Week 13

Handling Crashes and Performance

Oana Balmau

March 28, 2023

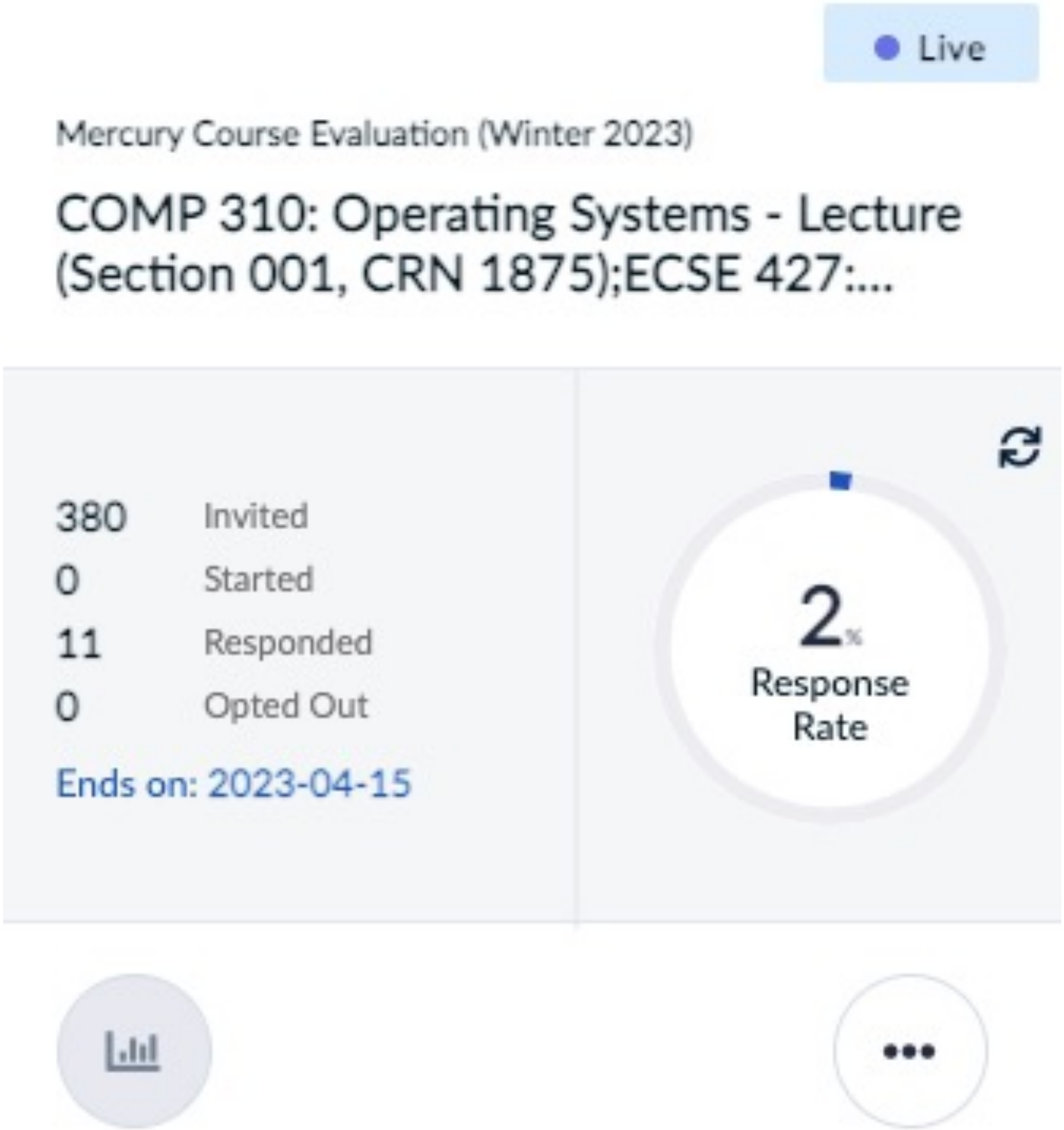
Class admin

	Allocation II				
Week 13 **File Systems**	mar 27 C Review: Advanced debugging	mar 28 **Handling Crashes & Performance (1/2)** **Optional reading: OSTEP Chapters 38, 43**	mar 29	mar 30 **Handling Crashes & Performance (2/2)** • ~~Grades released for Exercises Sheet~~ • Practice Exercises Sheet: File Systems	mar 31
Week 14 **Advanced Topics**	apr 3 No lab. Work on Assignment 3 ~~**Memory Management Assignment Due**~~	apr 4 **Advanced topics: Virtualization**	apr 5	apr 6 **Advanced topics: Operating Systems Research** (Invited Speaker: TBD) Grades released for Exercises Sheet	apr 7
Week 15 **Wrap-up**	apr 10 No Lab. Prepare for end-of-semester. **Memory Management Assignment Due**	apr 11 **End-of-semester Q&A– not recorded**	apr 12	apr 13 **End-of-semester Q&A — not recorded.** Last class!	apr 14 Grades released for Memory Management Assignment

Course evaluations

Please give us feedback!
https://www.mcgill.ca/mercury/

Resource on giving constructive feedback:
https://www.mcgill.ca/mercury/students/feedback



Course evaluations

Winter 21
Class size: 258
Evaluation: **3.2**/5 ⭐⭐⭐

✅ Good slides
✅ Good examples
❌ Unclear how book and lectures relate
❌ Inconsistent TA answers
❌ long waiting times on Ed
❌ Unclear assignments
❌ Too many assignments/quizzes
❌ Issues with C

Changes
- TA training and clear role assignment
- Changed the textbook + targeted readings
- No quizzes + drop 1 assignment
- Emphasize consistency and TA responsiveness
- Add practice exercises
- C labs

Winter 22
Class size: 287
Evaluation: **4.2**/5 ⭐⭐⭐⭐

✅ Good Slides
✅ Good examples
✅ Students appreciated low ED response time
✅ One of our TAs won SOCS Best TA
✅ Students liked the OSTEP book

❌ Want more exercises to prep exam
❌ Want to get a better idea of assignments expectations

Course evaluations

Winter 22
Class size: 287
Evaluation: **4.2**/5 ⭐⭐⭐⭐

✅ Good Slides
✅ Students appreciated low ED response time
✅ One of our TAs won SOCS Best TA
✅ Students liked the OSTEP book
❌ Want more exercises to prep exam
❌ Want to get a better idea of assignments expectations

Changes:

- Autograder.
- Explicit testcases.
- More in-class exercises.
- Exam-style graded exercises.
- Refining of C labs.

Winter 23
Class size: 380

What do you think?

RAID

Sometimes we want many disks

Why?

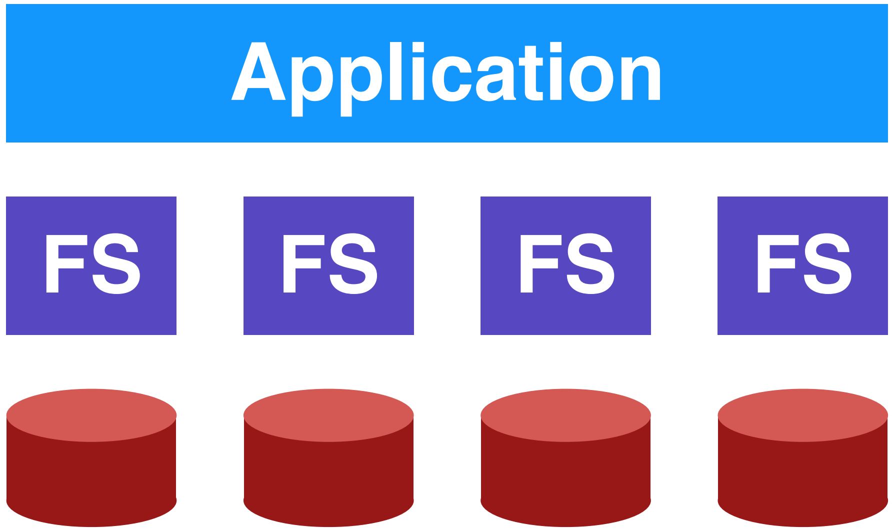
- Capacity
- Reliability
- Performance

Challenge: Most FS work with one disk

How to make a large, fast, reliable storage system?

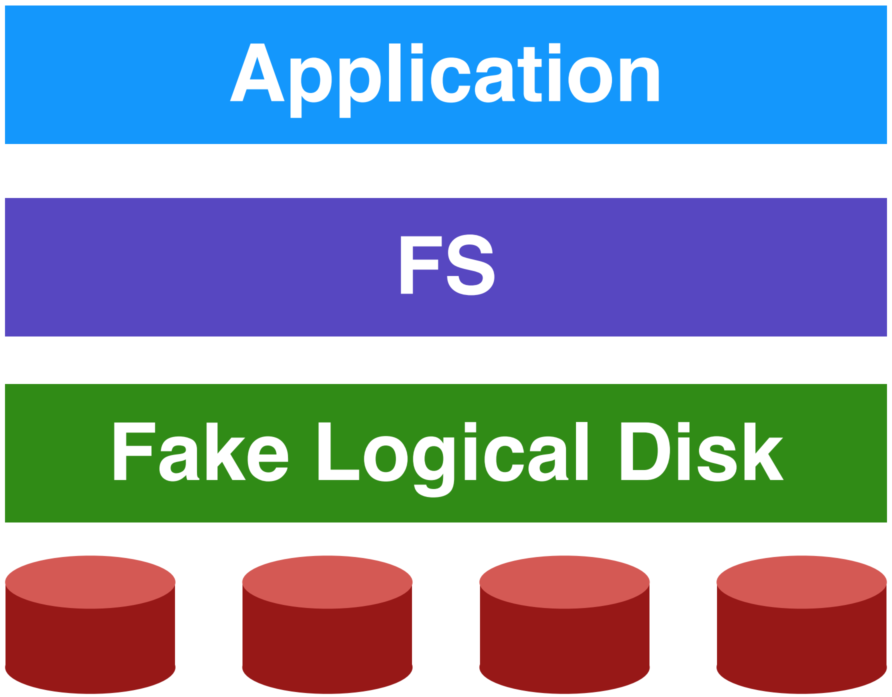
- What are the key techniques?
- What are trade-offs between different approaches?

Naïve Solution: Just a Bunch of Disks (JBOD)



Application is smart, stores different files on different file systems.

Better Solution: RAID



Create the illusion of one disk from many disks.

RAID

- *R*edundant *A*rray of *I*ndependent *D*isks

- Essential idea
  - Optimize I/O bandwidth through parallel I/O
  - Parallel I/O = I/O to multiple disks at once

RAID Format

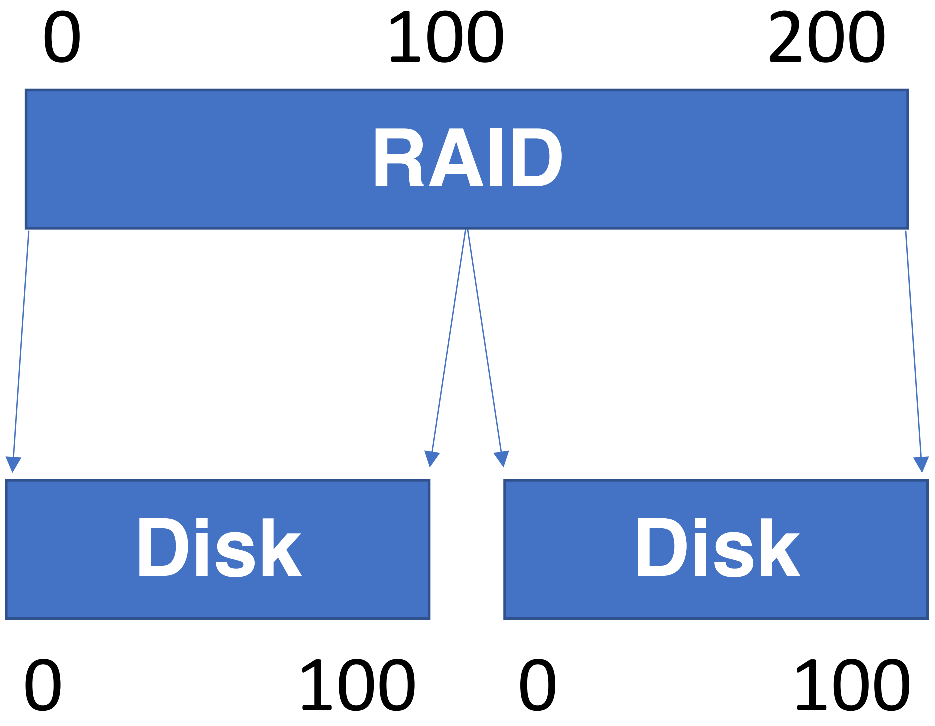
- Disks now cheap and small
- Many can go into a RAID box
- To OS: RAID box looks like disk
- Also possible: RAID in software

RAID - Two General Strategies

- **Mapping**
- **Redundancy**

Mapping

- Build fast, large disk from smaller ones



Striping

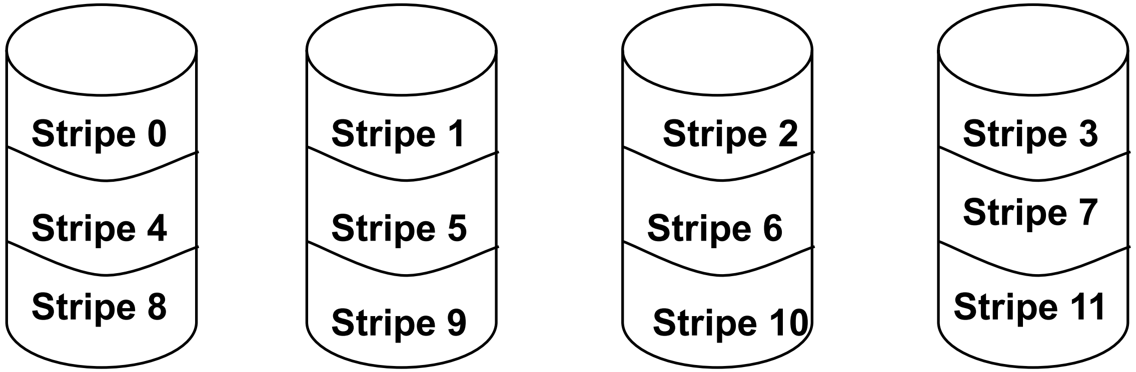
- A form of mapping
- Rather than put file on one disk
- Stripe it across a number of disks
    - File = Stripe0 | Stripe1 | Stripe2 …

      Stripe0 on disk0

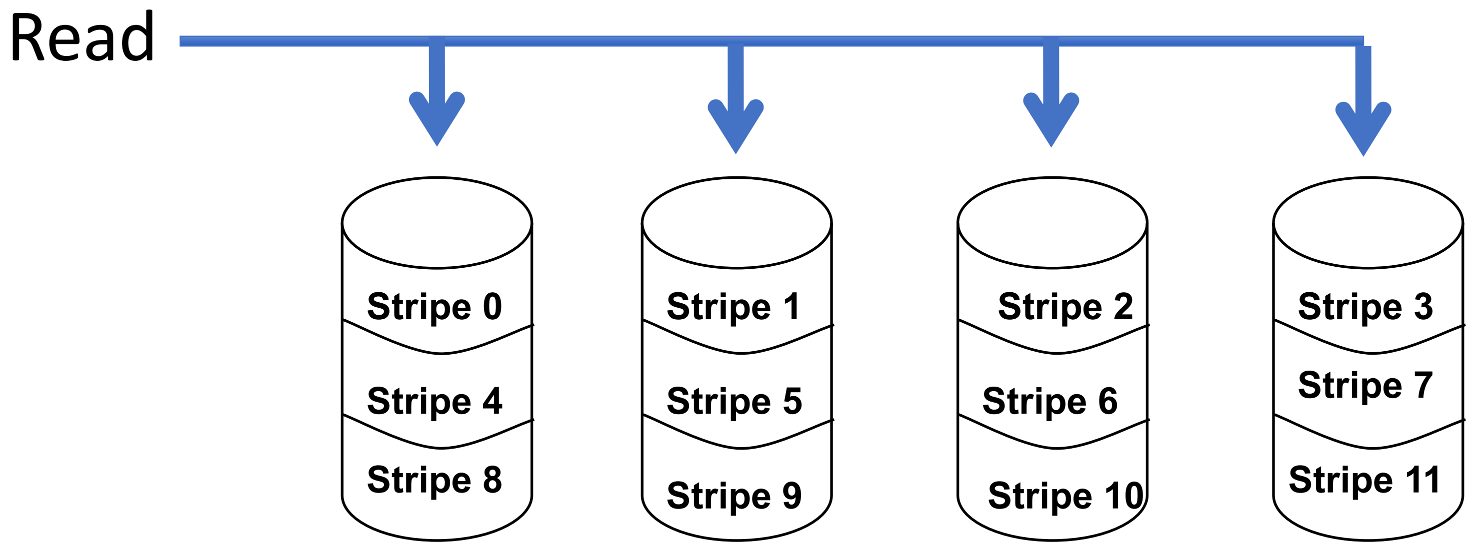
      Stripe1 on disk1

      …
- **Read and write in parallel ☺ ☺**

Striping

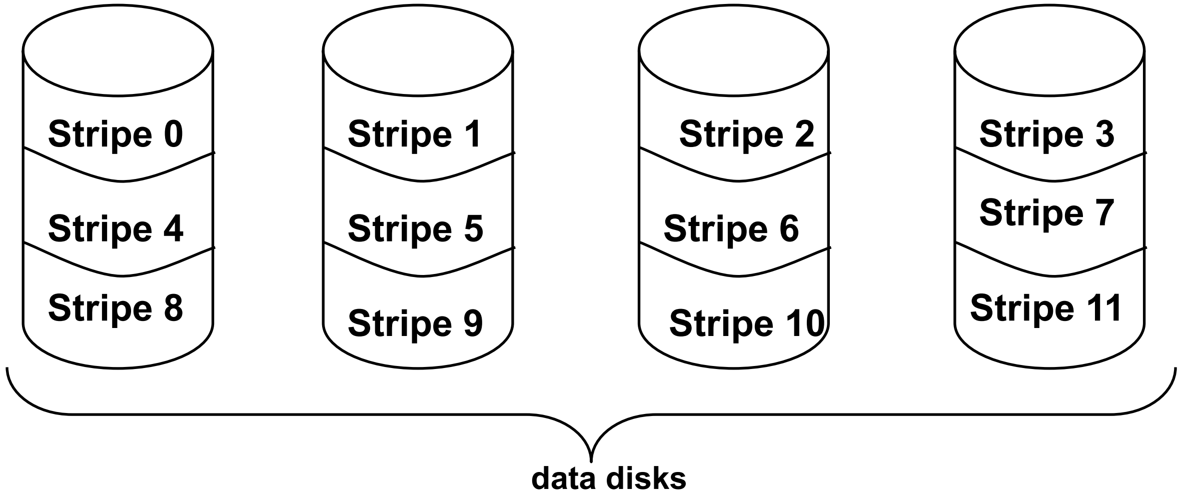

Stripe 0
Stripe 4
Stripe 8
Stripe 1
Stripe 5
Stripe 9
Stripe 2
Stripe 6
Stripe 10
Stripe 3
Stripe 7
Stripe 11

Striping Read/Write



RAID-0

- Uses striping
- Best possible read and write bandwidth
- Failure results in data loss



RAID-0 Analysis

What is **capacity?**	
How many disks can **fail**?	
Latency	
Throughput (seq, random)?	

N := number of disks
C := capacity of 1 disk
D := latency of one small I/O operation

S := sequential throughput of 1 disk
R := random throughput of 1 disk

RAID-0 Analysis

What is **capacity?**	**N*C**
How many disks can **fail**?	**0**
Latency	**D**
Throughput (seq, random)?	**N*S , N*R**

Buying more disks improves throughput, but not latency!

N := number of disks
C := capacity of 1 disk
D := latency of one small I/O operation

S := sequential throughput of 1 disk
R := random throughput of 1 disk

RAID-0 Analysis

What is **capacity?**	**N*C**
How many disks can **fail**?	**0**
Latency	**D**
Throughput (seq, random)?	**N*S , N*R**

Problem?

N := number of disks
C := capacity of 1 disk
D := latency of one small I/O operation

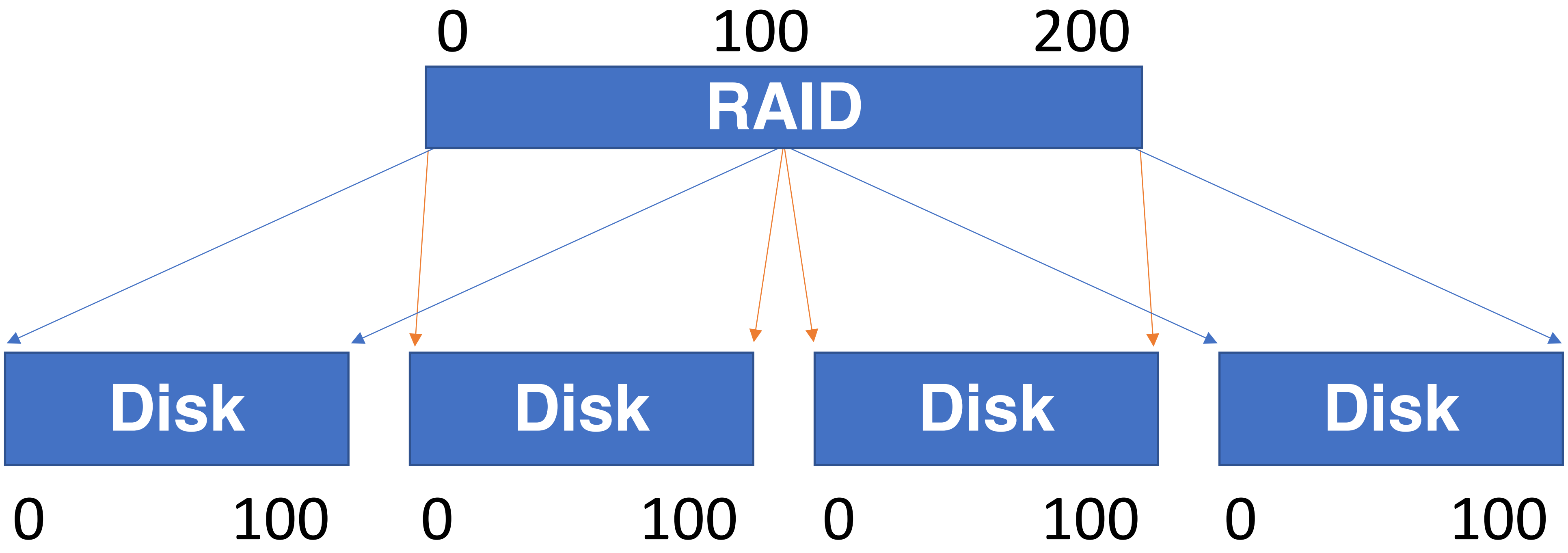
S := sequential throughput of 1 disk
R := random throughput of 1 disk

Problem with RAID-0

- One disk fails ➔ all data unavailable

Solution: Redundancy

- Store redundant data on different disks

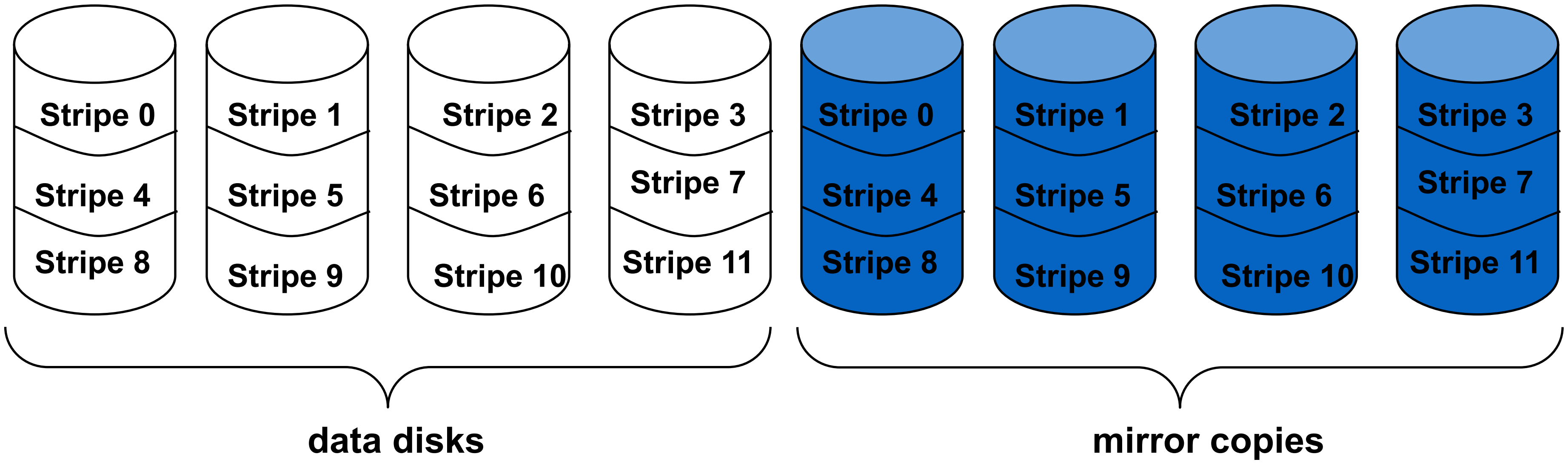


RAID Levels

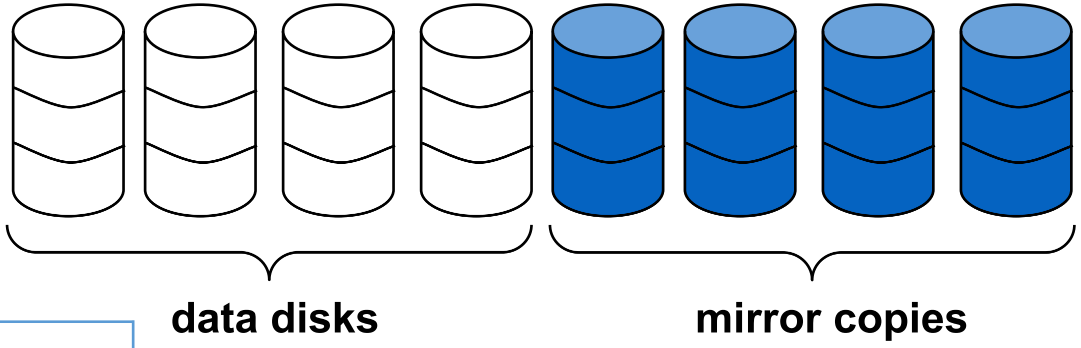
- Are redundancy levels
- What we have seen so far
  - RAID-0: No redundancy
- In reality:
  - RAID-1: Mirroring
  - RAID-2/3: not covered in this class
  - RAID-4: Parity disk
  - RAID-5: Distributed parity

RAID-1: Mirrored disks

- Write: to data and to mirror disk
- Read: from either data or mirror
- After crash: from surviving disk



RAID-1 Analysis

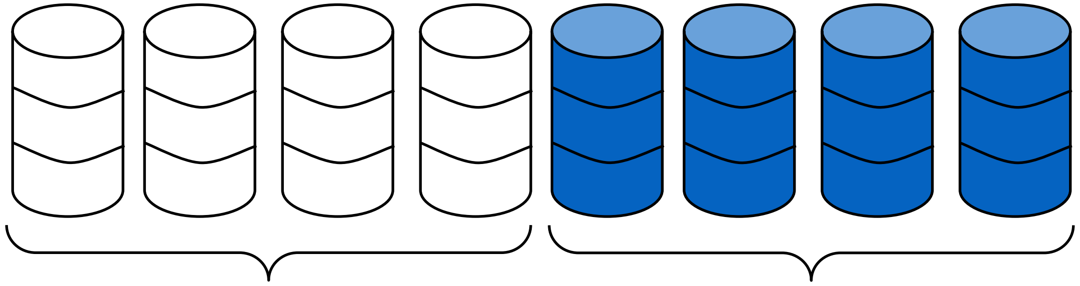


What is **capacity?**	
How many disks can **fail**?	
Latency	
Throughput (seq, random)?	

N := number of disks
C := capacity of 1 disk
D := latency of one small I/O operation

S := sequential throughput of 1 disk
R := random throughput of 1 disk

RAID-1 Analysis



What is **capacity?**	**N/2*C**
How many disks can **fail**?	**1 (or maybe N/2)**
Latency	**D**
Throughput (seq, random)?	

N := number of disks
C := capacity of 1 disk
D := latency of one small I/O operation

S := sequential throughput of 1 disk
R := random throughput of 1 disk

RAID-1 Analysis Throughput

Random reads	**N * R**
Random writes	**N/2 * R**
Sequential writes	**N/2 * S**
Sequential reads	**N/2 * S**

N := number of disks
C := capacity of 1 disk
D := latency of one small I/O operation

S := sequential throughput of 1 disk
R := random throughput of 1 disk

RAID-1 Analysis

For the same number of disks as RAID-0, **storage capacity is half!** ☹

What is **capacity?**	**N/2*C**
How many disks can **fail**?	**1 if unlucky (or maybe N/2)**
Latency	**D**
Throughput (seq, random)?	**N*R, N/2 * R ← rand** **N/2 * S ← seq**

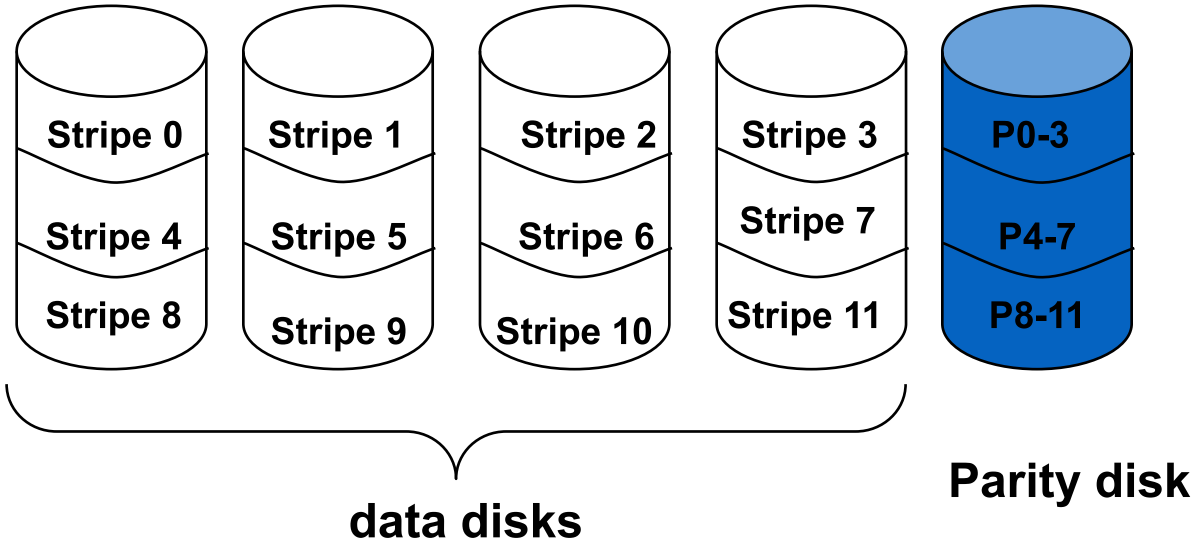
N := number of disks
C := capacity of 1 disk
D := latency of one small I/O operation

S := sequential throughput of 1 disk
R := random throughput of 1 disk

How to do better?

RAID-4

- N data disks + 1 parity disk



Parity

- A simple form of error detection and repair
- Not specific to RAID
- Also used in communications

Parity Idea

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

- Treat sectors across disks in a stripe as an equation.

- Data on bad disk is like an unknown in the equation.

Parity Example

Disk0 Disk1 Disk2 Disk3 Disk4

Stripe:

(parity)

Parity Example

	Disk0	Disk1	Disk2	Disk3	Disk4
Stripe:	5	3	0	1	
					(parity)

Parity Example

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

Parity Example

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

Parity Example

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

Parity Example

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

Parity Example 2

With XOR (exclusive OR) as the parity function

- 4 bits: x_0, x_1, x_2, x_3
- Parity $p = x_0$ XOR x_1 XOR x_2 XOR x_3
- If you lose one bit, say x_2
- Reconstruct as $x_2 = x_0$ XOR x_1 XOR x_3 XOR p

X	Y	X XOR Y
0	0	0
0	1	1
1	0	1
1	1	0

Parity Example 2

X	Y	X XOR Y
0	0	0
0	1	1
1	0	1
1	1	0

- 4 bits: $x_0x_1x_2x_3$ = 0101
- Parity p = x_0 XOR x_1 XOR x_2 XOR x_3 = 0
- If you lose one bit, say x_2
- Reconstruct as x_2 = x_0 XOR x_1 XOR x_3 XOR p =

  0 XOR 1 XOR 1 XOR 0 =

  0

→ X_2 = 0

RAID Parity Block

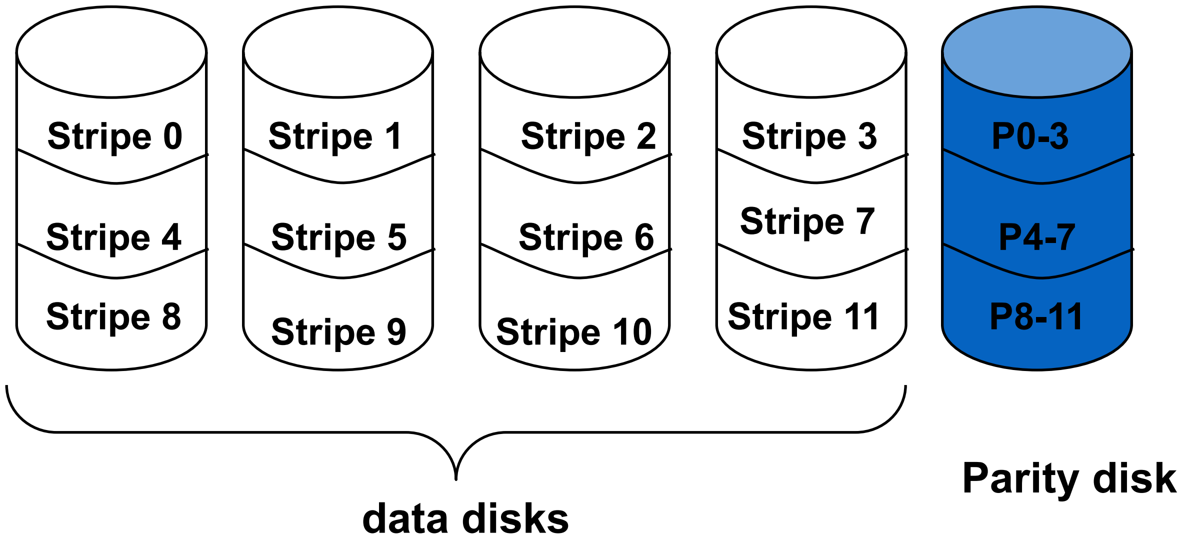
- Same idea at the disk block level

- Block on parity disk =

XOR of bits of data blocks at same position

RAID-4

- Read: read data disks
- Write: write data disks and parity disk
- Crash: recover from data and parity disk



RAID-4 Analysis

What is **capacity?**	
How many disks can **fail**?	
Latency	
Throughput (seq, random)?	

N := number of disks
C := capacity of 1 disk
D := latency of one small I/O operation

S := sequential throughput of 1 disk
R := random throughput of 1 disk

RAID-4 Analysis

What is **capacity?**	**(N-1)*C**
How many disks can **fail**?	**1**
Latency (read, write)	**D, 2*D (read and write parity disk)**
Throughput (seq, random)?	

N := number of disks
C := capacity of 1 disk
D := latency of one small I/O operation

S := sequential throughput of 1 disk
R := random throughput of 1 disk

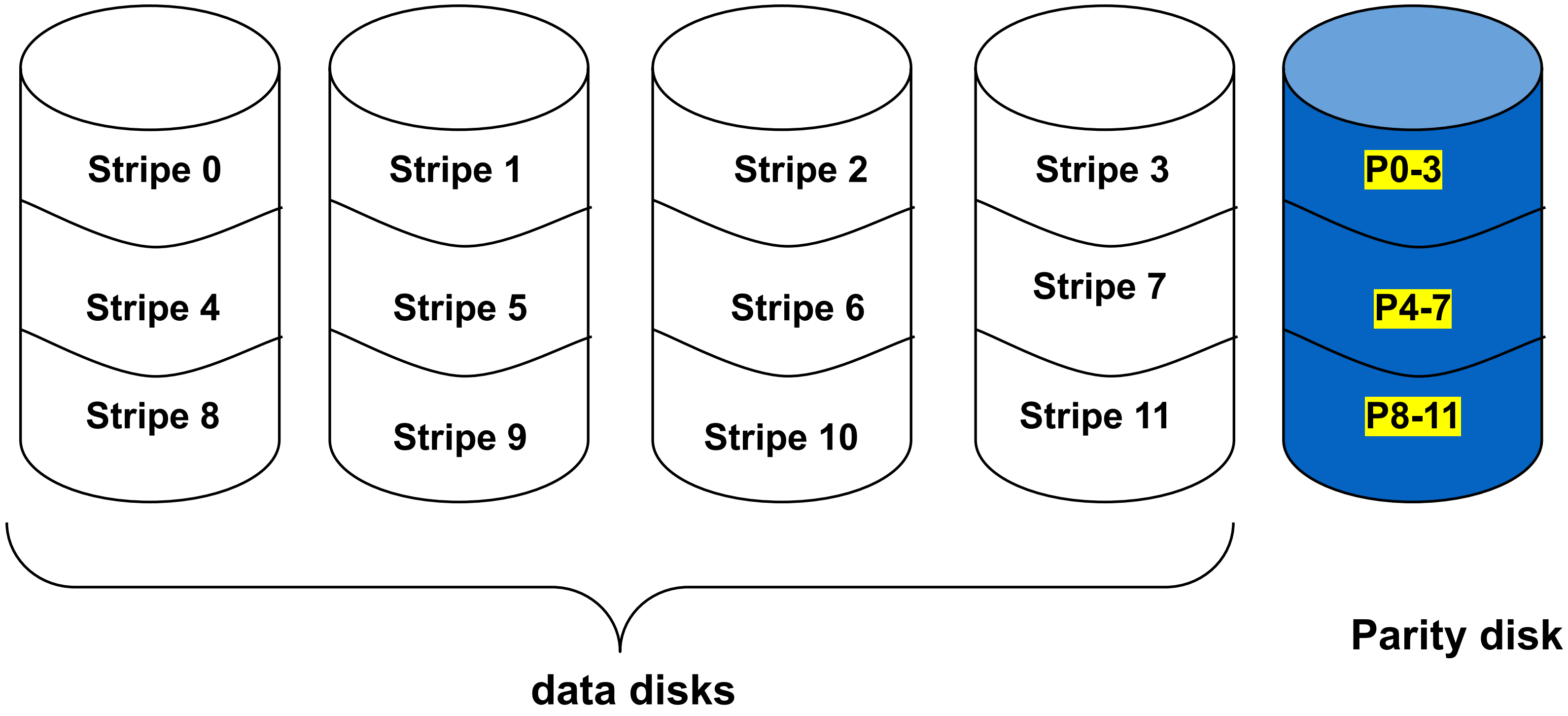
RAID-4 Analysis Throughput

Sequential reads	**(N-1) * S**
Sequential writes	**(N-1) * S**
Random reads	**(N-1) * R**
Random writes	**???**

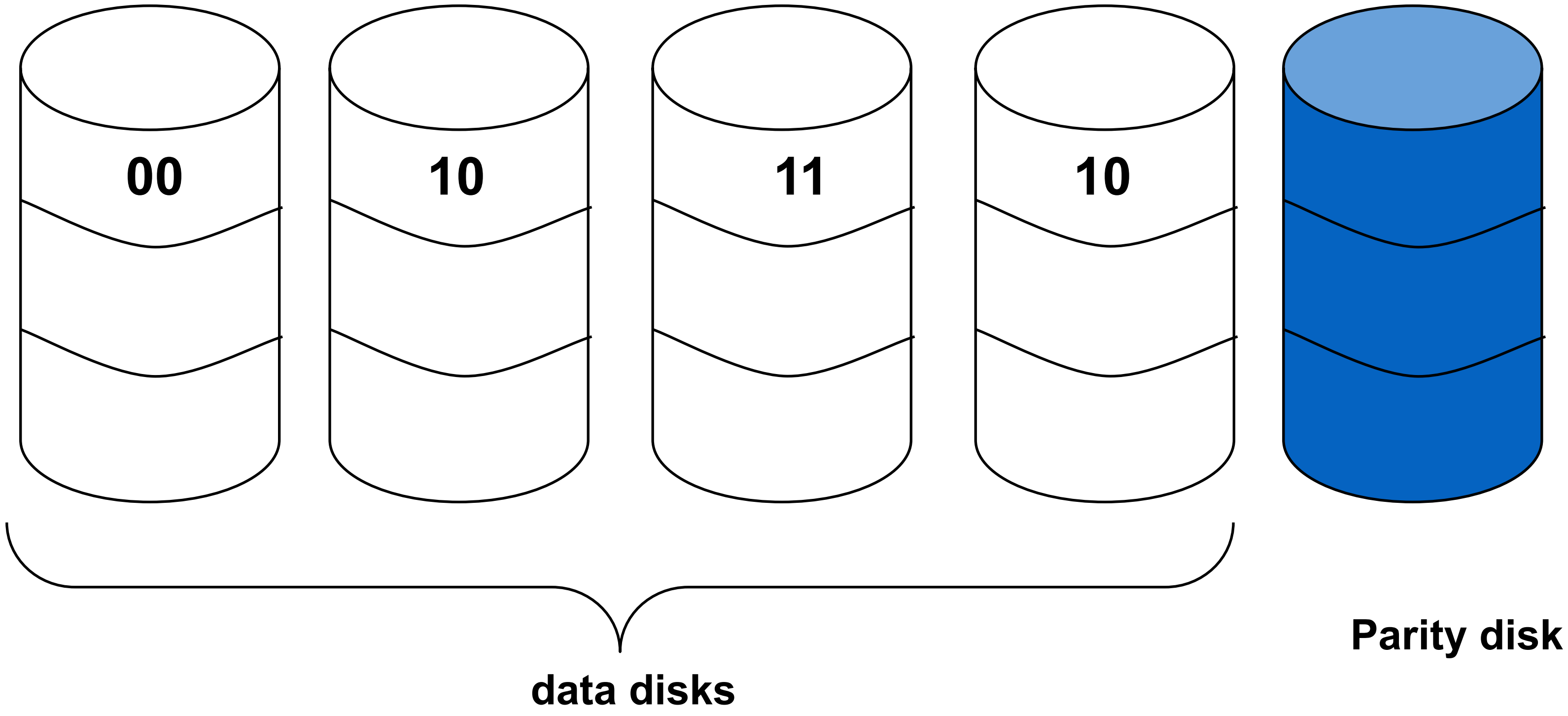
N := number of disks
C := capacity of 1 disk
D := latency of one small I/O operation

S := sequential throughput of 1 disk
R := random throughput of 1 disk

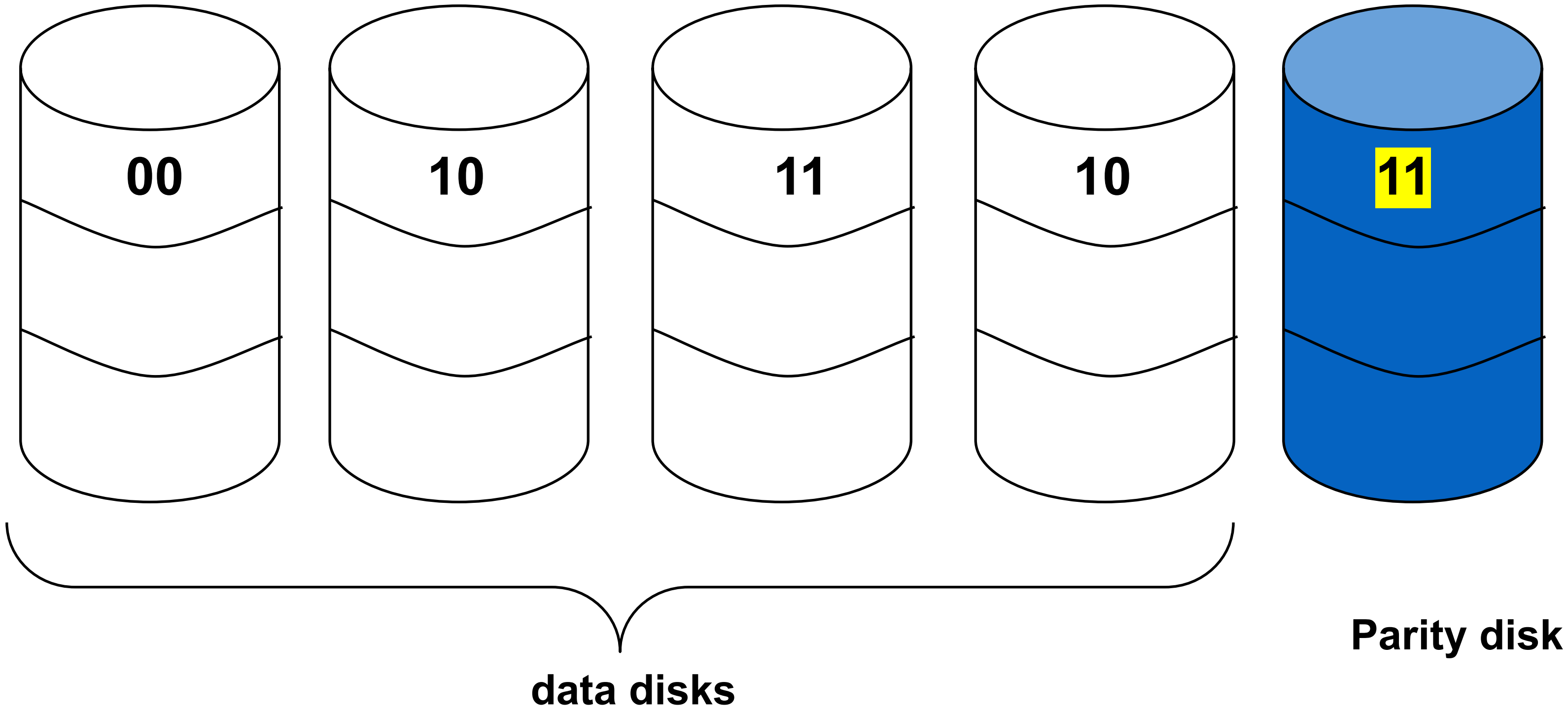
RAID-4 Random Writes Throughput Analysis



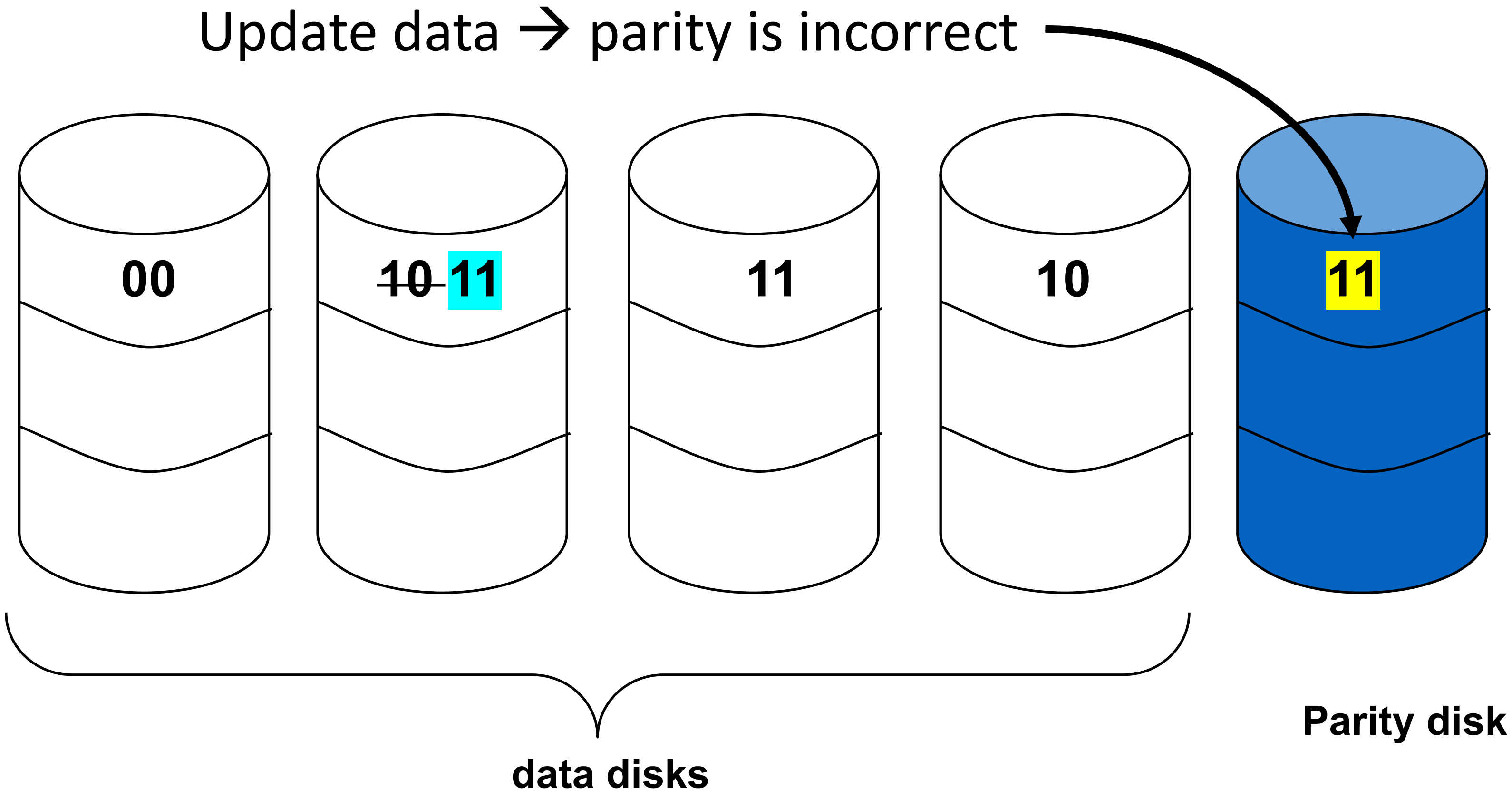
RAID-4 Random Writes Throughput Analysis



RAID-4 Random Writes Throughput Analysis



RAID-4 Random Writes Throughput Analysis



How is the parity updated?

2 methods:

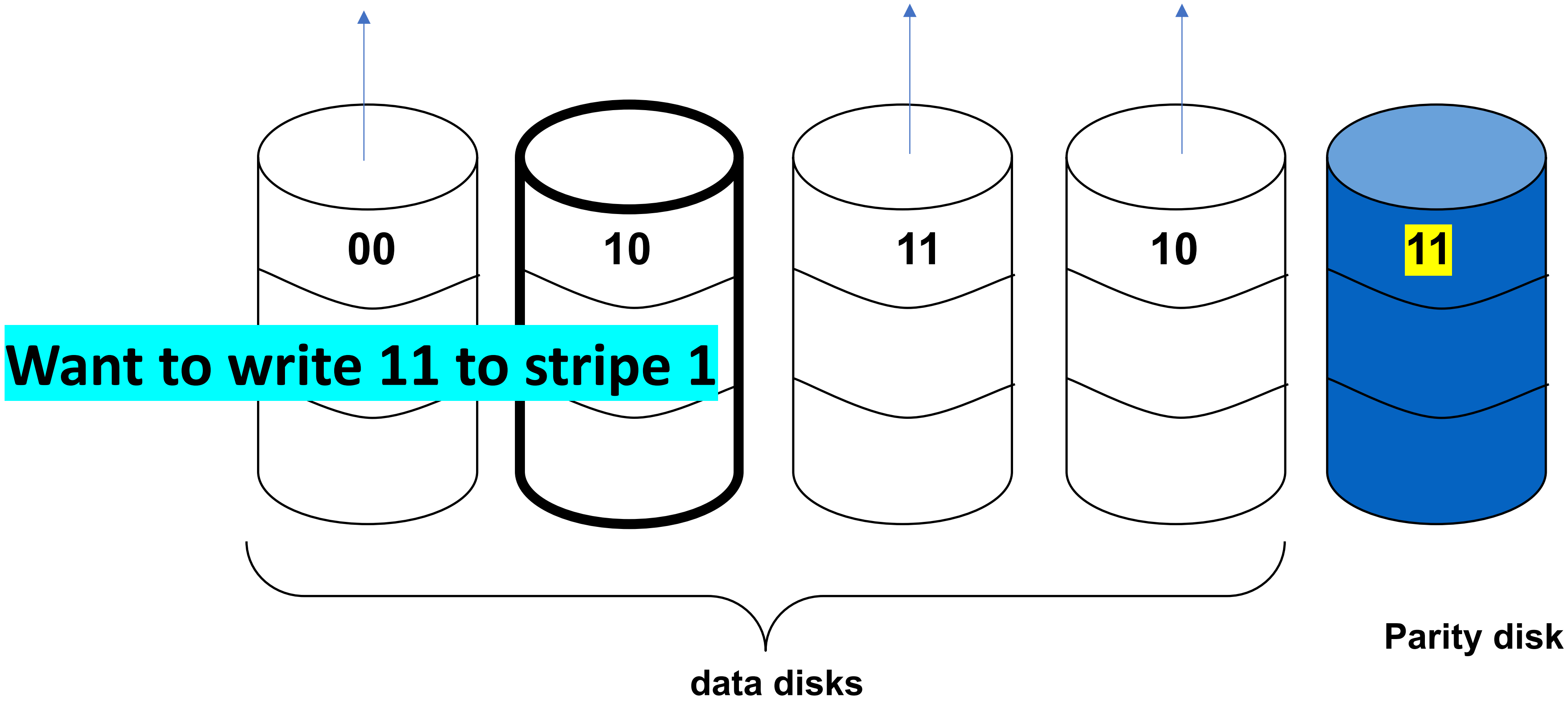
- Additive parity
- Subtractive parity

Additive parity

- Read all other data blocks in parallel
- XOR them with new block

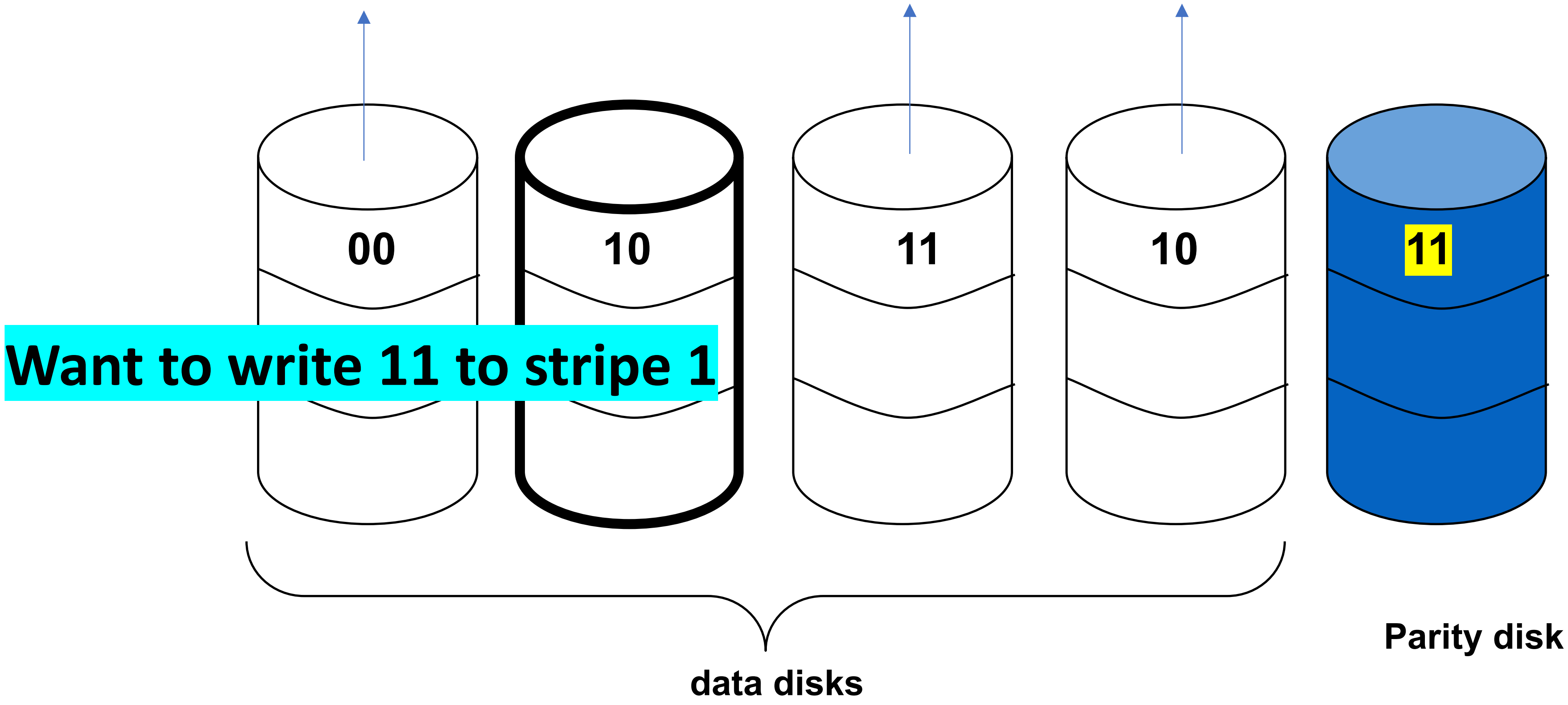
Additive parity

Read stripe 0, 2, 3 in parallel



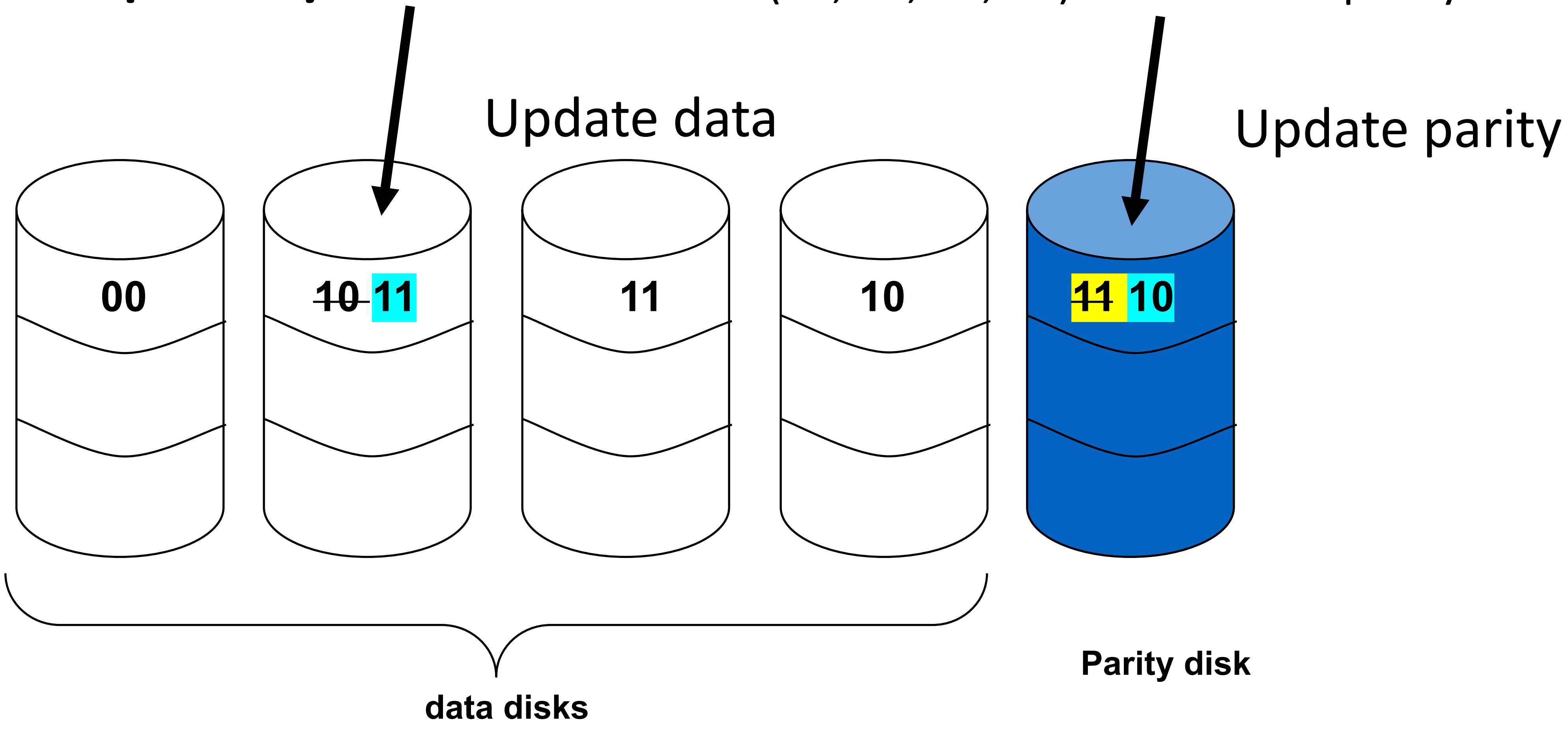
Additive parity

XOR (00, 11, 11, 10) = 10 ← new parity



Additive parity

XOR (00, 11, 11, 10) = 10 ← new parity



Additive Parity Performance

- 3 parallel read accesses to the data disks
  - Throughput = R

\+

- 2 parallel accesses to write the new parity plus new data
  - Throughput = R

→ RAID-4 Throughput for rand write = R/2

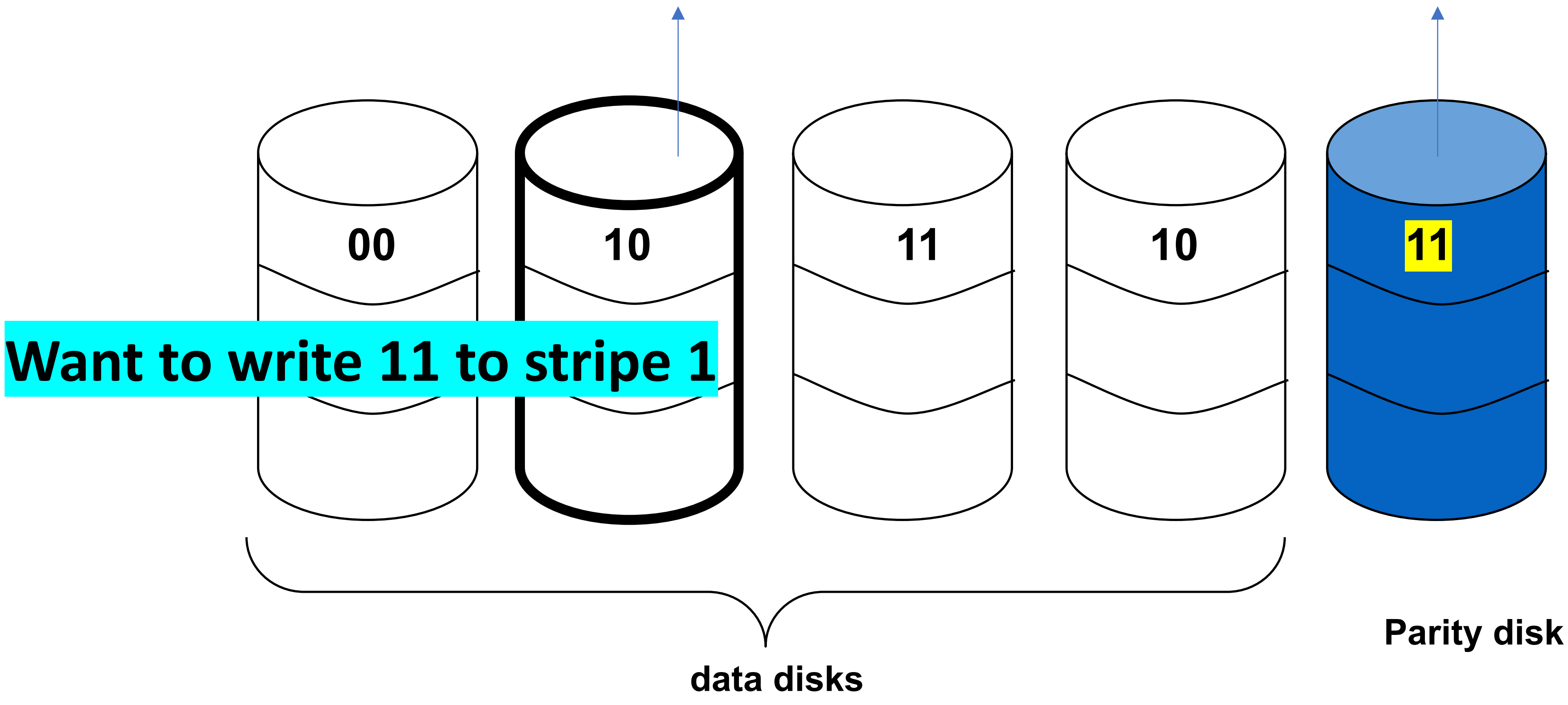
Subtractive parity

- Read old data & read old parity in parallel
- Compare new data with old data
- If new data == old data
  - Do nothing
- Else
  - Flip old parity bit to the opposite of its current state

$$\boxed{Par_{new} = (Data_{old} \text{ XOR } Data_{new}) \text{ XOR } Par_{old}}$$

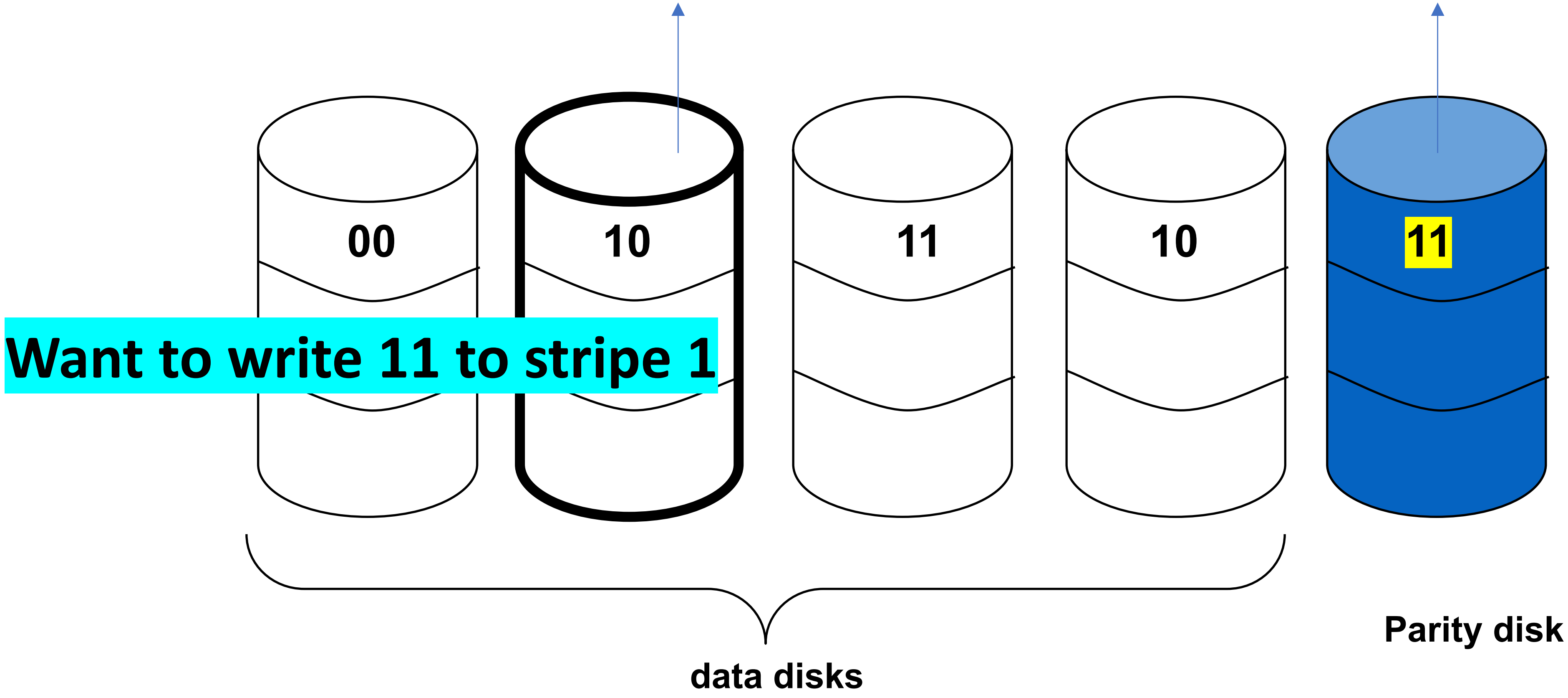
Subtractive parity

Read stripe 1 and parity in parallel



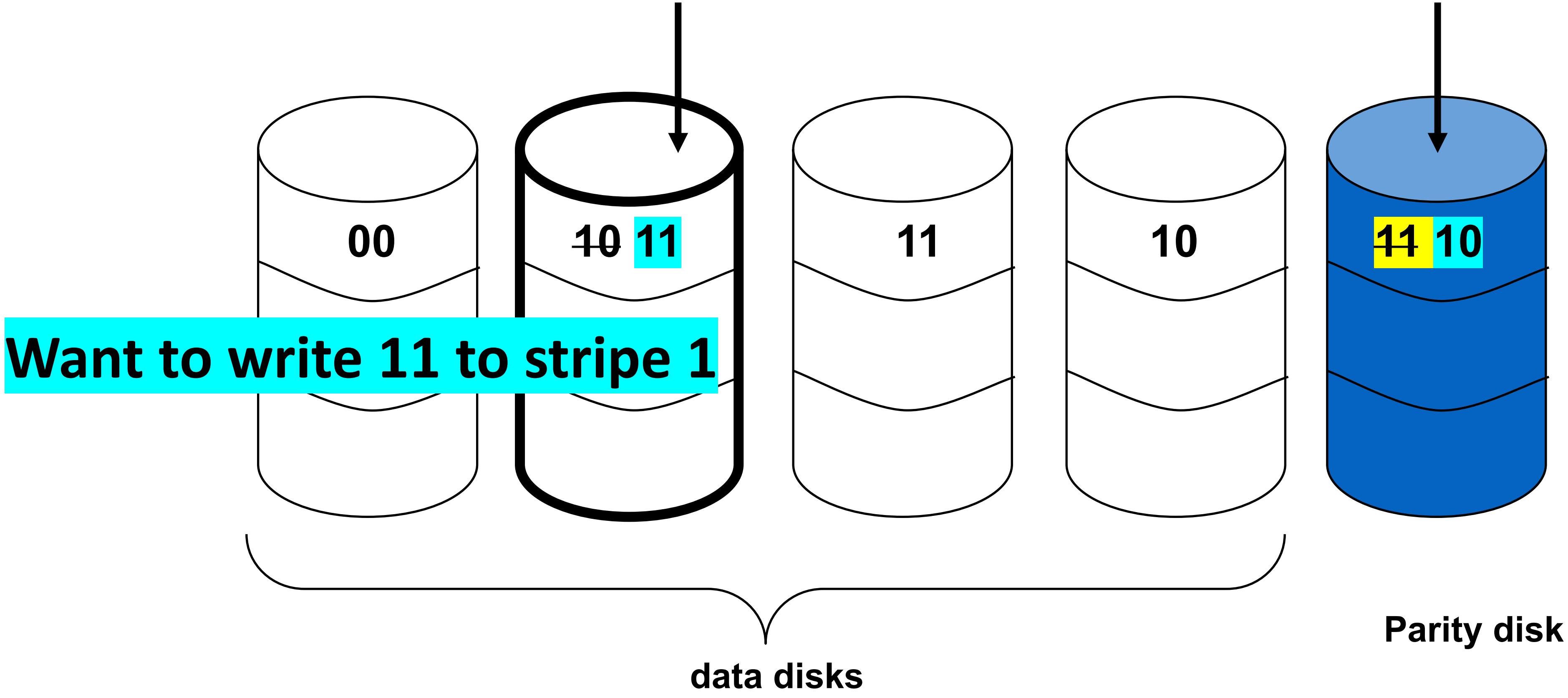
Subtractive parity

New parity = (11 XOR 10) XOR 11 =
= 01 XOR 11 = 10



Subtractive parity

Write new data and new parity in parallel



Subtractive Parity Performance

- 2 parallel read accesses to the data disk and parity disk
  - Throughput = R

\+

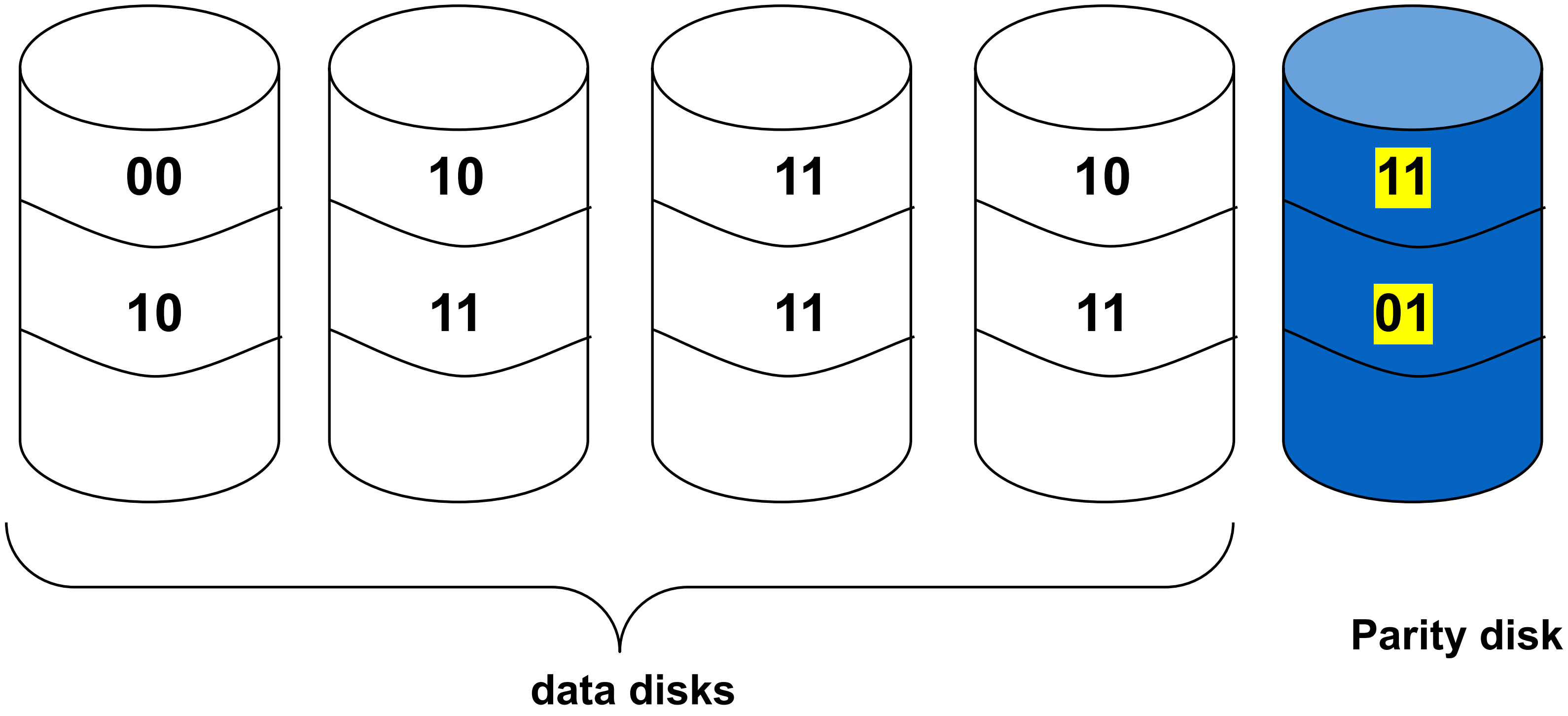
- 2 parallel accesses to write the new parity plus new data
  - Throughput = R

→ RAID-4 Throughput for rand write = R/2

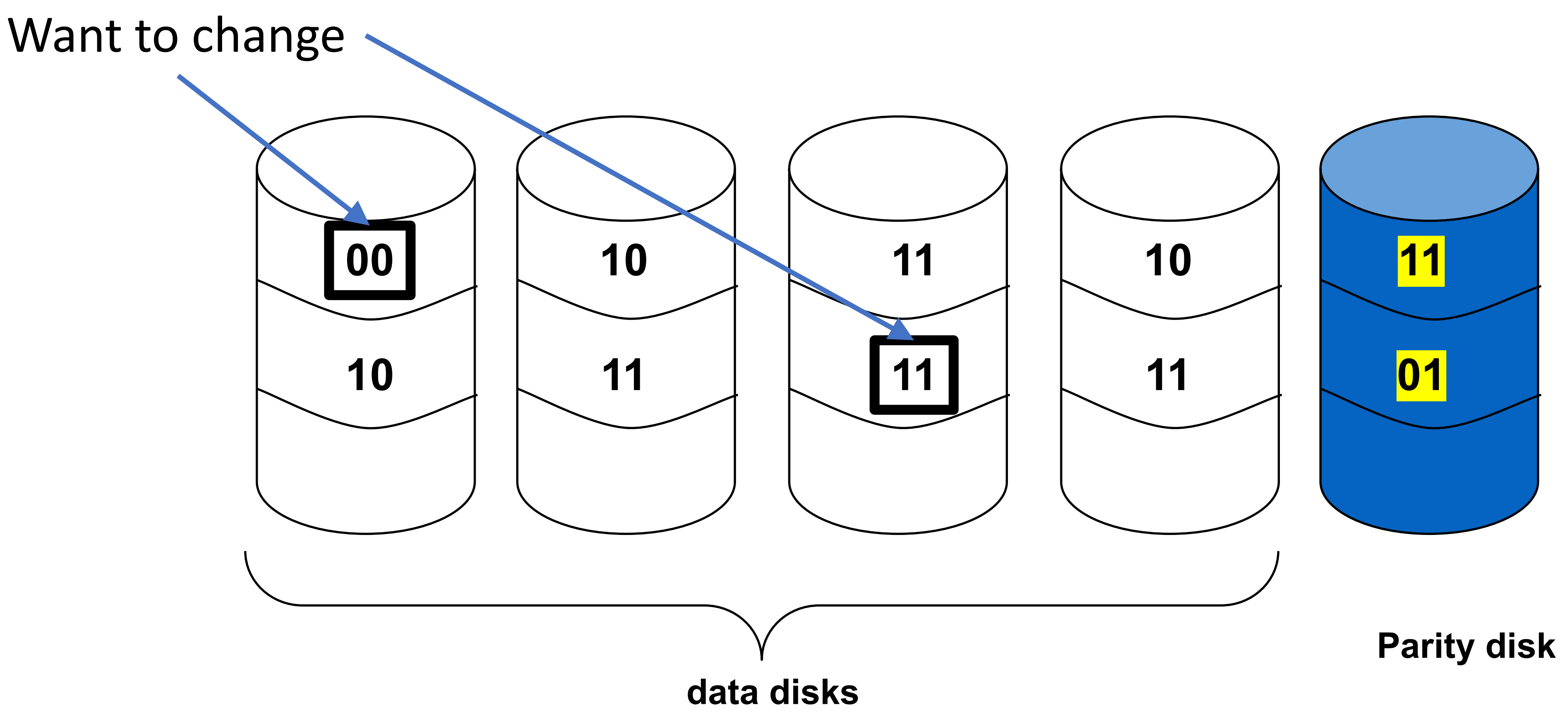
Issue with RAID-4

- What about concurrent random writes?
  - Every write **must** access parity disk.
  - Becomes bottleneck for write-heavy workload.

Concurrent random writes in RAID-4

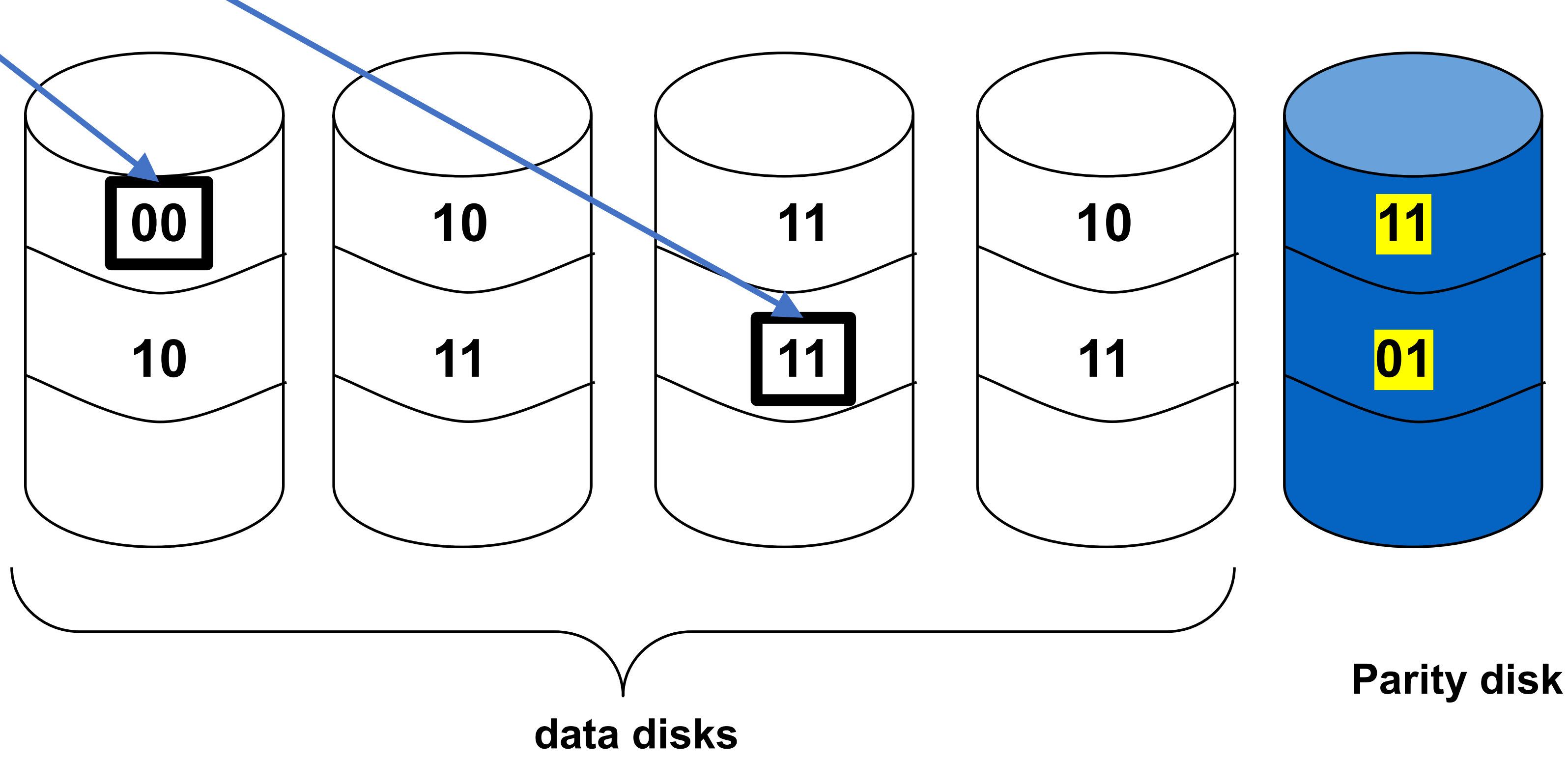


Concurrent random writes in RAID-4



Concurrent random writes in RAID-4

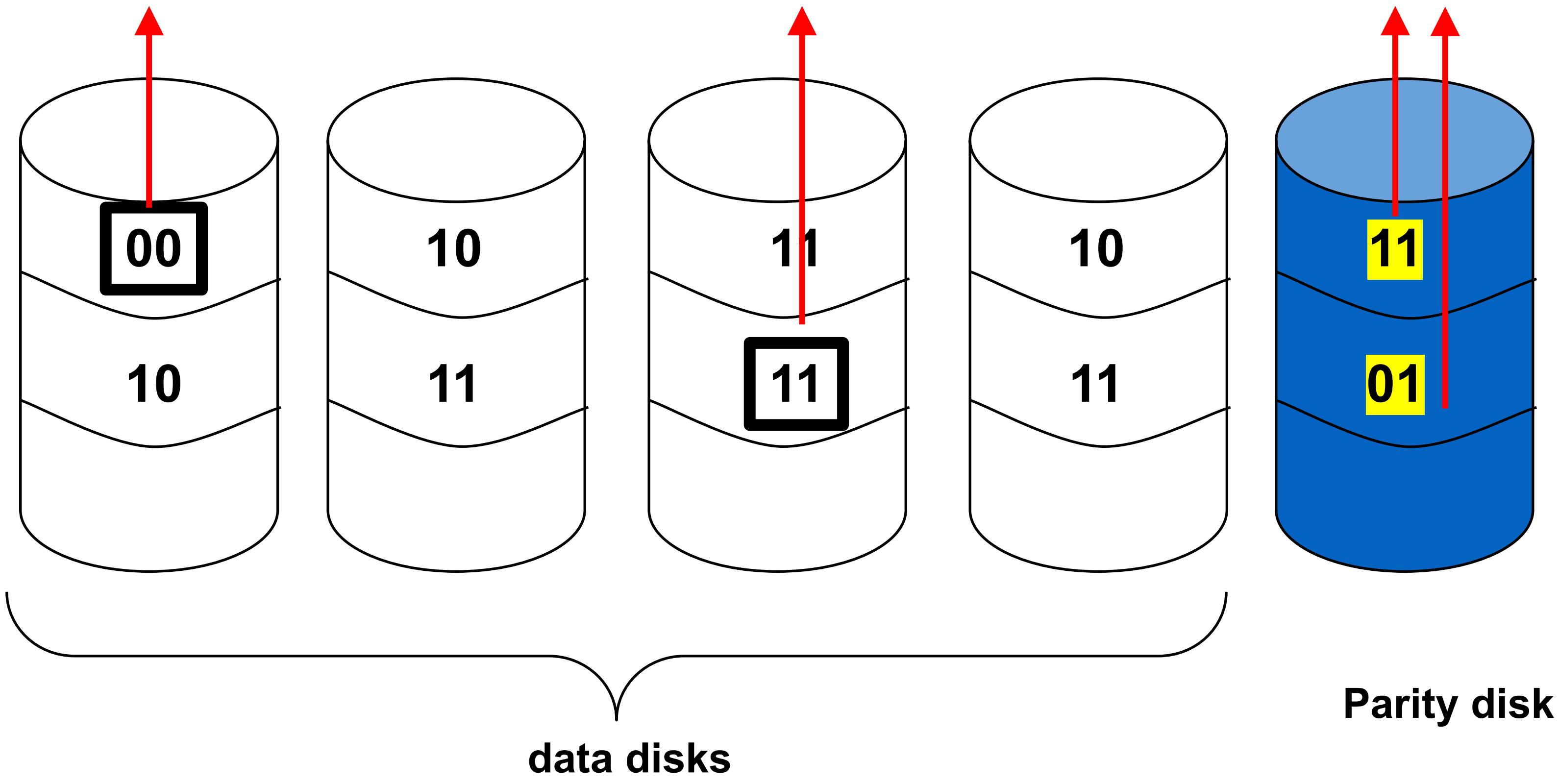
Want to change



Assume we use subtractive parity

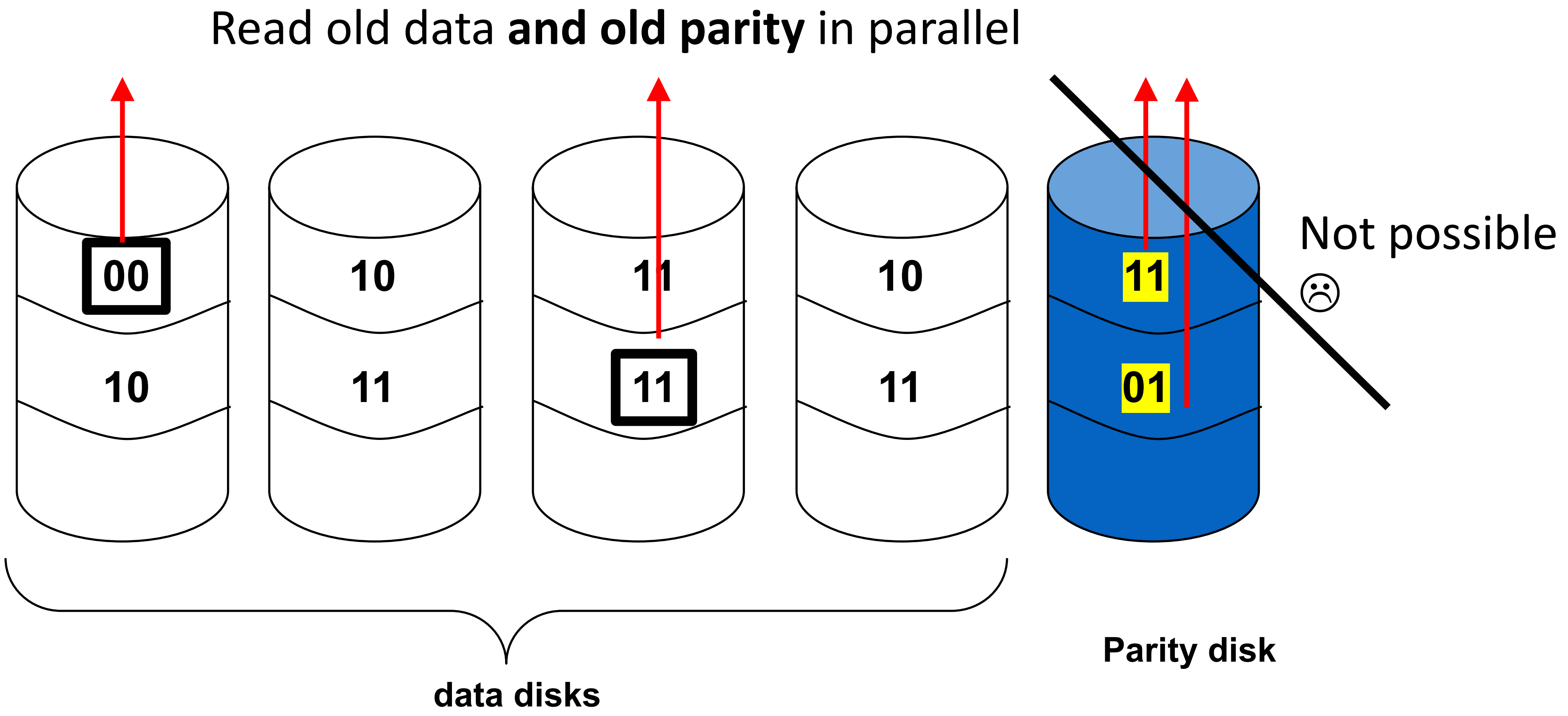
Concurrent random writes in RAID-4

Read old data and old parity in parallel



Assume we use subtractive parity

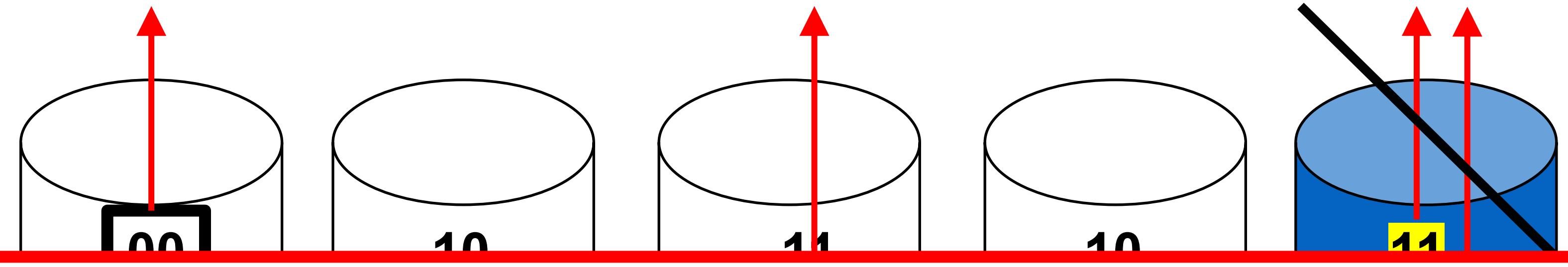
Concurrent random writes in RAID-4



Assume we use subtractive parity

Concurrent random writes in RAID-4

Read old data **and old parity** in parallel



Not possible

☹ Parity disk is a bottleneck for random writes in RAID-4

Issue with random write throughput

known as **small-write problem**

data disks

Parity disk

Assume we use subtractive parity

RAID-4 Analysis Throughput

Sequential reads	**(N-1) * S**
Sequential writes	**(N-1) * S**
Random reads	**(N-1) * R**
Random writes	**R/2 ☹**

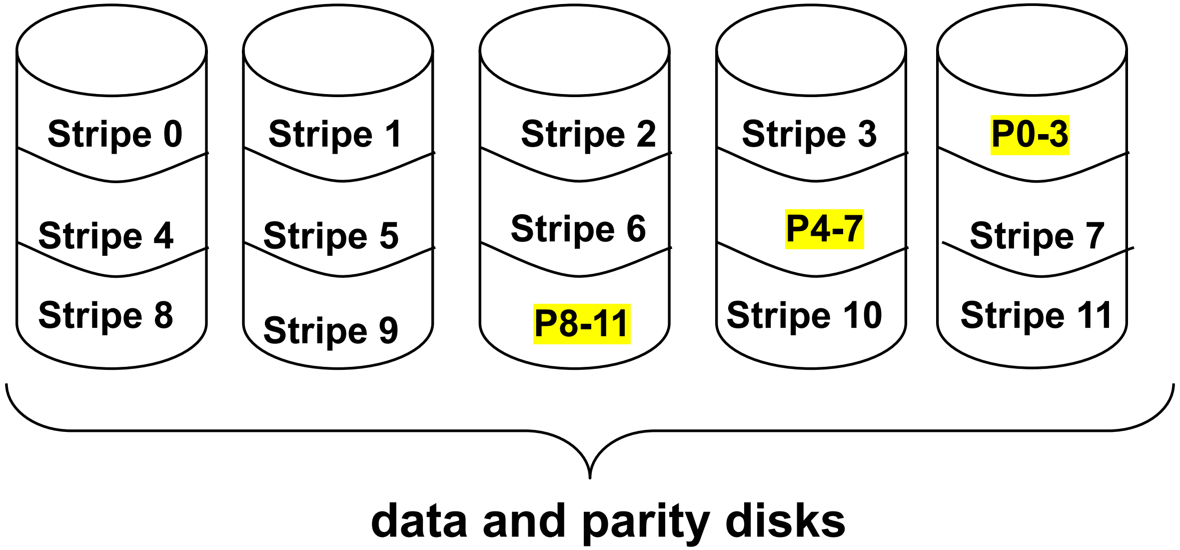
N := number of disks
C := capacity of 1 disk
D := latency of one small I/O operation

S := sequential throughput of 1 disk
R := random throughput of 1 disk

How to do better?

RAID-5

- Block interleaved **distributed parity**
- As RAID-4, but parity distributed over all disks
- Balances parity write load over disks



RAID-5 Analysis

	RAID 4	RAID 5
What is capacity?	**(N-1)*C**	
How many disks can fail?	**1**	
Latency	**D, 2*D**	

N := number of disks
C := capacity of 1 disk
D := latency of one small I/O operation

S := sequential throughput of 1 disk
R := random throughput of 1 disk

RAID-5 Analysis

	RAID 4	RAID 5
What is capacity?	**(N-1)*C**	**(N-1)*C**
How many disks can fail?	**1**	**1**
Latency	**D, 2*D**	**D, 2*D**

N := number of disks
C := capacity of 1 disk
D := latency of one small I/O operation

S := sequential throughput of 1 disk
R := random throughput of 1 disk

RAID-5 Analysis Throughput

	RAID 4	RAID 5
Sequential reads	**(N-1) * S**	
Sequential writes	**(N-1) * S**	
Random reads	**(N-1) * R**	
Random writes	**R/2**	

N := number of disks
C := capacity of 1 disk
D := latency of one small I/O operation

S := sequential throughput of 1 disk
R := random throughput of 1 disk

RAID-5 Analysis Throughput

	RAID 4	RAID 5
Sequential reads	**(N-1) * S**	**(N-1) * S**
Sequential writes	**(N-1) * S**	**(N-1) * S**
Random reads	**(N-1) * R**	
Random writes	**R/2**	

N := number of disks
C := capacity of 1 disk
D := latency of one small I/O operation

S := sequential throughput of 1 disk
R := random throughput of 1 disk

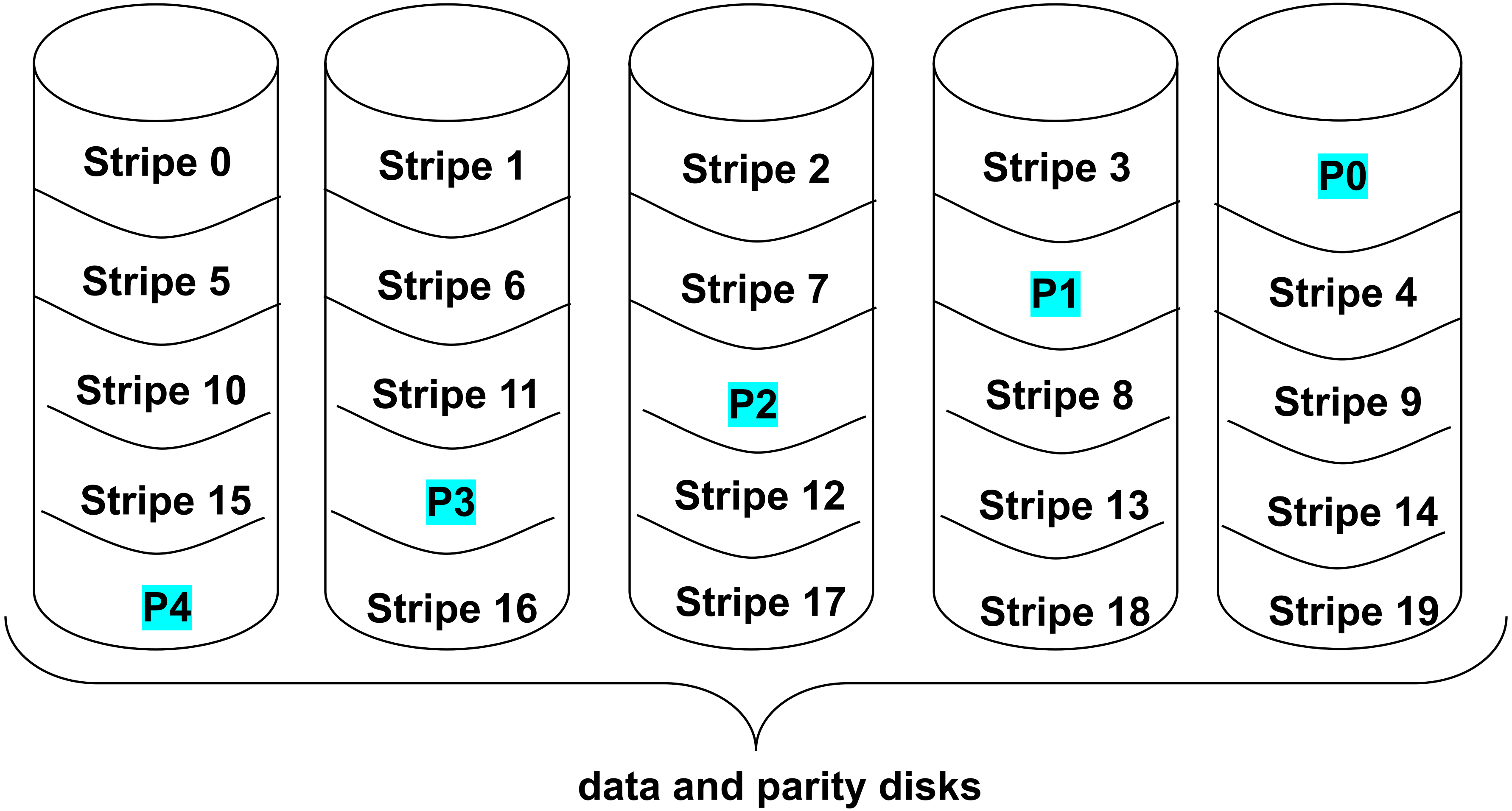
RAID-5 Analysis Throughput

	RAID 4	RAID 5
Sequential reads	**(N-1) * S**	**(N-1) * S**
Sequential writes	**(N-1) * S**	**(N-1) * S**
Random reads	**(N-1) * R**	**N*R**
Random writes	**R/2**	**N*R/4**

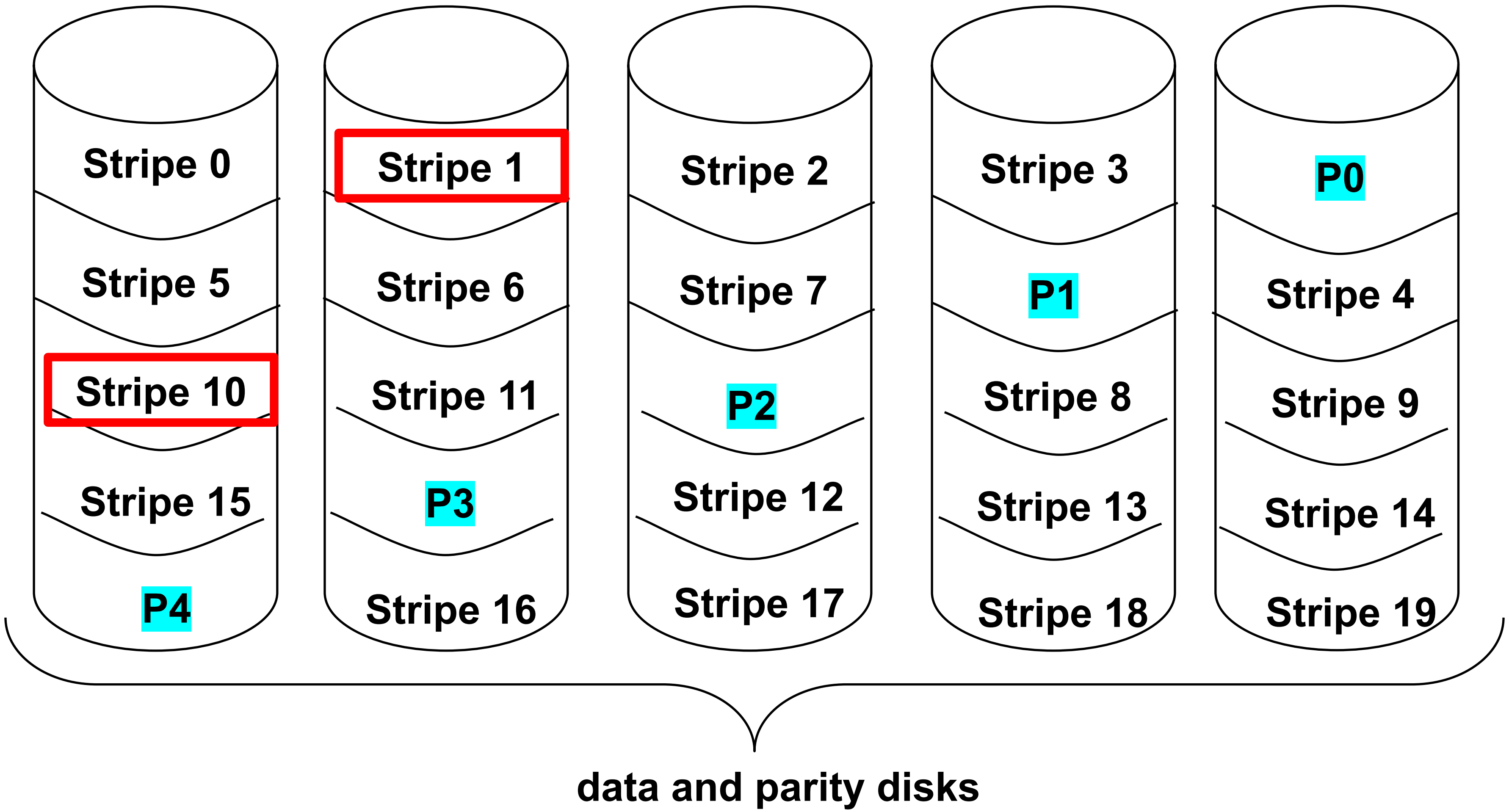
N := number of disks
C := capacity of 1 disk
D := latency of one small I/O operation

S := sequential throughput of 1 disk
R := random throughput of 1 disk

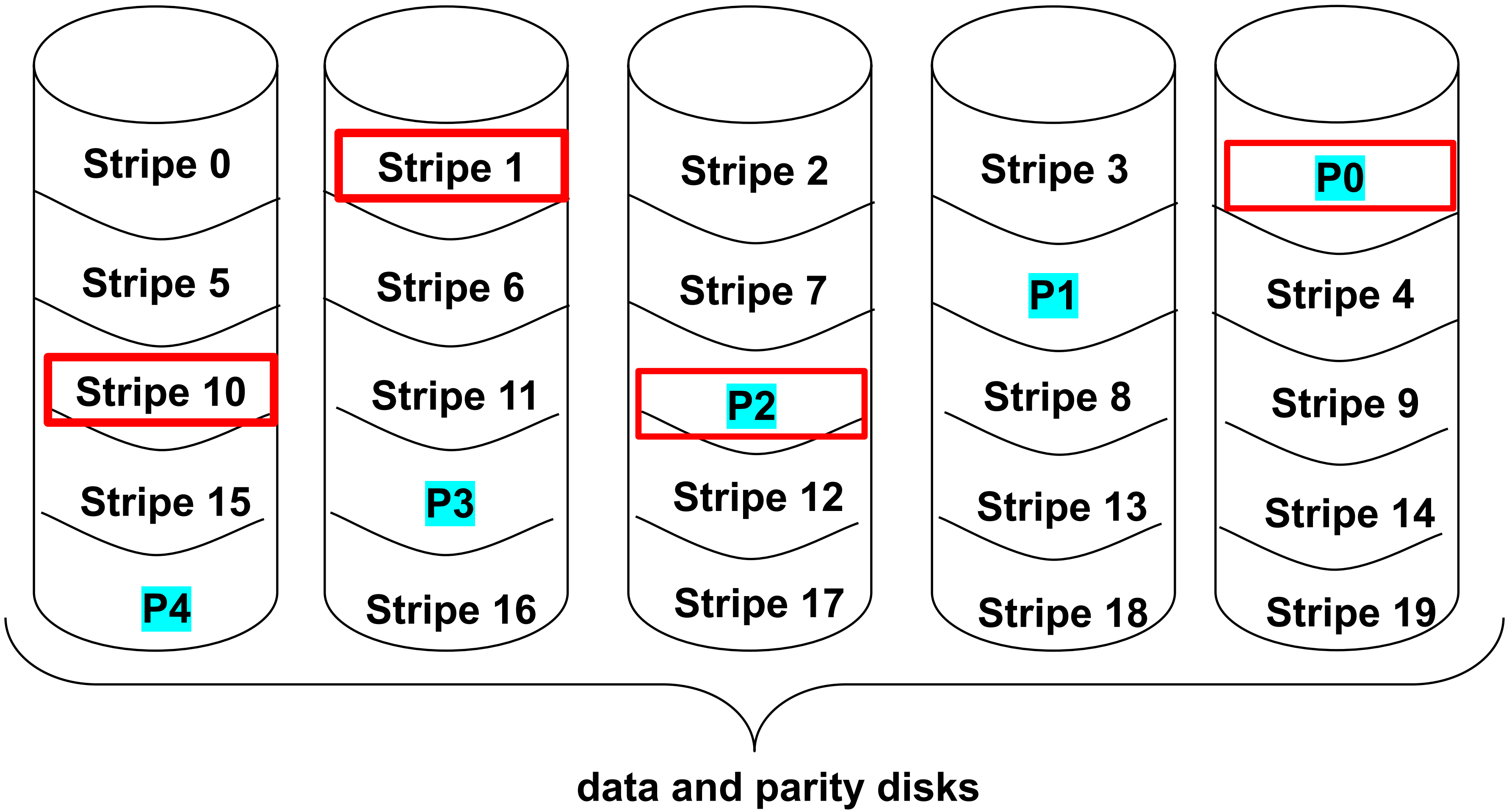
RAID-5 Random Writes



RAID-5 Random Writes



RAID-5 Random Writes



RAID-5 Random Writes: N/4 * R

- Why the factor of 4?
  - Assume subtractive parity

- Each write to RAID-5 needs 2 parallel reads and 2 parallel writes
  - Total: 4 random I/Os
- Assuming load is uniformly distributed across the disks & each operation takes 4 I/Os
  - The load on each disk for a write is multiplied by 4.

Let’s practice! (From Winter22 Final)

For a RAID system with a total of 5 disks, answer the following

for RAID-0 and RAID-5.

A. How much usable storage does the system have, if each individual disk has 2 GBytes of storage? Explain your answer.

B. For a workload consisting only of reads of a single stripe, evenly distributed, what is the throughput in stripe reads per second, assuming a single disk does 100 stripe reads per second?

C. For a workload consisting only of writes of a single stripe, evenly distributed, what is the throughput in stripe writes per second, assuming a single disk does 100 stripe writes per second?

RAID-0

A. How much usable storage does the system have, if each individual disk has 2 GBytes of storage? Explain your answer.

B. For a workload consisting only of reads of a single stripe, evenly distributed, what is the throughput in stripe reads per second, assuming a single disk does 100 stripe reads per second?

C. For a workload consisting only of writes of a single stripe, evenly distributed, what is the throughput in stripe writes per second, assuming a single disk does 100 stripe writes per second?

RAID-0

How much usable storage does the system have, if each individual disk has 2 GBytes of storage? Explain your answer.

2 * 5 = 10 GB

For a workload consisting only of reads of a single stripe, evenly distributed, what is the throughput in stripe reads per second, assuming a single disk does 100 stripe reads per second?

100 stripes/s * 5 = 500 stripes/s

For a workload consisting only of writes of a single stripe, evenly distributed, what is the throughput in stripe writes per second, assuming a single disk does 100 stripe writes per second?

100 stripes/s * 5 = 500 stripes/s

RAID-5

A. How much usable storage does the system have, if each individual disk has 2 GBytes of storage? Explain your answer.

B. For a workload consisting only of reads of a single stripe, evenly distributed, what is the throughput in stripe reads per second, assuming a single disk does 100 stripe reads per second?

C. For a workload consisting only of writes of a single stripe, evenly distributed, what is the throughput in stripe writes per second, assuming a single disk does 100 stripe writes per second?

RAID-5

How much usable storage does the system have, if each individual disk has 2 GBytes of storage? Explain your answer.

2 * 4 = 8 GB

For a workload consisting only of reads of a single stripe, evenly distributed, what is the throughput in stripe reads per second, assuming a single disk does 100 stripe reads per second?

5 * 100 stripes/s = 500 stripes/s

For a workload consisting only of writes of a single stripe, evenly distributed, what is the throughput in stripe writes per second, assuming a single disk does 100 stripe writes per second?

5 * 100/4 stripes/s = 125 stripes/s

Summary: RAID

- Disk bandwidth not improving very fast
- Disk size and cost improving fast
- Improve disk bandwidth
  - By parallel I/O
- Also improves reliability
  - Higher levels survive disk failures

RAID - Key Concepts

- Mirroring
- Striping
- Parity

Further Reading

Operating Systems: Three Easy Pieces by R. & A. Arpaci-Dusseau

Chapters 38, 43.

https://pages.cs.wisc.edu/~remzi/OSTEP/

Credits:

Some slides adapted from the OS courses of Profs. Remzi and Andrea Arpaci-Dusseau (University of Wisconsin-Madison), Prof. Willy Zwaenepoel (University of Sydney), and Prof. Youjip Won (Hanyang University).