Storage Systems (3)

Dr. Jun Zheng

CSE325 Principles of Operating Systems

12/2/2019

NEW MEXICO TECH
SCIENCE · ENGINEERING · RESEARCH · UNIVERSITY

Availability

- *Module availability* measures service as alternate between the 2 states of accomplishment and interruption (number between 0 and 1, e.g. 0.9)
- *Mean Time To Repair* (*MTTR*) measures Service Interruption
- *Mean Time Between Failures* (*MTBF*) = MTTF+MTTR
- *Module availability = MTTF / (MTTF + MTTR)*

Dependability of Disk Arrays

- with many more devices, dependability decreases: N devices generally have 1/Nth of the reliability of a single device.
  - Reliability metric: MTTF
  - 50,000 Hours ÷ 70 disks = 700 hours
  - Disk system MTTF: Drops from 6 years to 1 month!
- Result: disk array have many more faults than a small number of large disks

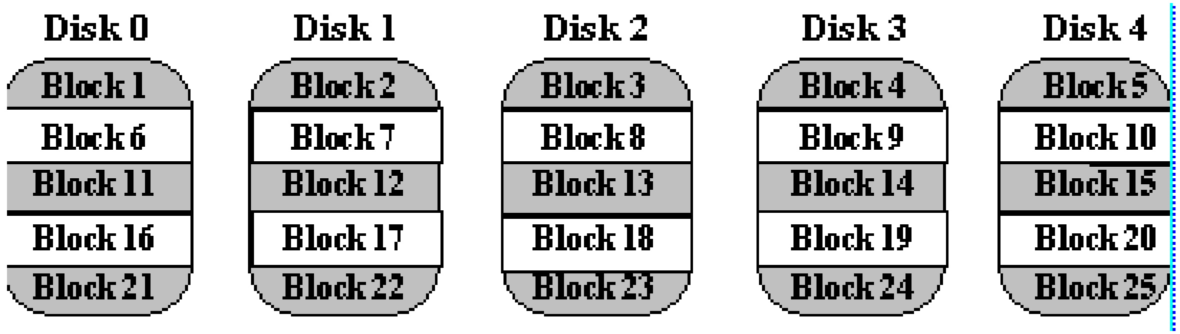
NEW MEXICO TECH
SCIENCE · ENGINEERING · RESEARCH · UNIVERSITY

RAID

- Add redundant disks to tolerate faults:
  - Dependability increases
  - If a single disk fails: the lost information is reconstructed from the redundant information
- **RAID: redundant array of inexpensive disks**
  - Spread the data over multiple disks: striping
  - If second disk fail while the first one is being repaired, cannot recover
  - Not a problem: MTTF of a disk is tens of years, while MTTR is hours -> redundancy makes the measured reliability of 100 disks much higher than that of a single disk
- Different RAID levels: 0 - 6

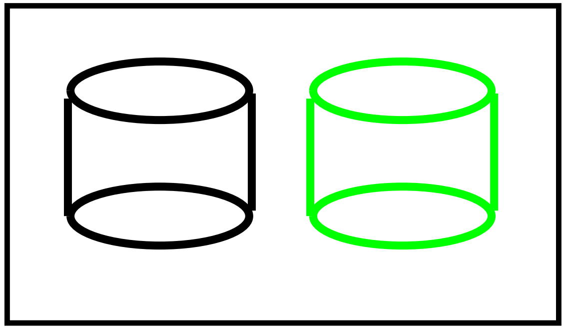
RAID 0 – No Redundancy

- Data are striped but there is no redundancy to tolerate disk failure
  - Data is divided into blocks and is spread in a fixed order among all the disks in the array.
- Improves the performance for large access because many disks operate in parallel
- No space overhead, no fault tolerance



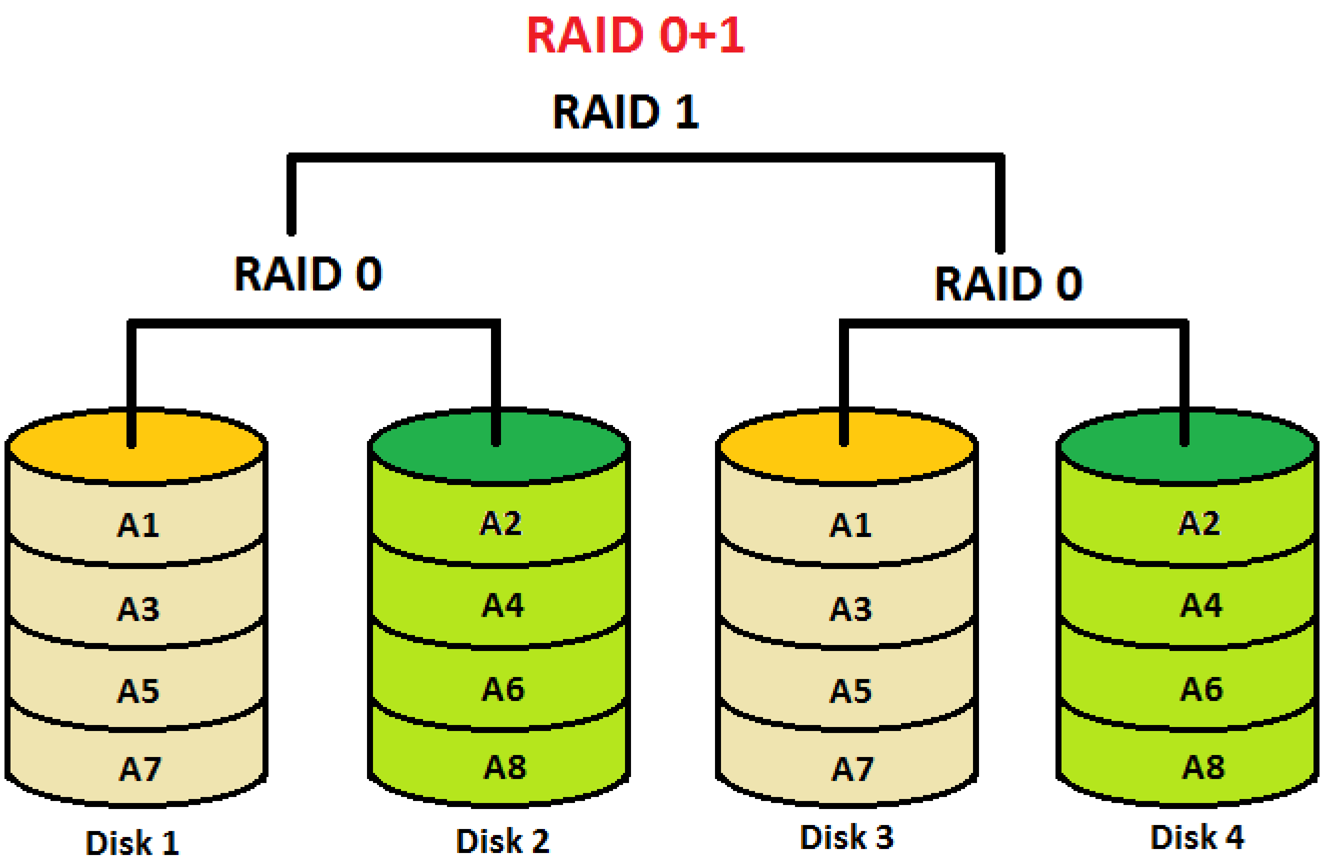
NEW MEXICO TECH
SCIENCE · ENGINEERING · RESEARCH · UNIVERSITY

RAID 1: Disk Mirroring/Shadowing



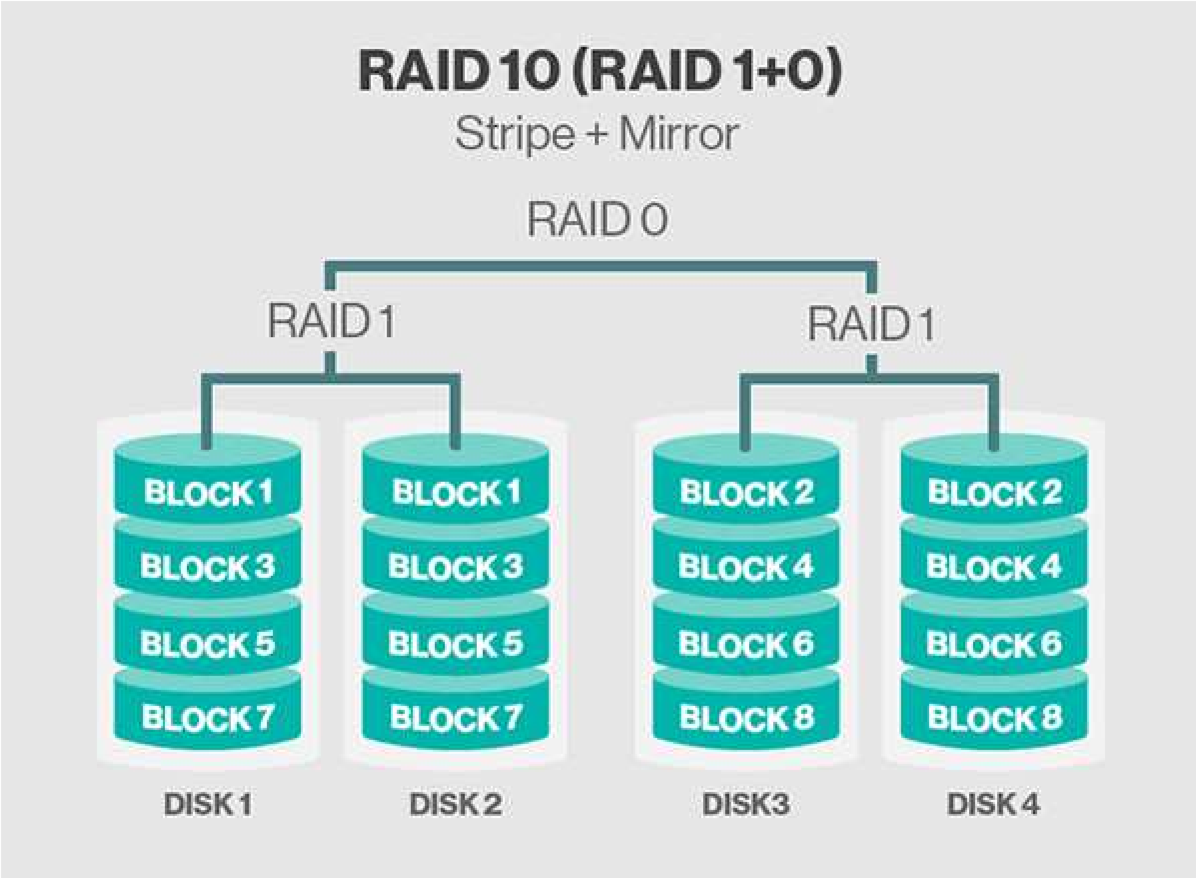
- The disk is fully duplicated onto its "mirror" (or more)
  - Very high availability can be achieved
- Bandwidth sacrifice on write:
  - Logical write = two physical writes
  - Reads may be optimized
- Most expensive solution: 100% capacity overhead
- (RAID 2 no longer used, so skip)

RAID 0+1



NEW MEXICO TECH
SCIENCE · ENGINEERING · RESEARCH · UNIVERSITY

RAID 1+0

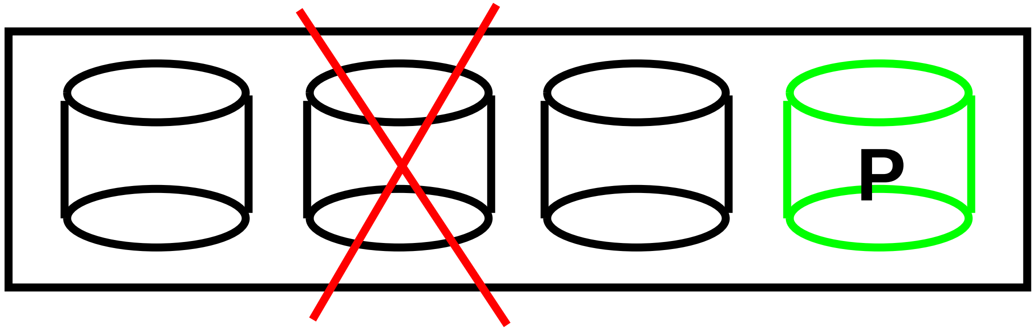


RAID 10 (RAID 1+0)
Stripe + Mirror
RAID 0
RAID 1
RAID 1
BLOCK 1
BLOCK 3
BLOCK 5
BLOCK 7
BLOCK 1
BLOCK 3
BLOCK 5
BLOCK 7
BLOCK 2
BLOCK 4
BLOCK 6
BLOCK 8
BLOCK 2
BLOCK 4
BLOCK 6
BLOCK 8
DISK 1
DISK 2
DISK3
DISK 4

RAID 3 – Parity Disk (Bit-Interleaved)

10010011
11001101
10010011
. . .

logical record



Striped physical records

1	1	1	1
0	1	0	1
1	0	1	0
0	0	0	0
0	1	0	1
0	1	0	1
1	0	1	0
1	1	1	1

P contains sum of other disks per stripe mod 2 ("parity")
If disk fails, subtract P from sum of other disks to find missing information

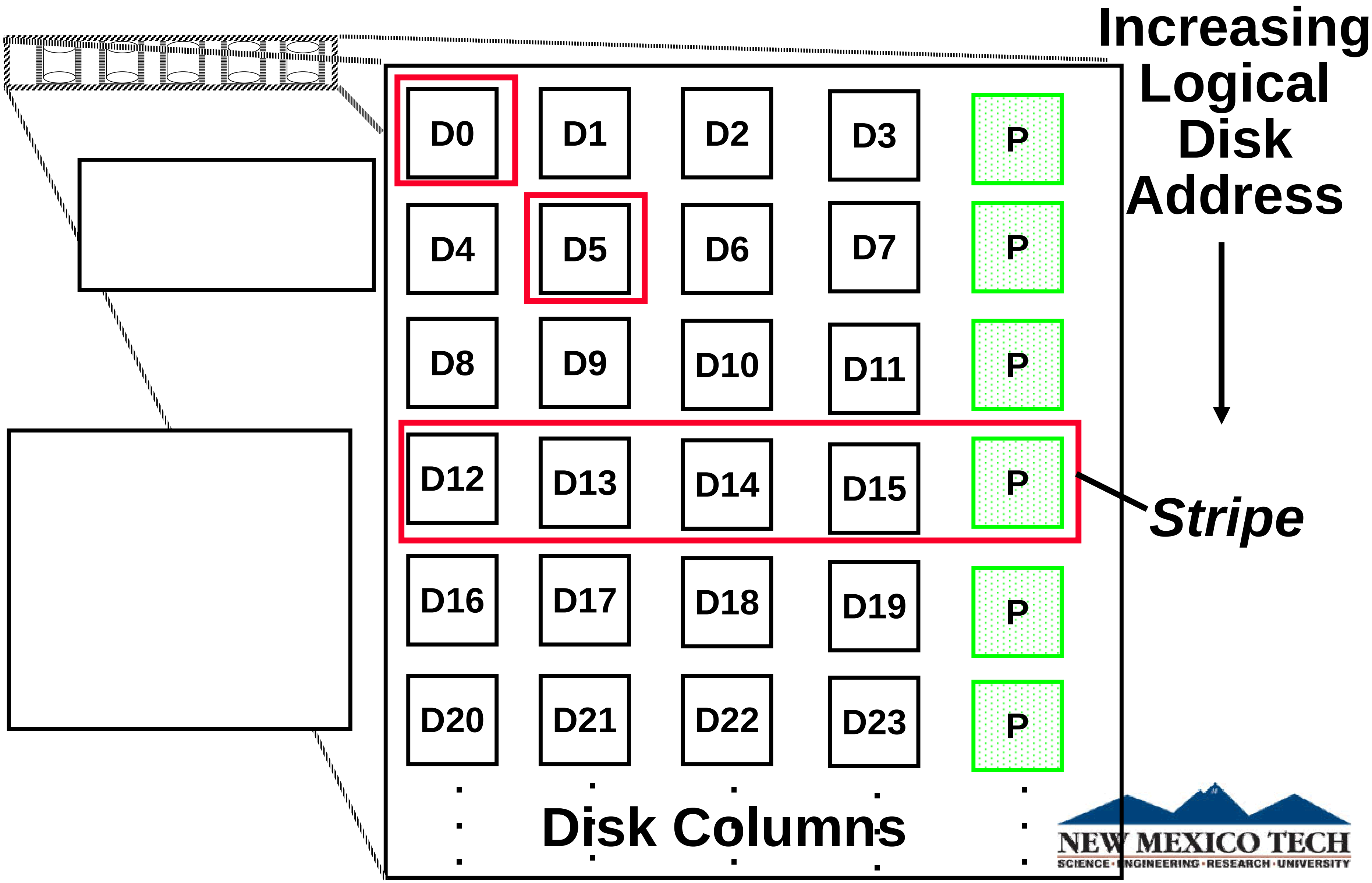
RAID 3 – Parity Disk

- Sum computed across recovery group to protect against hard disk failures, stored in P disk
- Logically, a single high capacity, high transfer rate disk: good for large transfers
- Wider arrays reduce capacity costs
  - 33% capacity cost for parity if 3 data disks and 1 parity disk

Inspiration for RAID 4 (Block-interleaved)

- RAID 3 relies on parity disk to discover errors on Read
- But every sector has an error detection field
- To catch errors on read, rely on error detection field vs. the parity disk
- Allows independent reads to different disks simultaneously

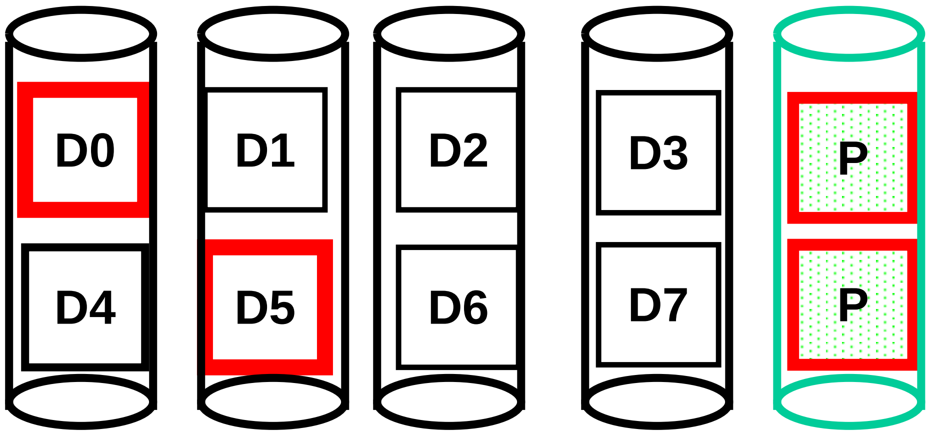
RAID 4 – High I/O Rate Parity



NEW MEXICO TECH
SCIENCE · ENGINEERING · RESEARCH · UNIVERSITY

Inspiration for RAID 5

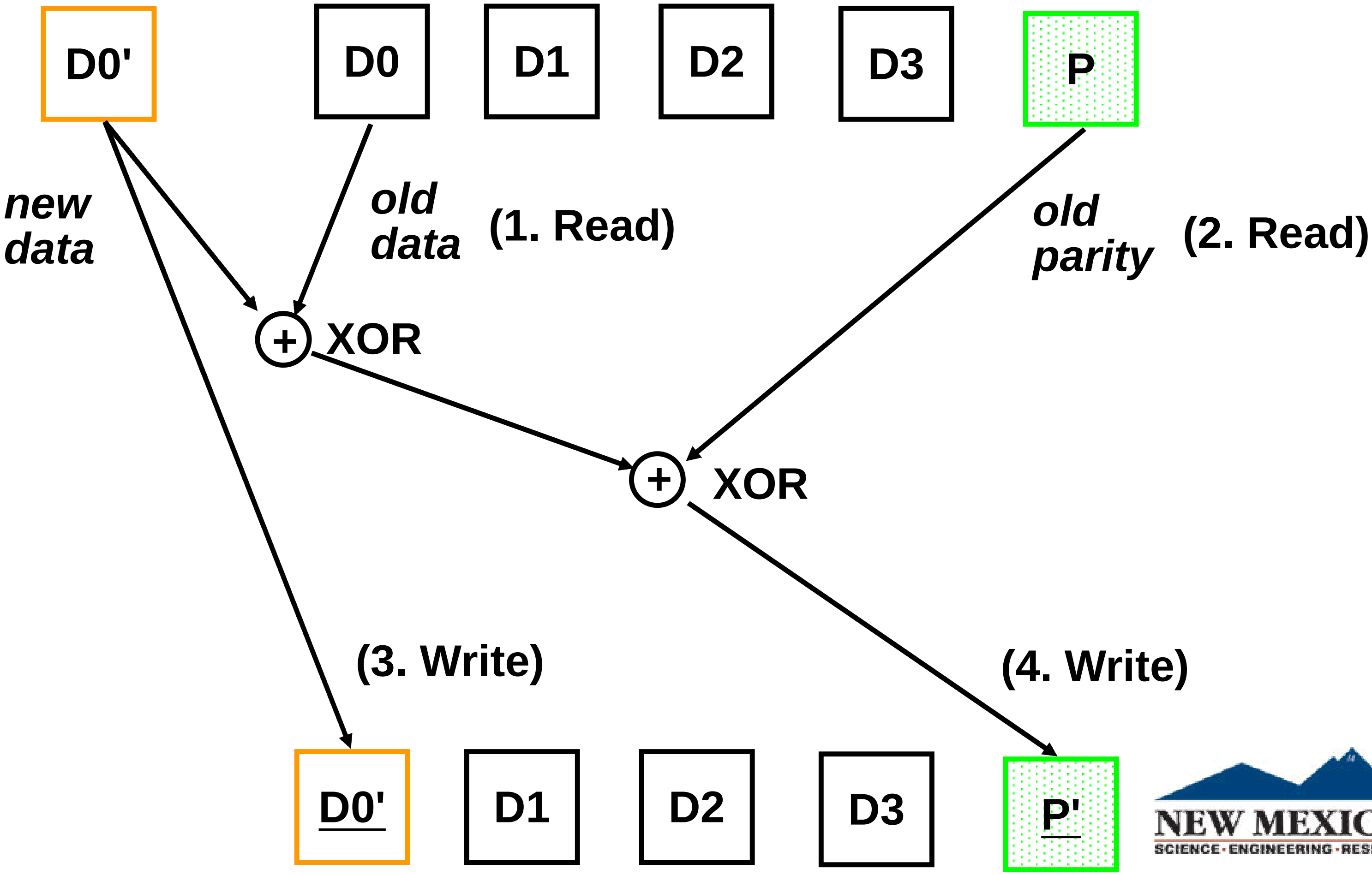
- RAID 4 works well for small reads
- Small writes (write to one disk):
  - Option 1: read other data disks, create new sum and write to Parity Disk
  - Option 2: since P has old sum, compare old data to new data, add the difference to P
- Small writes are limited by Parity Disk: Write to D0, D5 both also write to P disk



Problems of Disk Arrays: Small Writes

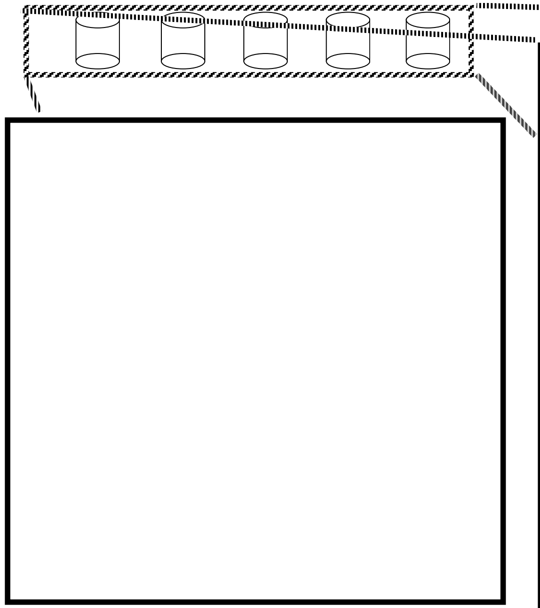
RAID-5: Small Write Algorithm

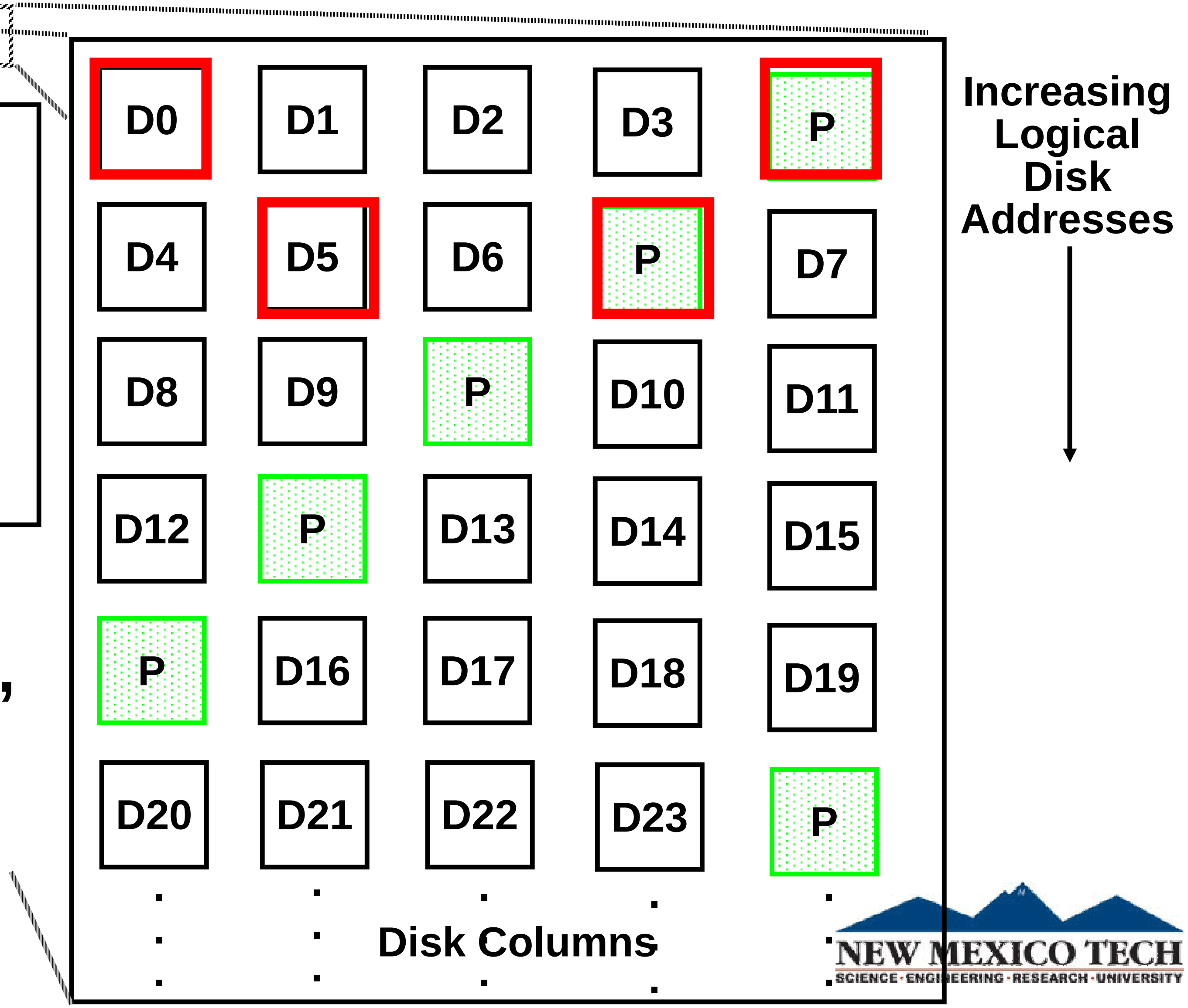
1 Logical Write = 2 Physical Reads + 2 Physical Writes



NEW MEXICO TECH
SCIENCE · ENGINEERING · RESEARCH · UNIVERSITY

RAID 5 – High I/O Interleaved Parity





Example: write to D0, D5 uses disks 0, 1, 3, 4