID
- Understand the principles of RAID configurations
- Understand how RAID impacts performance and reliability
- Understand failure & recovery constraints

Historic comparison

1956 first HDD IBM 350: $\sim$ 3.5 MB (enough to store one selfie!)
2015 first 10 TB disk: 1000s of times smaller, $\mathbf{3 \cdot 10^6} \times$ capacity

1000s of times cheaper!

Source: `https://www-03.ibm.com/ibm/history/exhibits/storage/storage_350.html`

Technology Trends and Drivers

1956 - 1980s

- Mainframe/Server Disk Drives
  - High capacity, large formats (e.g., 14”)
  - Expensive, low volume market
  - Somewhat slow evolution
- PCs used mostly floppy disks

(Early) 1990s

- Still two markets: Server & PC drives
- Most PCs use hard disks
- PC hard disk sales explode – high volume market
  - Drives costs lower
  - Drives disk technology faster

How to use PC disks to build server-class storage?

RAID

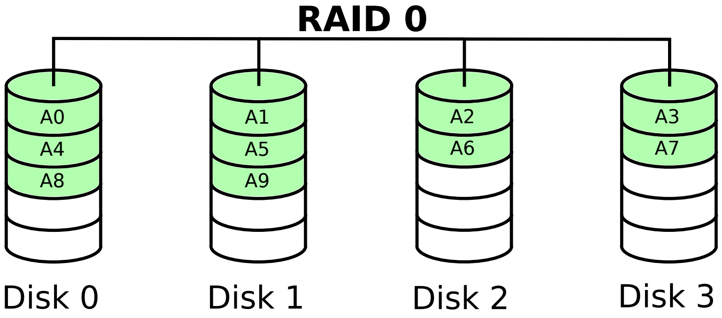
Redundant **A**rray of **I**ndependent **D**isks

- Compensate for loss of reliability, capacity, performance
- Use lots of cheap(er), (disposable ?) disks

Disk Problems and Solutions (recap)

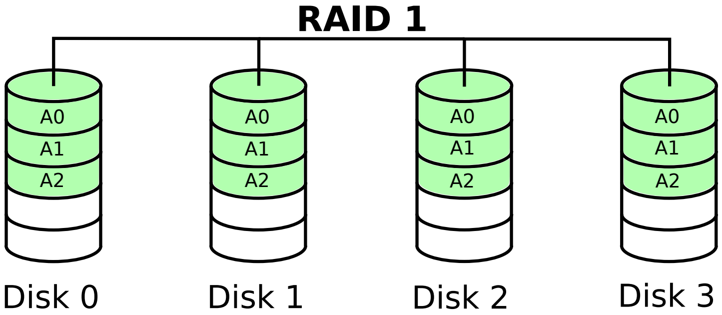
- Disks are too small
  - Fixed: use multiple disks
- Disks are too slow
  - Fixed: disk striping (RAID 0)
- **Disks are unreliable**
  - Fixed: disk mirroring (RAID 1)
  - Data redundancy

RAID 0 — Striping



Number of disks	***n***	**4**
Read (short ... long)	**1**× ... ***n***×	**1**× ... **4**×
Write (short ... long)	**1**× ... ***n***×	**1**× ... **4**×
Failure tolerance	**0** disks	**0** disks
Capacity efficiency	**1**	**100%**

RAID 1 — Mirroring



Number of disks	n	**4**
Read (short ... long)	$1\times$ ... $n\times$	$1\times$ ... $4\times$
Write (short ... long)	$1\times$ ... $1\times$	$1\times$ ... $1\times$
Failure tolerance	$n-1$ disks	**3** disks
Capacity efficiency	$1/n$	**25%**

Parity

- Old idea: first tape drive (1951) had a parity track
- Transverse redundancy check

- How does it really work ?

Parity

Assume we have three blocs $\boldsymbol{A_0}, \boldsymbol{A_1}, \boldsymbol{A_2}$:

$\boldsymbol{A_0}$	1	0	0	1	0	1	1	1	0	0	0
$\boldsymbol{A_1}$	0	1	1	1	1	0	0	0	1	0	1
$\boldsymbol{A_2}$	0	1	0	1	0	1	1	0	1	1	1
$\boldsymbol{A_p}$ (parity)	1	0	1	1	1	0	0	1	0	1	0

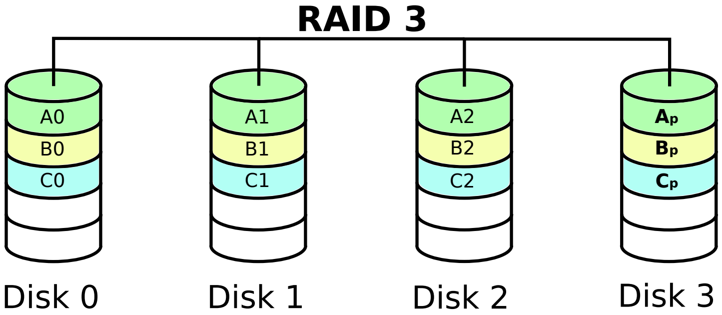
Where $\boldsymbol{A_p} = \boldsymbol{A_0} \oplus \boldsymbol{A_1} \oplus \boldsymbol{A_2}$.
And, importantly:

$$\boldsymbol{A_0} = \boldsymbol{A_p} \oplus \boldsymbol{A_1} \oplus \boldsymbol{A_2}$$

$$\boldsymbol{A_1} = \boldsymbol{A_0} \oplus \boldsymbol{A_p} \oplus \boldsymbol{A_2}$$

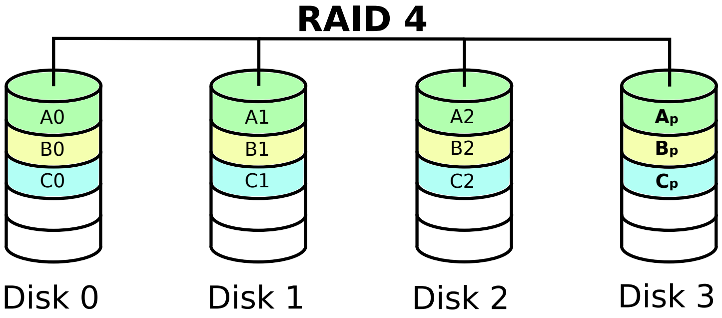
$$\boldsymbol{A_2} = \boldsymbol{A_p} \oplus \boldsymbol{A_1} \oplus \boldsymbol{A_0}$$

RAID 3 — Byte-Striping + Parity



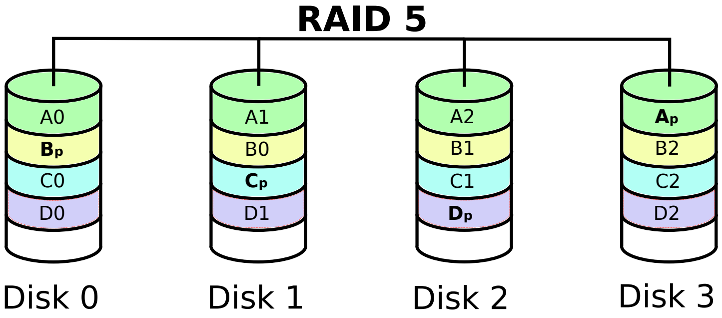
Number of disks	$n+1$	**4**
Read (short ... long)	$1\times$ ... $n\times$	$1\times$ ... $3\times$
Write (short ... long)	$1\times$ ... $n\times$	$1\times$ ... $3\times$
Failure tolerance	**1** disks	**1** disks
Capacity efficiency	$n/(n+1)$	**75%**

RAID 4 — Block-Striping + Parity



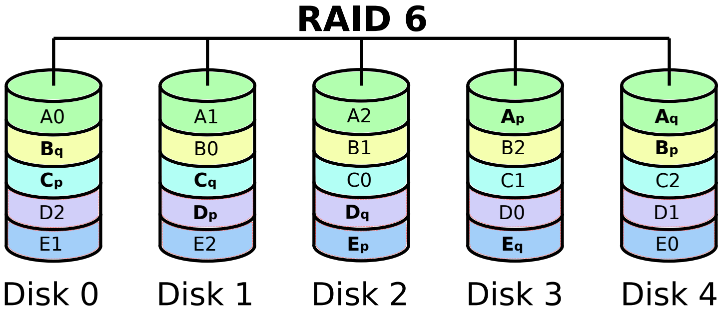
Number of disks	$n + 1$	**4**
Read (short ... long)	$1\times$ ... $n\times$	$1\times$ ... $3\times$
Write (short ... long)	$0.5\times$ (RMW) ... $n\times$	$0.5\times$ ... $3\times$
Failure tolerance	**1** disks	**1** disks
Capacity efficiency	$n/(n+1)$	**75%**

RAID 5 — Block-Striping + Distrib. Parity



Number of disks	$n+1$	**4**
Read (short ... long)	$1\times$ ... $n+1\times$	$1\times$ ... $4\times$
Write (short ... long)	$0.5\times$ (RMW) ... $n\times$	$0.5\times$ ... $3\times$
Failure tolerance	**1** disks	**1** disks
Capacity efficiency	$n/(n+1)$	**75%**

RAID 6 — Double Distributed Parity

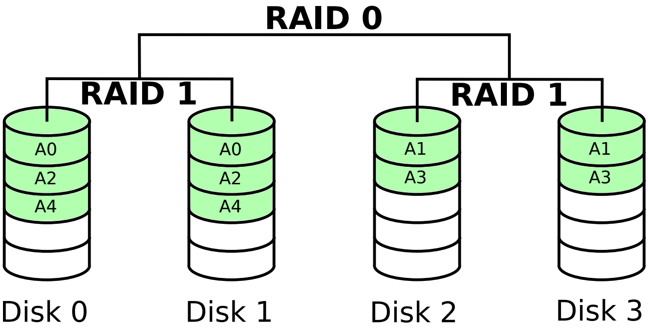


Number of disks	$n+2$	**5**
Read (short ... long)	$1\times$ ... $n+2\times$	$1\times$ ... $5\times$
Write (short ... long)	$0.5\times$ (RMW) ... $n\times$	$0.5\times$ ... $3\times$
Failure tolerance	**2** disks	**2** disks
Capacity efficiency	$n/(n+2)$	$3/5 = 60\%$

Nested RAID

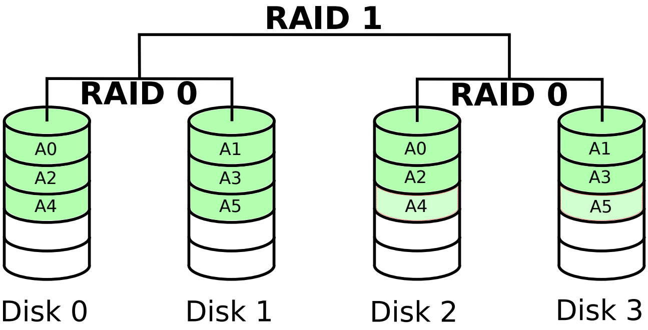
- Each raid configuration comes with tradeoffs
- Combine RAID configurations to alleviate shortcomings
- Multiple RAID layers

RAID 1+0 (aka. RAID 10)



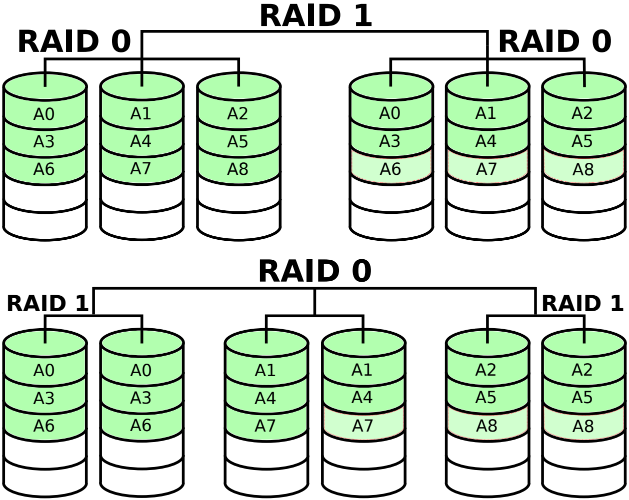
Number of disks	$m_{\text{[RAID 1]}} \cdot n_{\text{[RAID 0]}}$	$2 \cdot 2 = 4$
Read (short ... long)	$1\times$ ... $n \cdot m\times$	$1\times$ ... $4\times$
Write (short ... long)	$1\times$ ... $n\times$	$1\times$ ... $2\times$
Failure tolerance	$m - 1$ disks	1 disks
Capacity efficiency	$n/(m \cdot n) = 1/m$	$1/2 = 50\%$

RAID 01

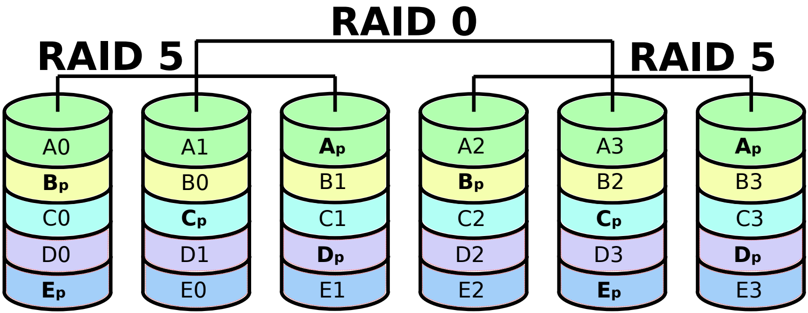


Number of disks	$m_{[\text{RAID 1}]} \cdot n_{[\text{RAID 0}]}$	$2 \cdot 2 = 4$
Read (short ... long)	$1\times$ ... $n \cdot m\times$	$1\times$ ... $4\times$
Write (short ... long)	$1\times$ ... $n\times$	$1\times$ ... $2\times$
Failure tolerance	$m - 1$ disks	1 disks
Capacity efficiency	$n/(m \cdot n) = 1/m$	$1/2 = 50\%$

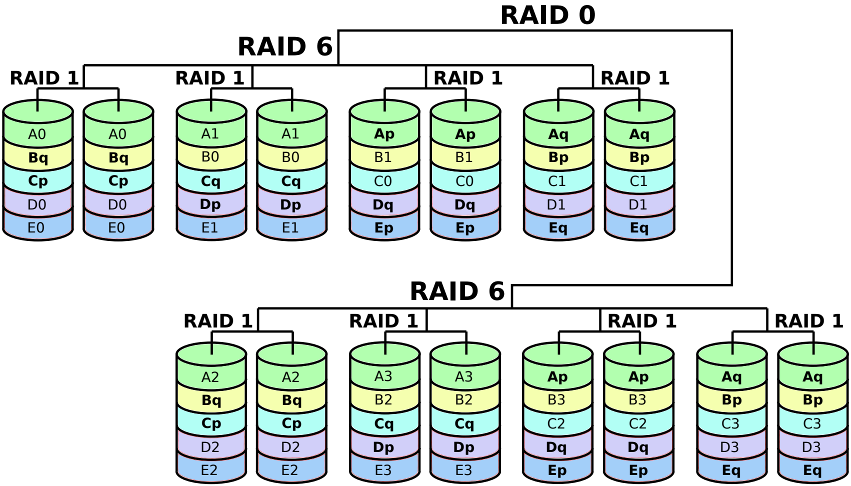
RAID 10 vs. 01 — different?



RAID 50


RAID 0
RAID 5
RAID 5
A0
Bp
C0
D0
Ep
A1
B0
Cp
D1
E0
Ap
B1
C1
Dp
E1
A2
Bp
C2
D2
E2
A3
B2
Cp
D3
Ep
Ap
B3
C3
Dp
E3

RAID 160


RAID 0
RAID 6
RAID 1
RAID 1
RAID 1
RAID 1
A0 Bq Cp D0 E0
A0 Bq Cp D0 E0
A1 B0 Cq Dp E1
A1 B0 Cq Dp E1
Ap B1 C0 Dq Ep
Ap B1 C0 Dq Ep
Aq Bp C1 D1 Eq
Aq Bp C1 D1 Eq
RAID 6
RAID 1
RAID 1
RAID 1
RAID 1
A2 Bq Cp D2 E2
A2 Bq Cp D2 E2
A3 B2 Cq Dp E3
A3 B2 Cq Dp E3
Ap B3 C2 Dq Ep
Ap B3 C2 Dq Ep
Aq Bp C3 D3 Eq
Aq Bp C3 D3 Eq

RAID Failure Mode Operation

What happens when a disk fails ?

RAID 0	Lose all data (hope there’s more than one RAID layer)
RAID 1	Business as usual, hot-swap the failed disk
RAID 2-6	Operate in degraded mode

- If data drive failed, every read must be reconstructed
- If parity drive failed, low performance impact

- Replace drive (hot-swapping: the system continues running)
- Rebuild the array (re-constitute the state of the lost drive)

RAID Recovery Limitations

Rebuilding a degraded array

- Sequentially
- How long ?
- On live system ?

Risk of failure during recovery

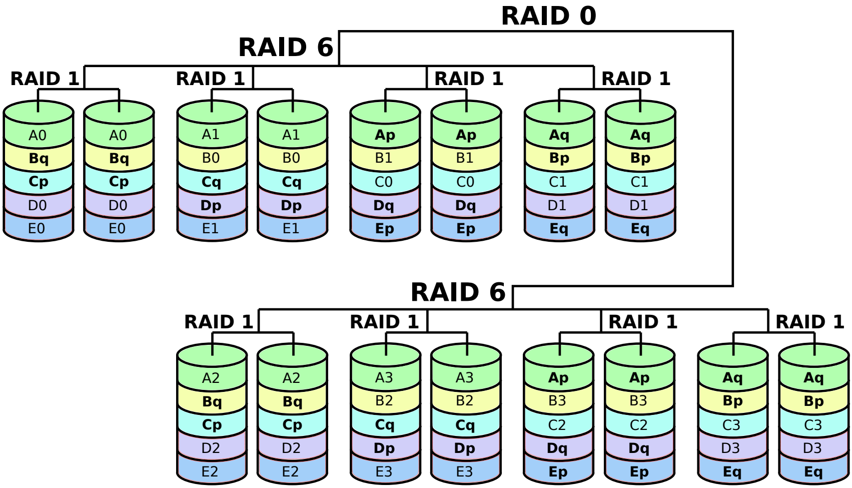
- Statistical distortion:

  higher risk of multiple failures within a narrow time frame

- RAID 5 risk advisory notice:

  Do not use for business-critical data! [Dell]

RAID 160



No performance degradation on disk failure!

Where to Implement RAID?

- Software – Operating System
  - Most OS now provide software RAID
  - E.g. Linux **md** (multiple devices) supports RAID 0, 1, 4, 5, 6 plus nestings
- Software – File System
  - E.g., ZFS
- Dedicated hardware (RAID controller)

Array Failure Rates (full data loss)

Failure rate of a disk drive: $\boldsymbol{r}$ (with **some** assumptions!)

Failure rate $\mathcal{R}$ of an array of $\boldsymbol{n}$ disks (RAID) where $\boldsymbol{k}$ disks can safely fail:

$$\mathcal{R} = \mathbf{1} - \left(\mathcal{P}(\mathbf{0}) + \mathcal{P}(\mathbf{1}) + \ldots + \mathcal{P}(\boldsymbol{k})\right)$$

where $\mathcal{P}(\boldsymbol{i})$ is the probability of precisely $\boldsymbol{i}$ disks failing:

$$\mathcal{P}(\boldsymbol{i}) = \binom{\boldsymbol{n}}{\boldsymbol{i}} \boldsymbol{r}^{\boldsymbol{i}} (\mathbf{1} - \boldsymbol{r})^{\boldsymbol{n}-\boldsymbol{i}}$$

Array Failure Rates for RAID configurations

RAID 0 $1-(1-r)^n$ (0 disks can safely fail)

RAID 1 r^n ($n-1$ disks can safely fail)

RAID 2 *It's complicated*

RAID 3-5 (1 disk can safely fail)

$$1-(1-r)^n-\binom{n}{1}r^1(1-r)^{n-1}$$

RAID 6 (2 disks can safely fail)

$$1-(1-r)^n-\binom{n}{1}r^1(1-r)^{n-1}-\binom{n}{2}r^2(1-r)^{n-2}$$

Array Failure Rate (Raid 6 example)

Failure rate:

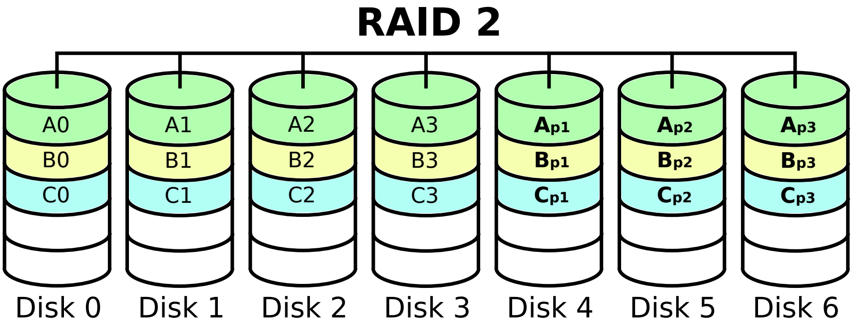
$$1 - (1-r)^n - \binom{n}{1} r^1 (1-r)^{n-1} - \binom{n}{2} r^2 (1-r)^{n-2}$$

Example drive **1%**/year failure rate with RAID 6 (3+2 drives):

$$1 - 0.99^5 - 5 \cdot 0.01 \cdot 0.99^4 - \frac{4 * 5}{2} \cdot 0.01^2 \cdot 0.99^3 = 0.0000098511$$

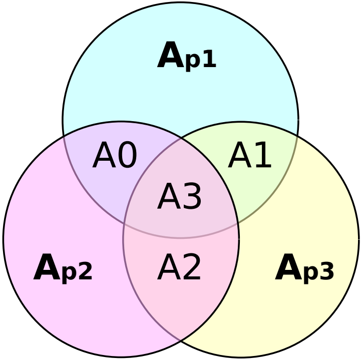
Rounding up, that's a **1%** failure rate of the RAID in **1000** years!

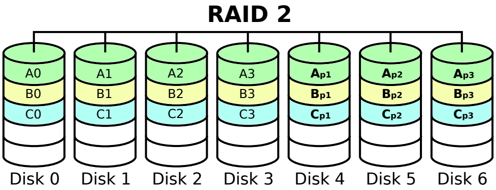
RAID 2 — Bit-Striping + Hamming Code



Number of disks	$2^k - 1$	$k = 3 \Longrightarrow 7$
Read (short ... long)	$1\times$ ... $2^k - k - 1\times$	$1\times$ ... $4\times$
Write (short ... long)	$1\times$ ... $2^k - k - 1\times$	$1\times$ ... $4\times$
Failure tolerance	1..2* disks	1..2* disks
Capacity efficiency	$\frac{2^k - k - 1}{2^k - 1}$	$4/7 \Longrightarrow 57\%$

Hamming Codes





- RAID 2 no longer used, but...
- Hamming code error correction
- ECC