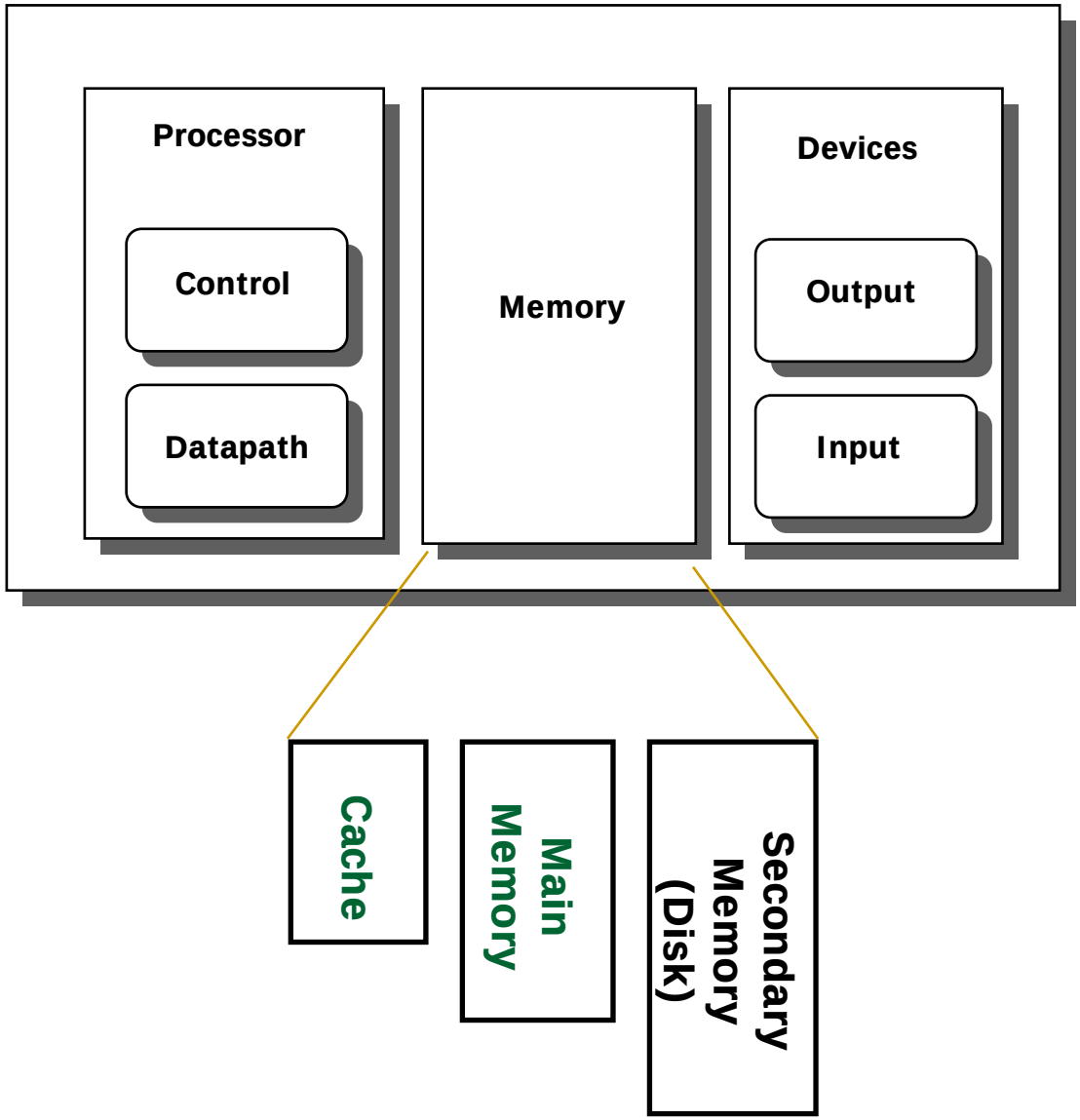
COMP303
Computer Architecture
Lecture 17

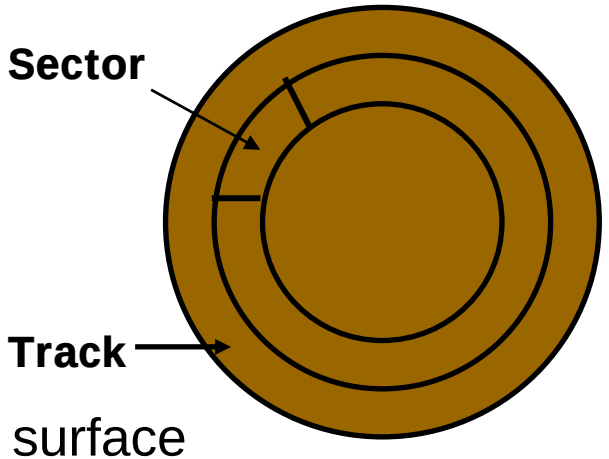
Storage

Review: Major Components of a Computer



Magnetic Disk

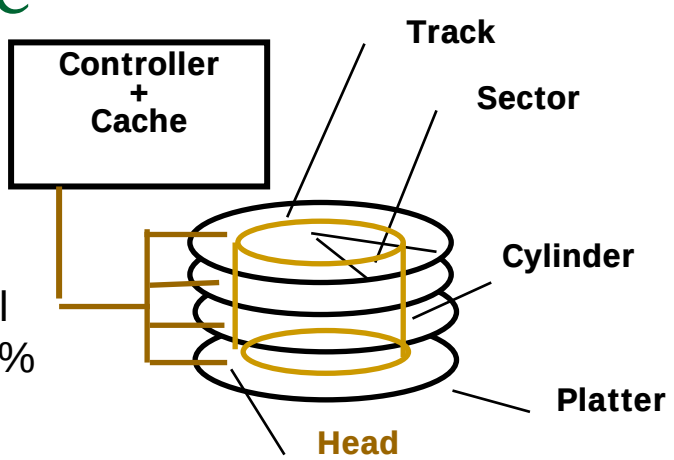
- Purpose
  - Long term, nonvolatile storage
  - Lowest level in the memory hierarchy
    - slow, large, inexpensive
- General structure
  - A rotating platter coated with a magnetic surface
  - A moveable read/write head to access the information on the disk



- Typical numbers
  - 1 to 4 platters (each with 2 recordable surfaces) per disk of 1” to 3.5” 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 13 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 (70 to 125 MB/s are typical disk transfer rates in 2008)
   - the disk controller's "cache" takes advantage of spatial locality in disk accesses
     - cache transfer rates are much faster (e.g., 375 MB/s → 3 Gbit/sec)
4. Controller time: the overhead the disk controller imposes in performing a disk I/O access (typically < .2 ms)

Disk Interface Standards

- Higher-level disk interfaces have a microprocessor disk controller that can lead to performance optimizations
  - ATA (Advanced Technology Attachment) – An interface standard for the connection of storage devices such as hard disks, solid-state drives, and CD-ROM drives. Parallel ATA has been largely replaced by serial ATA (SATA).
  - SCSI (Small Computer Systems Interface) – A set of standards (commands, protocols, and electrical and optical interfaces) for physically connecting and transferring data between computers and peripheral devices. Most commonly used for hard disks and tape drives.

- In particular, disk controllers have SRAM disk caches which support fast access to data that was recently read and often also include prefetch algorithms to try to anticipate demand

Magnetic Disk Examples (www.seagate.com)

Feature	Seagate ST31000340NS	Seagate ST973451SS	Seagate ST9160821AS
Disk diameter (inches)	3.5	2.5	2.5
Capacity (GB)	1000	73	160
# of surfaces (heads)	4	2	2
Rotation speed (RPM)	7,200	15,000	5,400
Transfer rate (MB/sec)	105	79-112	44
Minimum seek (ms)	0.8r-1.0w	0.2r-0.4w	1.5r-2.0w
Average seek (ms)	8.5r-9.5w	2.9r-3.3w	12.5r-13.0w
MTTF (hours@25^{o}C)	1,200,000	1,600,000	??
Dim (inches), Weight (lbs)	1x4x5.8, 1.4	0.6x2.8x3.9, 0.5	0.4x2.8x3.9, 0.2
GB/cu.inch, GB/watt	43, 91	11, 9	37, 84
Power: op/idle/sb (watts)	11/8/1	8/5.8/-	1.9/0.6/0.2
Price in 2008, $/GB	~$0.3/GB	~$5/GB	~$0.6/GB

Disk Latency & Bandwidth Milestones

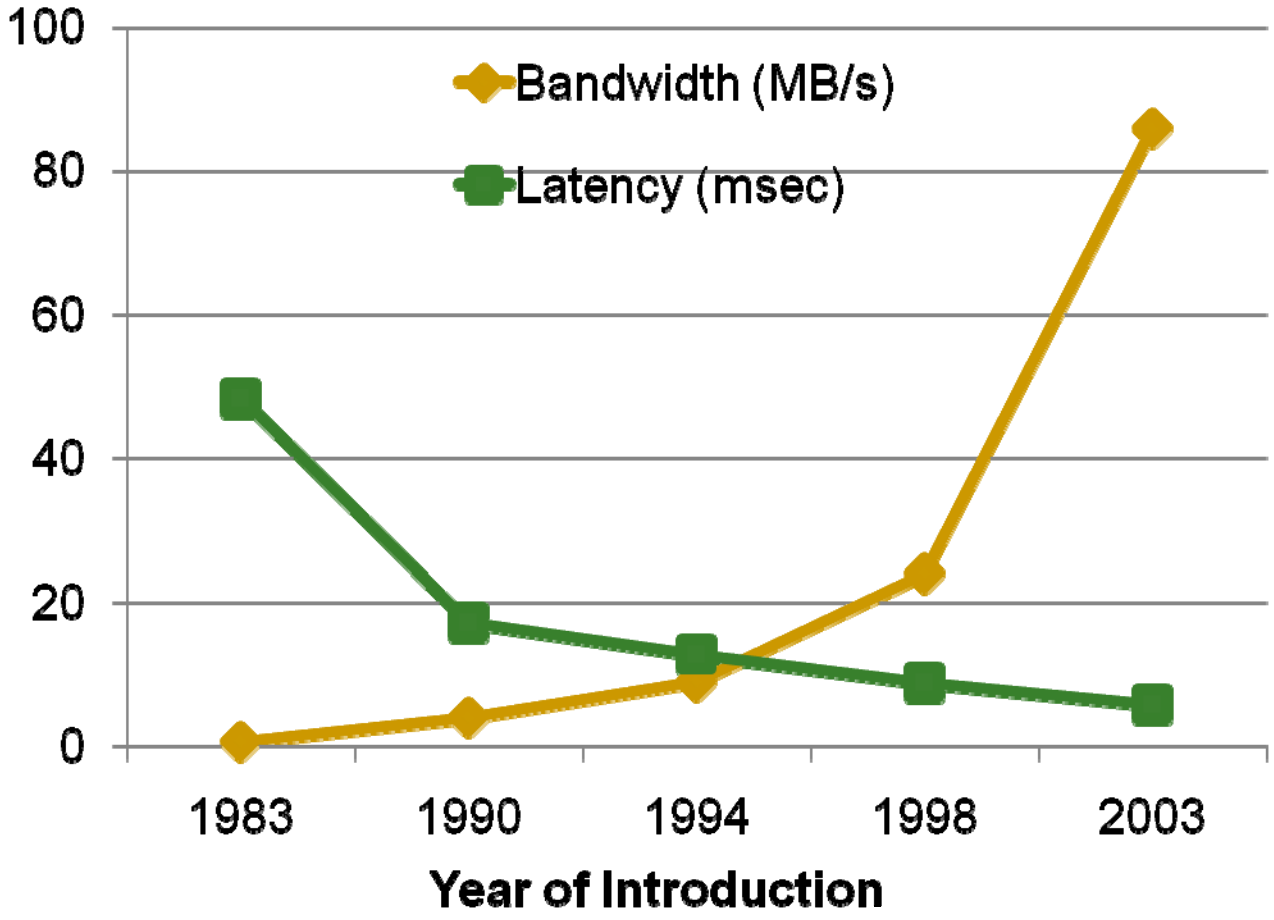
	CDC Wren	SG ST41	SG ST15	SG ST39	SG ST37
RSpeed (RPM)	3600	5400	7200	10000	15000
Year	1983	1990	1994	1998	2003
Capacity (Gbytes)	0.03	1.4	4.3	9.1	73.4
Diameter (inches)	5.25	5.25	3.5	3.0	2.5
Interface	ST-412	SCSI	SCSI	SCSI	SCSI
Bandwidth (MB/s)	0.6	4	9	24	86
Latency (msec)	48.3	17.1	12.7	8.8	5.7

Patterson, CACM Vol 47, #10, 2004

- Disk latency is one average seek time plus the rotational latency.
- Disk bandwidth is the peak transfer time of formatted data from the media (not from the cache).

Latency & Bandwidth Improvements

- In the time that the disk bandwidth doubles the latency improves by a factor of only 1.2 to 1.4



Flash Storage

- Flash memory is the first credible challenger to disks. It is semiconductor memory that is nonvolatile like disks, but has latency 100 to 1000 times faster than disk and is smaller, more power efficient, and more shock resistant.
  - In 2008, the price of flash is $4 to $10 per GB or about 2 to 10 times higher than disk and 5 to 10 times lower than DRAM.
  - Flash memory bits wear out (unlike disks and DRAMs), but wear leveling can make it unlikely that the write limits of the flash will be exceeded

Feature	Kingston	Transend	RiDATA
Capacity (GB)	8	16	32
Bytes/sector	512	512	512
Transfer rates (MB/sec)	4	20r-18w	68r-50w
MTTF	>1,000,000	>1,000,000	>4,000,000
Price (2008)	~ $30	~ $70	~ $300

Dependability, Reliability, Availability

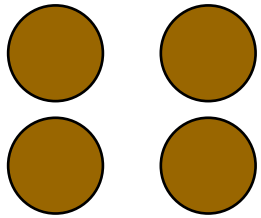
- Reliability – measured by the mean time to failure (MTTF). Service interruption is measured by mean time to repair (MTTR)
- Availability – a measure of service accomplishment

  Availability = MTTF/(MTTF + 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 (hardware)
     - Fault detection versus fault correction
     - Permanent faults versus transient faults

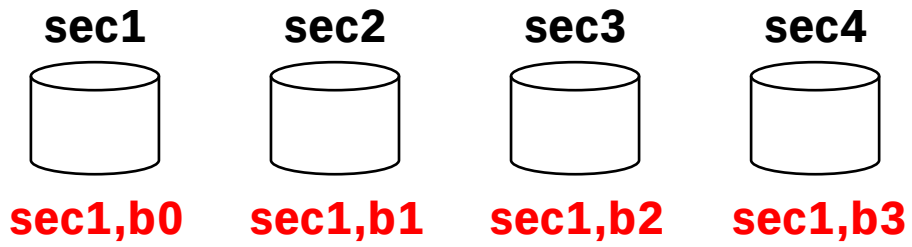
RAIDs: Disk Arrays

Redundant Array of
Inexpensive Disks



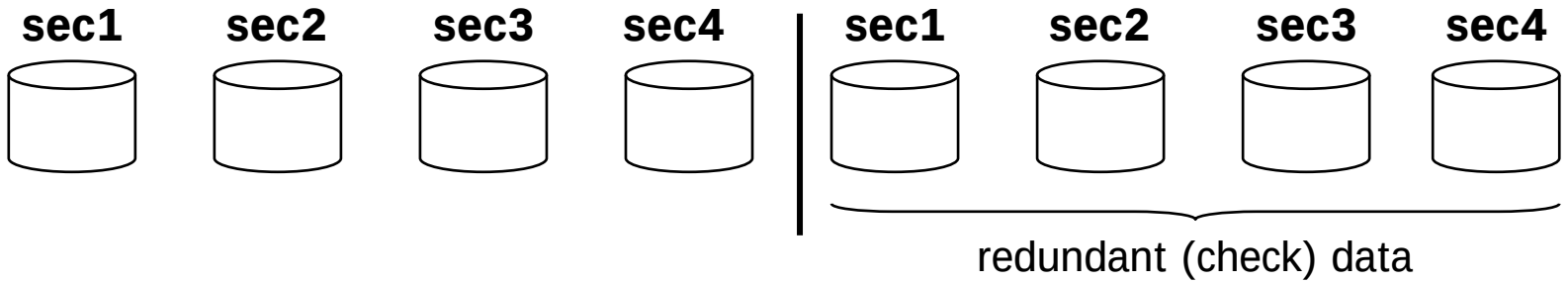
- Arrays of small and inexpensive disks
  - Increase potential throughput by having many disk drives
    - Data is spread over multiple disk
    - Multiple accesses are made to several disks at a time
- Reliability is lower than a single disk
- But availability can be improved by adding redundant disks (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)



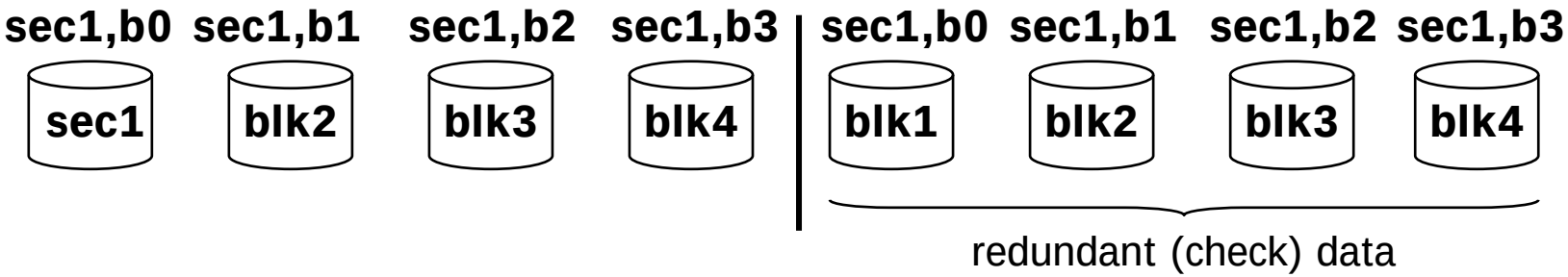
- Multiple smaller disks as opposed to one big disk
  - Spreading the sector over multiple disks – striping – means that multiple blocks can be accessed in parallel increasing the performance
    - A 4 disk system gives four times the throughput of a 1 disk system
  - Same cost as one *big* disk – assuming 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)



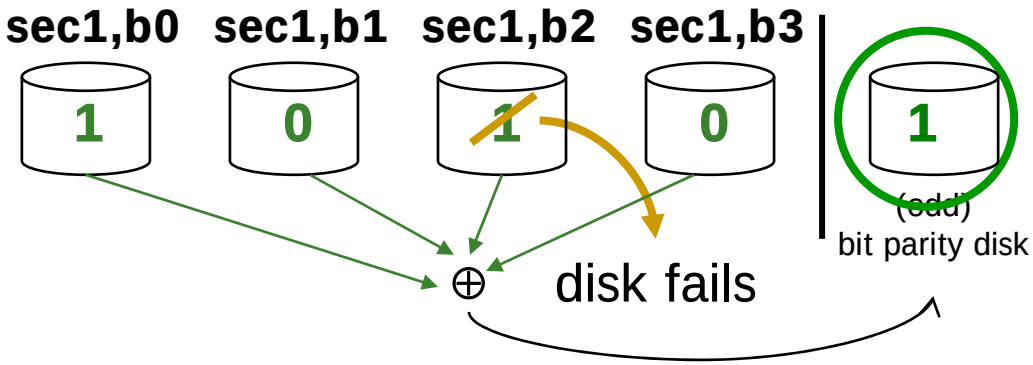
- Uses twice as many disks as RAID 0 (e.g., 8 smaller disks with the second set of 4 duplicating the first set) so there are always two copies of the data
  - # redundant disks = # of data disks so twice the cost of one big disk
    - writes have to be made to both sets of disks, so writes would be only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



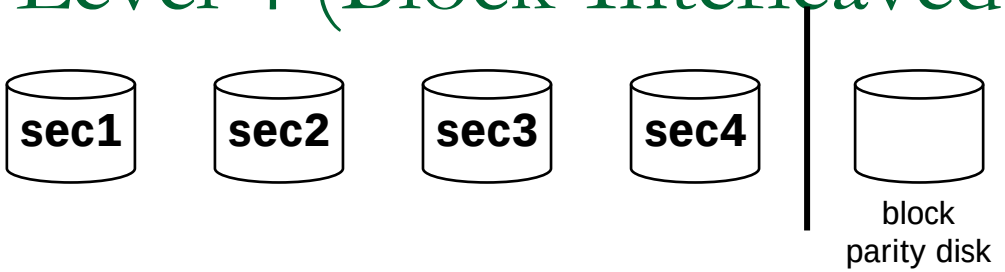
- Combines the best of RAID 0 and RAID 1, data is striped across four disks and mirrored to four disks
  - Four times the throughput (due to striping)
  - # redundant disks = # of data disks so twice the cost of one big disk
    - writes have to be made to both sets of disks, so writes would be only 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* (single) disk failure, since the data can be reconstructed
    - reads require reading all the operational data disks as well as the parity disk to calculate the missing data that was stored on the failed disk

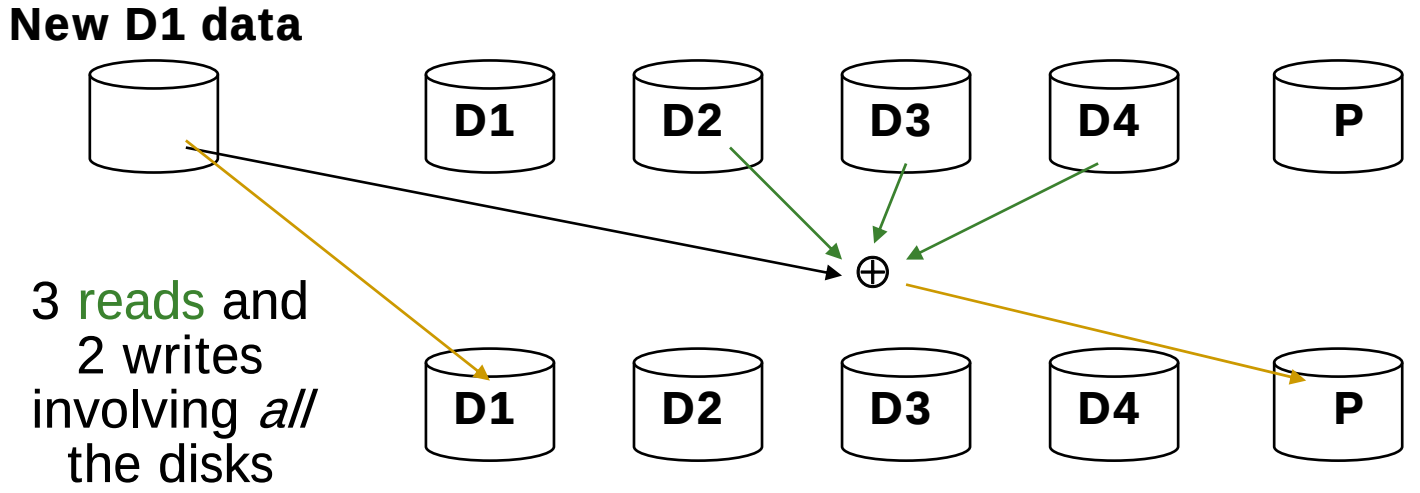
RAID: Level 4 (Block-Interleaved Parity)



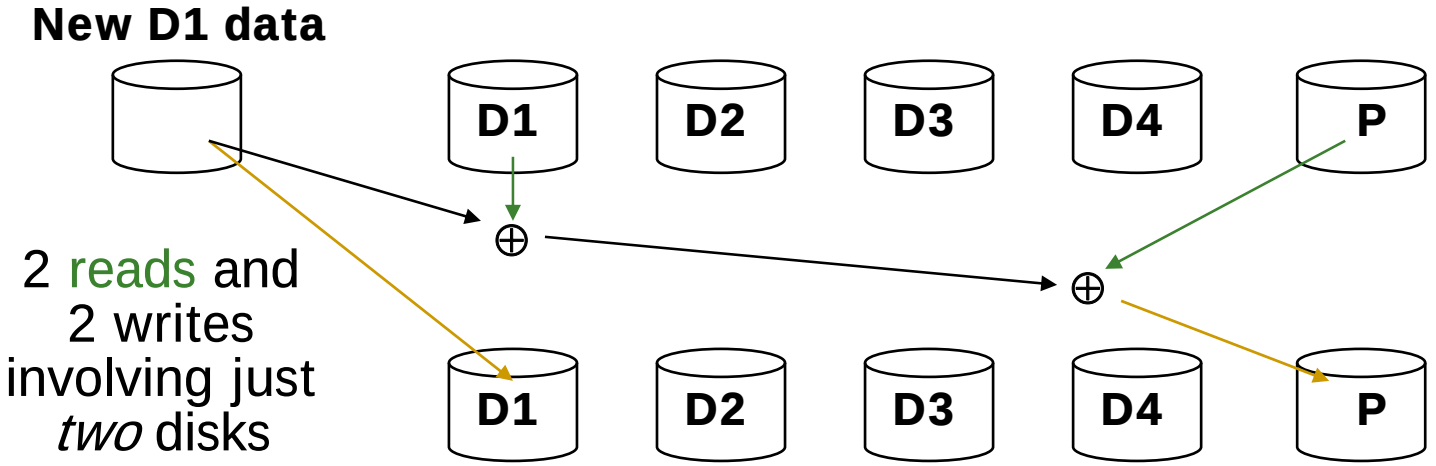
- Cost of higher availability still only 1/N but the parity is stored as blocks associated with sets of data blocks
  - Four times the throughput (striping)
  - # redundant disks = 1 × # of protection groups
  - Supports "small reads" and "small writes" (reads and writes that go to just one (or a few) 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

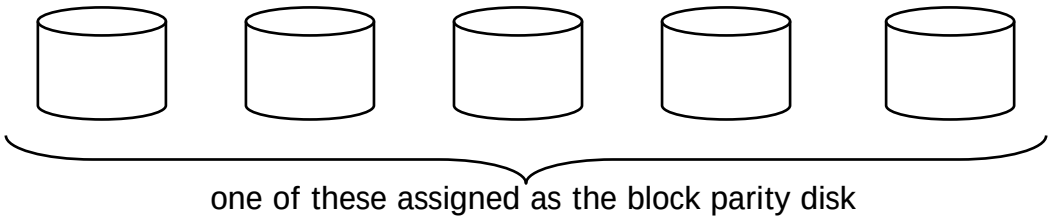
- RAID 3 writes



- RAID 4 *small* writes

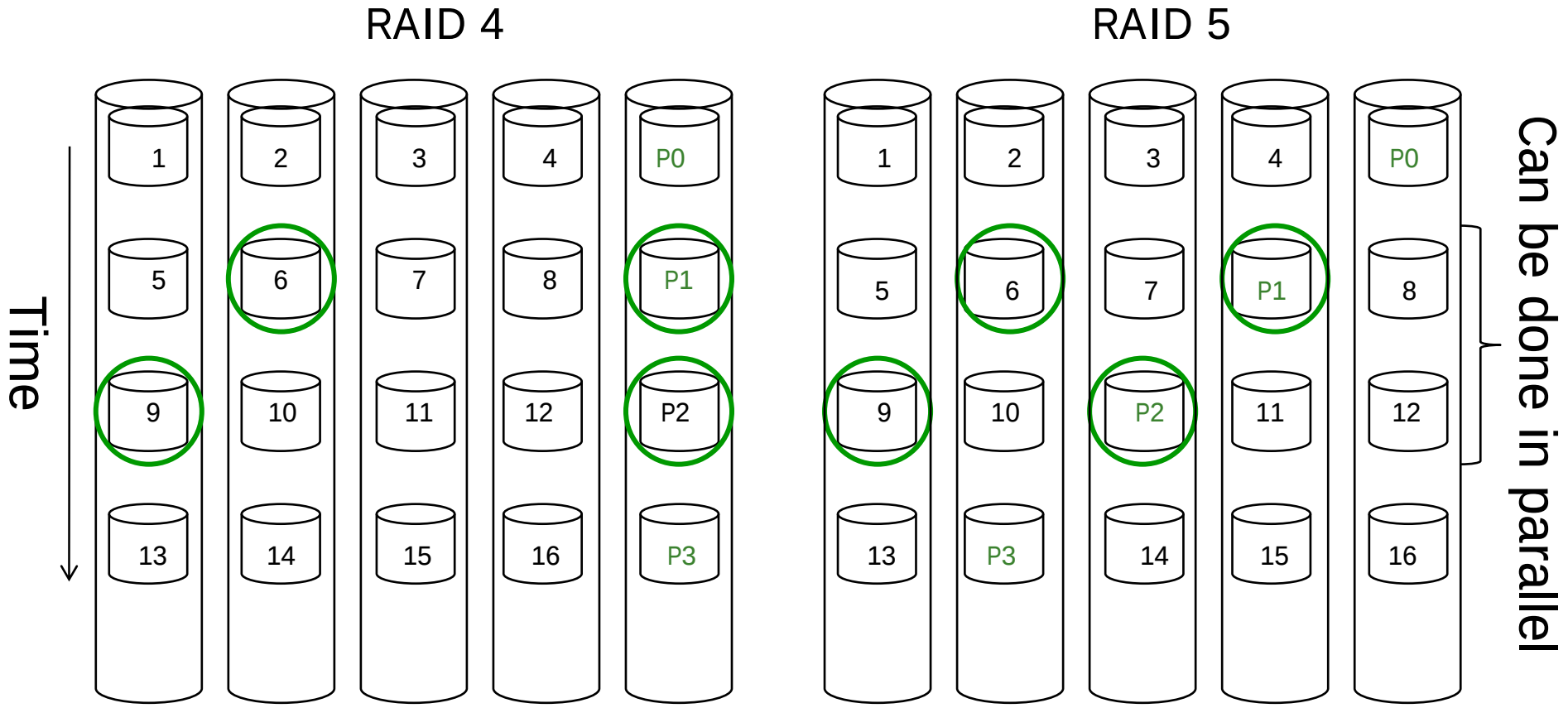


RAID: Level 5 (Distributed Block-Interleaved Parity)



- 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

Summary

- Four components of disk access time:
  - Seek Time: advertised to be 3 to 14 ms but lower in real systems
  - Rotational Latency: 5.6 ms at 5400 RPM and 2.0 ms at 15000 RPM
  - Transfer Time: 30 to 80 MB/s
  - Controller Time: typically less than .2 ms
- RAIDS can be used to improve availability
  - RAID 1 and RAID 5 – widely used in servers, one estimate is that 80% of disks in servers are RAIDs
  - RAID 0+1 (mirroring) – EMC, Tandem, IBM
  - RAID 3 – Storage Concepts
  - RAID 4 – Network Appliance
- RAIDS have enough redundancy to allow continuous operation, but not hot swapping