Disks & RAIDs

Slides Courtesy of Mary Jane Irwin (www.cse.psu.edu/~mji)

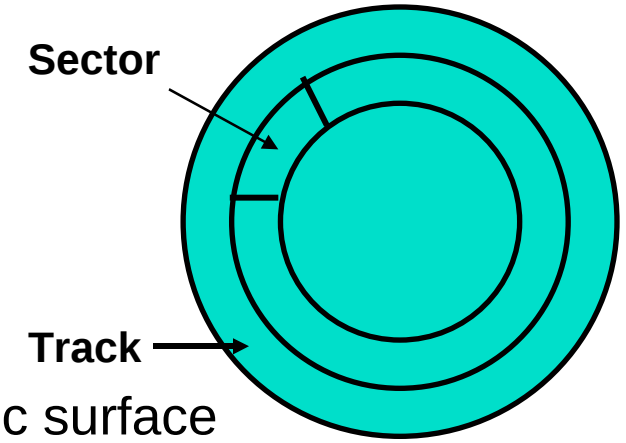
www.cse.psu.edu/~cg431

With Modifications by Nathan Sprague

[Adapted from *Computer Organization and Design*, Patterson & Hennessy, © 2005]

Magnetic Disk

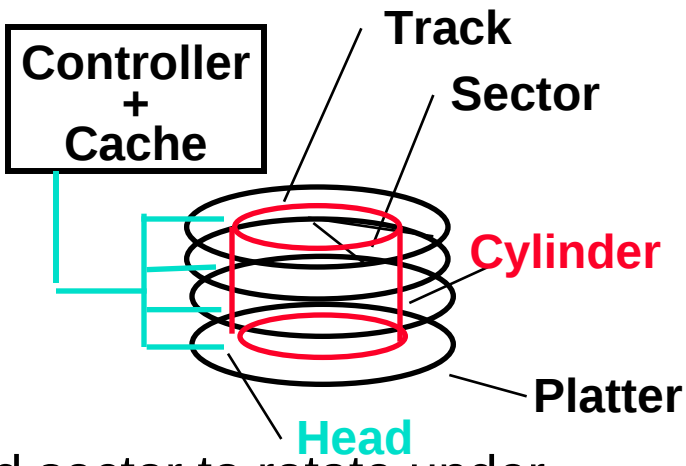
- Purpose
  - Long term, nonvolatile storage
  - Lowest level in the memory hierarchy
    - slow, large, inexpensive



- General structure
  - A rotating platter coated with a magnetic surface
  - A moveable read/write head to access the information on the disk
- Typical numbers
  - 1 to 5 (1 or 2 surface) platters per disk of 1” to 5.25” in diameter
  - Rotational speeds of 5,400 to 15,000 RPM
  - 10,000 to 50,000 tracks per surface
    - cylinder - all the tracks under the head at a given point on all surfaces
  - 100 to 500 sectors per track
    - the smallest unit that can be read/written (typically 512B)

Magnetic Disk Characteristic

- Disk read/write components



1. Seek time: position the head over the proper track (3 to 14 ms avg)
   - due to locality of disk references the actual average seek time may be only 25% to 33% of the advertised number
2. Rotational latency: wait for the desired sector to rotate under the head (½ of 1/RPM converted to ms)
   - 0.5/5400RPM = 5.6ms to 0.5/15000RPM = 2.0ms
3. Transfer time: transfer a block of bits (one or more sectors) under the head to the disk controller's cache (30 to 100 MB/s are typical disk transfer rates)
   - the disk controller's "cache" takes advantage of spatial locality in disk accesses
     - cache transfer rates are much faster (e.g., 320 MB/s)
4. Controller time: the overhead the disk controller imposes in performing a disk I/O access (typically < .2 ms)

Typical Disk Access Time

- The average time to read or write a 512B sector for a disk rotating at 10,000RPM with average seek time of 6ms, a 50MB/sec transfer rate, and a 0.2ms controller overhead

  Avg disk read/write = 6.0ms + 0.5/(10000RPM/(60sec/minute))+ 0.5KB/(50MB/sec) + 0.2ms = 6.0 + 3.0 + 0.01 + 0.2 = 9.21ms

  If the measured average seek time is 25% of the advertised average seek time, then

  Avg disk read/write = 1.5 + 3.0 + 0.01 + 0.2 = 4.71ms

- The rotational latency is usually the largest component of the access time

Solid State Drives

- Flash memory – non-volatile storage that uses transistor based memory cells.
  - nor flash – fast reads, slow writes
    - Frequently used for EEPROM
  - nand flash – slower reads, faster writes, higher density
    - Used for drives and general storage
- Limited to approximately 100,000 – 1,000,000 write cycles
  - Drives incorporate load-leveling logic.
- No rotational or seek latency.
- More expensive than traditional drives (2009)
  - SSD: About $3.00/GB (120GB drive)
  - Disk: About $.10/GB (1TB drive)
- Let's compare some data sheets...

Dependability, Reliability, Availability

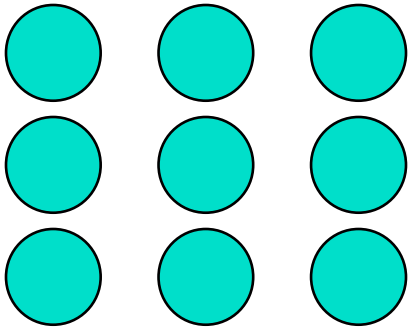
- Reliability – measured by the mean time to failure (MTTF).
- Service interruption is measured by mean time to repair (MTTR)
- Availability – a measure of service accomplishment

$$\text{Availability} = \text{MTTF}/(\text{MTTF} + \text{MTTR})$$

- To increase MTTF, either improve the quality of the components or design the system to continue operating in the presence of faulty components
  1. Fault avoidance: preventing fault occurrence by construction
  2. Fault tolerance: using redundancy to correct or bypass faulty components

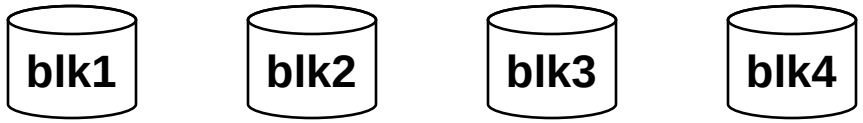
RAIDs: Disk Arrays

Redundant Array of Inexpensive Disks



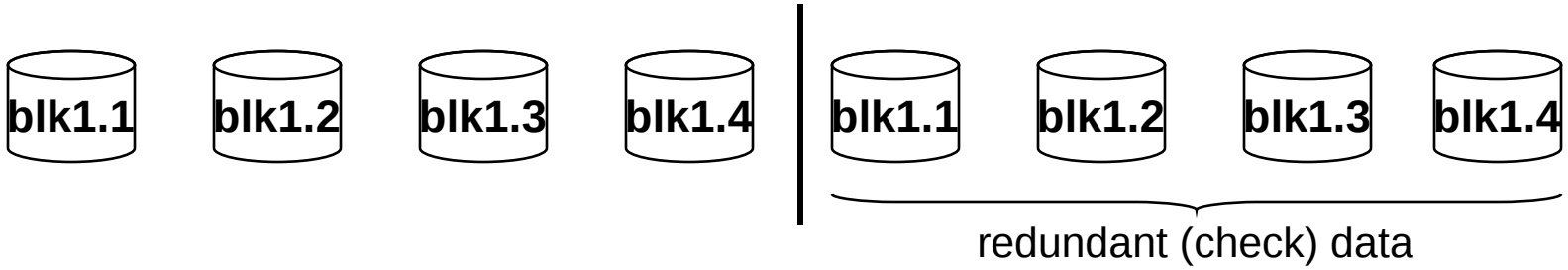
- Arrays of small and inexpensive disks
  - Increase potential throughput by having many disk drives
    - Data is spread over multiple disk
    - Multiple accesses are made to several disks at a time
- Reliability is lower than a single disk
- But availability can be improved by adding redundant disks (RAID)
  - Lost information can be reconstructed from redundant information
  - MTTR: mean time to repair is in the order of hours
  - MTTF: mean time to failure of disks is tens of years

RAID: Level 0 (No Redundancy: Striping)



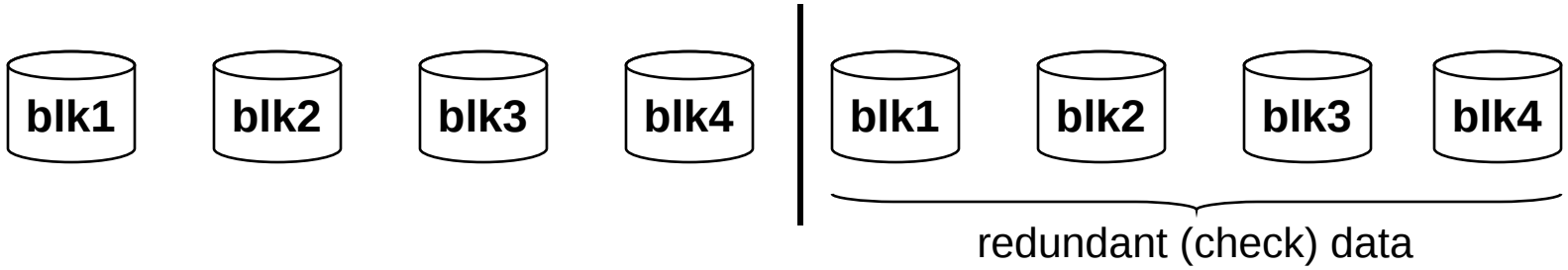
- Multiple smaller disks as opposed to one big disk
  - Spreading the blocks over multiple disks – striping – means that multiple blocks can be accessed in parallel increasing the performance
    - A 4 disk system gives four times the throughput of a 1 disk system
  - Same cost as one *big* disk – assuming 4 small disks cost the same as one big disk
- No redundancy, so what if one disk fails?
  - Failure of one or more disks is more likely as the number of disks in the system increases

RAID: Level 1 (Redundancy via Mirroring)



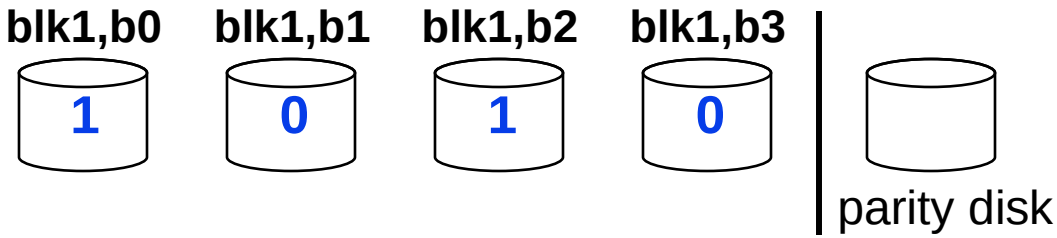
- Uses twice as many disks as RAID 0 (e.g., 8 smaller disks with second set of 4 duplicating the first set) so there are always two copies of the data
  - # redundant disks = # of data disks so twice the cost of one big disk
    - writes have to be made to both sets of disks, so writes would be only 1/2 the performance of RAID 0
- What if one disk fails?
  - If a disk fails, the system just goes to the "mirror" for the data

RAID: Level 0+1 (Striping with Mirroring)



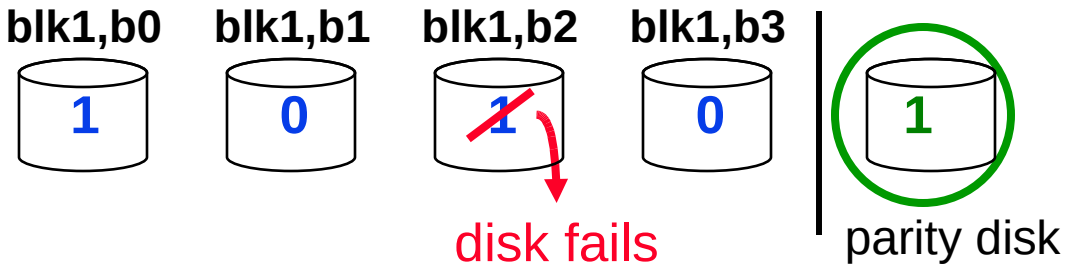
- Combines the best of RAID 0 and RAID 1, data is striped across four disks and mirrored to four disks
  - Four times the throughput (due to striping)
  - # redundant disks = # of data disks so twice the cost of one big disk
    - writes have to be made to both sets of disks, so writes would be only 1/2 the performance of RAID 0
- What if one disk fails?
  - If a disk fails, the system just goes to the "mirror" for the data

RAID: Level 3 (Bit-Interleaved Parity)



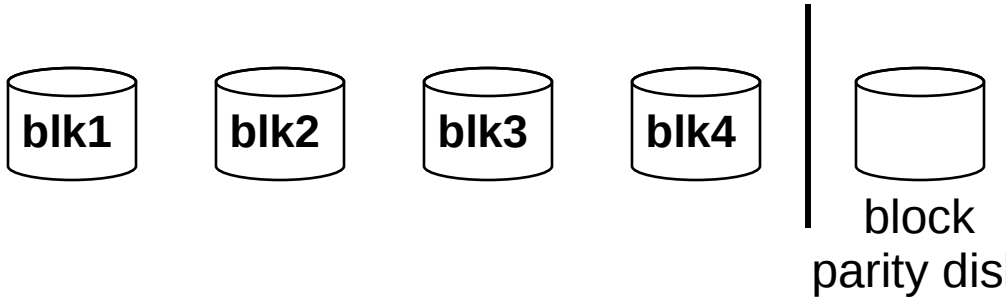
- Cost of higher availability is reduced to 1/N where N is the number of disks in a protection group
  - # redundant disks = 1 × # of protection groups
    - writes require writing the new data to the data disk as well as computing the parity, meaning reading the other disks, so that the parity disk can be updated
- Can tolerate *limited* disk failure, since the data can be reconstructed

RAID: Level 3 (Bit-Interleaved Parity)



- Cost of higher availability is reduced to 1/N where N is the number of disks in a protection group
  - # redundant disks = 1 × # of protection groups
    - writes require writing the new data to the data disk as well as computing the parity, meaning reading the other disks, so that the parity disk can be updated
- Can tolerate *limited* disk failure, since the data can be reconstructed

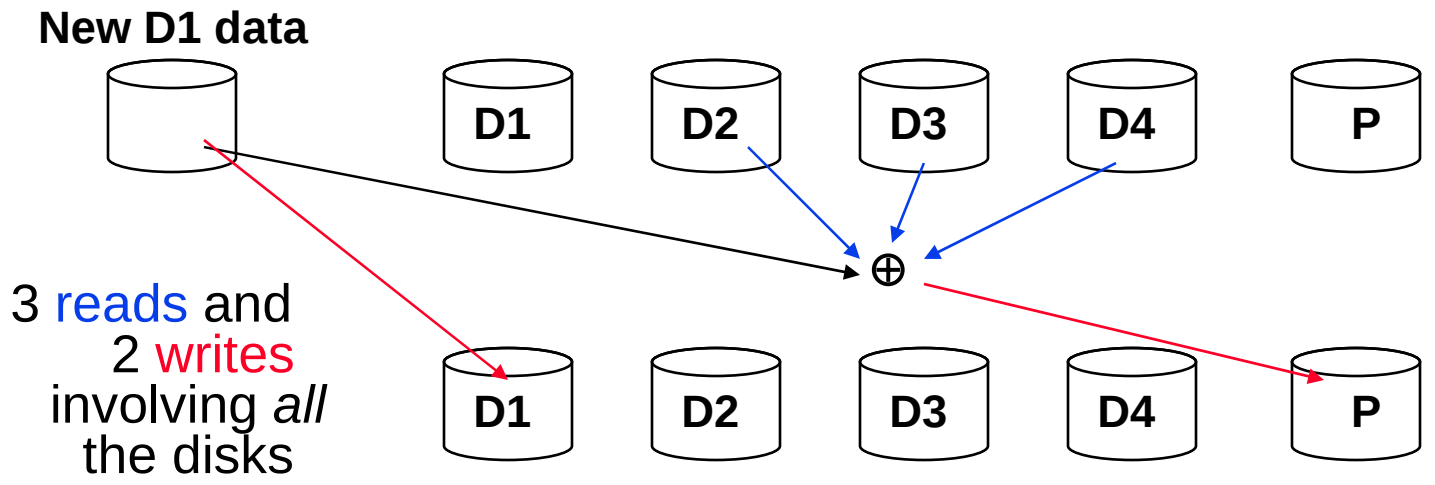
RAID: Level 4 (Block-Interleaved Parity)



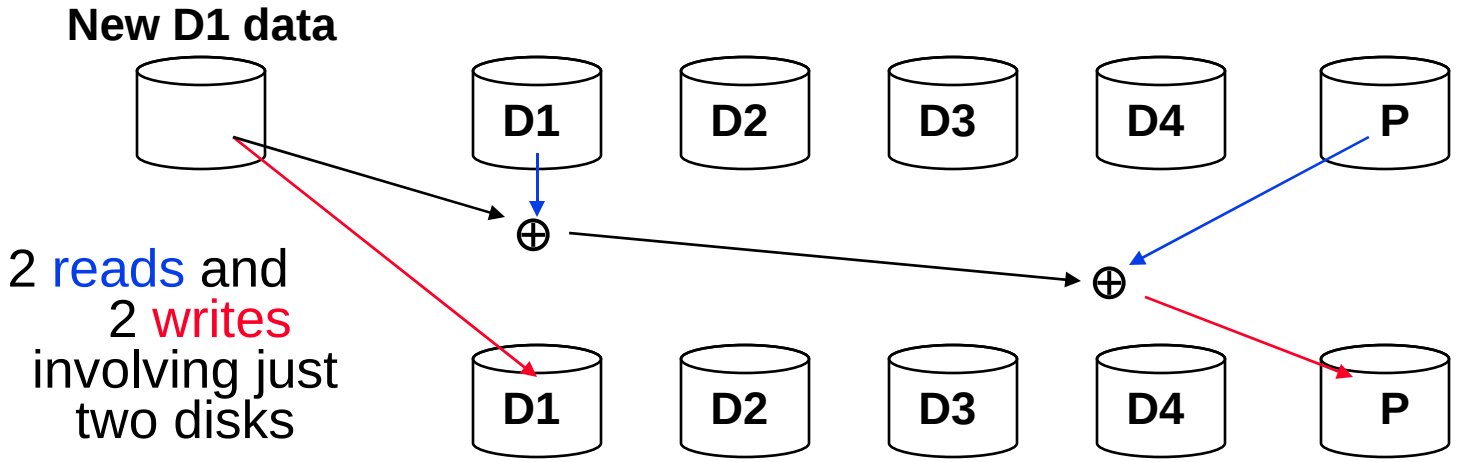
- Cost of higher availability still only 1/N but the parity is stored as blocks associated with sets of data blocks
  - Four times the throughput (striping)
  - # redundant disks = 1 × # of protection groups
  - Supports "small reads" and "small writes" (reads and writes that go to just one (or a few) data disk in a protection group)
    - by watching which bits change when writing new information, need only to change the corresponding bits on the parity disk
    - the parity disk must be updated on every write, so it is a bottleneck for back-to-back writes
- Can tolerate *limited* disk failure, since the data can be reconstructed

Small Writes

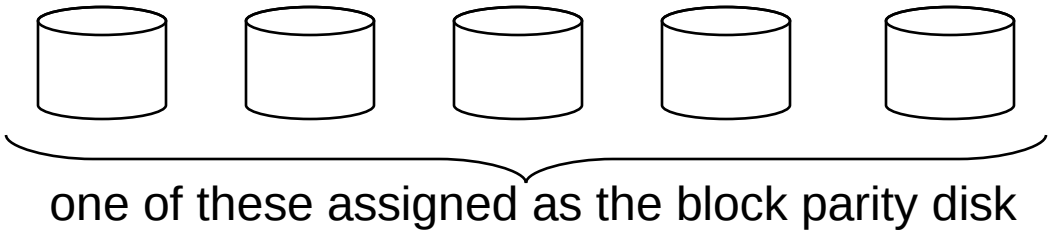
- RAID 3 small writes



- RAID 4 small writes

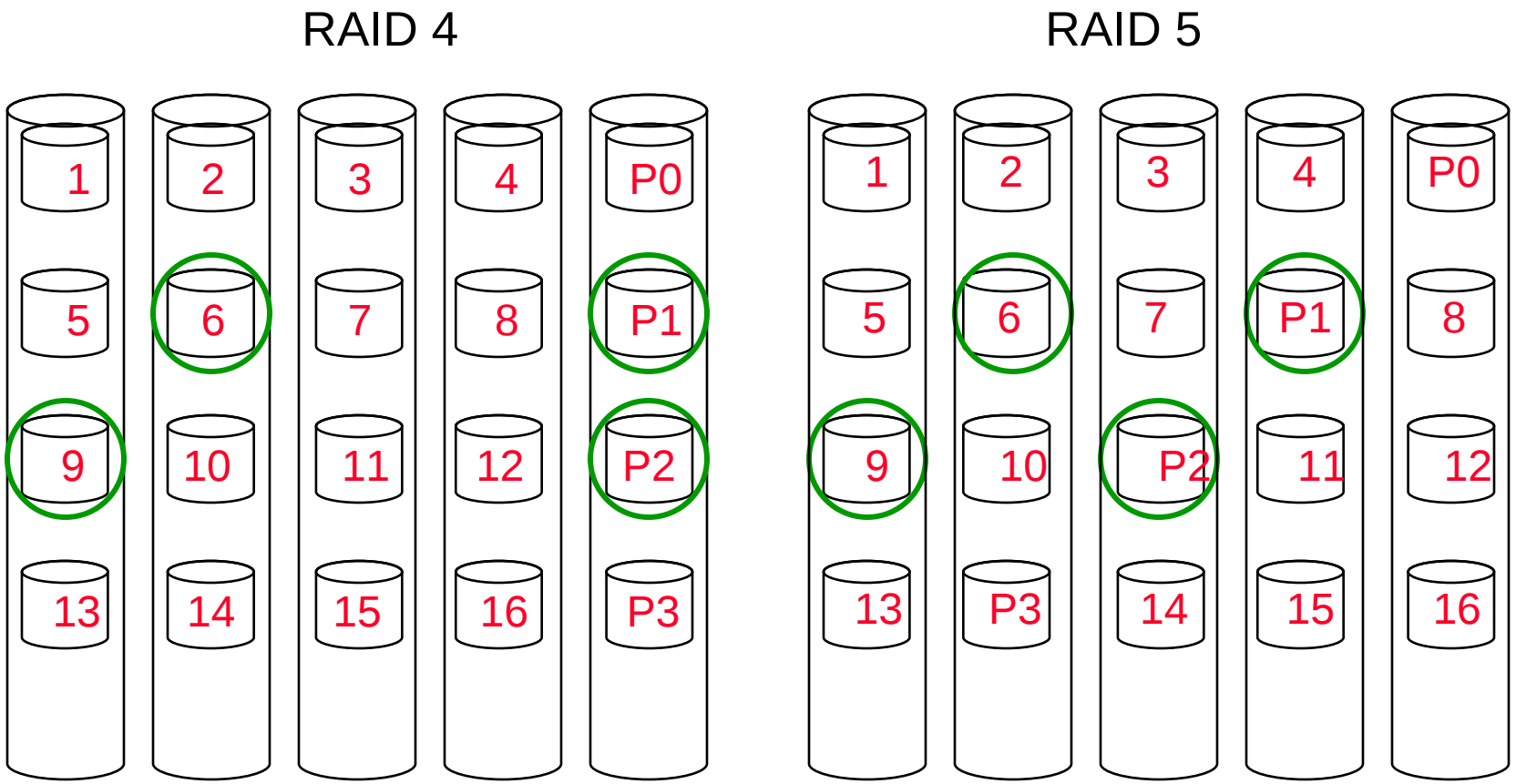


RAID: Level 5 (Distributed Block-Interleaved Parity)



- Cost of higher availability still only 1/N but the parity block can be located on any of the disks so there is no single bottleneck for writes
  - Still four times the throughput (striping)
  - # redundant disks = 1 × # of protection groups
  - Supports "small reads" and "small writes" (reads and writes that go to just one (or a few) data disk in a protection group)
  - Allows multiple simultaneous writes as long as the accompanying parity blocks are not located on the same disk
- Can tolerate *limited* disk failure, since the data can be reconstructed

Distributing Parity Blocks



- By distributing parity blocks to all disks, some small writes can be performed in parallel