Instructor Notes

Patterson, D., Gibson, G., and Katz, R.,
A Case for Redundant Arrays of Inexpensive Disks (RAID)
SIGMOD, June 1988.

Very influential
- OS, Arch, DB

John Wilkes, Richard Golding, Carl Staelin, and Tim Sullivan
The HP AutoRAID Hierarchical Storage System
ACM Transactions on Computer Systems, Vol. 14, No, 1, February 1996

1 What were some of the reasons why RAIDs were needed?

- CPUs getting faster @ faster rate than disks

Moore's law: CPU perf doubles every 18 months

disks: seek + rot: 10%/year

bw: 40%/year

→ Apps dominated by I/O perf over time (Amdahl's Law)

- Soln: Use small, inexpensive disks
  - cheaper: commodity
  - better performance — parallelism

improves BANDWIDTH

not individual latency

2 What is the main challenge that must be addressed when using a RAID? Can you derive the MTTF of a RAID given $MTTF_{disk}$, G (number of data disks in a group), C (number of check disks in a group), MTTR, and n_G (number of groups)?

Reliability: $MTTF_{array} = \frac{MTTF_{disk}}{\#\ disks}$ → horrible scaling

RAID - add redundancy so can recover from 1 failure

① Redundancy within GROUP of disks

$$MTTF_{group} = \frac{MTTF_{disk}}{G+C} \cdot \frac{1}{\text{prob of failure before reparing}} \} P$$

$$P = \frac{MTTR}{MTTF_{disk}/(G+C-1)}$$

$$MTTF_{RAID} = \frac{MTTF_{Group}}{N_G}$$

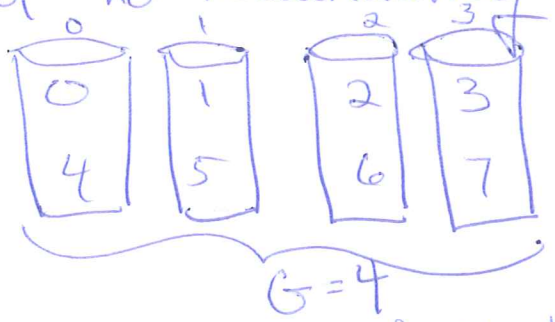
$$\frac{(MTTF_{Disk})^2}{(G+C)N_G * (G+C-1) * MTTR}$$

Make $MTTF_R$ larger?

- MTTFdisk larger - fewer disks
- MTTR shorter ⇒ hot spare

3 What is RAID-0? Draw the location of 8 blocks over 4 disks. For a given block address above RAID-0, how can the disk and disk offset be calculated? What read **bandwidth** can RAID-0 deliver as a function of D (total number of data disks)? Write bandwidth? Major drawback of RAID-0?

- striping w/ no redundancy



C=0

- export linear array of block #s, static mapping
- simple calculation:

  disk: b % G

  offset on disk: b / G

Read bw: D · single disk bw

Write : D

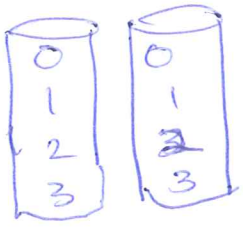
Drawback: Reliability

4 What is RAID-1? Default assumes G=C=1. What is RAID-10? What is RAID-01? What read bandwidth can RAID-1 deliver as a function of D? What write bandwidth? What is the major drawback of RAID-1?

G = 1

Mirroring

Default:



10 + 01: Combine mirroring + striping

RAID-01 : mirror of stripes

RAID-10 : array of mirrors

00 11 22 33

Which is better? 1 disk? both ok

What happens if lose 2 disks? only RAID-10 okay

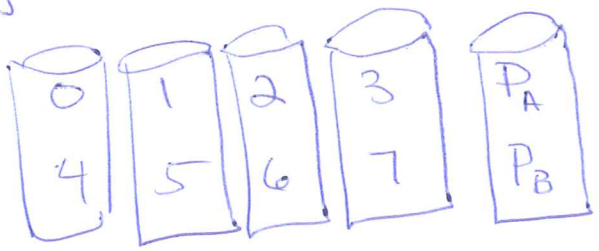
Read BW: 2D (read 1@ random)

Write BW: D (update both)

Drawback: Wasting capacity, 2x

5 What is RAID-4? Draw the location of 8 blocks. What happens on a read? What is its bandwidth? What happens on a large write? What is its bandwidth? What happens on a small write? What is its bandwidth? What is the major drawback of RAID-4?

- Striping w/ dedicated parity disk (simple X-OR)



C = 1
G = 4

- Read: normal - read after static calculation

BW: D

- Large write: must update parity disk - but in parallel

BW: D (update all in parallel + parity)

- Small write:

0 1 2 3 - what should P_A be?

↓
1'

option A: read 0, 2, 3 calc new P_A

option B: read old 1, old P_A, $P_A' = 1 \oplus P_A \oplus 1'$

bw: 2 reads + 2 writes ~~(sequential)~~ ⇒ 1 read + 1 write to parity disk

determined by speed of single parity disk

$N_G/2 = D/2G$

Drawback: parity disk is bottleneck for small writes

6 What is RAID-5? Draw the standard mapping of 15 blocks across 5 disks. What seems bad about this for large reads? Draw left symmetric mapping. What is the bandwidth improvement for small reads? How many operations must occur for a small write?

striping w/ rotating parity block

overlap

standard

0	1	2	3	P
4	5	6	P	7
8	9	P	10	11
12	P	13	14	15

dk

left-symmetric

0	1	2	3	P
5	6	7	P	4
10	11	P	8	9
15	P	12	13	14

large read? disk contention

large: same as RAID-4

small: extra disk useful for additional small reads

small read: how many disks are there?

$$D + C \cdot N_G = D + C\left(\frac{D}{G}\right) = D\left(1 + \frac{C}{G}\right) = D(1 + r_c)$$

small write: 2 read + 2 write = 4 ops

$$\frac{D(1 + r_c)}{4}$$

7. Why is RAID-6 sometimes desired? What does it basically entail?

Handles 2 errors, specifically designed to handle discovered "latent sector errors" during reconstruction from failed disk.

Add 2nd parity block per stripe

(2nd disk no worse than hot spare often used w/ RAID-5)

8 Why is LFS a good file system for RAID-5?

- No small writes, all large sequential writes

- Can build up all of data in log
  + then calculate parity + write out

Conclusions:

RAID-1 best for workloads w/ small writes

RAID-5 best for read intensive + large writes + compromising space

9 What was the motivation for AutoRAID? What data is stored with RAID-1? What data is stored with RAID-5? How is data moved between the different RAID levels? Policy?

- Different RAID levels appropriate for different workloads
- Hard to configure RAIDs for best case
  - sensitive to chunk size, group size
- World be nice to be able to add more disks over time

RAID-1: hot data, frequently changed

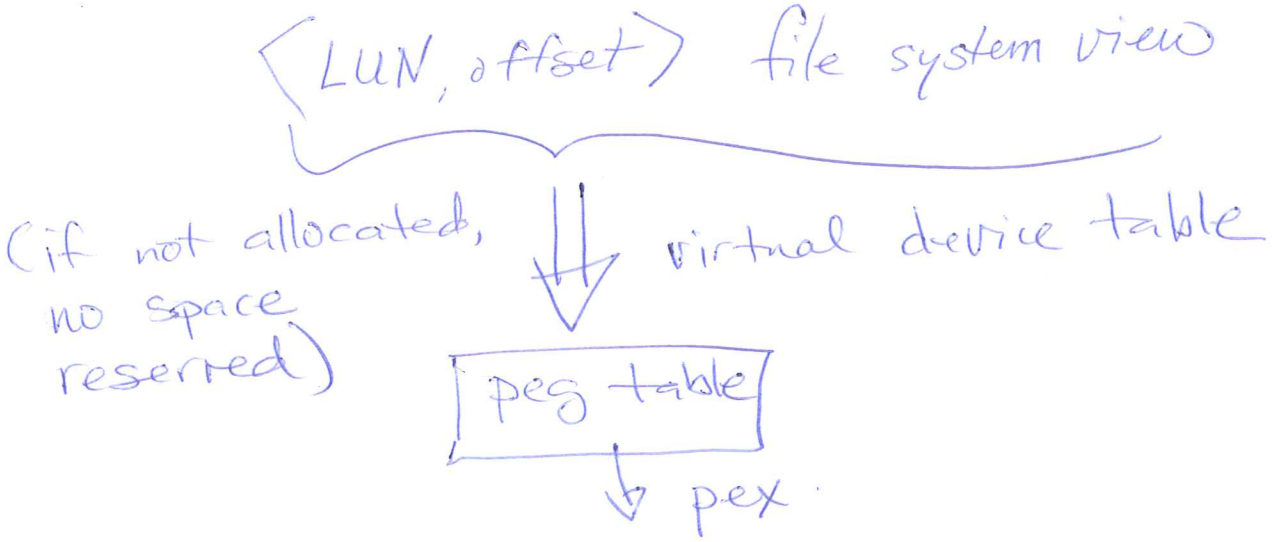
RAID-5: inactive data

Policy? Start all in RAID-1; when full, migrate some to RAID-5 (saves space → more room in RAID-1)

Newly active (write burst) to RAID-1

Background: move from RAID-1 to RAID-5

10 How is data mapped to disks? (Figs 3 and 4)



11 To perform a READ operation, how is the data in AutoRAID found? (Fig 5).

- might be in controller's cache memory
- do lookups in Virt. device table + peg tables

12 What happens on a WRITE operation? Must the host wait for this WRITE to complete? How are small writes to RAID-5 avoided?

- Keep in controller NVRAM
- No waiting @ host!
- Small write: 1st priority to have available room in RAID-1 PEGs
- Otherwise: add to log and flush out to RAID-5 PEGs ~~in one~~ sequentially (can calculate XOR as you add more data)

Automatically moving data between RAID levels makes sense..

Easy to use + admin