416 Distributed Systems

RAID, Feb 26 2018

Thanks to Greg Ganger and Remzi Arapaci-Dusseau for slides

Outline

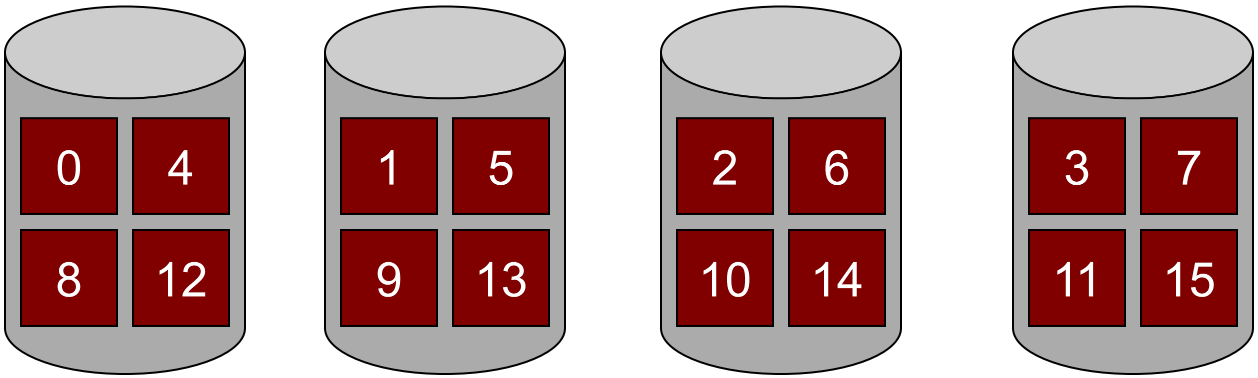
- Using multiple disks
  - Why have multiple disks?
  - problem and approaches
- RAID levels and performance

RAID Taxonomy

- Redundant Array of Inexpensive Independent Disks
  - Constructed by UC-Berkeley researchers in late 80s (Garth)
- RAID 0 – Coarse-grained Striping with no redundancy
- RAID 1 – Mirroring of independent disks
- RAID 2 – Fine-grained data striping plus Hamming code disks
  - Uses Hamming codes to detect and correct multiple errors
  - Originally implemented when drives didn't always detect errors
  - Not used in real systems
- RAID 3 – Fine-grained data striping plus parity disk
- RAID 4 – Coarse-grained data striping plus parity disk
- RAID 5 – Coarse-grained data striping plus striped parity
- RAID 6 – Coarse-grained data striping plus 2 striped codes

RAID-0: Striping

- Stripe blocks across disks in a “chunk” size
  - How to pick a reasonable chunk size?

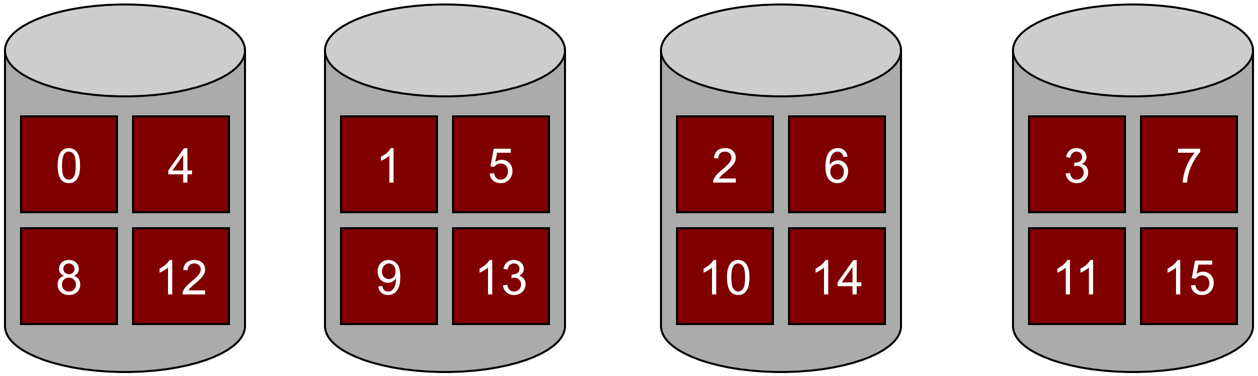


How to calculate where chunk # lives?

Disk #:

Offset within disk:

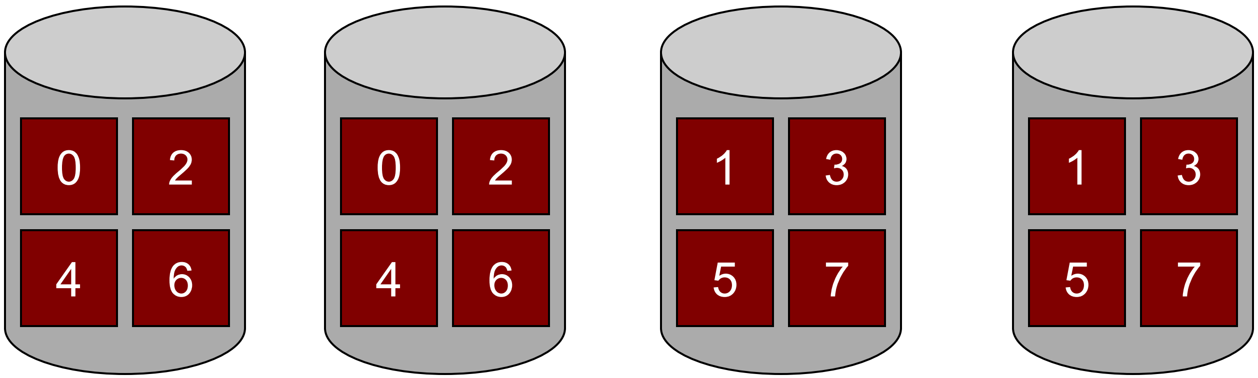
RAID-0: Striping



- Evaluate for D disks
- Performance: How much faster than 1 disk? (best case)
- Reliability: More or less reliable than 1 disk?

RAID-1: Mirroring

- Motivation: Handle disk failures
- Put copy (mirror or replica) of each chunk on another disk

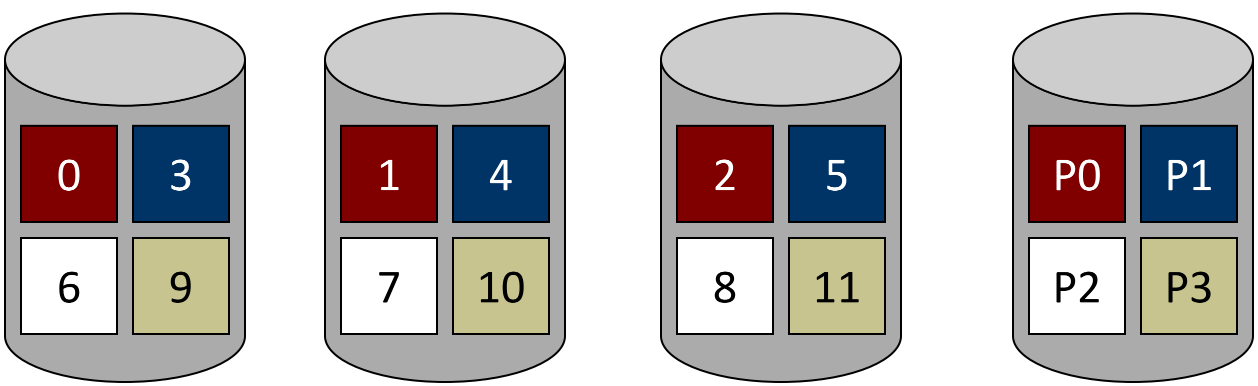


- Capacity
- Reliability
- Performance

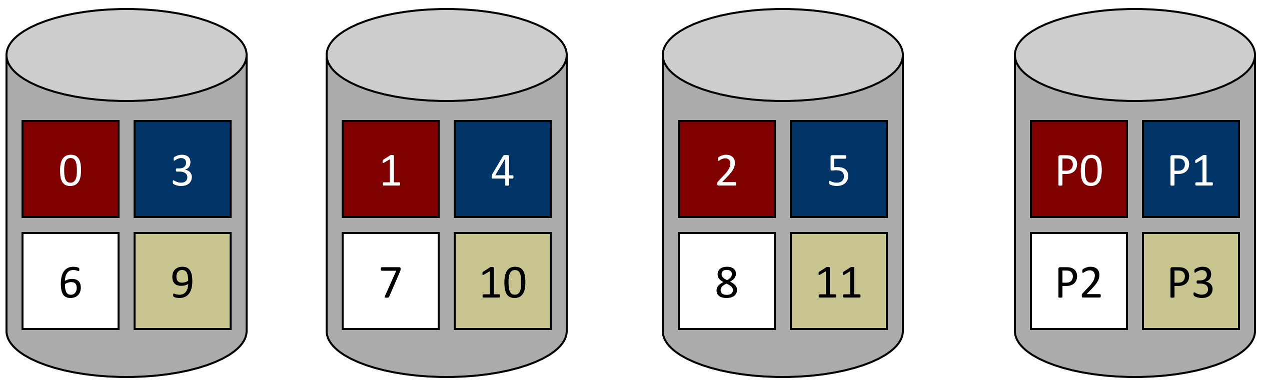
RAID-4: Parity

- Motivation: Improve capacity
- Idea: Allocate parity block to encode info about blocks
  - Parity checks all other blocks in stripe across other disks
- Parity block = XOR over others (gives "even" parity)
  - Example: 0 1 0 → Parity value?
- How do you recover from a failed disk?
  - Example: x 0 0 and parity of 1
  - What is the failed value?

A	B	XOR
0	0	0
0	1	1
1	0	1
1	1	0

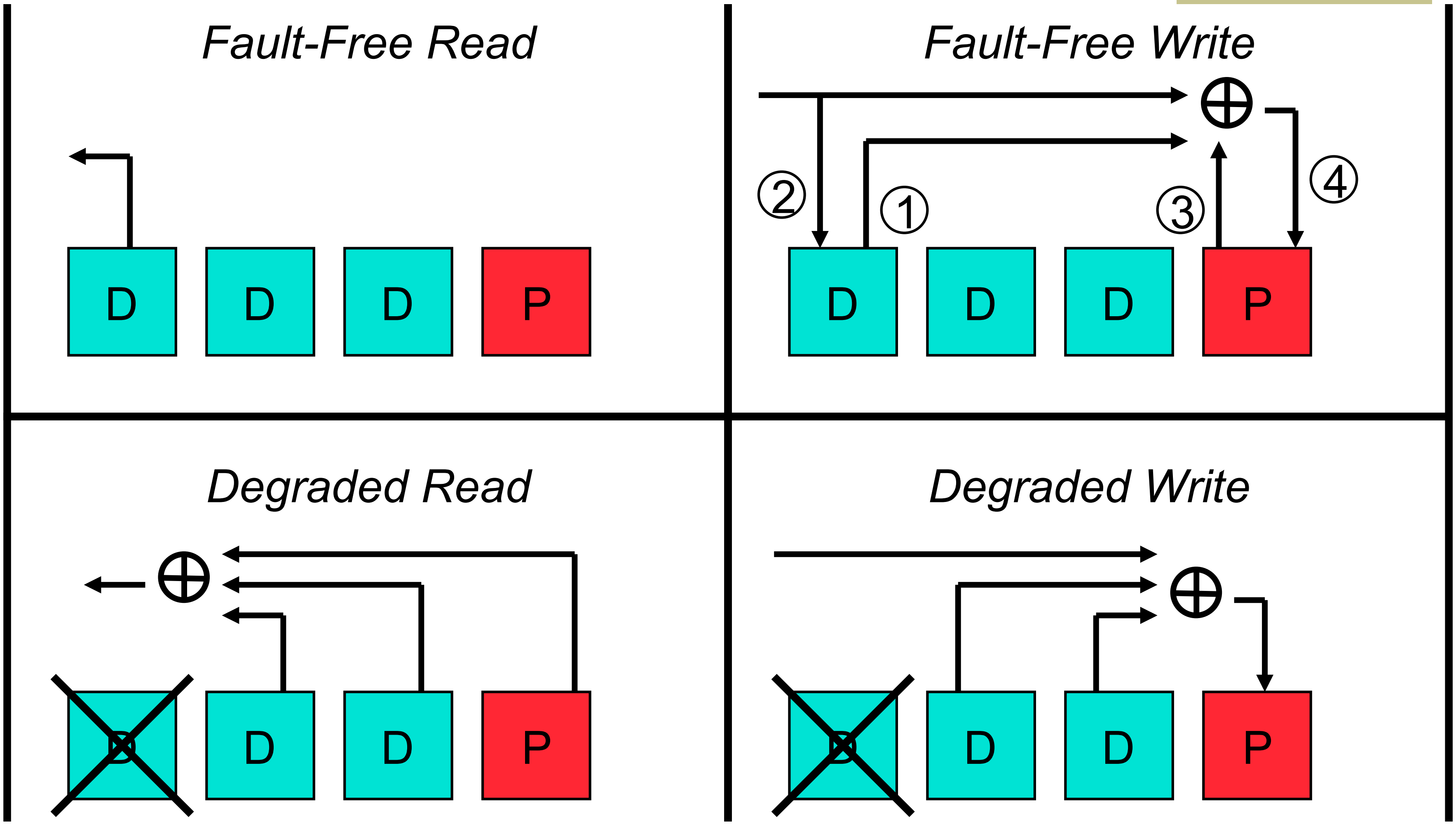


RAID-4: Parity



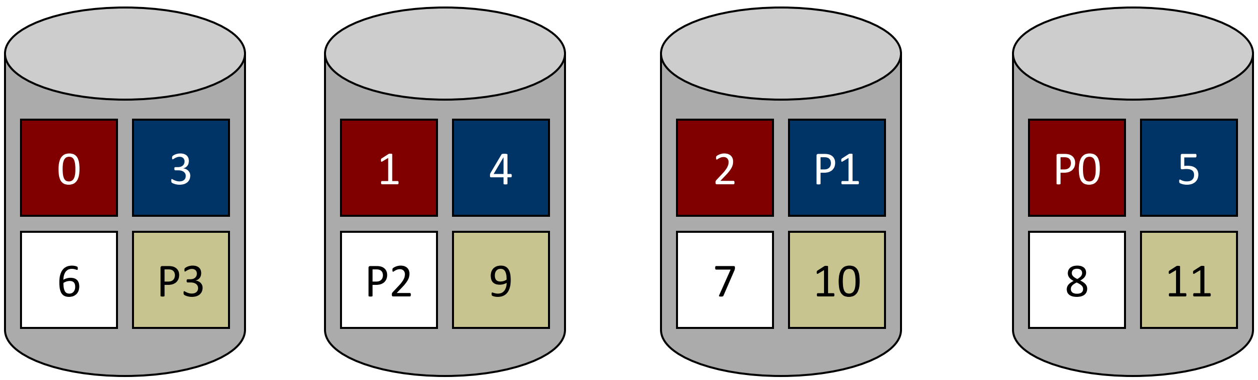
- Capacity:
- Reliability:
- Performance:
  - Reads
  - Writes: How to update parity block?
    - Two ways:
      - Use parity disk
      - Re-compute parity from non-parity disks
    - (Parity disk is the bottleneck)

Updating and using the parity



RAID-5: Rotated/Striped Parity

Rotate location of parity across all disks



- Capacity:
- Reliability:
- Performance:
  - Reads:
  - Writes:
  - Still requires 4 I/Os per write, but not always to same parity disk

Comparison

N: number of disks
S: throughput of 1 disk sequential read/write
R: throughput of 1 disk random read/write
D: delay to read/write from 1 disk

	RAID-0	RAID-1	RAID-4	RAID-5
Capacity	N	$N/2$	$N-1$	$N-1$
Reliability	0	1 (for sure) $\frac{N}{2}$ (if lucky)	1	1
Throughput				
Sequential Read	$N \cdot S$	$(N/2) \cdot S$	$(N-1) \cdot S$	$(N-1) \cdot S$
Sequential Write	$N \cdot S$	$(N/2) \cdot S$	$(N-1) \cdot S$	$(N-1) \cdot S$
Random Read	$N \cdot R$	$N \cdot R$	$(N-1) \cdot R$	$N \cdot R$
Random Write	$N \cdot R$	$(N/2) \cdot R$	$\frac{1}{2} \cdot R$	$\frac{N}{4} R$
Latency				
Read	D	D	D	D
Write	D	D	$2D$	$2D$

Table 38.7: **RAID Capacity, Reliability, and Performance**

Comparison

N: number of disks
S: throughput of 1 disk sequential read/write
R: throughput of 1 disk random read/write
D: delay to read/write from 1 disk

	RAID-0	RAID-1	RAID-4	RAID-5
Capacity	N	$N/2$	$N-1$	$N-1$
Reliability	0	1 (for sure) $\frac{N}{2}$ (if lucky)	1	1
Throughput				
Sequential Read	$N \cdot S$	$(N/2) \cdot S$	$(N-1) \cdot S$	$(N-1) \cdot S$
Sequential Write	$N \cdot S$	$(N/2) \cdot S$	$(N-1) \cdot S$	$(N-1) \cdot S$
Random Read	$N \cdot R$	$N \cdot R$	$(N-1) \cdot R$	$N \cdot R$
Random Write	$N \cdot R$	$(N/2) \cdot R$	$\frac{1}{2} \cdot R$	$\frac{N}{4} R$
Latency				
Read	D	D	D	D
Write	D	D	$2D$	$2D$

Table 38.7: **RAID Capacity, Reliability, and Performance**

Comparison

N: number of disks
S: throughput of 1 disk sequential read/write
R: throughput of 1 disk random read/write
D: delay to read/write from 1 disk

	RAID-0	RAID-1	RAID-4	RAID-5
Capacity	N	$N/2$	$N-1$	$N-1$
Reliability	0	1 (for sure) $\frac{N}{2}$ (if lucky)	1	1
Throughput				
Sequential Read	$N \cdot S$	$(N/2) \cdot S$	$(N-1) \cdot S$	$(N-1) \cdot S$
Sequential Write	$N \cdot S$	$(N/2) \cdot S$	$(N-1) \cdot S$	$(N-1) \cdot S$
Random Read	$N \cdot R$	$N \cdot R$	$(N-1) \cdot R$	$N \cdot R$
Random Write	$N \cdot R$	$(N/2) \cdot R$	$\frac{1}{2} \cdot R$	$\frac{N}{4} R$
Latency				
Read	D	D	D	D
Write	D	D	$2D$	$2D$

Table 38.7: **RAID Capacity, Reliability, and Performance**

Comparison

N: number of disks
S: throughput of 1 disk sequential read/write
R: throughput of 1 disk random read/write
D: delay to read/write from 1 disk

	RAID-0	RAID-1	RAID-4	RAID-5
Capacity	N	$N/2$	$N-1$	$N-1$
Reliability	0	1 (for sure) $\frac{N}{2}$ (if lucky)	1	1
Throughput				
Sequential Read	$N \cdot S$	$(N/2) \cdot S$	$(N-1) \cdot S$	$(N-1) \cdot S$
Sequential Write	$N \cdot S$	$(N/2) \cdot S$	$(N-1) \cdot S$	$(N-1) \cdot S$
Random Read	$N \cdot R$	$N \cdot R$	$(N-1) \cdot R$	$N \cdot R$
Random Write	$N \cdot R$	$(N/2) \cdot R$	$\frac{1}{2} \cdot R$	$\frac{N}{4} R$
Latency				
Read	D	D	D	D
Write	D	D	$2D$	$2D$

Table 38.7: **RAID Capacity, Reliability, and Performance**

Comparison

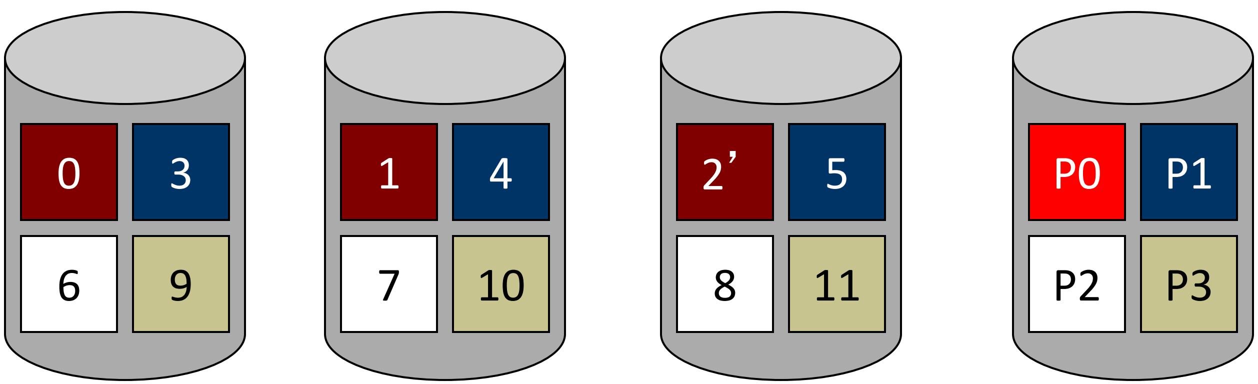
N: number of disks
S: throughput of 1 disk sequential read/write
R: throughput of 1 disk random read/write
D: delay to read/write from 1 disk

	RAID-0	RAID-1	RAID-4	RAID-5
Capacity	N	$N/2$	$N-1$	$N-1$
Reliability	0	1 (for sure) $\frac{N}{2}$ (if lucky)	1	1
Throughput				
Sequential Read	$N \cdot S$	$(N/2) \cdot S$	$(N-1) \cdot S$	$(N-1) \cdot S$
Sequential Write	$N \cdot S$	$(N/2) \cdot S$	$(N-1) \cdot S$	$(N-1) \cdot S$
Random Read	$N \cdot R$	$N \cdot R$	$(N-1) \cdot R$	$N \cdot R$
Random Write	$N \cdot R$	$(N/2) \cdot R$	$\frac{1}{2} \cdot R$	$\frac{N}{4} R$
Latency				
Read	D	D	D	D
Write	D	D	$2D$	$2D$

Table 38.7: **RAID Capacity, Reliability, and Performance**

Advanced Issues

- What happens if more than one fault?
  - Example: One disk fails plus "latent sector error" on another
  - RAID-5 cannot handle two faults
  - Solution: RAID-6: add multiple parity blocks
- Why is NVRAM useful?
  - Example: What if update 2, don't update P0 before power failure (or crash), and then disk 1 fails?
  - NVRAM solution: Use to store blocks updated in same stripe
    - If power failure, can replay all writes in NVRAM
  - Software RAID solution: Perform parity scrub over entire disk



Conclusions

- RAID turns multiple disks into a larger, faster, more reliable disk
- RAID-0: Striping
  Good when performance and capacity really matter, but reliability doesn't
- RAID-1: Mirroring
  Good when reliability and write performance matter, but capacity (cost) doesn't
- RAID-4: Parity disk
- RAID-5: Rotating parity
  Good when capacity and cost matter or workload is read-mostly
  - Good compromise choice