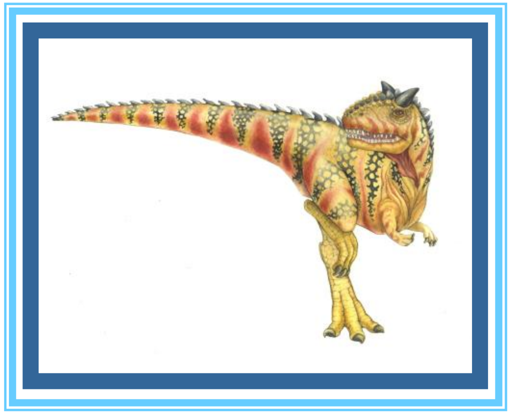
Mass-Storage Systems

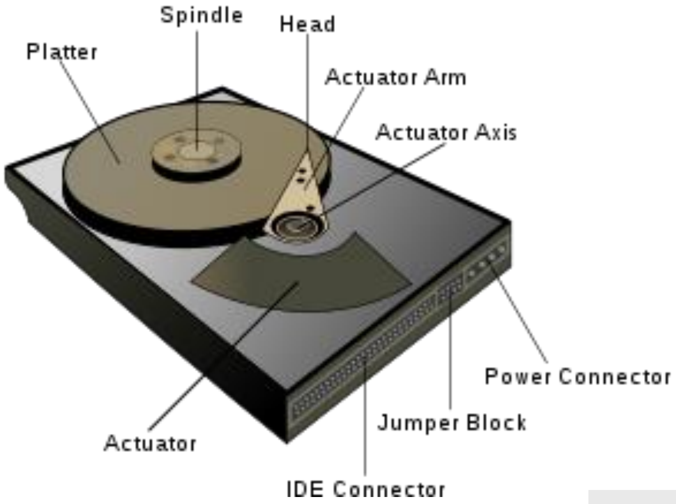


Mass-Storage Systems: What to Learn

- Structure of mass-storage devices and the resulting effects on the uses of the devices
  - Hard Disk Drive
  - SSD (Solid-state drive)
  - Hybrid Disk
- Performance characteristics and management of mass-storage devices
  - Disk Scheduling
- RAID –(originally redundant array of inexpensive disks; now commonly **redundant array of independent disks**) is a data storage virtualization technology that combines multiple disk drive components into a single logical unit for the purposes of data redundancy or performance improvement.
  - improve performance/reliability

Mass Storage: HDD and SSD

- Most popular: Magnetic hard disk drives

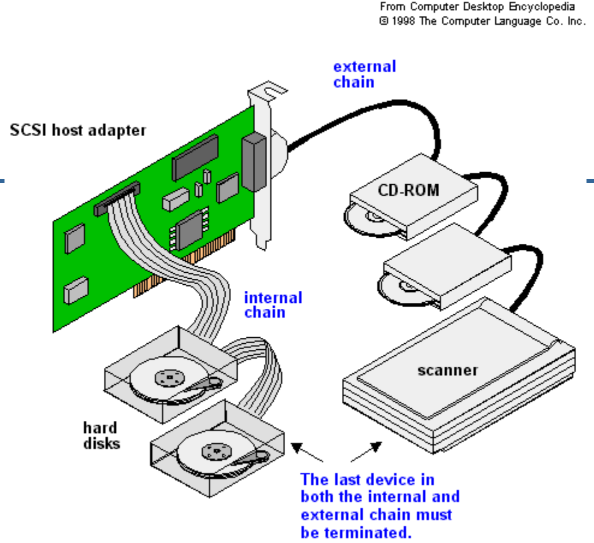


- Solid state drives: (SSD)

Magnetic Tape

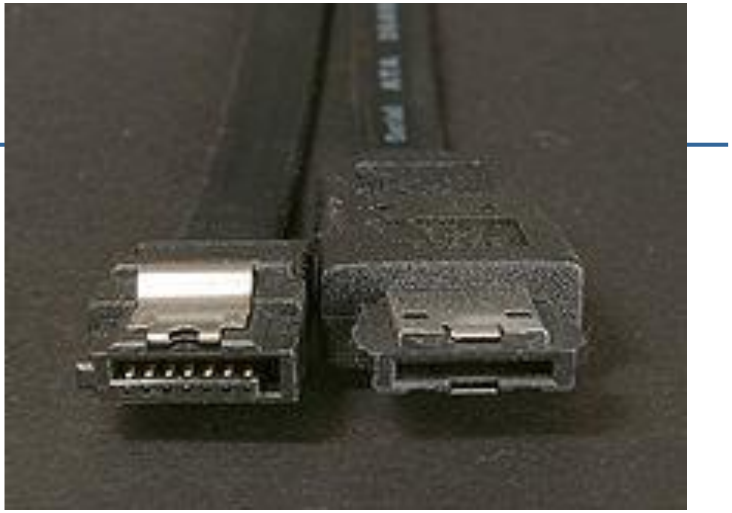
- Relatively permanent and holds large quantities of data
- Random access ~1000 times slower than disk
- Mainly used for backup, storage of infrequently-used data, transfer medium between systems
- 20-1.5TB typical storage
- Common technologies are 4mm, 8mm, 19mm, LTO-2 and SDLT

Disk Attachment


From Computer Desktop Encyclopedia
© 1998 The Computer Language Co. Inc.
external chain
SCSI host adapter
CD-ROM
internal chain
scanner
hard disks
The last device in both the internal and external chain must be terminated.

- Drive attached to computer via **I/O bus**
- USB-Universal Serial Bus
- SATA **Serial Advanced Technology Attachment** (replacing ATA, PATA, EIDE (**Enhanced Integrated Drive Electronics**))
- SCSI
  - itself is a bus, up to 16 devices on one cable, **SCSI initiator** requests operation and **SCSI targets** perform tasks
- FC (Fiber Channel) is high-speed serial architecture
  - Can be switched fabric with 24-bit address space – the basis of **storage area networks (SAN**s**)** in which many hosts attach to many storage units

- SATA connectors

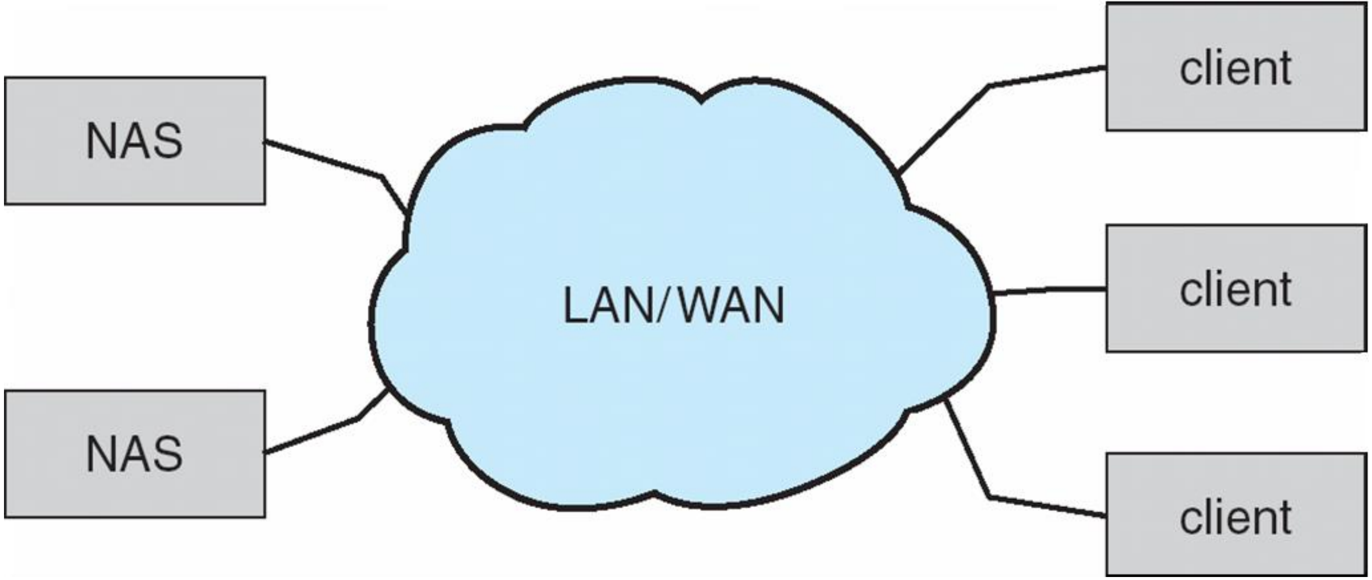


- SCSI

- FC with SAN-switch

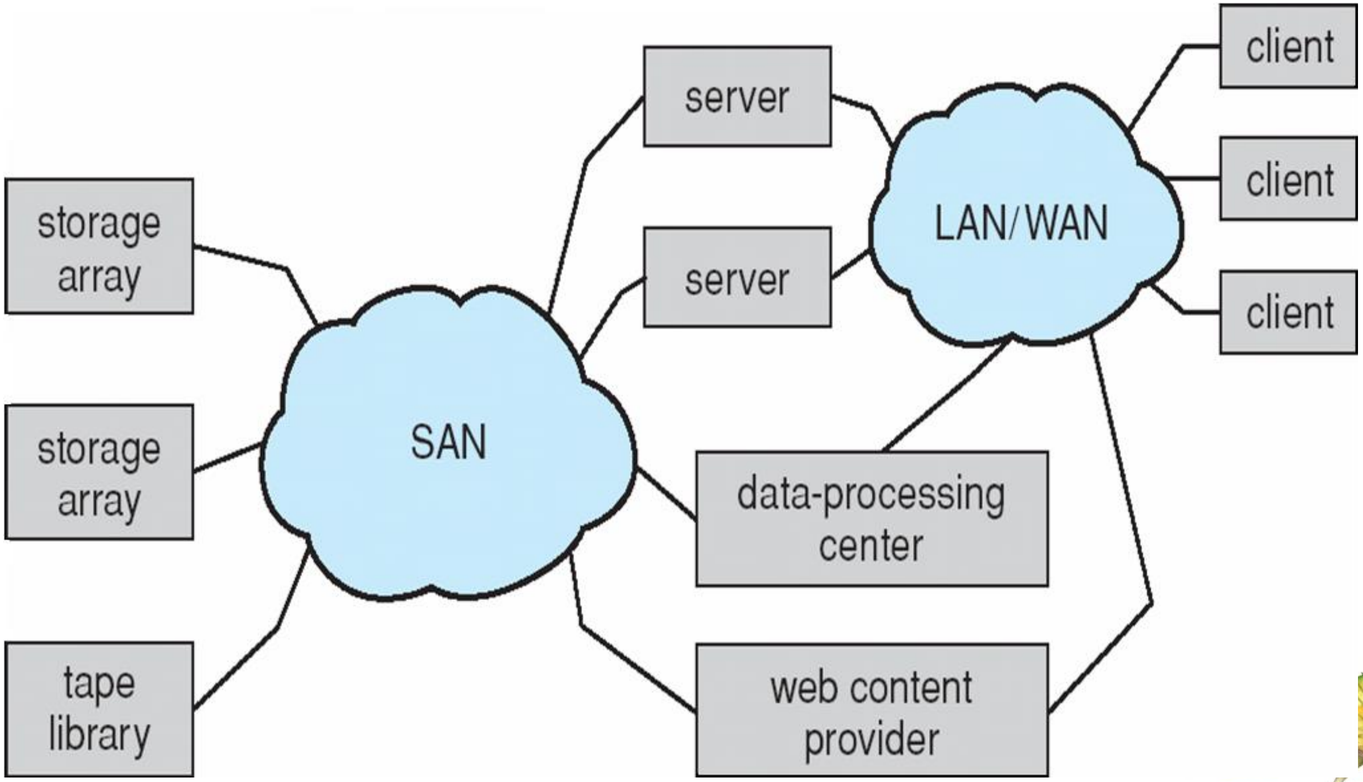
Network-Attached Storage

- Network-attached storage (**NAS**) is storage made available over a network rather than over a local connection (such as a bus)
- NFS(**Network File System**) and CIFS (**Common Internet File System**)are common protocols
- Implemented via remote procedure calls (RPCs) between host and storage
- New **Iscsi (** Internet Small **Computer** System Interface**)** protocol uses IP network to carry the SCSI protocol



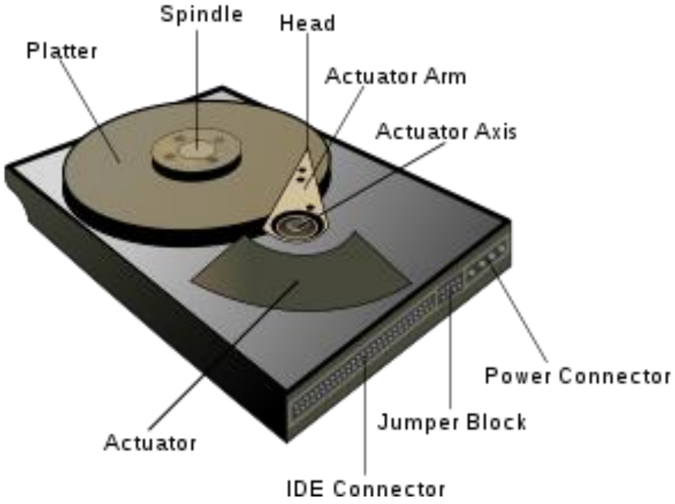
Storage Area Network (SAN)

- Special/dedicated network for accessing block level data storage
- Multiple hosts attached to multiple storage arrays - flexible



Performance characteristics of disks

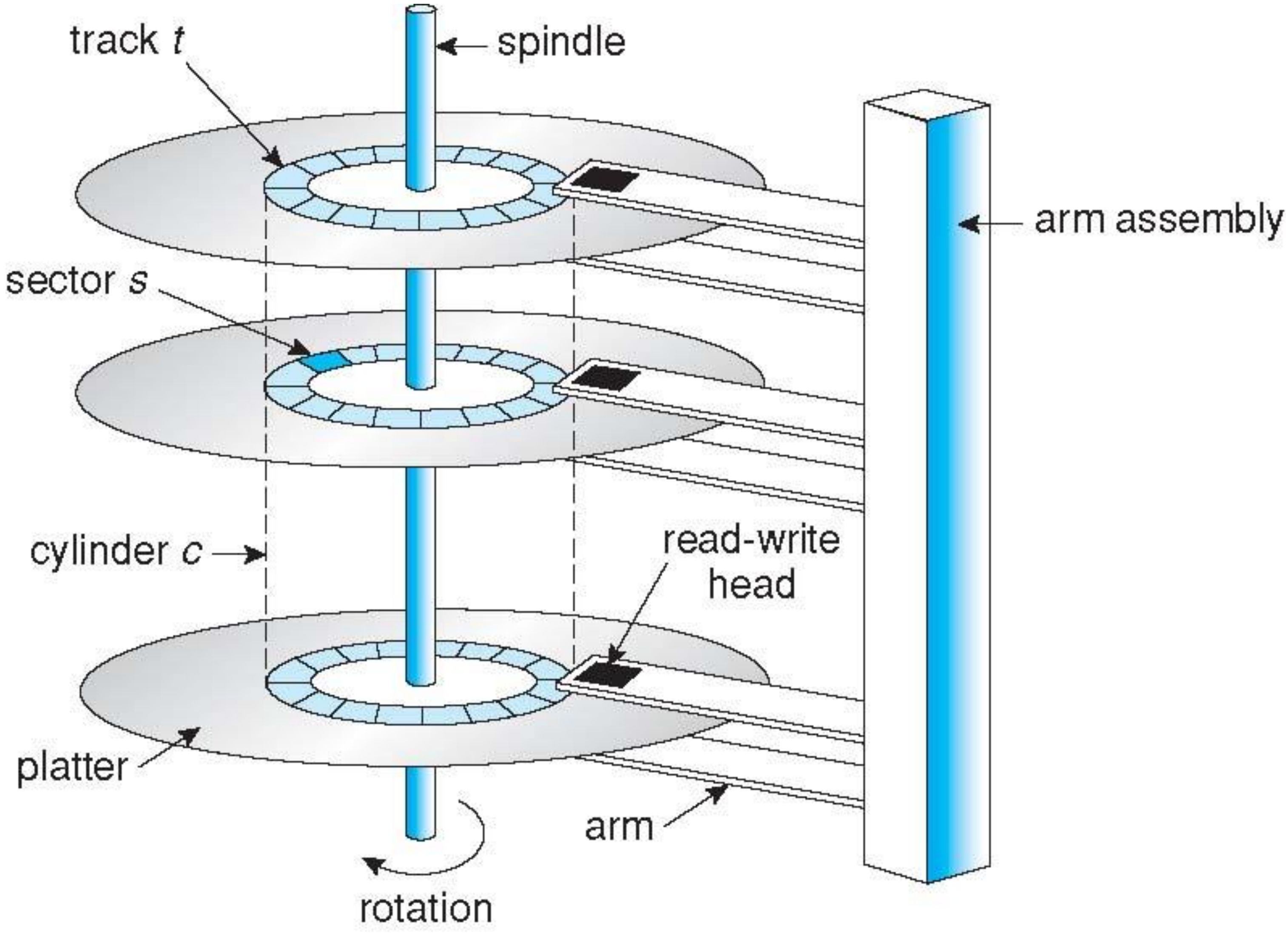
- Drives rotate at 60 to 200 times per second



- **Positioning time** is
  - time to move disk arm to desired cylinder (**seek time**)
  - plus time for desired sector to rotate under the disk head (**rotational latency**)

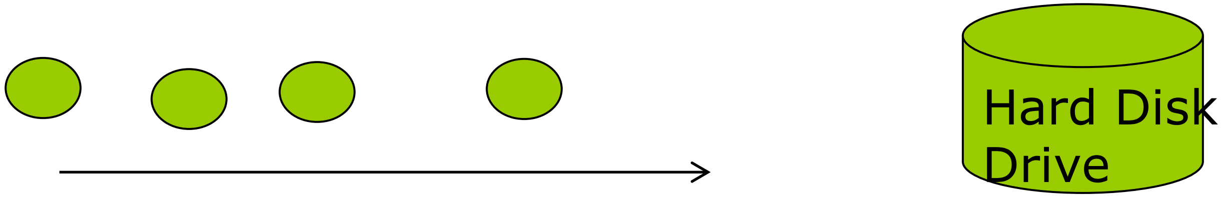
- **Transfer rate**: data flow speed between drive and computer
  - *Sustained bandwidth*: “average data transfer rate during a large transfer– that is the, number of bytes divided by transfer time”
    - data rate without positioning time
  - *Effective bandwidth*: average transfer rate including positioning time

Moving-head Disk Mechanism



Disk Scheduling: Objective

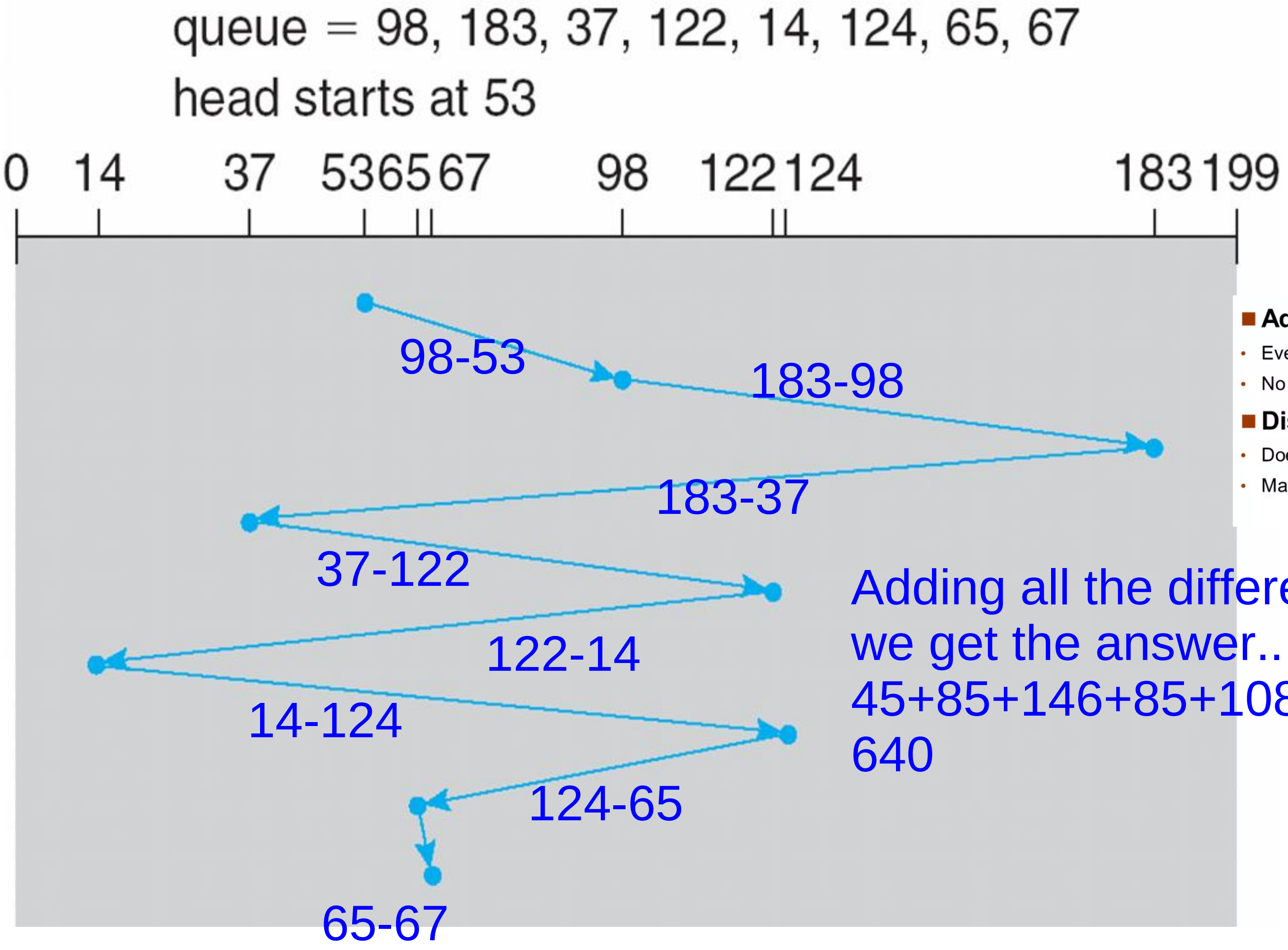
- Given a set of IO requests



- Coordinate disk access of multiple I/O requests for faster performance and reduced seek time.
  - Seek time ≈ seek distance
  - Measured by total head movement in terms of cylinders from one request to another.

FCFS (First Come First Serve)

total head movement: 640 cylinders for executing all requests

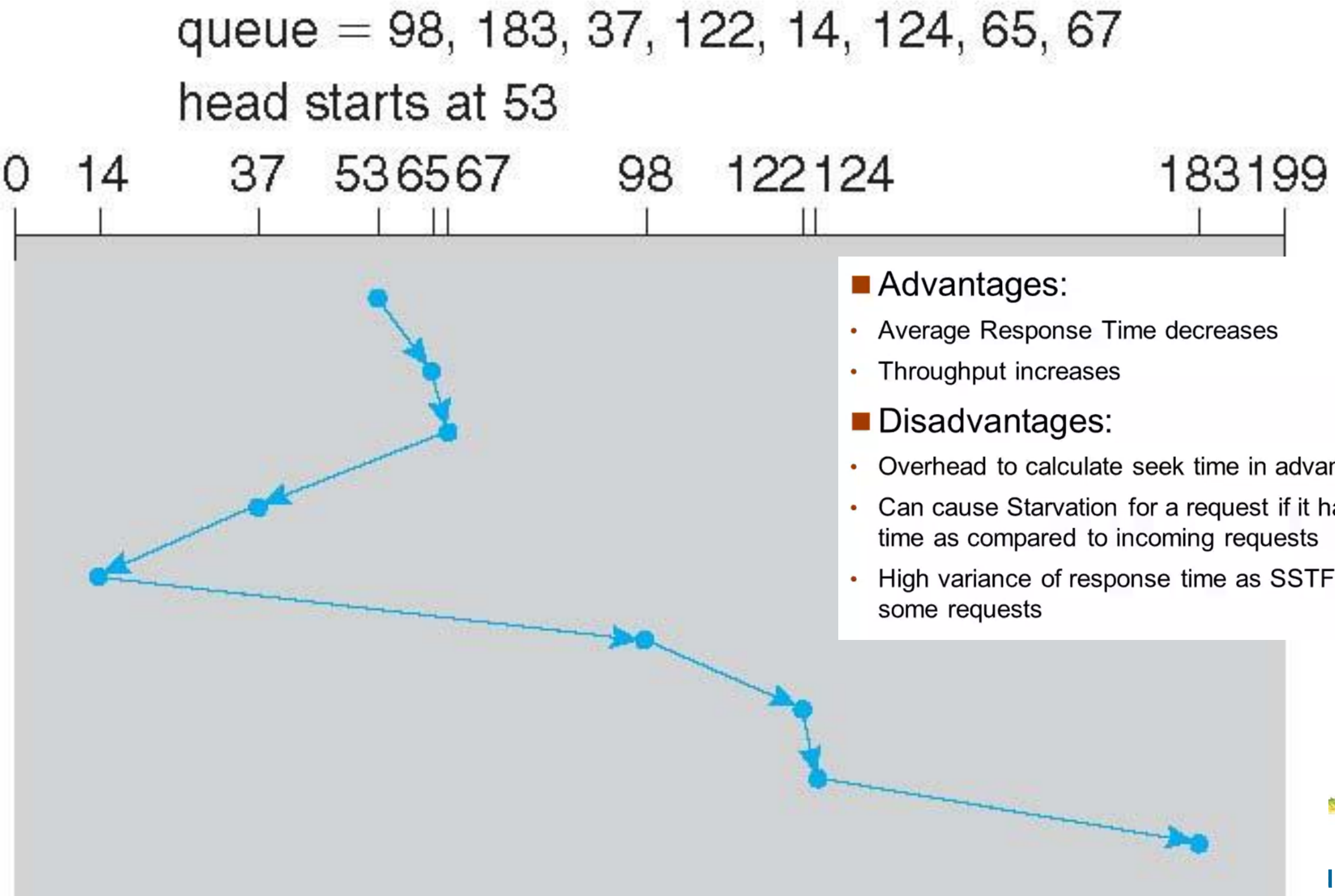


- **Advantages:**
  - Every request gets a fair chance
  - No indefinite postponement
- **Disadvantages:**
  - Does not try to optimize seek time
  - May not provide the best possible service

Adding all the differences we get the answer...
45+85+146+85+108+110+59+2 = 640

SSTF (Shortest Seek Time First)

- Selects the request with the minimum seek time from the current head position
- total head movement: 236 cylinders

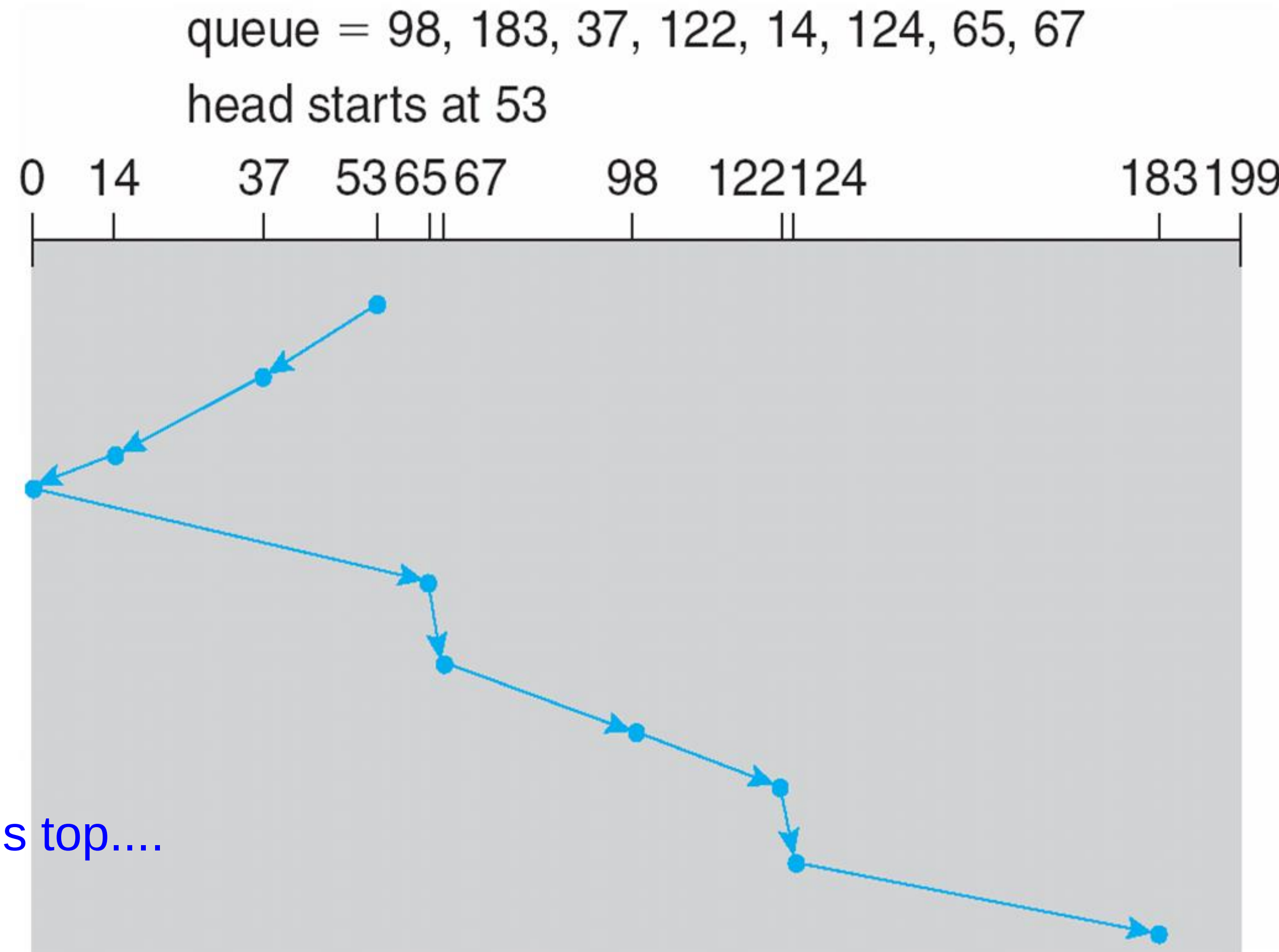


- **Advantages:**
  - Average Response Time decreases
  - Throughput increases
- **Disadvantages:**
  - Overhead to calculate seek time in advance
  - Can cause Starvation for a request if it has higher seek time as compared to incoming requests
  - High variance of response time as SSTF favours only some requests

SCAN: Elevator algorithm

- The disk arm starts at one end of the disk, and moves toward the other end, servicing requests until it gets to the other end of the disk, where the head movement is reversed and servicing continues.
- total head movement : 208 cylinders

Go to one side first then go to other side...



queue = 98, 183, 37, 122, 14, 124, 65, 67

head starts at 53

Here movement towards down first

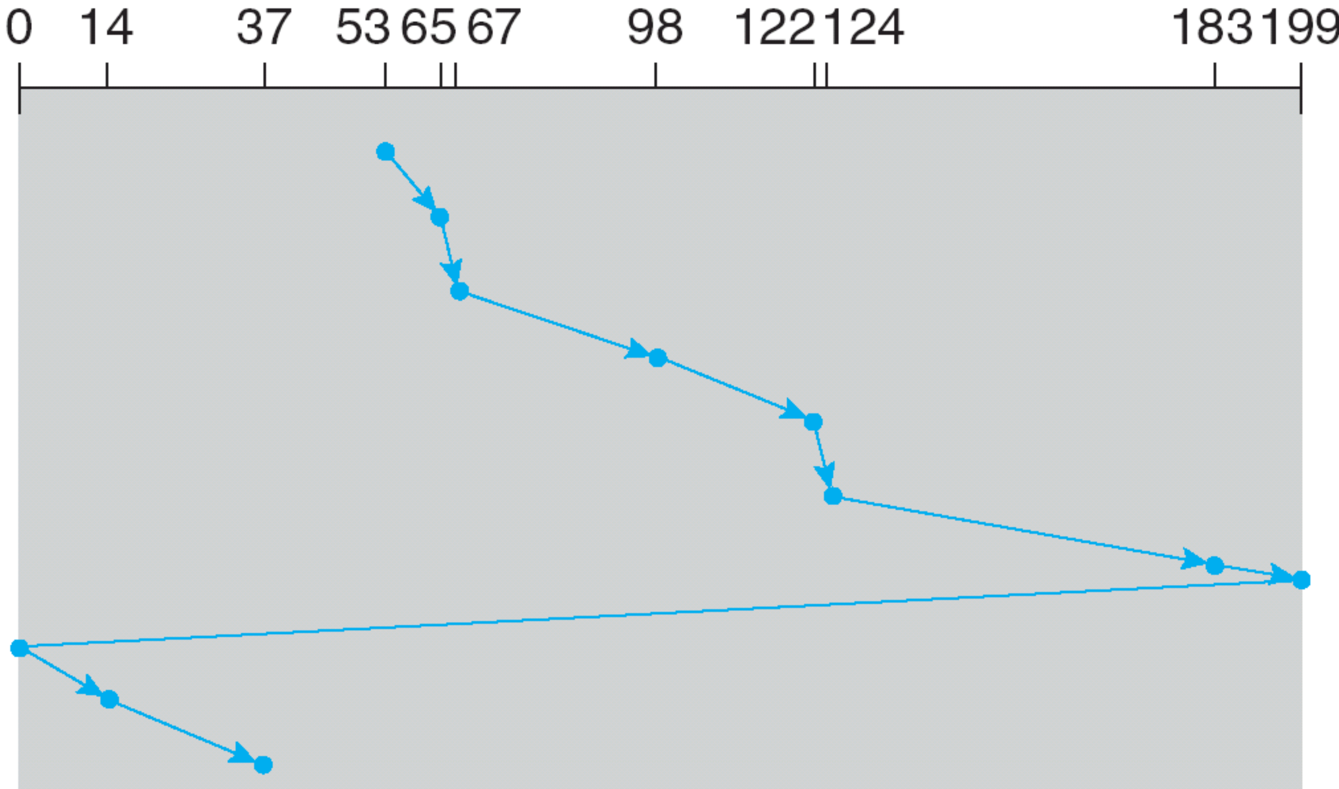
Then movement towards top....

C-SCAN (Circular-SCAN)

- Provides a more uniform wait time than SCAN by treating cylinders as a circular list.
- The head moves from one end of the disk to the other, servicing requests as it goes. When it reaches the other end, it immediately returns to the beginning of the disk, without servicing any requests on the return trip

queue = 98, 183, 37, 122, 14, 124, 65, 67

head starts at 53

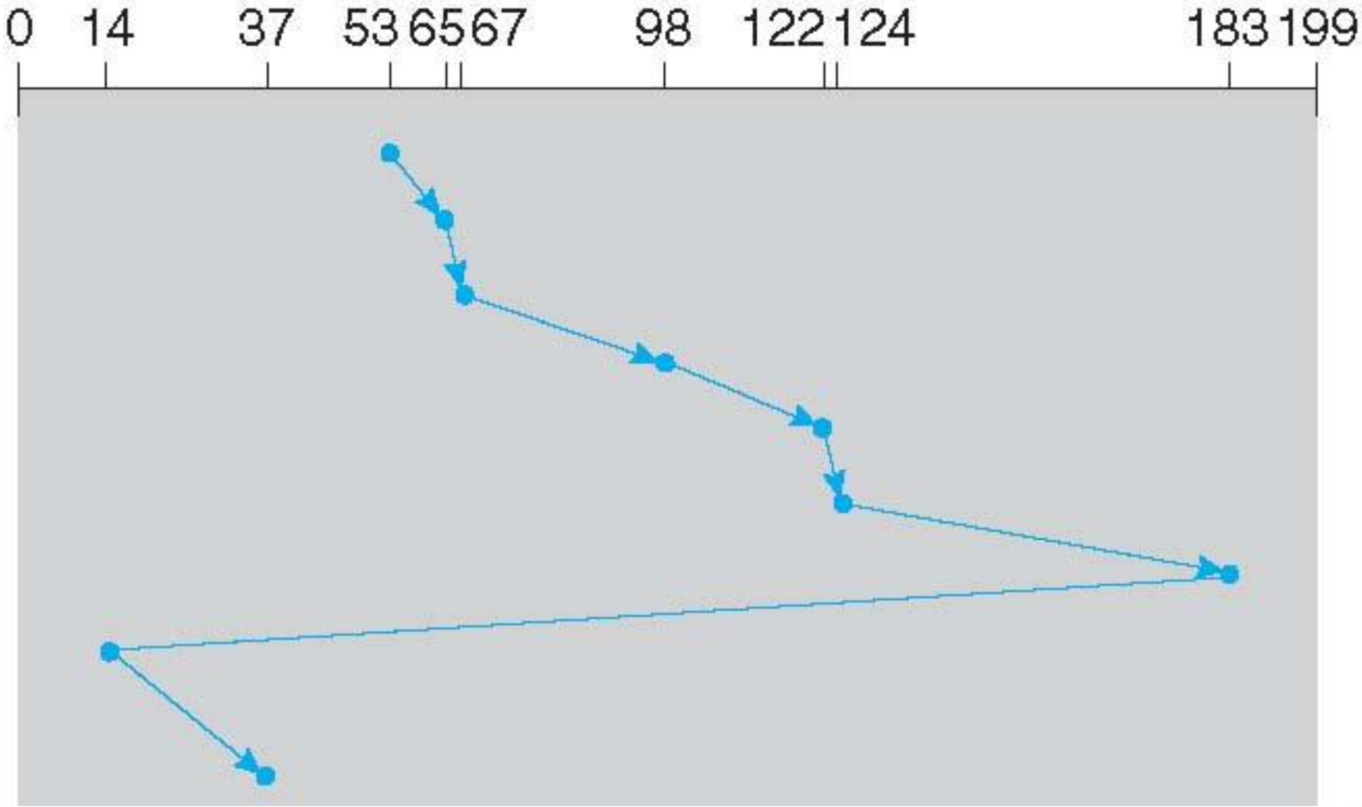


C-LOOK: A version of C-Scan

- Arm only goes as far as the last request in each direction, then reverses direction immediately, without first going all the way to the end of the disk

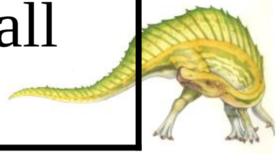
queue = 98, 183, 37, 122, 14, 124, 65, 67

head starts at 53



Scheduling Algorithms

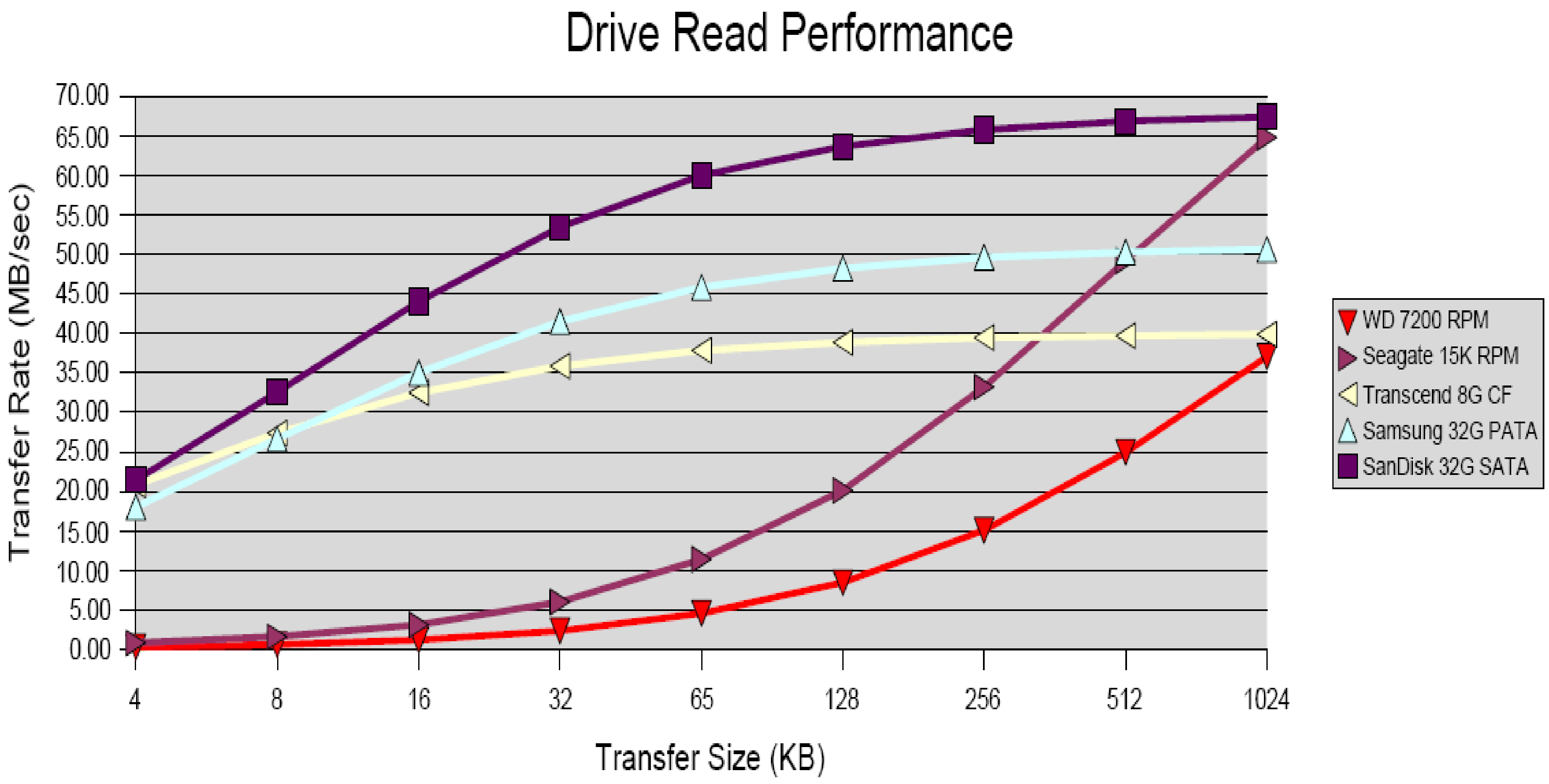
Algorithm Name	Description
FCFS	First-come first-served
SSTF	Shortest seek time first; process the request that reduces next seek time
SCAN (aka Elevator)	Move head from end to end (has a current direction)
C-SCAN	Only service requests in one direction (circular SCAN)
LOOK	Similar to SCAN, but donot go all the way to the end of the disk.
C-LOOK	Circular LOOK. Similar to C-SCAN, but donot go all the way to the end of the disk.



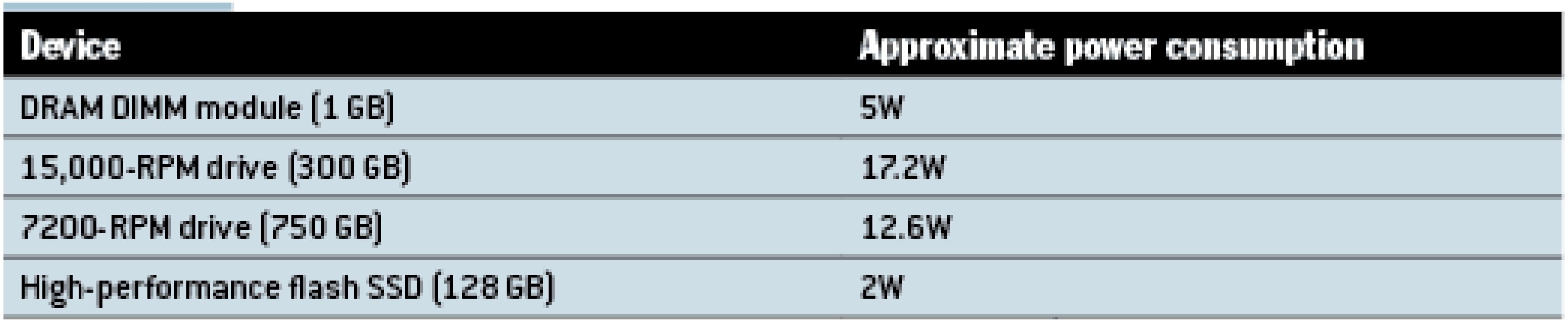
Selecting a Disk-Scheduling Algorithm

- Either SSTF or C-LOOK is a reasonable choice for the default algorithm
  - SSTF is common with its natural appeal (but it may lead to starvation issue).
  - C-LOOK is fair and efficient
  - SCAN and C-SCAN perform better for systems that place a heavy load on the disk
- Performance depends on the number and types of requests

Drive read performance



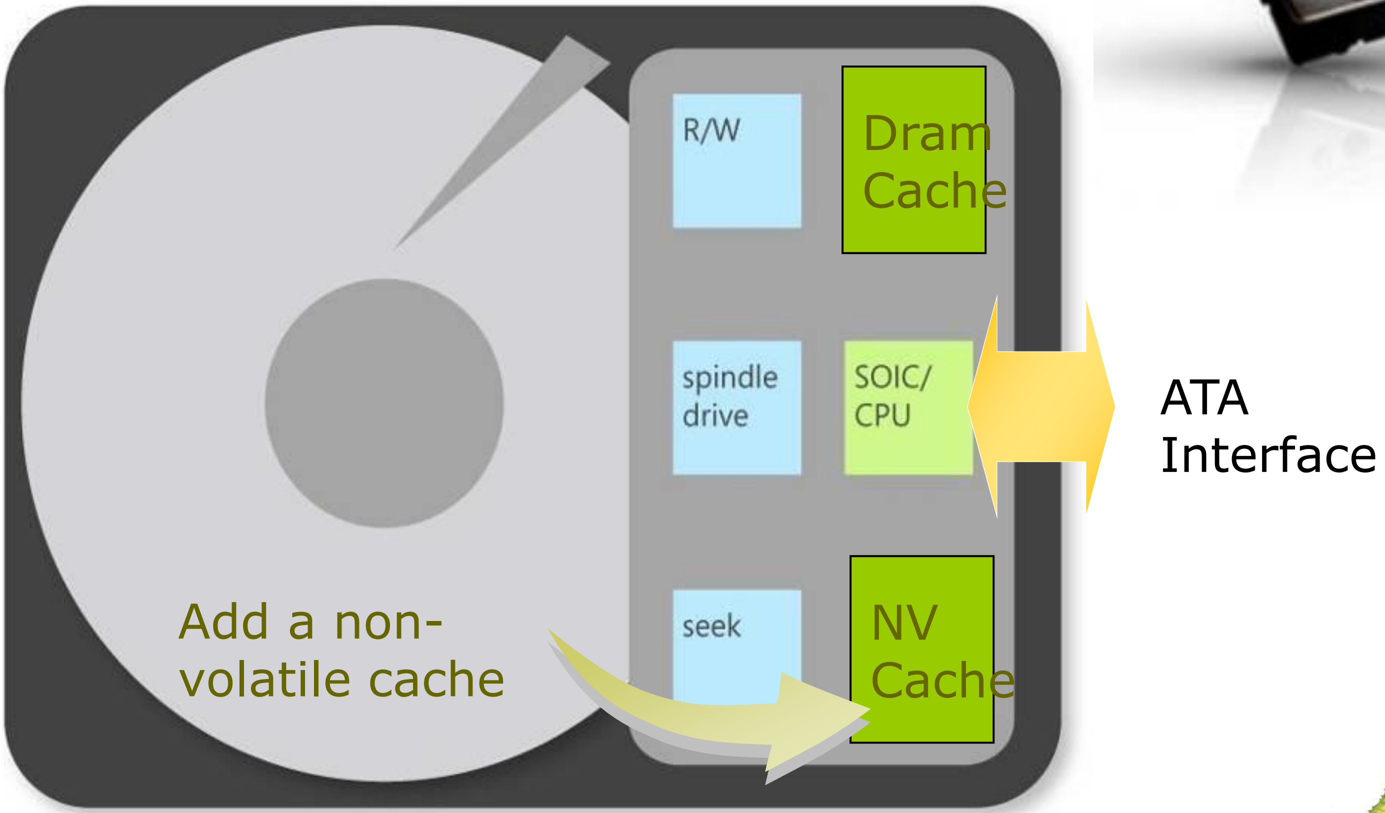
Power consumption



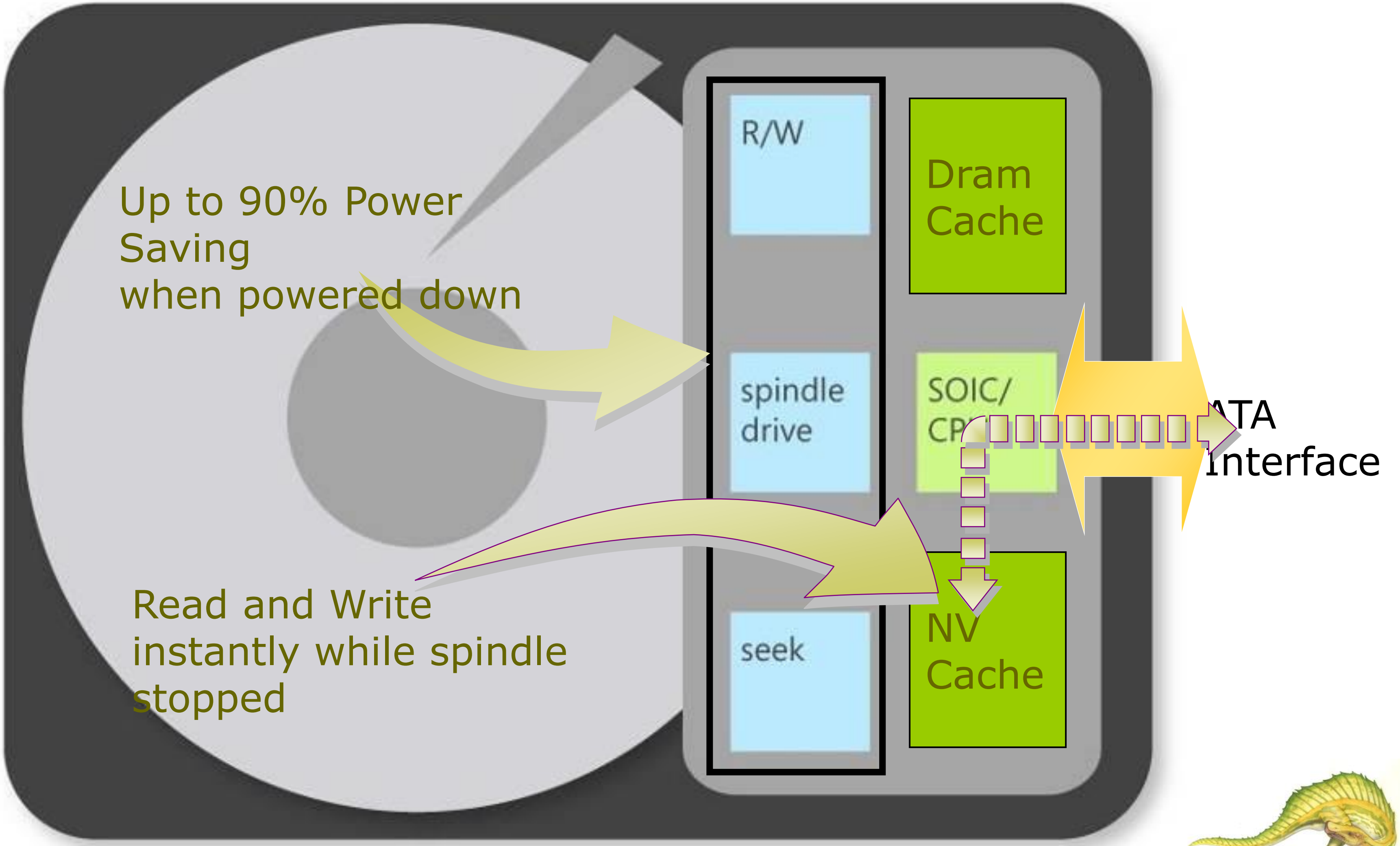
Device	Approximate power consumption
DRAM DIMM module (1 GB)	5W
15,000-RPM drive (300 GB)	17.2W
7200-RPM drive (750 GB)	12.6W
High-performance flash SSD (128 GB)	2W

Hybrid Disk Drive

- A hybrid disk uses a small SSD as a buffer for a larger drive
- All dirty blocks can be flushed to the actual hard drive based on:
  - Time, Threshold, Loss of power/computer shutdown



Hybrid Disk Drive Benefits



Up to 90% Power Saving
when powered down

Read and Write instantly while spindle stopped

RAID (Redundant Array of Inexpensive Disks)

- Multiple disk drives provide reliability via **redundancy**.

Increases the **mean time to failure**

- Hardware RAID with RAID controller vs software RAID

RAID (Cont.)

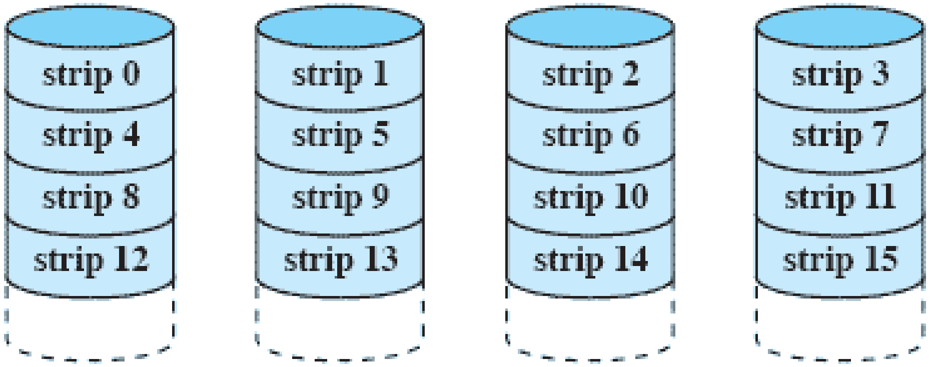
- RAID
  - multiple disks work cooperatively
  - Improve reliability by storing redundant data
  - Improve performance with disk **striping** (use a group of disks as one storage unit)
- RAID is arranged into six different levels
  - **Mirroring** (**RAID 1**) keeps duplicate of each disk
  - Striped mirrors (**RAID 1+0**) or mirrored stripes (**RAID 0+1**) provides high performance and high reliability
  - **Block interleaved parity** (**RAID 4, 5, 6**) uses much less redundancy

RAID

- Redundant Array of Independent Disks
- Set of physical disk drives viewed by the operating system as a single logical drive
- Data are distributed across the physical drives of an array
- Redundant disk capacity is used to store parity information which provides recoverability from disk failure

Disk striping is the process of dividing a body of data into blocks and spreading the data blocks across multiple storage devices, such as hard disks or solid-state drives (SSDs).

RAID 0 - Stripped

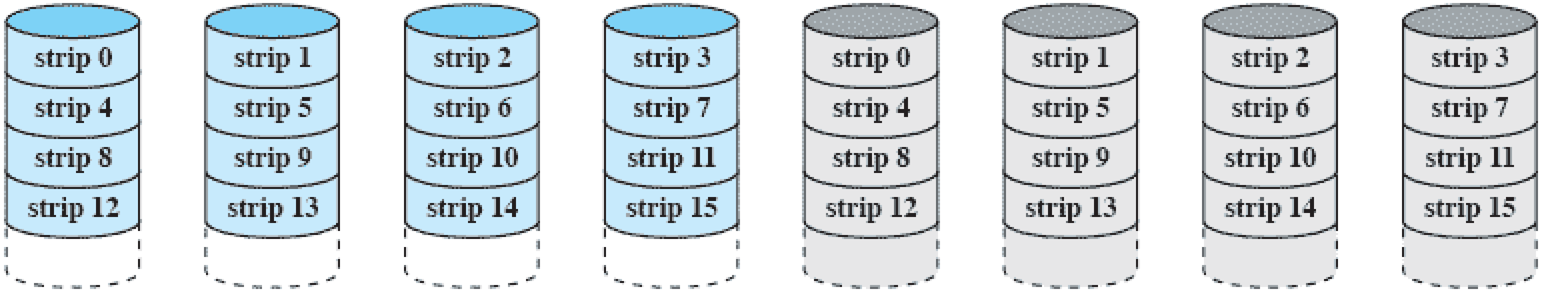


(a) RAID 0 (non-redundant)

- Not a true RAID – no redundancy
- This configuration has striping but no redundancy of data
- Disk failure is catastrophic
- Very fast due to parallel read/write

RAID 1 - Mirrored

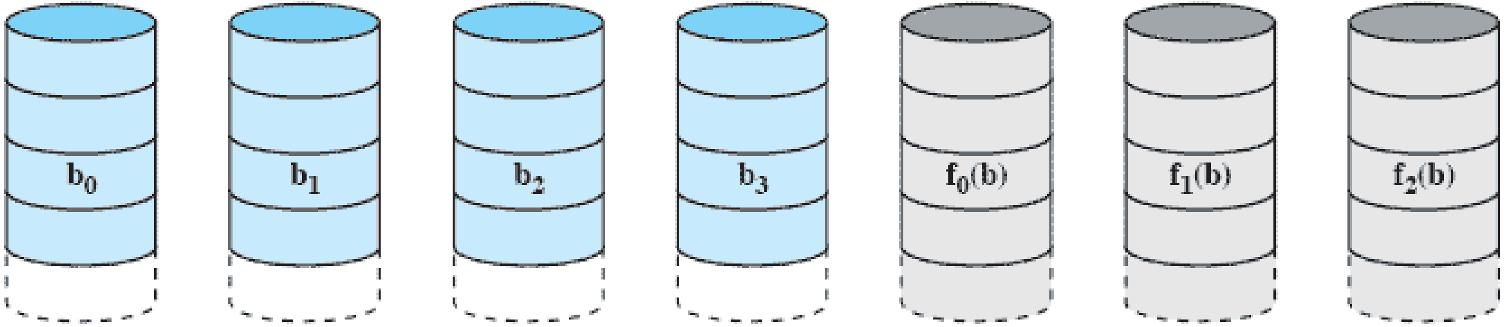
- Redundancy through duplication instead of parity.
- Read requests can made in parallel.
- Simple recovery from disk failure



(b) RAID 1 (mirrored)

RAID 2
(Using Hamming code)

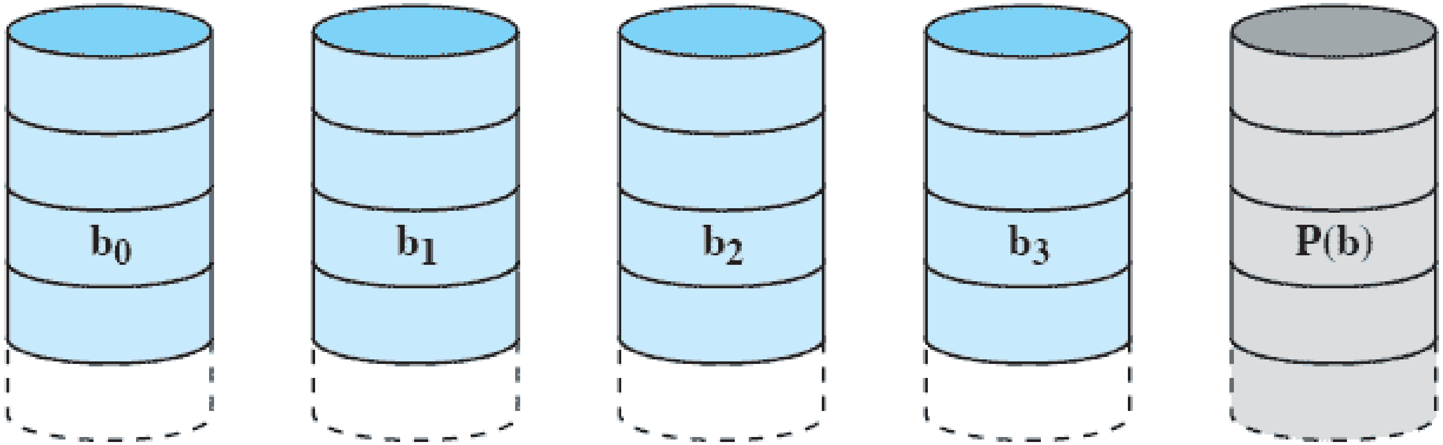
- Synchronised disk rotation
- Data stripping is used (extremely small)
- Hamming code used to correct single bit errors and detect double-bit errors



(c) RAID 2 (redundancy through Hamming code)

RAID 3
bit-interleaved parity

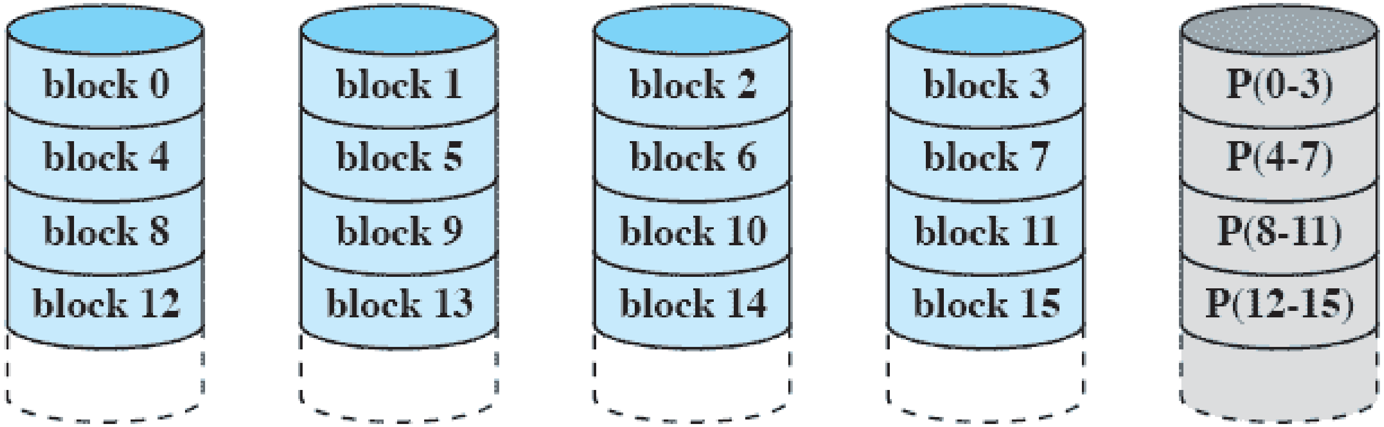
- Similar to RAID-2 but uses all parity bits stored on a single drive



(d) RAID 3 (bit-interleaved parity)

RAID 4
Block-level parity

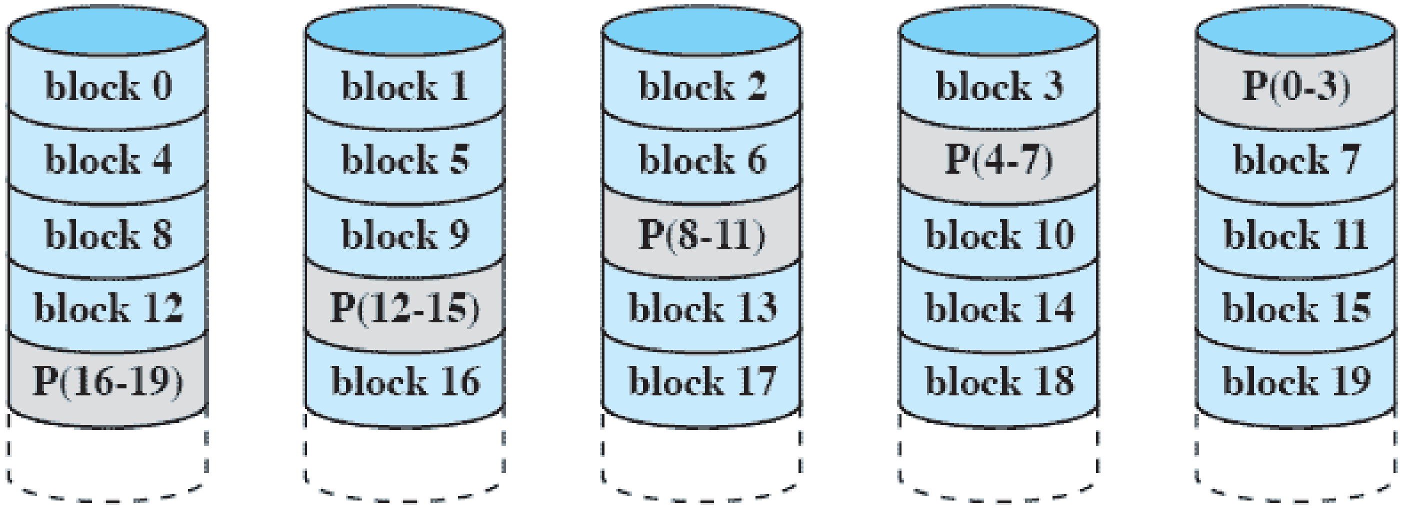
- A bit-by-bit parity strip is calculated across corresponding strips on each data disk
- The parity bits are stored in the corresponding strip on the parity disk.



(e) RAID 4 (block-level parity)

RAID 5
Block-level Distributed parity

- Similar to RAID-4 but distributing the parity bits across all drives

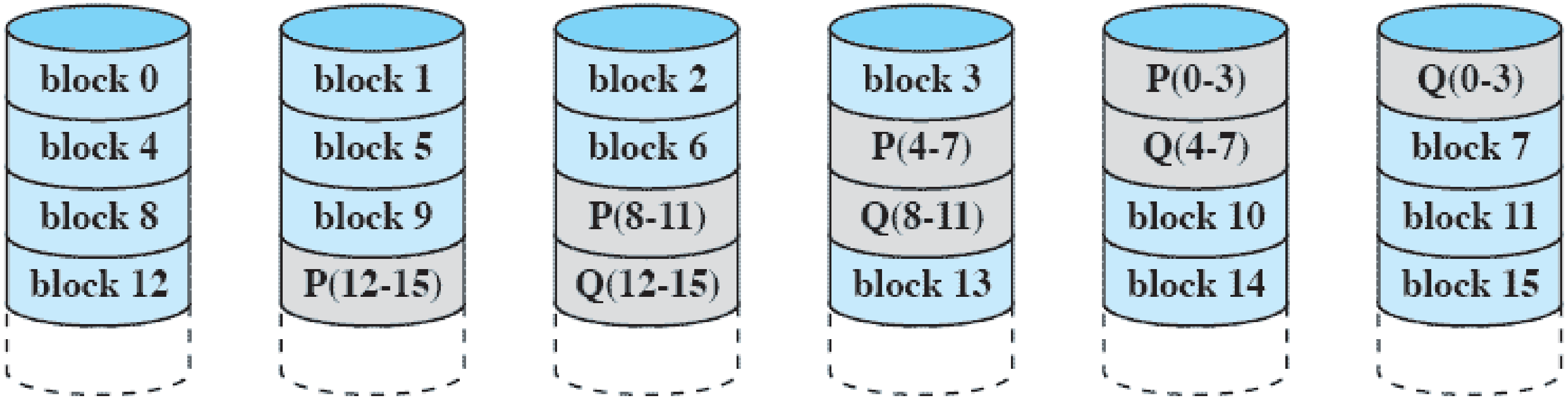


block 0	block 1	block 2	block 3	P(0-3)
block 4	block 5	block 6	P(4-7)	block 7
block 8	block 9	P(8-11)	block 10	block 11
block 12	P(12-15)	block 13	block 14	block 15
P(16-19)	block 16	block 17	block 18	block 19

(f) RAID 5 (block-level distributed parity)

RAID 6
Dual Redundancy

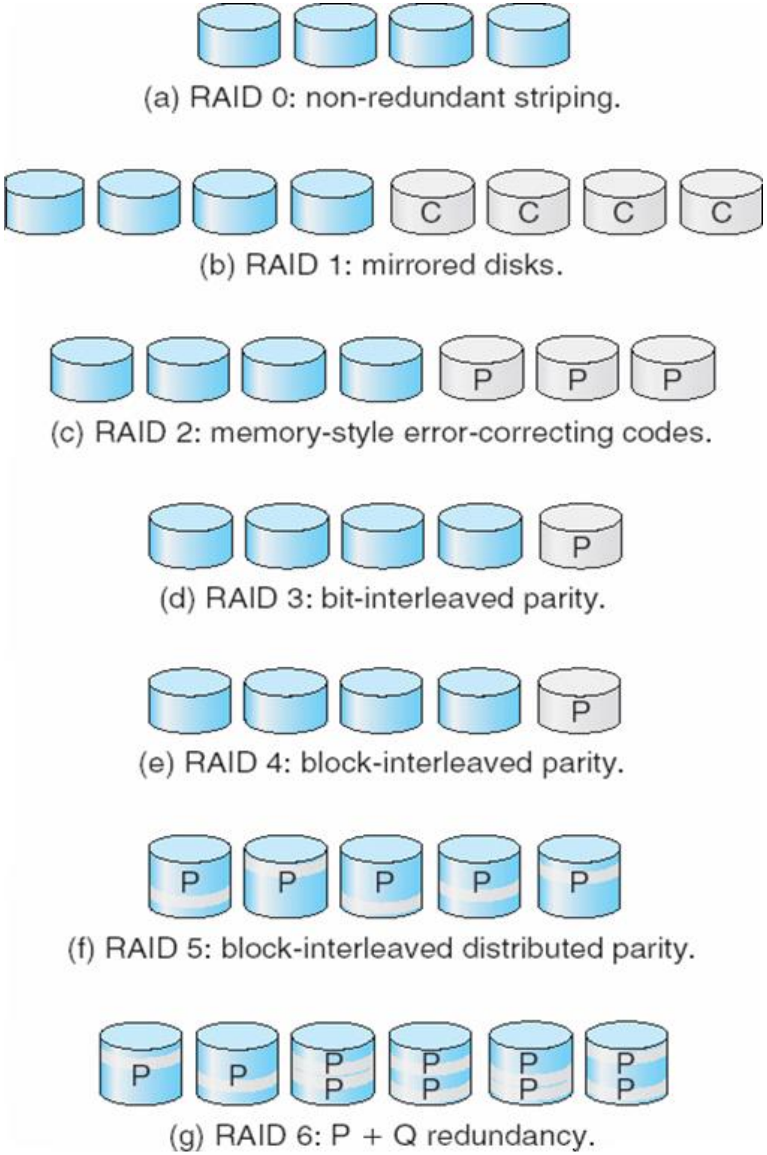
- Two different parity calculations are carried out
  - stored in separate blocks on different disks.
- Can recover from two disks failing



block 0	block 1	block 2	block 3	P(0-3)	Q(0-3)
block 4	block 5	block 6	P(4-7)	Q(4-7)	block 7
block 8	block 9	P(8-11)	Q(8-11)	block 10	block 11
block 12	P(12-15)	Q(12-15)	block 13	block 14	block 15

(g) RAID 6 (dual redundancy)

6 RAID Levels



(a) RAID 0: non-redundant striping.

(b) RAID 1: mirrored disks.

(c) RAID 2: memory-style error-correcting codes.

(d) RAID 3: bit-interleaved parity.

(e) RAID 4: block-interleaved parity.

(f) RAID 5: block-interleaved distributed parity.

(g) RAID 6: P + Q redundancy.