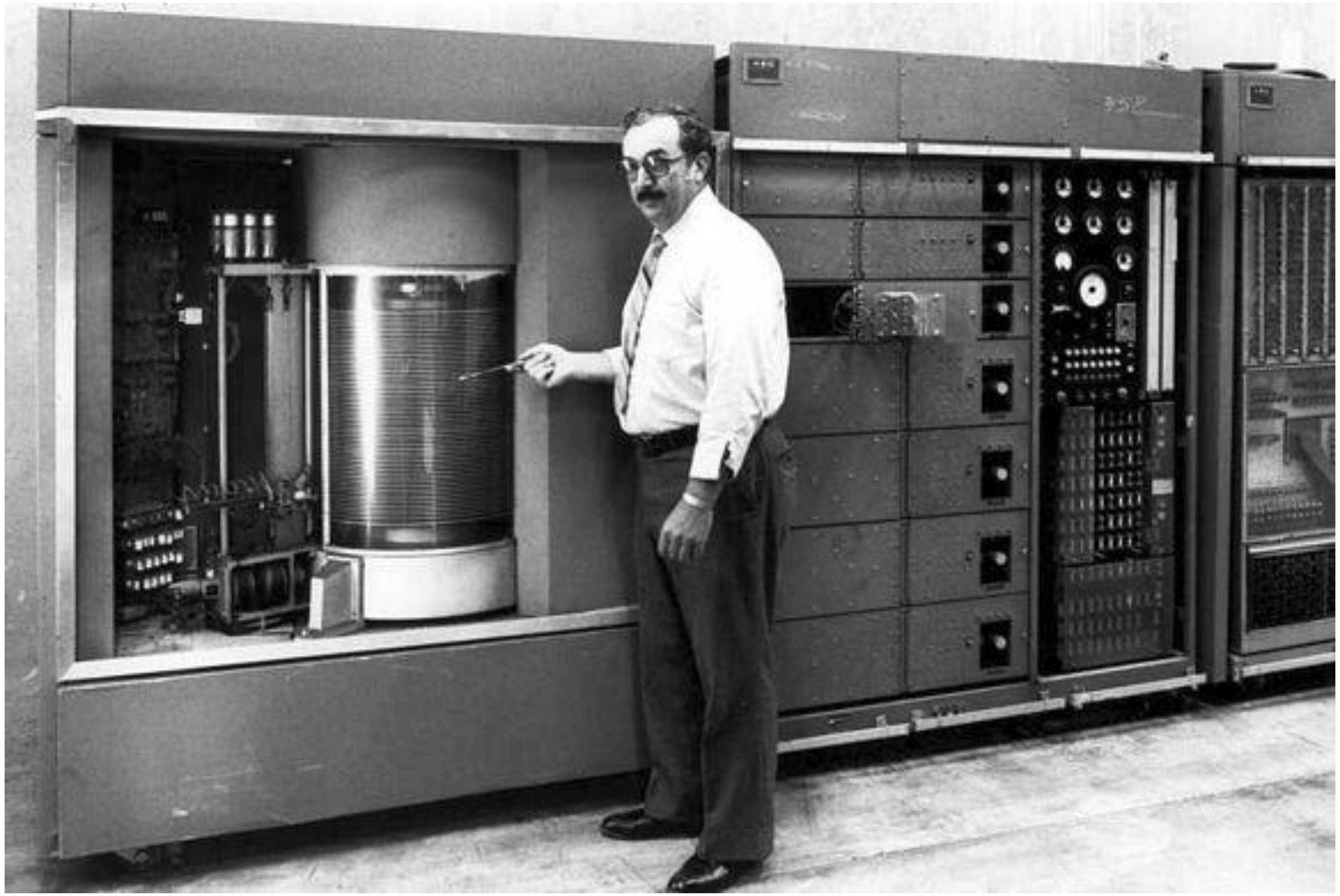
Redundant Array of Inexpensive Disks (RAID)

Modern Operating Systems, by Andrew Tanenbaum Chap. 5.4

Operating Systems: Three Easy Pieces (a.k.a. the OSTEP book) Chap. 38

IBM Model 350 disk storage system

- Introduced in 1956
- 5M (7 bit) characters
- 50 x 24" platters
- Access time: < 1 sec.!

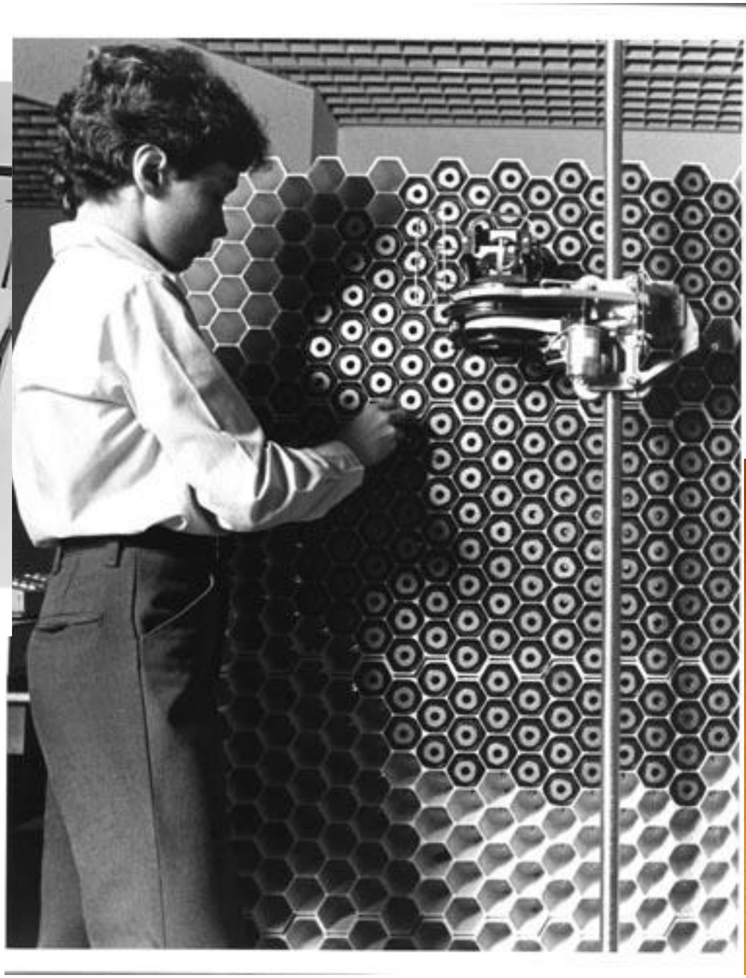


IBM "Winchester" (3340) Disk Drives

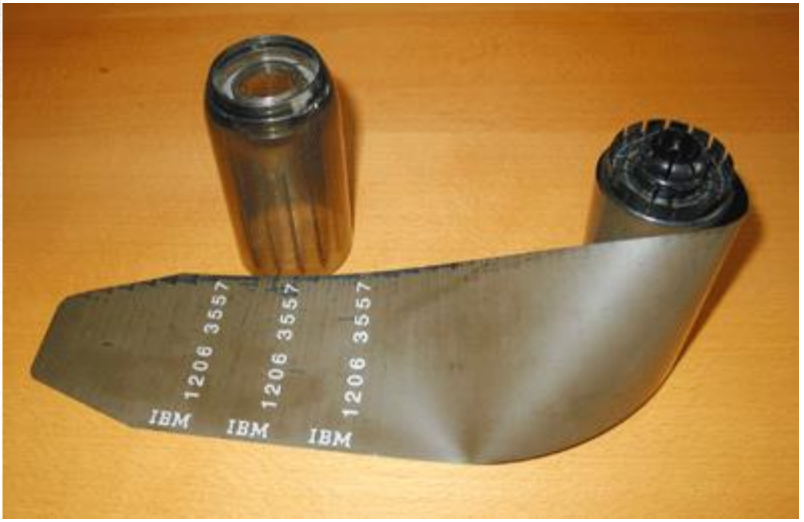
- Two 30M replaceable drives
- "Single Large Expensive Disk" (SLED)

Mainframe "Raised Floor"

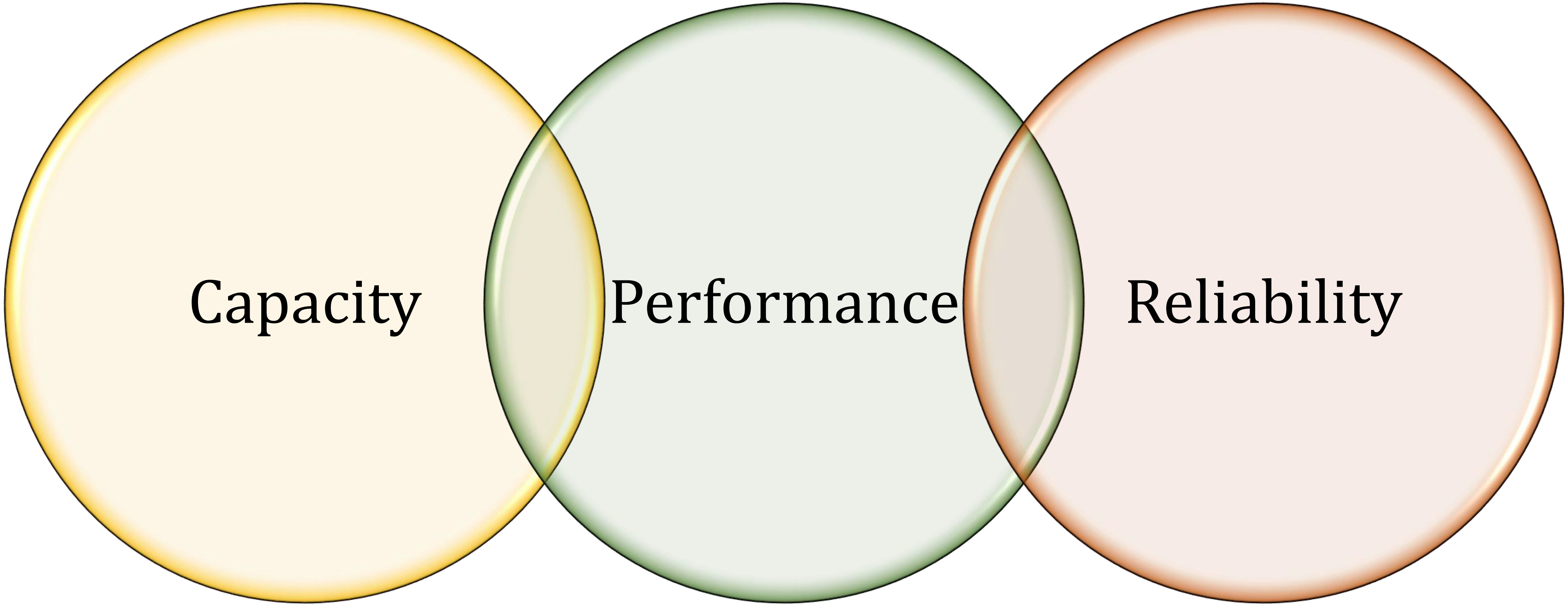
IBM 3850 "Mass Storage Device"



- Almost 10K cartridges
- On request, loaded onto a hard disk drive



Disk Drive Trade-Offs

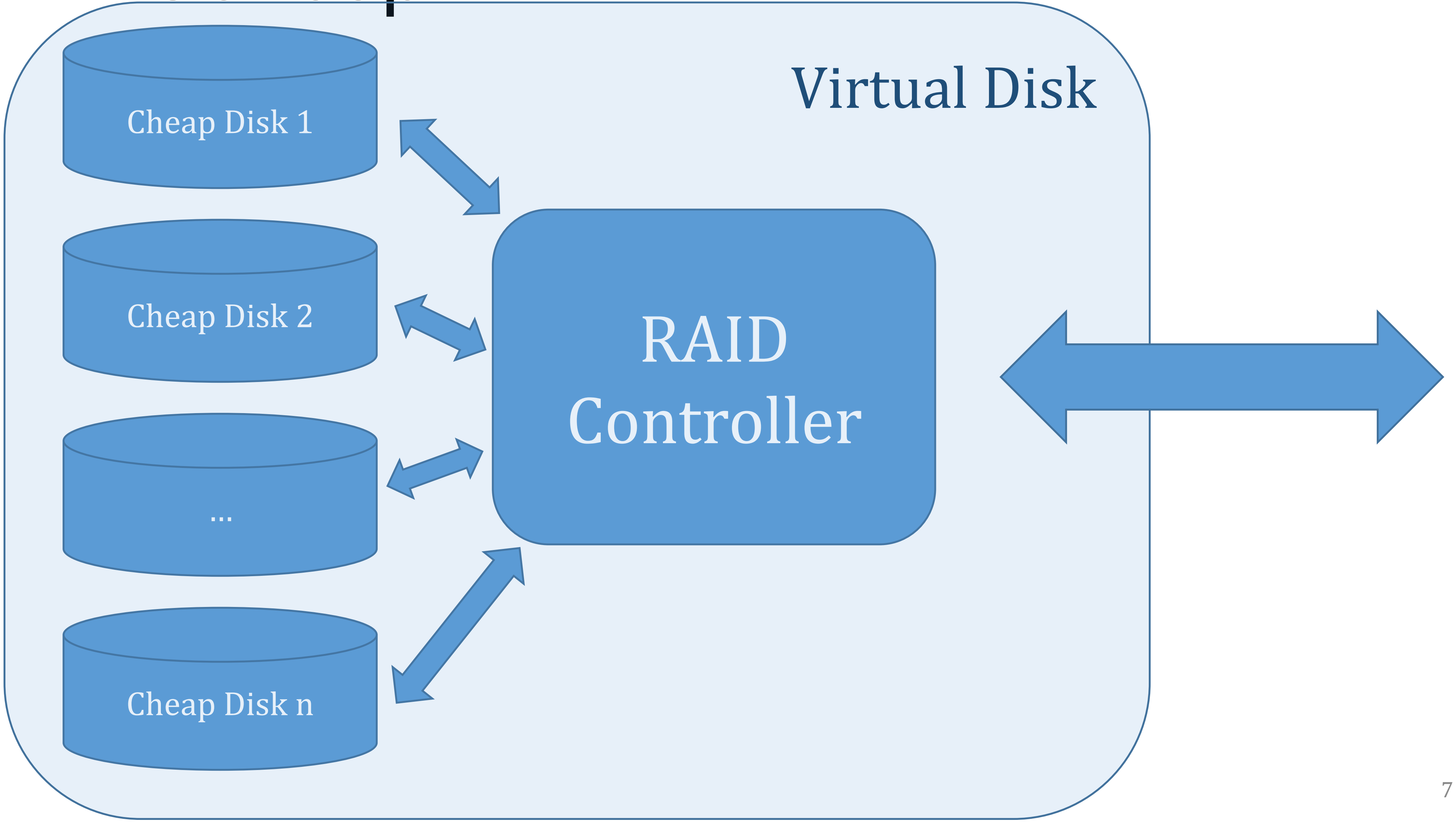


Capacity

Performance

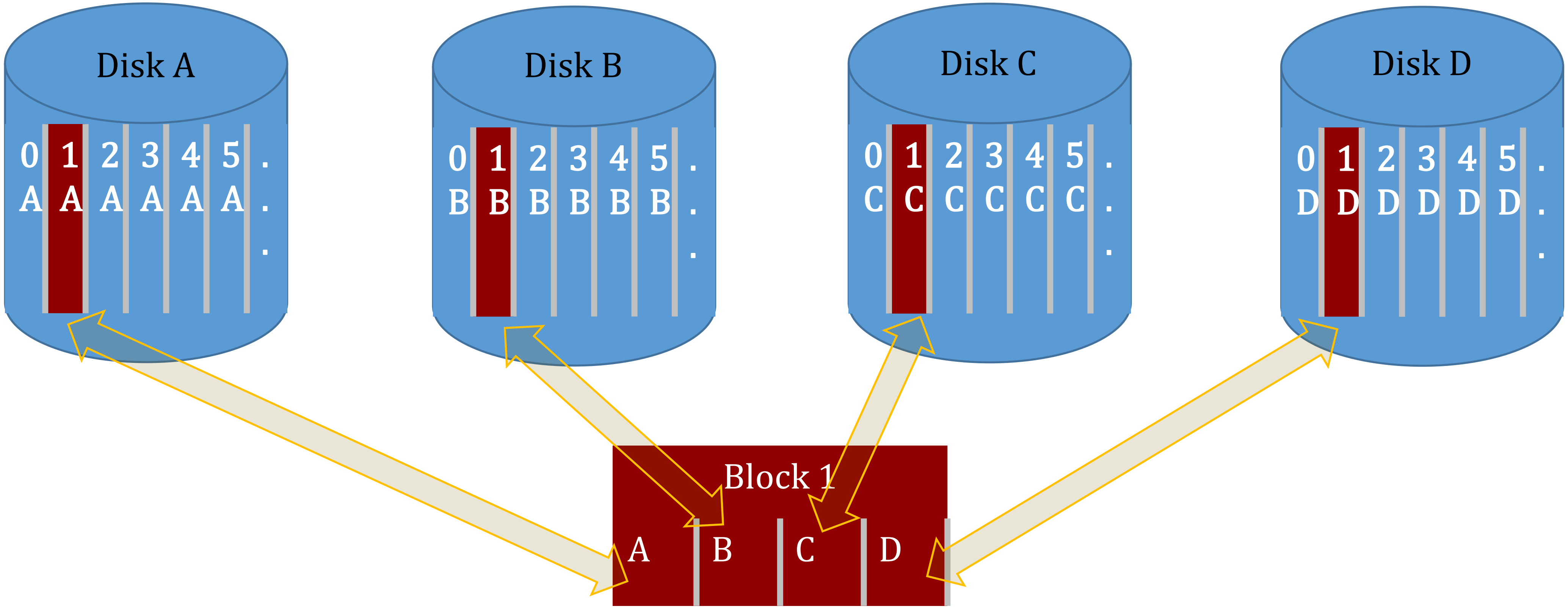
Reliability

RAID Concept



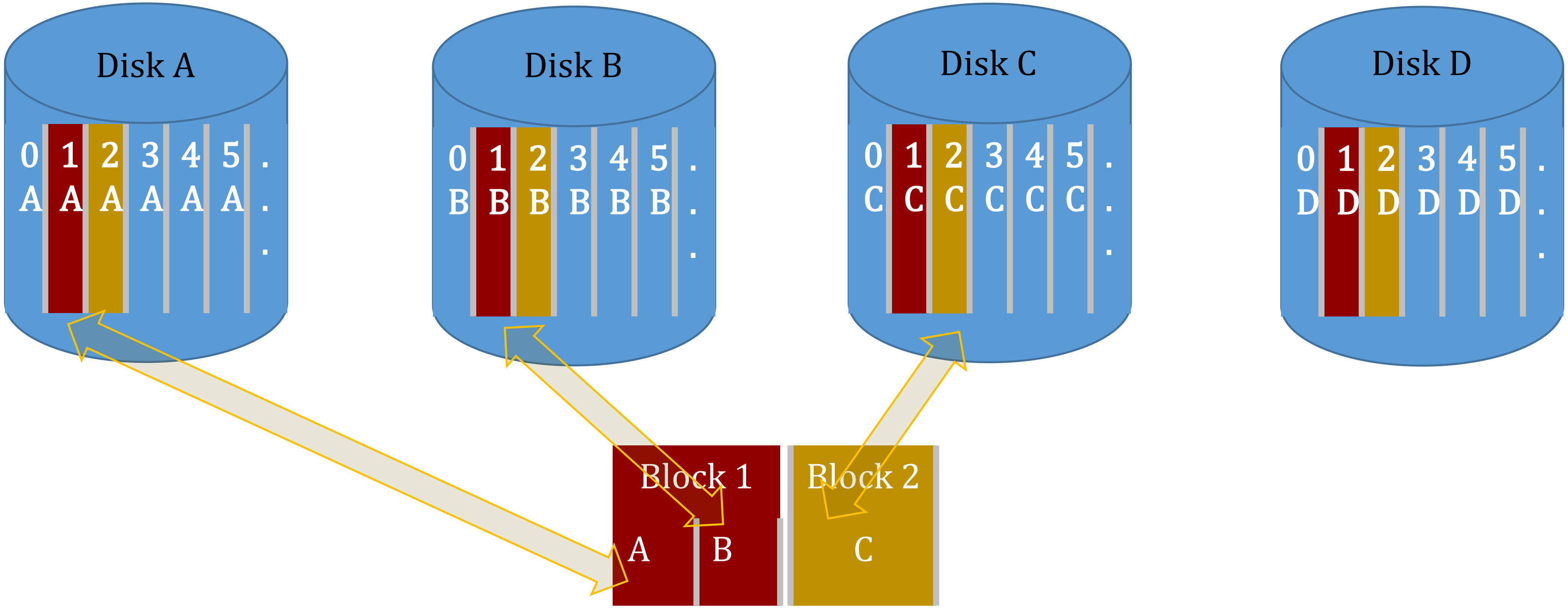
RAID level 0

- "Striping" – spread a large "block" of data over multiple drives



RAID level 0 – Overlapped Requests

- For smaller blocks, can overlap requests

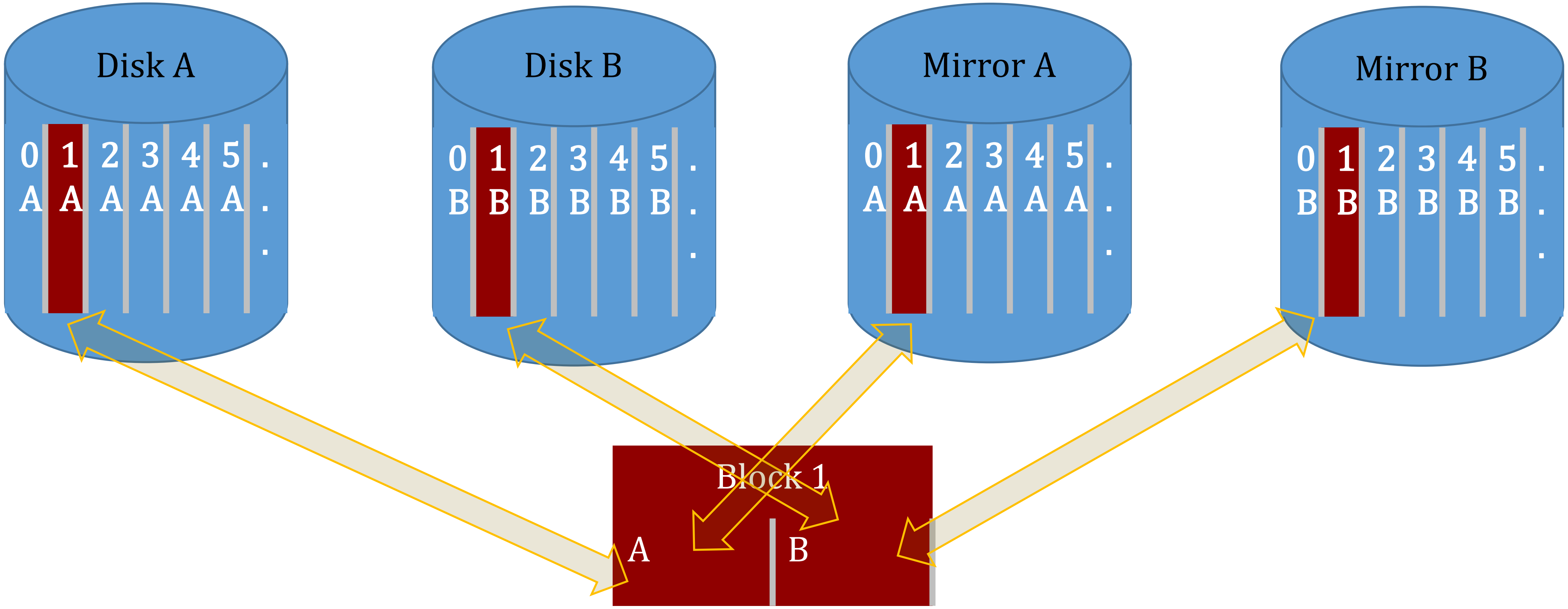


RAID level 0 Trade-Offs

Performance	NX speed improvement with parallel read for NX size blocks! Assuming N RAID disk drives If R/W block is not multiple of disk block size, not realized
Capacity	Increased because of multiple disks All disk space is used for data – full utilization!
Reliability	NX decrease! Measure reliability as "Mean Time to Failure" (MTF) If MTF of one disk is 30,000 hours, then MTF of N disks is $30{,}000/N$

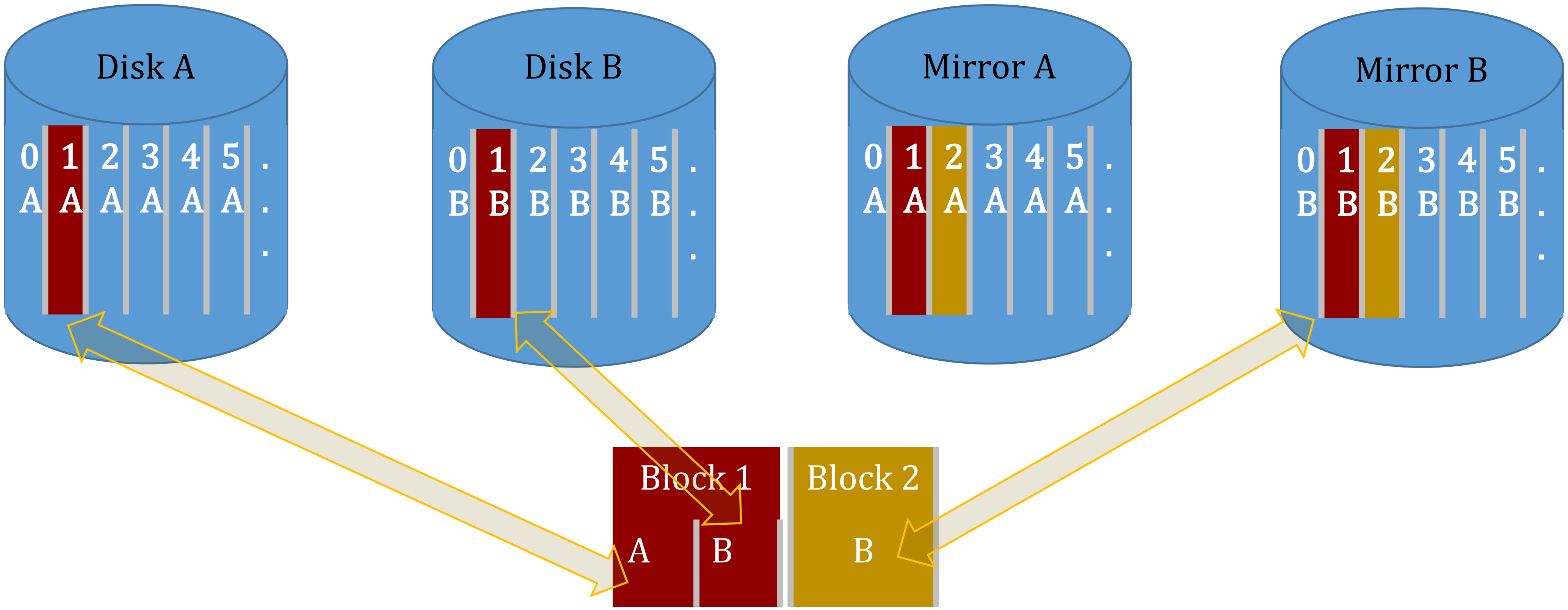
RAID level 1

- "Mirroring" – Two copies of each disk block



RAID level 1 – Overlapped Requests

- "Mirroring" – Two copies of each disk block

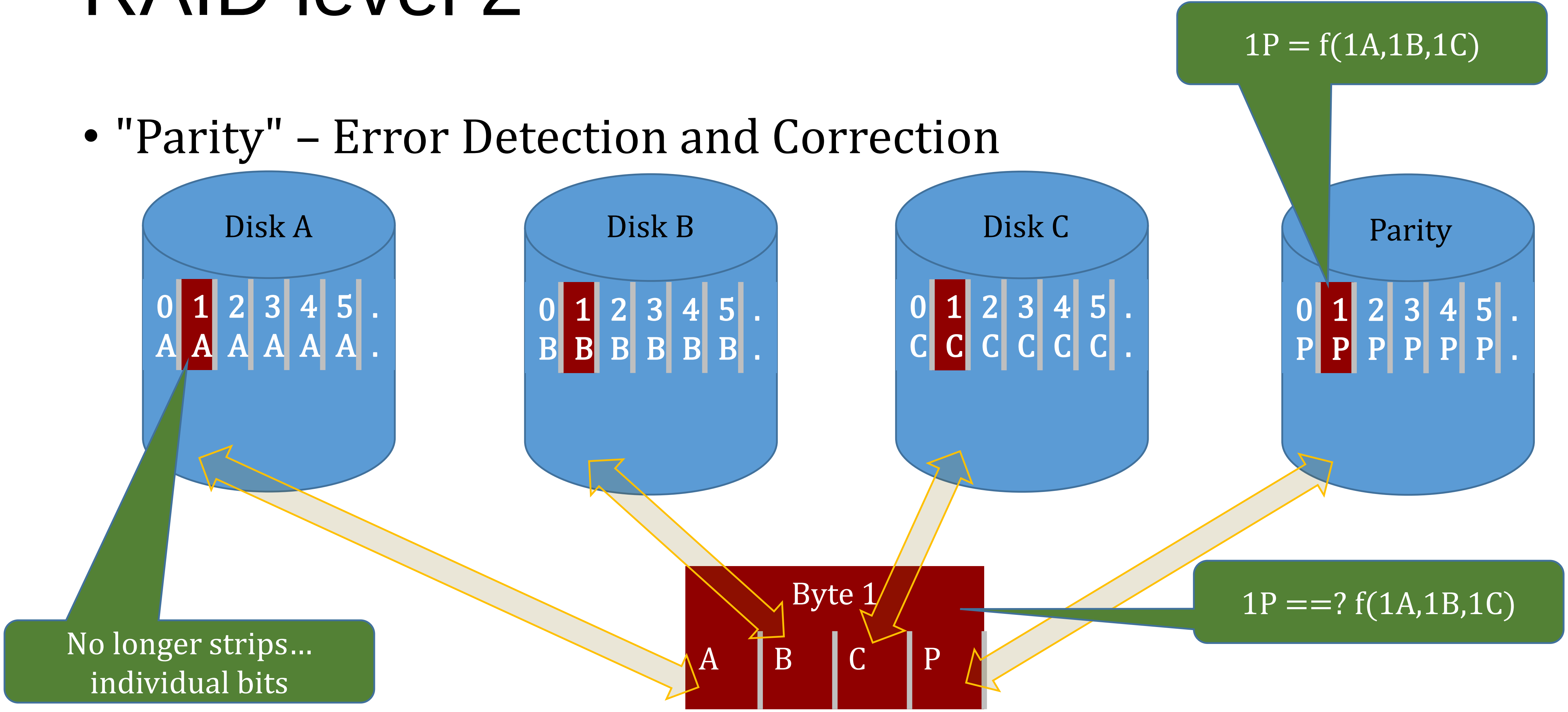


RAID level 1 Trade-Offs

Performance	$N/2$ X speed improvement with parallel read for $N/2$ X size blocks! If R/W block is not multiple of disk block size, not realized But, overlaps with Mirror disks allows more speed
Capacity	Increased because of multiple disks But half the capacity of level 0
Reliability	Increased! (MTF more than squared) If MTF of one disk is 30,000 hours, then MTF of $N/2$ disks is $60{,}000/N$ But on failure, copy from Mirror! Need 2 disks to fail simultaneously to lose data

RAID level 2

- "Parity" – Error Detection and Correction



1P = f(1A,1B,1C)

1P ==? f(1A,1B,1C)

No longer strips… individual bits

Error Detection

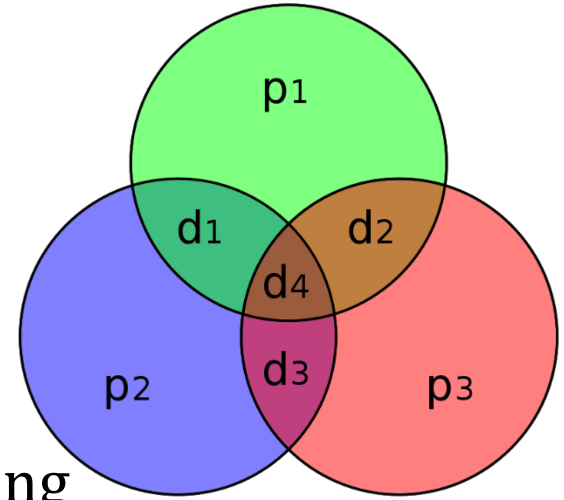
- Error Detection
  - Simplest scheme is parity – even or odd number of 1 bits
  - Also hash functions like checksum, or cyclic redundancy check (CRC)
  - If $1P_W$!= $1P_R$ then an error occurred… at least one bit is wrong!

A	B	XOR(A,B)	Parity
0	0	0	Even
0	1	1	Odd
1	0	1	Odd
1	1	0	Even

- Parity
  - Overhead depends on number of bits XORed
  - Can detect all single-bit errors
  - Cannot detect many multi-bit errors!

Error Correction

- Simplest is Hamming(7,4)
  - 4 data bits: d_1,d_2,d_3,d_4 and 3 parity bits: p_1,p_2,p_3
  - p_1=XOR(d_1,d_2,d_4); p_2=XOR(d_1,d_3,d_4); p_3=XOR(d_2,d_3,d_4)
  - Compare parity bits to determine WHICH data bit is wrong
  - Can correct all single bit errors
  - Can detect all two bit errors
  - Parity bits are almost as big as data bits!
- Simple RAID level 2 has 4 data & 3 parity drives
  - Losing 1 data drive doesn't stop reads!



p_1?	p_2?	p_3?	Error
0	0	0	d_4
0	0	1	d_1
0	1	0	d_2
1	0	0	d_3
… multi-bit errors …			
1	1	1	∅

RAID level 2 Trade-Offs

Performance	(*N-P*) X speed improvement for *N* disk drives with *P* parity bits Better than level 1, but not as good as a perfect level 0 No overlapped reads/writes Requires synchronized disk reads at a bit level!
Capacity	Increased because of multiple disks Less, *(N-P)/N,* than the capacity of level 0
Reliability	Increased! Depends on sophistication of error detection or correction Error detection increases reliability but doesn't help MTF at least you KNOW there's a problem Error correction increases MTF

Most modern disk drives have built-in error detection/correction, so RAID level 2 is rarely used anymore.

XOR features

A	C	XOR(A,C)
0	0	0=A
1	0	1=A
0	1	1=!A
1	1	0=!A

- XOR with 0 does not change value
- XOR with 1 inverts value

- The result of XOR identifies where A and B are different

A	1100
B	1010
R	0110

Where are A and B different?

A	1100
R	0110
B	1010

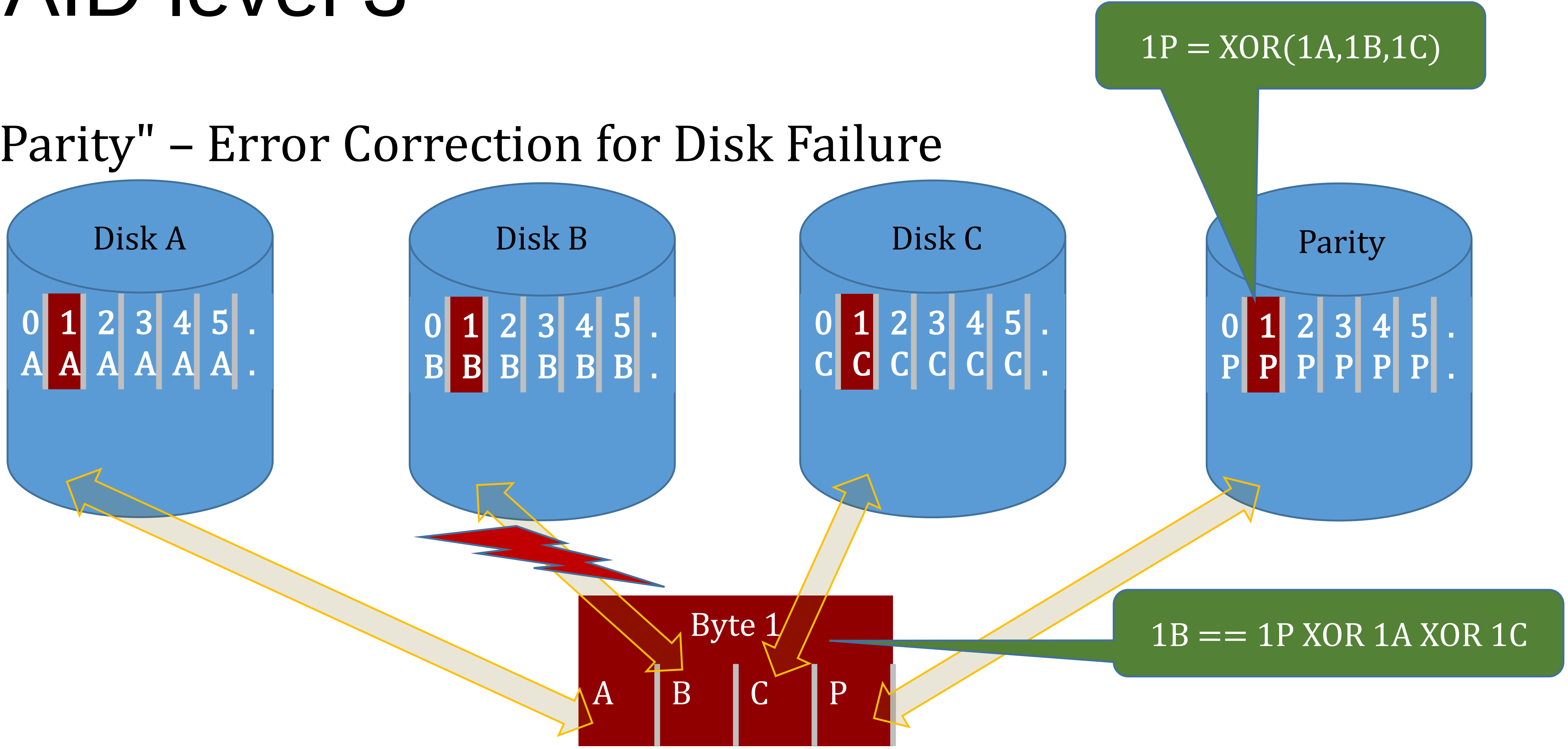
Flip the bits in A where A is different from B

B	1010
R	0110
A	1100

Flip the bits in B where A is different from B

RAID level 3

- "Parity" – Error Correction for Disk Failure

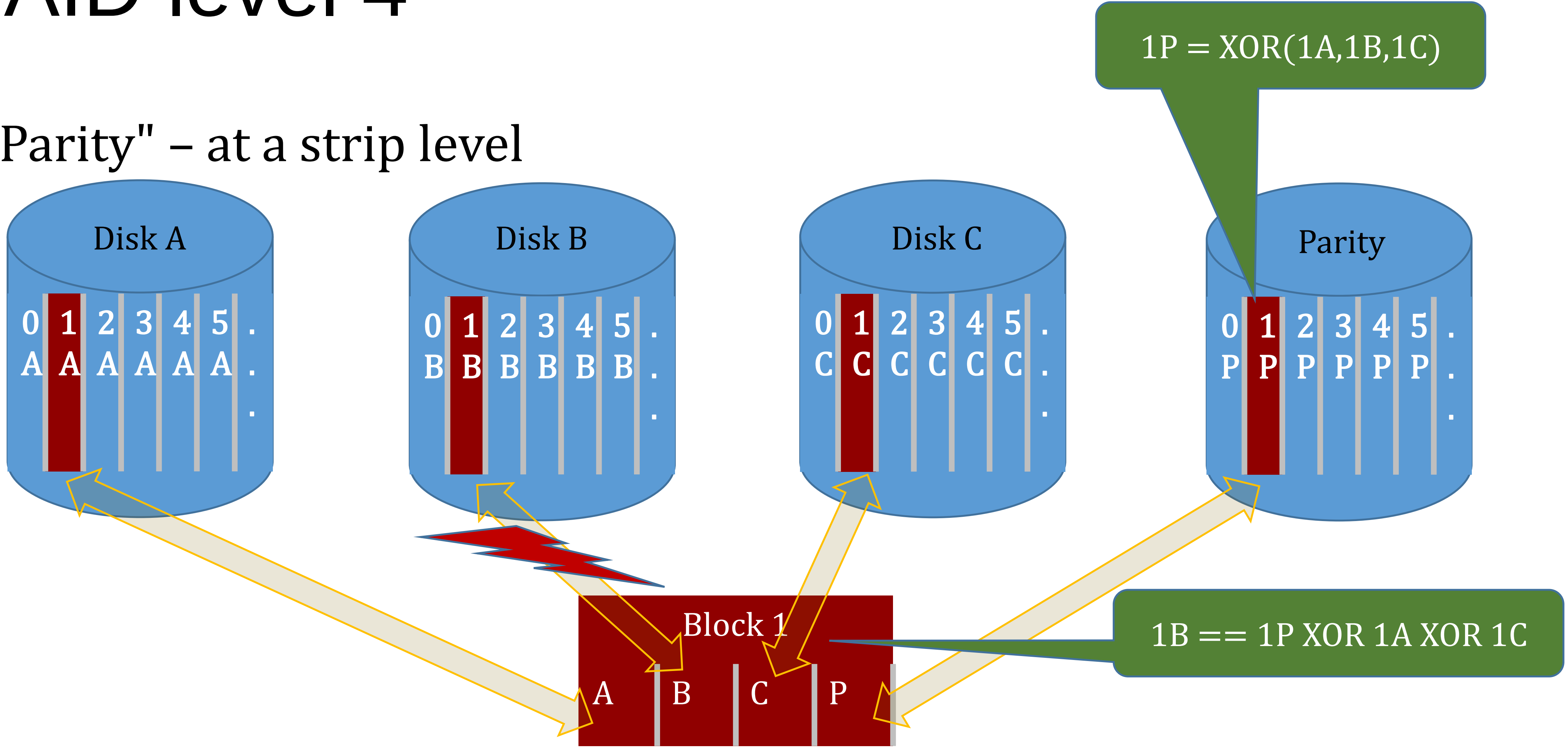


RAID level 3 Trade-Offs

Performance	Better than level 2 because only 1 parity drive required But, still has all the level 2 restrictions synchronized read, no overlapped reads/writes
Capacity	Better than level 2 because only 1 parity drive required
Reliability	Increased! Can live with a single disk drive failure Assumes disk drive failure identifies itself No help with random bit level errors (but drive may do that)

RAID level 4

- "Parity" – at a strip level



Small Update Problem

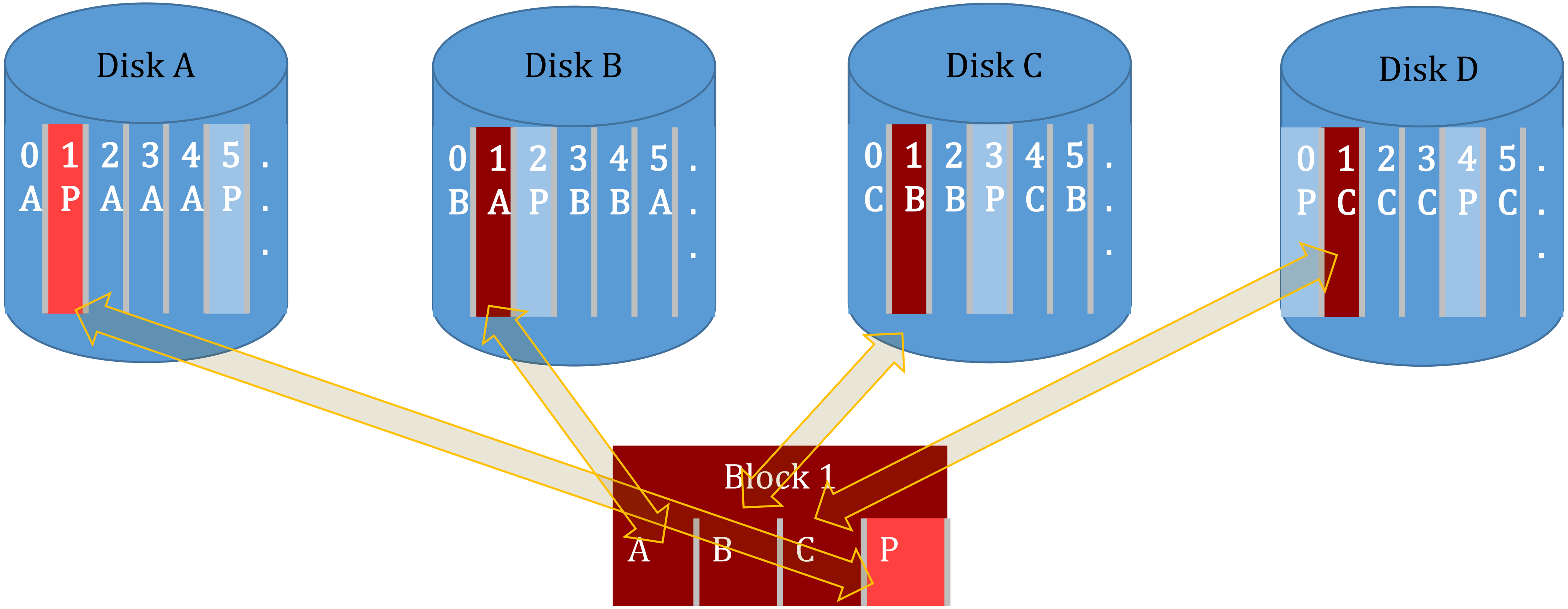
- Suppose you update 1 slot in Block 1; Slot 1C …
  - Compute $1P_{new}$ = XOR($1A_{old}$, $1B_{old}$, $1C_{new}$)
  - Needs 2 Reads to get $1A_{old}$ and $1B_{old}$ and 2 writes to write $1C_{new}$ and $1P_{new}$
  - Reads and Writes may occur in parallel
  - However, Reads and Writes prevent overlapped requests
    - Read must read N-2 drives to recover old data
    - Write must write to parity drive.. no other parity read or write allowed
- Every small update requires 1 parallel read and 1 parallel write that prevents overlapped requests

RAID level 4 Trade-Offs

Performance	Similar to level 0 for (*N*-1) strip reads/writes Almost level 0 Synchronized reads no longer required Contention for parity disk drive prevents overlapped R/W
Capacity	Same as level 2
Reliability	Same as level 3

RAID level 5

- Distribute Parity Strip over all disks



Small Update – Distributed Parity

- Suppose you update 1 slot in Block 1; Slot 1C …
  - Compute $1P_{new} = XOR(1P_{old}, 1C_{old}, 1C_{new})$
  - Needs 2 Reads to get $1P_{old}$ and $1C_{old}$ and 2 writes to write $1C_{new}$ and $1P_{new}$
  - Reads and Writes may occur in parallel
  - However, Reads and Writes no longer prevent overlapped requests
    - Read must read only 2 drives to recover old data
    - No longer a single parity drive – Overlapped parity slot on different disk

- Every small update requires 1 parallel read and 1 parallel write but no longer prevents overlapped requests

RAID level 5 Trade-Offs

Performance	Same as level 4 except… Overlapped Reads/Writes allowed
Capacity	Same as level 2
Reliability	Same as level 3

Note: There is a RAID level 6 which uses 2 parity drives to increase reliability, but no new concepts.

RAID level comparison

		RAID 0	RAID 1	RAID 2	RAID 3	RAID 4	RAID 5
Performance	Parallel Read	N	N	N-P	N-1	N-1	N-1
	Parallel Write	N	N/2	N-P	N-1	1 or (N-1)*	(N-1)/2
	Synced Drives	no	no	yes	yes	no	no
Capacity	Overhead	0	N/2	P	1	1	1
Reliability	Fault Tolerance	None	1-disk (some 2)	1-disk 2-disk det.	1-disk	1-disk	1-disk

Conclusions

- Original purpose: Take advantage of commodity drives
  - smaller and cheaper than conventional disk drives
  - Nobody does this anymore – very large disks are very cheap now
- Today: Improve performance and reliability
  - Fault tolerant storage, No backup required, High throughput
- RAID: Good solution for small installations
  - Cheap drives & controllers
  - Prefer RAID level 3 for simplicity, level 5 for parallelism
  - Add Non-Volatile RAM to improve write performance