

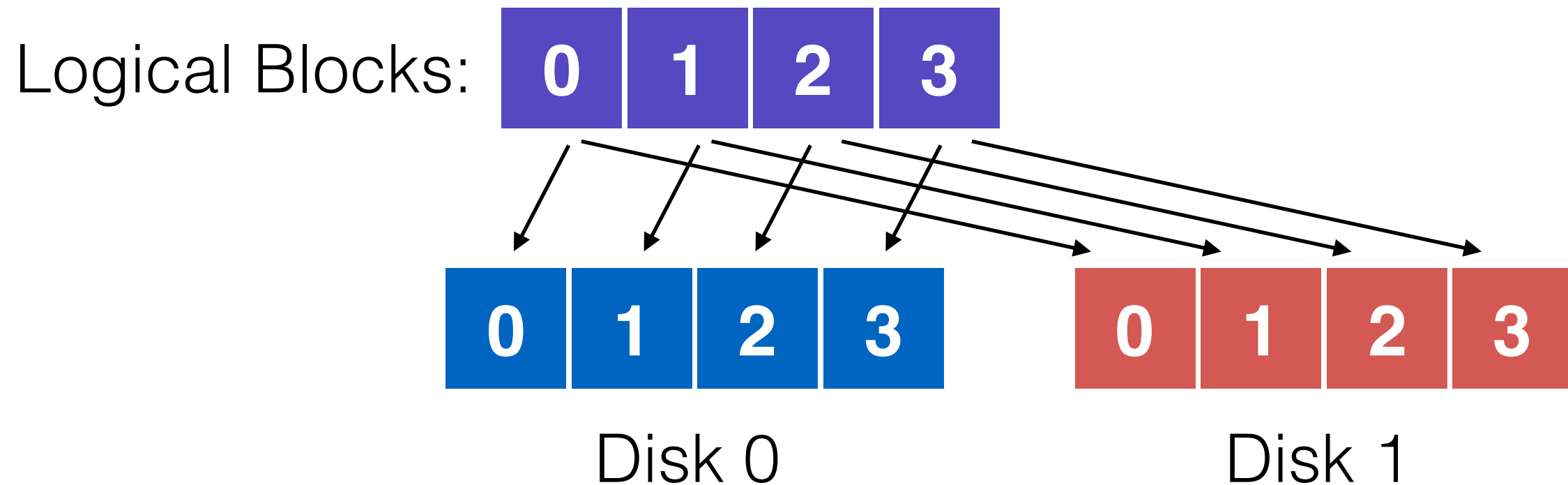
Operating Systems

RAID continued

Nipun Batra

RAID-1: Mirroring

Keep two copies of all data.



Assumptions

Assume disks are **fail-stop**.

- they work or they don't
- we know when they don't

Tougher Errors:

- latent sector errors
- silent data corruption

2 disks

Disk 0	Disk 1
0	0
1	1
2	2
3	3

4 disks

Disk 0	Disk 1	Disk 2	Disk 3
0	0	1	1
2	2	3	3
4	4	5	5
6	6	7	7

4 disks

Disk 0	Disk 1	Disk 2	Disk 3
0	0	1	1
2	2	3	3
4	4	5	5
6	6	7	7

How many disks can fail?

RAID-1: Analysis

RAID-1: Analysis

- What is capacity?

RAID-1: Analysis

- What is capacity?
 - $N/2 * C$

RAID-1: Analysis

- What is capacity?
 - $N/2 * C$
- How many disks can fail?

RAID-1: Analysis

- What is capacity?
 - $N/2 * C$
- How many disks can fail?
 - 1 or $N/2$ (best case)

RAID-1: Analysis

- What is capacity?
 - $N/2 * C$
- How many disks can fail?
 - 1 or $N/2$ (best case)
- Throughput?

RAID-1: Analysis

- What is capacity?
 - $N/2 * C$
- How many disks can fail?
 - 1 or $N/2$ (best case)
- Throughput?
 - Sequential write — $(N/2)*S$

RAID-1: Analysis

- What is capacity?
 - $N/2 * C$
- How many disks can fail?
 - 1 or $N/2$ (best case)
- Throughput?
 - Sequential write — $(N/2)*S$
 - Sequential read — $(N/2)*S$

RAID-1: Analysis

- What is capacity?
 - $N/2 * C$
- How many disks can fail?
 - 1 or $N/2$ (best case)
- Throughput?
 - Sequential write — $(N/2)*S$
 - Sequential read — $(N/2)*S$
 - Random write — $(N/2)*R$

RAID-1: Analysis

- What is capacity?
 - $N/2 * C$
- How many disks can fail?
 - 1 or $N/2$ (best case)
- Throughput?
 - Sequential write — $(N/2)*S$
 - Sequential read — $(N/2)*S$
 - Random write — $(N/2)*R$
 - Random read — $(N*R)$

RAID-1: Analysis

- What is capacity?
 - $N/2 * C$
- How many disks can fail?
 - 1 or $N/2$ (best case)
- Throughput?
 - Sequential write — $(N/2)*S$
 - Sequential read — $(N/2)*S$
 - Random write — $(N/2)*R$
 - Random read — $(N)*R$
- Latency

RAID-1: Analysis

- What is capacity?
 - $N/2 * C$
- How many disks can fail?
 - 1 or $N/2$ (best case)
- Throughput?
 - Sequential write — $(N/2)*S$
 - Sequential read — $(N/2)*S$
 - Random write — $(N/2)*R$
 - Random read — $(N*R)$
- Latency
 - **D**

RAID-1: Analysis

- What is capacity?
 - $N/2 * C$
- How many disks can fail?
 - 1 or $N/2$ (best case)
- Throughput?
 - Sequential write — $(N/2)*S$
 - Sequential read — $(N/2)*S$
 - Random write — $(N/2)*R$
 - Random read — $(N*R)$
- Latency
 - D

Disk 0	Disk 1	Disk 2	Disk 3
0	0	1	1
2	2	3	3
4	4	5	5
6	6	7	7

RAID-1: Analysis

- What is capacity?
 - $N/2 * C$
- How many disks can fail?
 - 1 or $N/2$ (best case)
- Throughput?
 - Sequential write — $(N/2)*S$
 - Sequential read — $(N/2)*S$
 - Random write — $(N/2)*R$
 - Random read — $(N)*R$
- Latency
 - D

Disk 0	Disk 1	Disk 2	Disk 3
0	0	1	1
2	2	3	3
4	4	5	5
6	6	7	7

RAID-1: Analysis

- What is capacity?
 - $N/2 * C$
- How many disks can fail?
 - 1 or $N/2$ (best case)
- Throughput?
 - Sequential write — $(N/2)*S$
 - Sequential read — $(N/2)*S$
 - Random write — $(N/2)*R$
 - Random read — $(N*R)$
- Latency
 - D

Disk 0	Disk 1	Disk 2	Disk 3
0	0	1	1
2	2	3	3
4	4	5	5
6	6	7	7

RAID-1: Analysis

- What is capacity?
 - $N/2 * C$
- How many disks can fail?
 - 1 or $N/2$ (best case)
- Throughput?
 - Sequential write — $(N/2)*S$
 - Sequential read — $(N/2)*S$
 - Random write — $(N/2)*R$
 - Random read — $(N*R)$
- Latency
 - D

Disk 0	Disk 1	Disk 2	Disk 3
0	0	1	1
2	2	3	3
4	4	5	5
6	6	7	7

RAID-1: Analysis

- What is capacity?
 - $N/2 * C$
- How many disks can fail?
 - 1 or $N/2$ (best case)
- Throughput?
 - Sequential write — $(N/2)*S$
 - Sequential read — $(N/2)*S$
 - Random write — $(N/2)*R$
 - Random read — $(N*R)$
- Latency
 - D

Disk 0	Disk 1	Disk 2	Disk 3
0	0	1	1
2	2	3	3
4	4	5	5
6	6	7	7

RAID-1: Analysis

- What is capacity?
 - $N/2 * C$
- How many disks can fail?
 - 1 or $N/2$ (best case)
- Throughput?
 - Sequential write — $(N/2)*S$
 - Sequential read — $(N/2)*S$
 - Random write — $(N/2)*R$
 - Random read — $(N)*R$
- Latency
 - D

Disk 0	Disk 1	Disk 2	Disk 3
0	0	1	1
2	2	3	3
4	4	5	5
6	6	7	7

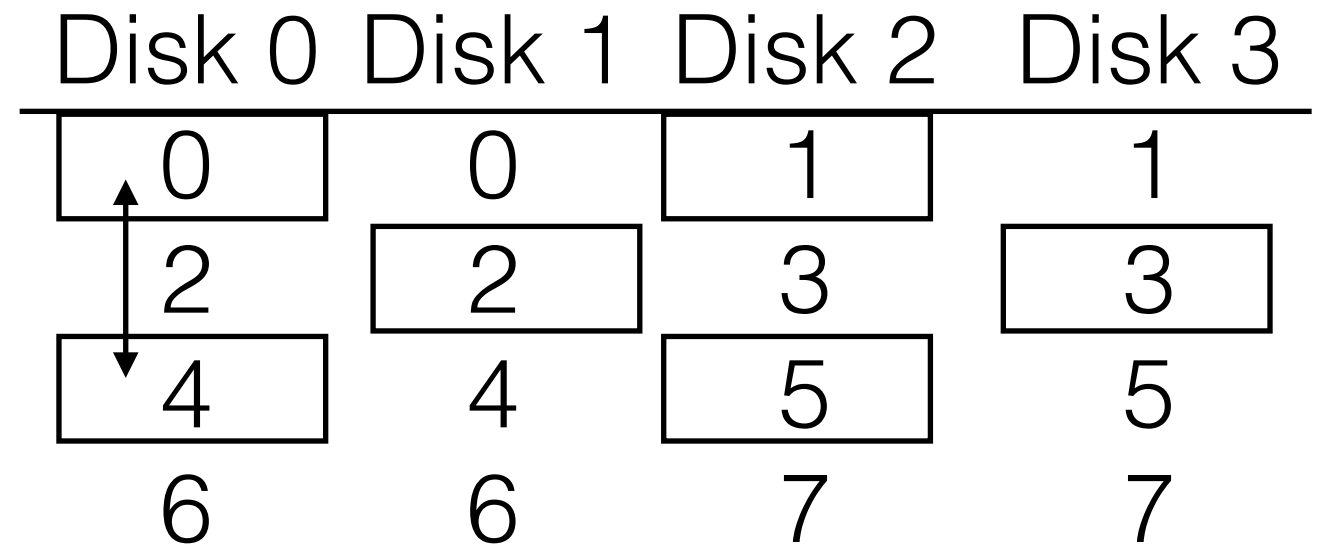
RAID-1: Analysis

- What is capacity?
 - $N/2 * C$
- How many disks can fail?
 - 1 or $N/2$ (best case)
- Throughput?
 - Sequential write — $(N/2)*S$
 - Sequential read — $(N/2)*S$
 - Random write — $(N/2)*R$
 - Random read — $(N)*R$
- Latency
 - D

Disk 0	Disk 1	Disk 2	Disk 3
0	0	1	1
2	2	3	3
4	4	5	5
6	6	7	7

RAID-1: Analysis

- What is capacity?
 - $N/2 * C$
- How many disks can fail?
 - 1 or $N/2$ (best case)
- Throughput?
 - Sequential write — $(N/2)*S$
 - Sequential read — $(N/2)*S$
 - Random write — $(N/2)*R$
 - Random read — $(N)*R$
- Latency
 - D



Crashes

	Disk0	Disk1
0	A	A
1	B	B
2	C	C
3	D	D

Crashes

	Disk0	Disk1
0	A	A
1	B	B
2	C	C
3	D	D

write(A) to 2

Crashes

	Disk0	Disk1
0	A	A
1	B	B
2	A	C
3	D	D

write(A) to 2

Crashes

	Disk0	Disk1
0	A	A
1	B	B
2	A	A
3	D	D

write(A) to 2

Crashes

	Disk0	Disk1
0	A	A
1	B	B
2	A	A
3	D	D

Crashes

	Disk0	Disk1
0	A	A
1	B	B
2	A	A
3	D	D

write(T) to 3

Crashes

	Disk0	Disk1
0	A	A
1	B	B
2	A	A
3	D	T

write(T) to 3

Crashes

	Disk0	Disk1
0	A	A
1	B	B
2	A	A
3	D	T

CRASH!!!

Crashes

	Disk0	Disk1
0	A	A
1	B	B
2	A	A
3	D	T

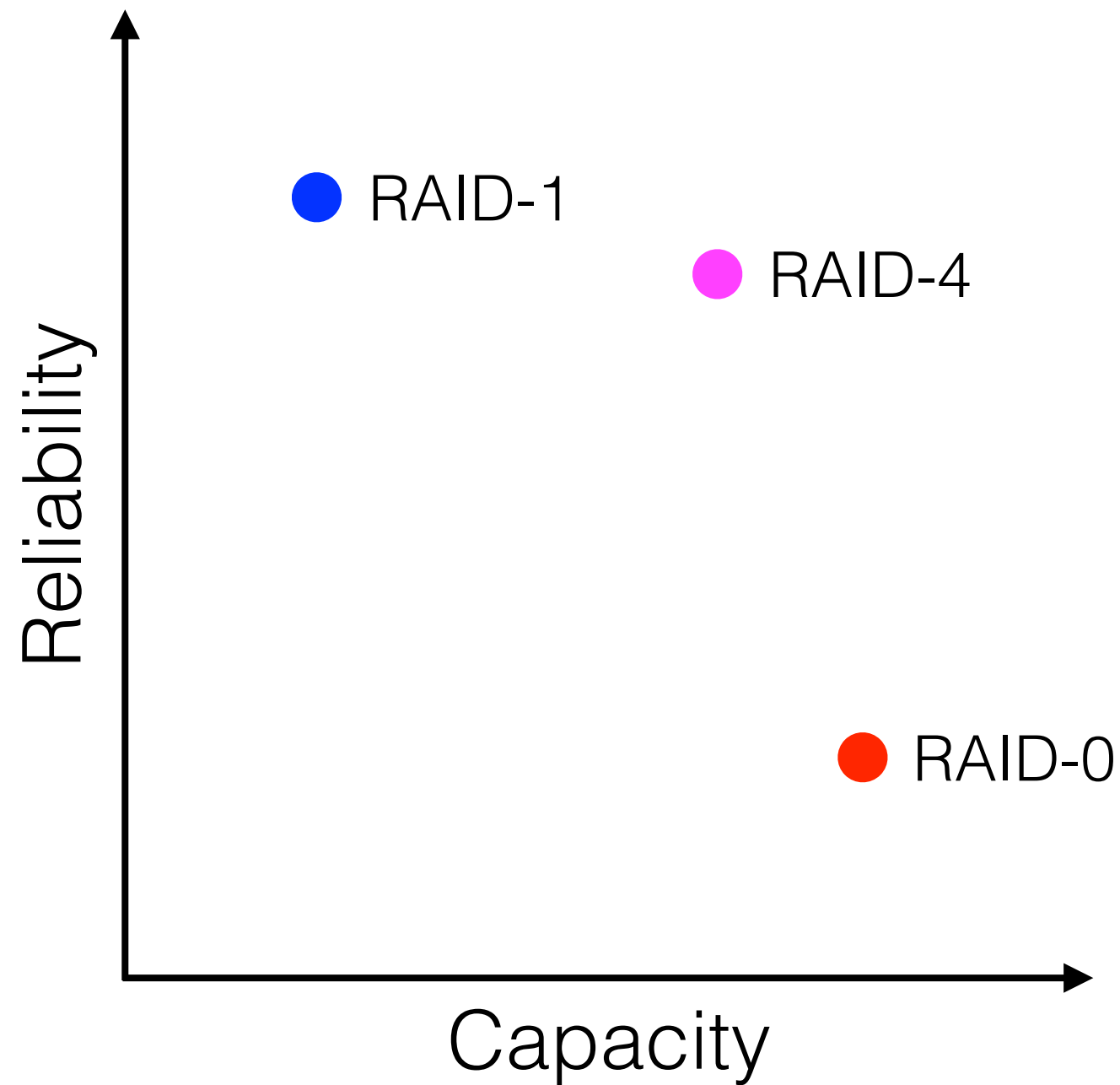
after reboot, how to
tell which data is right?

H/W Solution

Problem: Consistent-Update Problem

Use non-volatile RAM in RAID controller.

RAID-4 compared to RAID-1 and RAID-0



Strategy

Strategy

- Use parity disk.

Strategy

- Use parity disk.

Strategy

- Use parity disk.
- In algebra, if an equation has N variables, and $N-1$ are known, you can often solve for the unknown.

Strategy

- Use parity disk.
- In algebra, if an equation has N variables, and $N-1$ are known, you can often solve for the unknown.

Strategy

- Use parity disk.
- In algebra, if an equation has N variables, and $N-1$ are known, you can often solve for the unknown.
- Treat the sectors across disks in a stripe as an equation.

Strategy

- Use parity disk.
- In algebra, if an **equation** has N variables, and $N-1$ are know, you can often **solve for the unknown**.
- Treat the sectors across disks in a stripe as an equation.

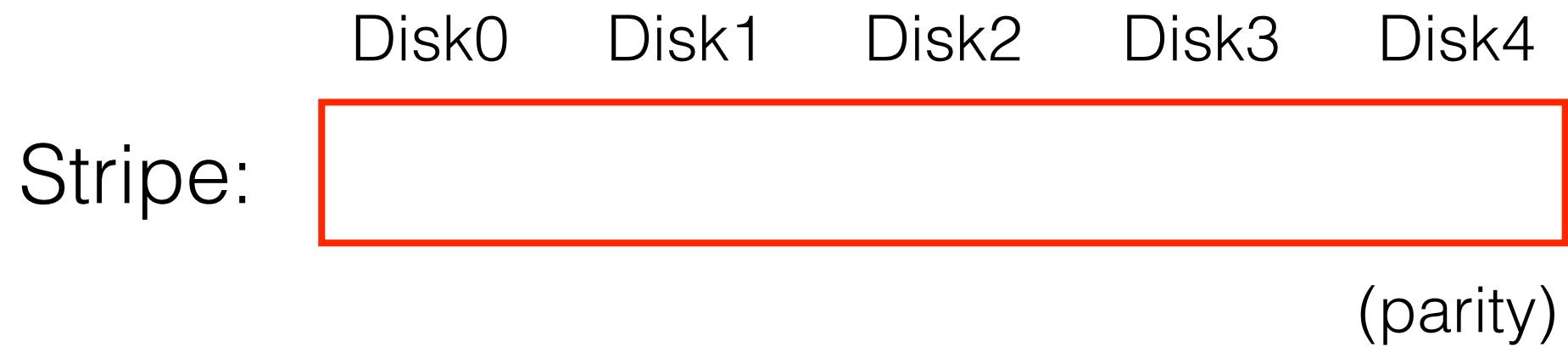
Strategy

- Use parity disk.
- In algebra, if an **equation** has N variables, and $N-1$ are known, you can often **solve for the unknown**.
- Treat the sectors across disks in a stripe as an equation.
- A **failed disk** is like an unknown in the equation.

Example

	Disk0	Disk1	Disk2	Disk3	Disk4
Stripe:	<div></div>				

Example



Example

	Disk0	Disk1	Disk2	Disk3	Disk4
Stripe:	5	3	0	1	

(parity)

Example

	Disk0	Disk1	Disk2	Disk3	Disk4
Stripe:	5	3	0	1	9

(parity)

Example

	Disk0	Disk1	Disk2	Disk3	Disk4
Stripe:	5	X	0	1	9

(parity)

Example

	Disk0	Disk1	Disk2	Disk3	Disk4
Stripe:	5	3	0	1	9

(parity)

Example

	Disk0	Disk1	Disk2	Disk3	Disk4
Stripe:	2	1	1	X	5

(parity)

Example

	Disk0	Disk1	Disk2	Disk3	Disk4
Stripe:	2	1	1	1	5

(parity)

Example

	Disk0	Disk1	Disk2	Disk3	Disk4
Stripe:	3	0	1	2	X

(parity)

Example

	Disk0	Disk1	Disk2	Disk3	Disk4
Stripe:	3	0	1	2	6

(parity)

Parity Functions

Which functions could we use to compute parity?

	Disk0	Disk1	Disk2	Disk3	Disk4
Stripe:	0	1	0	1	$\text{XOR}(0,1,0,1)=0$
					(parity)

	Disk0	Disk1	Disk2	Disk3	Disk4
Stripe:	00	01	10	11	$(\text{XOR}(0,0,1,1), \text{XOR}(0,1,0,1))=00$
					(parity)

RAID-4: Analysis

RAID-4: Analysis

- What is capacity?

RAID-4: Analysis

- What is capacity?
 - $(N-1) * C$

RAID-4: Analysis

- What is capacity?
 - $(N-1) * C$
- How many disks can fail?

RAID-4: Analysis

- What is capacity?
 - $(N-1) * C$
- How many disks can fail?
 - 1

RAID-4: Analysis

- What is capacity?
 - $(N-1) * C$
- How many disks can fail?
 - 1
- Throughput?

RAID-4: Analysis

- What is capacity?
 - $(N-1) * C$
- How many disks can fail?
 - 1
- Throughput?
 - Sequential write — $(N-1)*S$

RAID-4: Analysis

- What is capacity?
 - $(N-1) * C$
- How many disks can fail?
 - 1
- Throughput?
 - Sequential write — $(N-1)*S$
 - Sequential read — $(N-1)*S$

RAID-4: Analysis

- What is capacity?
 - $(N-1) * C$
- How many disks can fail?
 - 1
- Throughput?
 - Sequential write — $(N-1)*S$
 - Sequential read — $(N-1)*S$
 - Random read — $(N-1)*R$

RAID-4: Analysis

- What is capacity?
 - $(N-1) * C$
- How many disks can fail?
 - 1
- Throughput?
 - Sequential write — $(N-1)*S$
 - Sequential read — $(N-1)*S$
 - Random read — $(N-1)*R$
 - Random write?

RAID-4: Analysis for Random Write ...

	Disk0	Disk1	Disk2	Disk3	Disk4
Stripe:	0	1	0	1	$\text{XOR}(0,1,0,1)=0$
					(parity)

RAID-4: Analysis for Random Write ...

	Disk0	Disk1	Disk2	Disk3	Disk4
Stripe:	0	1	0	1	$\text{XOR}(0,1,0,1)=0$
					(parity)

- Want to: **Write** 0 to Disk 1

RAID-4: Analysis for Random Write ...

	Disk0	Disk1	Disk2	Disk3	Disk4
Stripe:	0	1	0	1	$\text{XOR}(0,1,0,1)=0$
					(parity)

- Want to: **Write** 0 to Disk 1
- **Read** old value of Disk 1

RAID-4: Analysis for Random Write ...

	Disk0	Disk1	Disk2	Disk3	Disk4
Stripe:	0	1	0	1	$\text{XOR}(0,1,0,1)=0$
					(parity)

- Want to: **Write** 0 to Disk 1
- **Read** old value of Disk 1
- **Read** old value of parity

RAID-4: Analysis for Random Write ...

	Disk0	Disk1	Disk2	Disk3	Disk4
Stripe:	0	1	0	1	$\text{XOR}(0,1,0,1)=0$
					(parity)

- Want to: **Write** 0 to Disk 1
- **Read** old value of Disk 1
- **Read** old value of parity
- If New value of Disk 1 == Old value of Disk 1, Do nothing

RAID-4: Analysis for Random Write ...

	Disk0	Disk1	Disk2	Disk3	Disk4
Stripe:	0	1	0	1	$\text{XOR}(0,1,0,1)=0$
					(parity)

- Want to: **Write** 0 to Disk 1
- **Read** old value of Disk 1
- **Read** old value of parity
- If New value of Disk 1 == Old value of Disk 1, Do nothing
- Else, **Write** new flipped parity and **Write** new value to Disk 1

RAID-4: Analysis for Random Write ...

	Disk0	Disk1	Disk2	Disk3	Disk4
Stripe:	0	1	0	1	$\text{XOR}(0,1,0,1)=0$
					(parity)

- Want to: **Write** 0 to Disk 1
- **Read** old value of Disk 1
- **Read** old value of parity
- If New value of Disk 1 == Old value of Disk 1, Do nothing
- Else, **Write** new flipped parity and **Write** new value to Disk 1
- Each random write, needs 2 reads and 2 writes

RAID-4: Analysis for Random Write ...

	Disk0	Disk1	Disk2	Disk3	Disk4
Stripe:	0	1	0	1	$\text{XOR}(0,1,0,1)=0$
					(parity)

- Want to: **Write** 0 to Disk 1
- **Read** old value of Disk 1
- **Read** old value of parity
- If New value of Disk 1 == Old value of Disk 1, Do nothing
- Else, **Write** new flipped parity and **Write** new value to Disk 1
- Each random write, needs 2 reads and 2 writes
- Assume we get 2 writes: Disk 0 and Disk 1

RAID-4: Analysis for Random Write ...

	Disk0	Disk1	Disk2	Disk3	Disk4
Stripe:	0	1	0	1	$\text{XOR}(0,1,0,1)=0$
					(parity)

- Want to: **Write** 0 to Disk 1
- **Read** old value of Disk 1
- **Read** old value of parity
- If New value of Disk 1 == Old value of Disk 1, Do nothing
- Else, **Write** new flipped parity and **Write** new value to Disk 1
- Each random write, needs 2 reads and 2 writes
- Assume we get 2 writes: Disk 0 and Disk 1
 - Both wait to read and write Parity Disk

RAID-4: Analysis for Random Write ...

	Disk0	Disk1	Disk2	Disk3	Disk4
Stripe:	0	1	0	1	$\text{XOR}(0,1,0,1)=0$
					(parity)

- Want to: **Write** 0 to Disk 1
- **Read** old value of Disk 1
- **Read** old value of parity
- If New value of Disk 1 == Old value of Disk 1, Do nothing
- Else, **Write** new flipped parity and **Write** new value to Disk 1
- Each random write, needs 2 reads and 2 writes
- Assume we get 2 writes: Disk 0 and Disk 1
 - Both wait to read and write Parity Disk
 - R/2 throughput (**independent of N**)

RAID-4: Analysis for Random Write ...

	Disk0	Disk1	Disk2	Disk3	Disk4
Stripe:	0	1	0	1	$\text{XOR}(0,1,0,1)=0$
					(parity)

- Want to: **Write** 0 to Disk 1
- **Read** old value of Disk 1
- **Read** old value of parity
- If New value of Disk 1 == Old value of Disk 1, Do nothing
- Else, **Write** new flipped parity and **Write** new value to Disk 1
- Each random write, needs 2 reads and 2 writes
- Assume we get 2 writes: Disk 0 and Disk 1
 - Both wait to read and write Parity Disk
 - R/2 throughput (**independent of N**)
- Latency for random write is 2D (2 parallel reads and 2 parallel writes)

RAID-5 (Improve Random Write Performance)

Disk0	Disk1	Disk2	Disk3	Disk4
-	-	-	-	P
-	-	-	P	-
-	-	P	-	-
...				

RAID-5: Analysis

0a) What is capacity? **$(N-1) * C$**

0b) How many disks can fail? **1**

0c) Throughput? **???**

0d) Latency? **D for read and $2*D$ for write**

RAID-5: Throughput

What is steady-state throughput for

- sequential reads?
- sequential writes?
- random reads?
- random writes?

RAID-5: Throughput

What is steady-state throughput for

- sequential reads? $(N-1) * S$
- sequential writes? $(N-1) * S$
- random reads? $N * R$
- random writes? $N * R / 4$

RAID-5: Throughput

What is steady-state throughput for

- sequential reads? $(N-1) * S$
- sequential writes? $(N-1) * S$
- random reads? $N * R$
- random writes? $N * R / 4$

$$\begin{array}{l} (N-1) * R \\ R/2 \end{array}$$

RAID-4

All RAID

	Reliability	Capacity
RAID-0	0	$C \cdot N$
RAID-1	1	$C \cdot N / 2$
RAID-4	1	$N - 1$
RAID-5	1	$N - 1$

All RAID

	Read Latency	Write Latency
RAID-0	D	D
RAID-1	D	D
RAID-4	D	2D
RAID-5	D	2D

All RAID

	Read Latency	Write Latency
RAID-0	D	D
RAID-1	D	D
RAID-4	D	2D
RAID-5	D	2D

but RAID-5 can
do more in parallel

All RAID

	Seq Read	Seq Write	Rand Read	Rand Write
RAID-0	$N * S$	$N * S$	$N * R$	$N * R$
RAID-1	$N/2 * S$	$N/2 * S$	$N * R$	$N/2 * R$
RAID-4	$(N-1)*S$	$(N-1)*S$	$(N-1)*R$	$R/2$
RAID-5	$(N-1)*S$	$(N-1)*S$	$N * R$	$N/4 * R$

All RAID

	Seq Read	Seq Write	Rand Read	Rand Write
RAID-0	$N * S$	$N * S$	$N * R$	$N * R$
RAID-1	$N/2 * S$	$N/2 * S$	$N * R$	$N/2 * R$
RAID-4	$(N-1) * S$	$(N-1) * S$	$(N-1) * R$	$R/2$
RAID-5	$(N-1) * S$	$(N-1) * S$	$N * R$	$N/4 * R$

RAID-5 is strictly better than RAID-4

All RAID

	Seq Read	Seq Write	Rand Read	Rand Write
RAID-0	$N * S$	$N * S$	$N * R$	$N * R$
RAID-1	$N/2 * S$	$N/2 * S$	$N * R$	$N/2 * R$
RAID-5	$(N-1) * S$	$(N-1) * S$	$N * R$	$N/4 * R$

All RAID

	Seq Read	Seq Write	Rand Read	Rand Write
RAID-0	$N * S$	$N * S$	$N * R$	$N * R$
RAID-1	$N/2 * S$	$N/2 * S$	$N * R$	$N/2 * R$
RAID-5	$(N-1) * S$	$(N-1) * S$	$N * R$	$N/4 * R$

RAID-0 is always fastest and has best capacity.
(but at cost of reliability)

All RAID

	Seq Read	Seq Write	Rand Read	Rand Write
RAID-0	$N * S$	$N * S$	$N * R$	$N * R$
RAID-1	$N/2 * S$	$N/2 * S$	$N * R$	$N/2 * R$
RAID-5	$(N-1) * S$	$(N-1) * S$	$N * R$	$N/4 * R$

RAID-5 better than RAID-1 for sequential.

All RAID

	Seq Read	Seq Write	Rand Read	Rand Write
RAID-0	$N * S$	$N * S$	$N * R$	$N * R$
RAID-1	$N/2 * S$	$N/2 * S$	$N * R$	$N/2 * R$
RAID-5	$(N-1) * S$	$(N-1) * S$	$N * R$	$N/4 * R$

All RAID

	Seq Read	Seq Write	Rand Read	Rand Write
RAID-0	$N * S$	$N * S$	$N * R$	$N * R$
RAID-1	$N/2 * S$	$N/2 * S$	$N * R$	$N/2 * R$
RAID-5	$(N-1) * S$	$(N-1) * S$	$N * R$	$N/4 * R$

RAID-1 better than RAID-4 for random write.

Summary

Many **engineering tradeoffs** with RAID.
(capacity, reliability, different types of performance).

H/W RAID controllers can handle crashes easier.

Transparent, deployable solutions are popular.