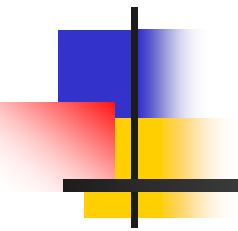
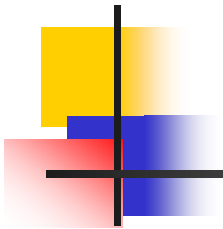


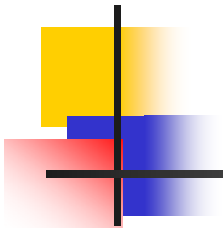
Mass-Storage Structure

Chapter 12



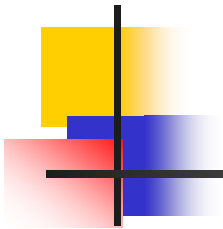
Objectives

- Describe the physical structure of secondary and tertiary storage devices and the resulting effects on the uses of the devices
- Explain the performance characteristics of mass-storage devices
- Discuss operating-system services provided for mass storage, including RAID and HSM



Chapter Outline

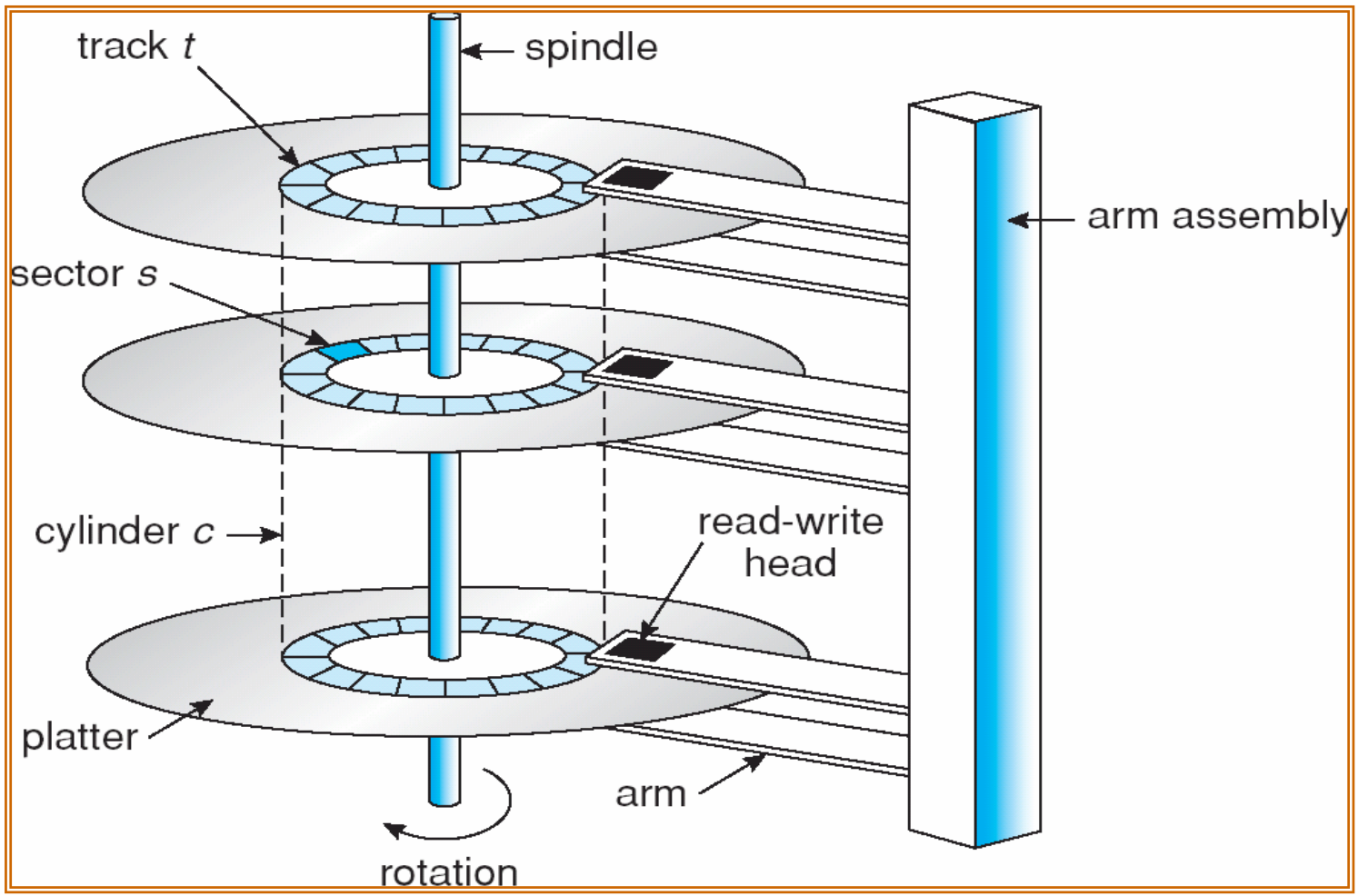
- Overview of Mass Storage Structure
- Disk Structure
- Disk Attachment
- Disk Scheduling
- Disk Management
- Swap-Space Management
- RAID Structure

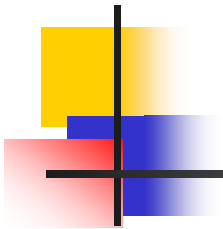


- Overview of Mass Storage Structure

- Magnetic disks provide bulk of secondary storage of modern computers
  - Drives rotate at 60 to 200 times per second
  - **Transfer rate** is rate at which data flow between drive and computer
  - **Positioning time** (**random-access time**) is time to move disk arm to desired cylinder (**seek time**) and time for desired sector to rotate under the disk head (**rotational latency**)
  - **Head crash** results from disk head making contact with the disk surface.
- Disks can be removable
- Drive attached to computer via **I/O bus**
  - Busses vary, including **EIDE, ATA, SATA, USB, Fibre Channel, SCSI**

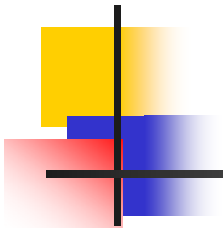
-- Moving-head Disk Mechanism





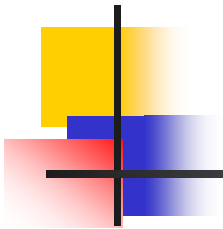
- Disk Structure

- Disk drives are addressed as large 1-dimensional arrays of *logical blocks*, where the logical block is the smallest unit of transfer.
- The 1-dimensional array of logical blocks is mapped into the sectors of the disk sequentially.
  - Sector 0 is the first sector of the first track on the outermost cylinder.
  - Mapping proceeds in order through that track, then the rest of the tracks in that cylinder, and then through the rest of the cylinders from outermost to innermost.



- Disk Attachment

- Computers access storages in 2 ways
  - Via I/O ports (or host-attached storage)
  - Network attached
  - Storage-area network

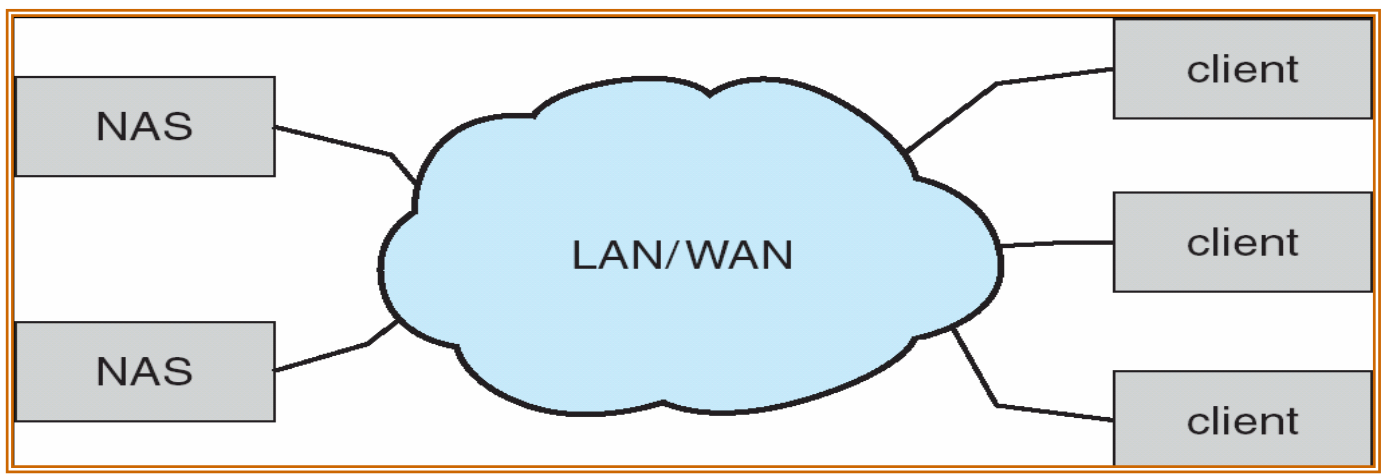


-- Host-attached Storages

- Host-attached storage accessed through I/O ports talking to I/O busses.
- There are many I/O bus architectures. Examples:
  - IDE
  - ATA
  - SATA
  - SCSI
  - Fiber Channel

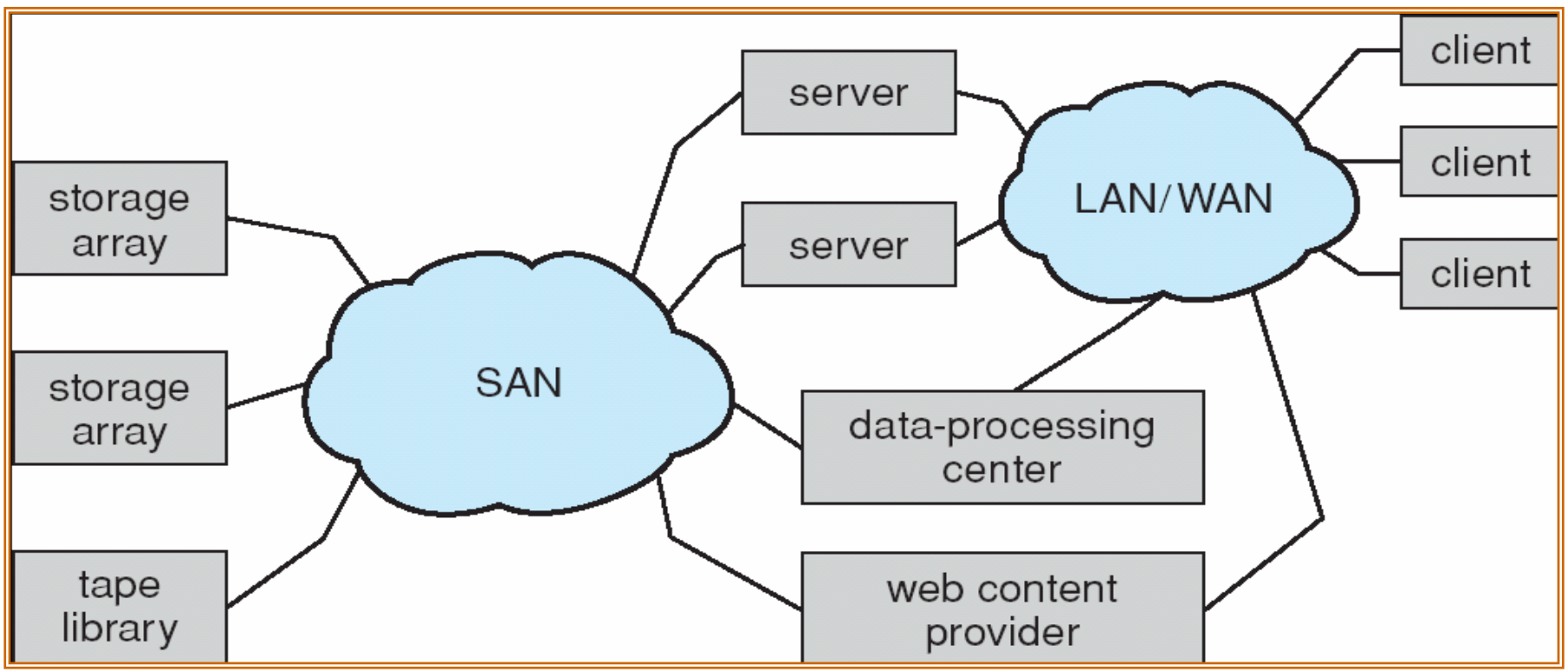
-- Network-Attached Storage

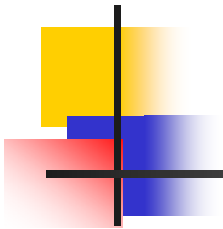
- Network-attached storage (**NAS**) is storage made available over a network rather than over a local connection (such as a bus)
- NFS and CIFS are common protocols
- Implemented via remote procedure calls (RPCs) between host and storage
- New iSCSI protocol uses IP network to carry the SCSI protocol



-- Storage Area Network

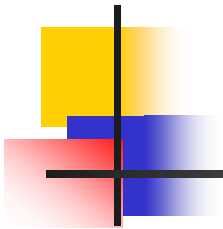
- Common in large storage environments (and becoming more common)
- Multiple hosts attached to multiple storage arrays - flexible





- Disk Scheduling ...

- The operating system is responsible for using hardware efficiently — for the disk drives, this means having a fast access time and disk bandwidth.
- Access time has two major components
  - **Seek time** is the time for the disk are to move the heads to the cylinder containing the desired sector.
  - **Rotational latency** is the additional time waiting for the disk to rotate the desired sector to the disk head.
- Minimize seek time
- Seek time ≈ seek distance
- Disk bandwidth is the total number of bytes transferred, divided by the total time between the first request for service and the completion of the last transfer.



... - Disk Scheduling

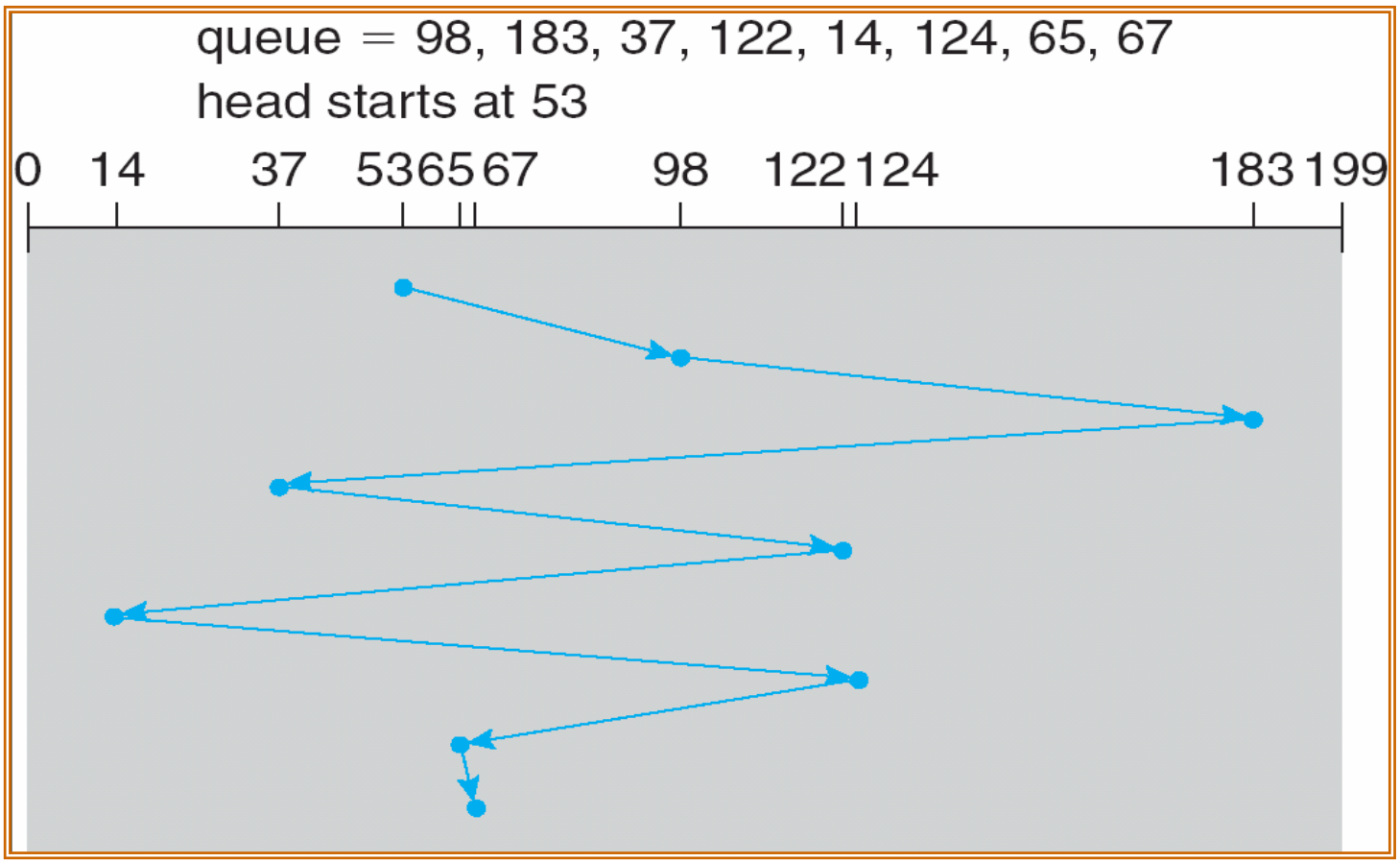
- Several algorithms exist to schedule the servicing of disk I/O requests.
  - FCFS
  - SSTF
  - SCAN
  - C-SCAN
  - LOOK
  - C-LOOK
- We illustrate them with a request queue (0-199).

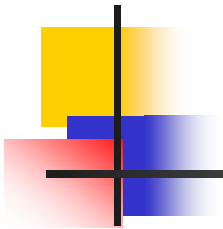
  98, 183, 37, 122, 14, 124, 65, 67

  Head pointer 53

-- FCFS

Illustration shows total head movement of 640 cylinders.

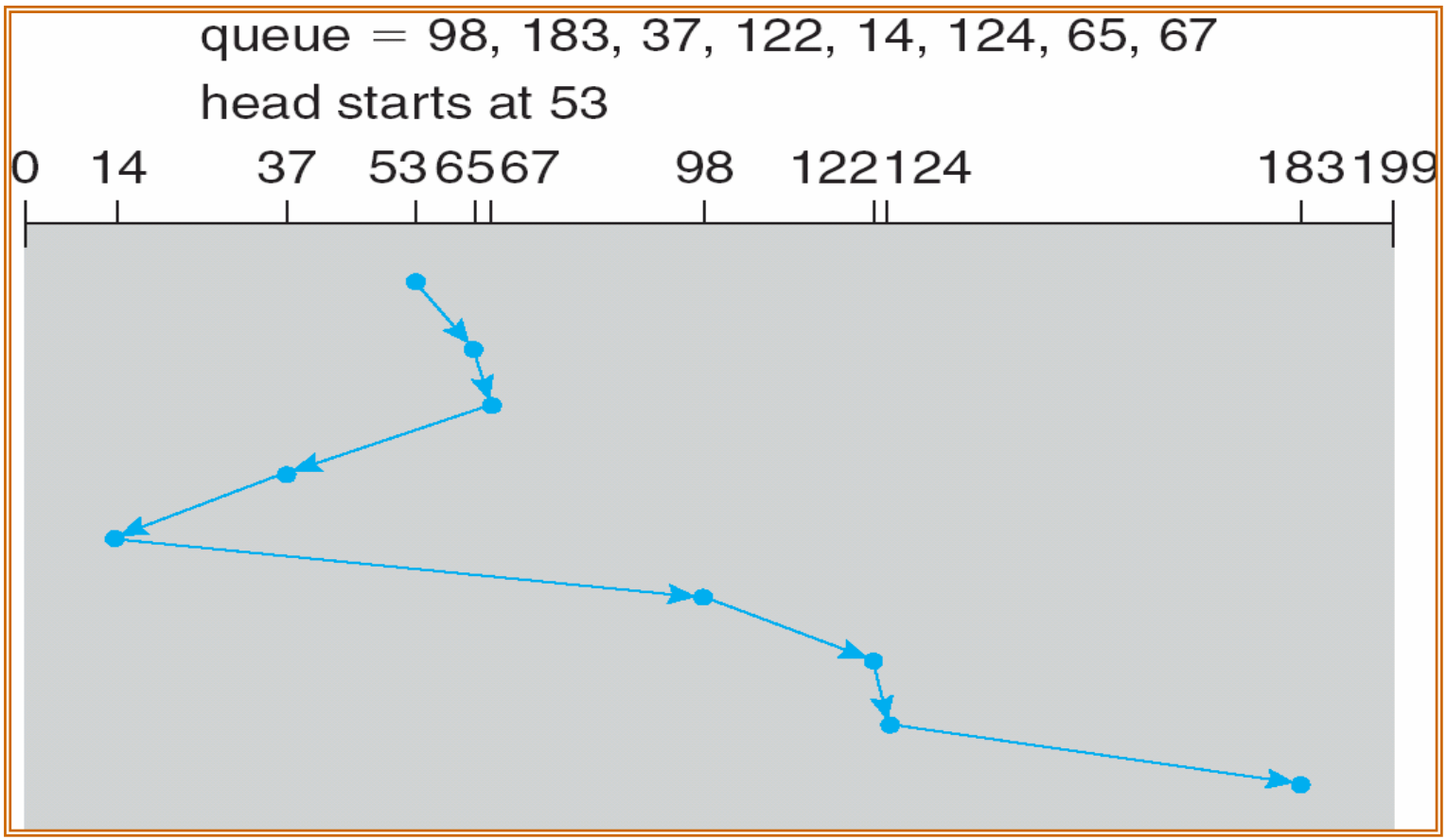


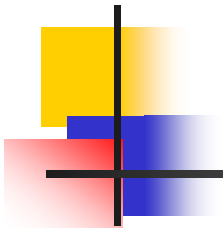


-- SSTF ...

- Selects the request with the minimum seek time from the current head position.
- SSTF scheduling is a form of SJF scheduling; may cause starvation of some requests.
- Illustration shows total head movement of 236 cylinders.

... -- SSTF

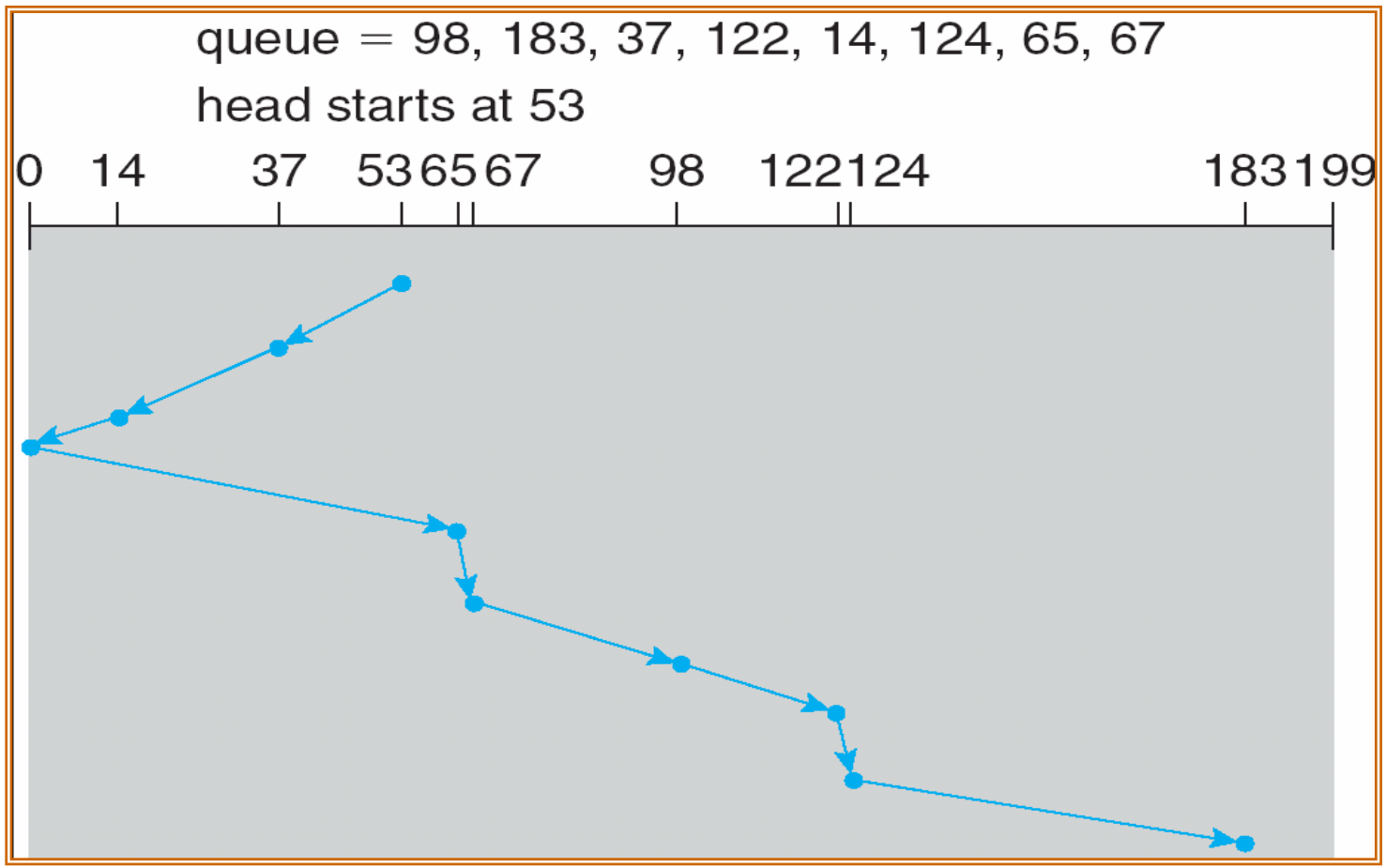

queue = 98, 183, 37, 122, 14, 124, 65, 67
head starts at 53
0
14
37
53
65
67
98
122
124
183
199

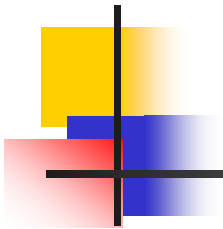


-- SCAN ...

- 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.
- Sometimes called the **elevator algorithm**.
- Illustration shows total head movement of 208 cylinders.

... -- SCAN

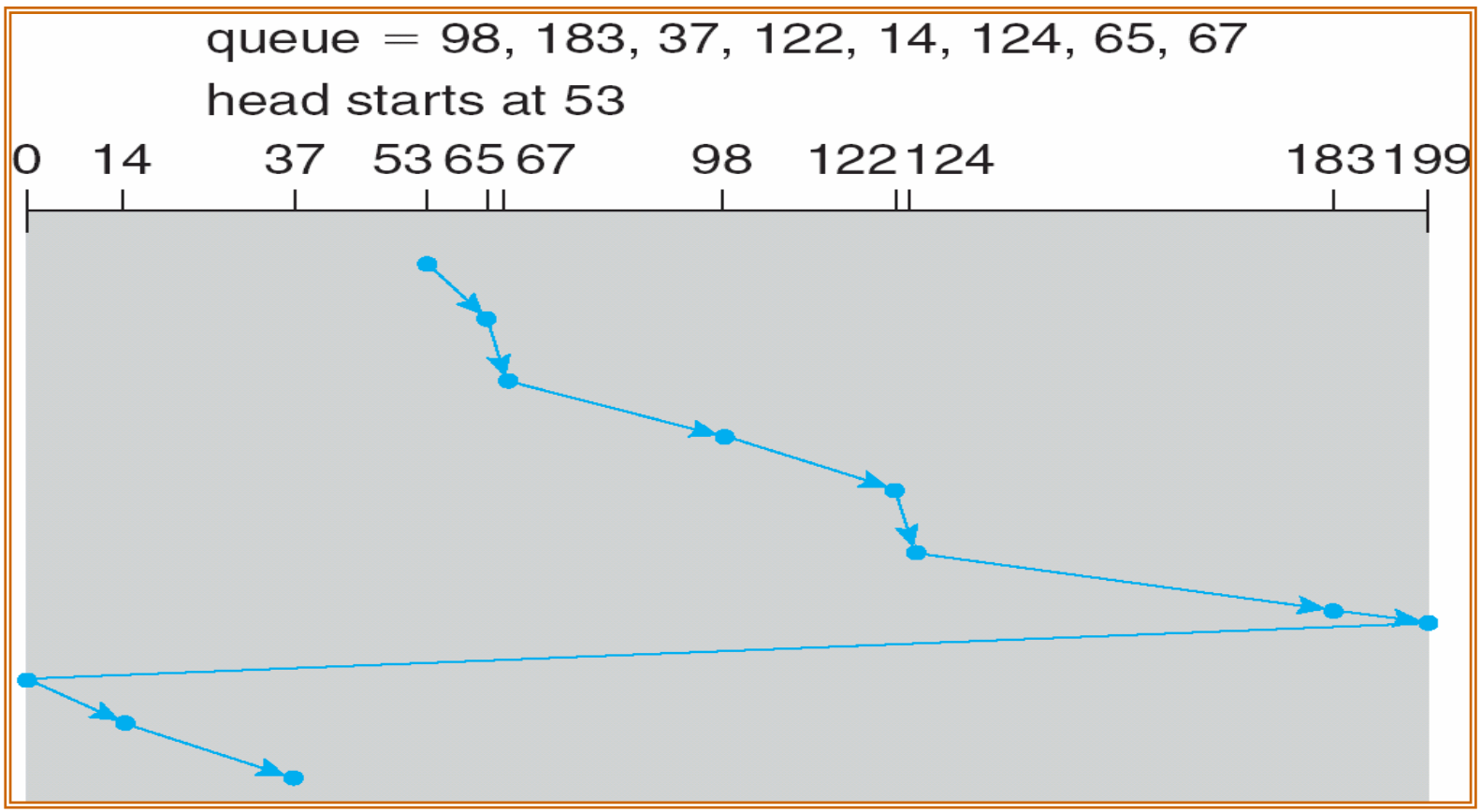

queue = 98, 183, 37, 122, 14, 124, 65, 67
head starts at 53
0
14
37
53
65
67
98
122
124
183
199

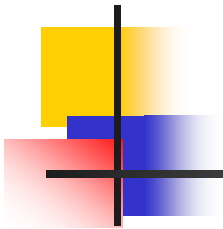


-- C-SCAN ...

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

... -- C-SCAN

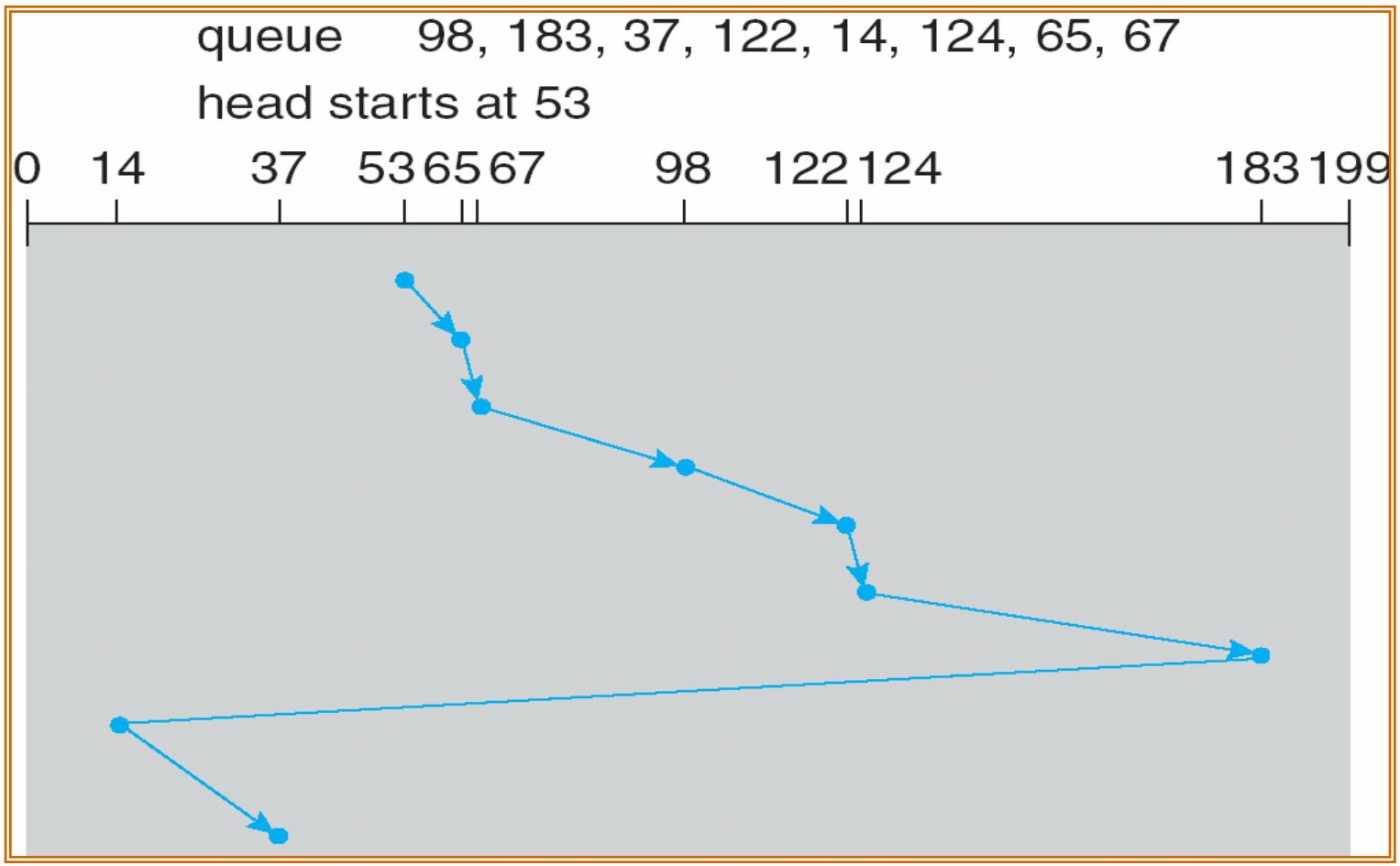

queue = 98, 183, 37, 122, 14, 124, 65, 67
head starts at 53
0
14
37
53
65
67
98
122
124
183
199

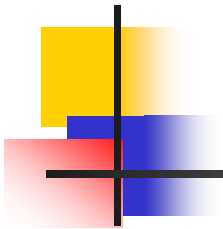


-- C-LOOK ...

- 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.

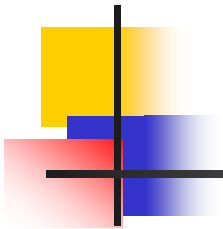
... -- C-LOOK


queue 98, 183, 37, 122, 14, 124, 65, 67
head starts at 53
0 14 37 53 65 67 98 122 124 183 199



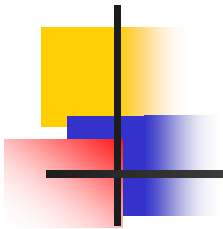
-- Selecting a Disk-Scheduling Algorithm

- SSTF is common and has a natural appeal
- 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.
- Requests for disk service can be influenced by the file-allocation method.
- The disk-scheduling algorithm should be written as a separate module of the operating system, allowing it to be replaced with a different algorithm if necessary.
- Either SSTF or LOOK is a reasonable choice for the default algorithm.



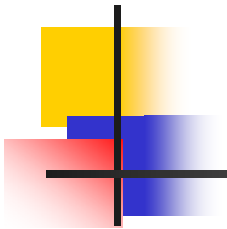
- Disk Management

- **Low-level formatting**, or **physical formatting** — Dividing a disk into sectors that the disk controller can read and write.
- To use a disk to hold files, the operating system still needs to record its own data structures on the disk.
  - **Partition** the disk into one or more groups of cylinders.
  - **Logical formatting** or “making a file system”.
- Boot block initializes system.
  - The bootstrap is stored in ROM.
  - *Bootstrap loader* program.
- Methods such as **sector sparing** used to handle bad blocks.



- Swap-Space Management

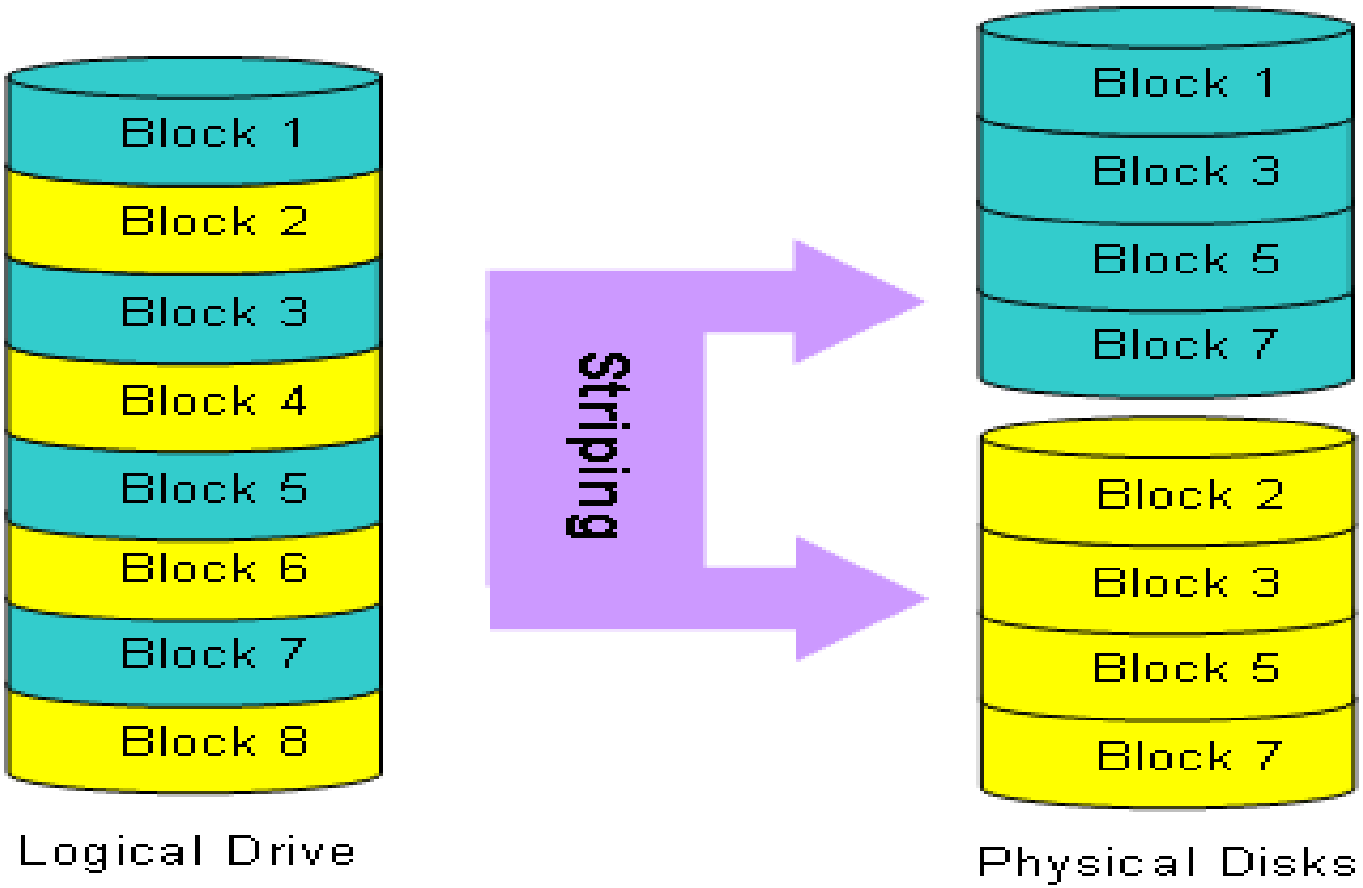
- Swap-space — Virtual memory uses disk space as an extension of main memory.
- Swap-space can be carved out of the normal file system,or, more commonly, it can be in a separate disk partition.
- Swap-space management
  - 4.3BSD allocates swap space when process starts; holds *text segment* (the program) and *data segment.*
  - Kernel uses *swap maps* to track swap-space use.
  - Solaris 2 allocates swap space only when a page is forced out of physical memory, not when the virtual memory page is first created.



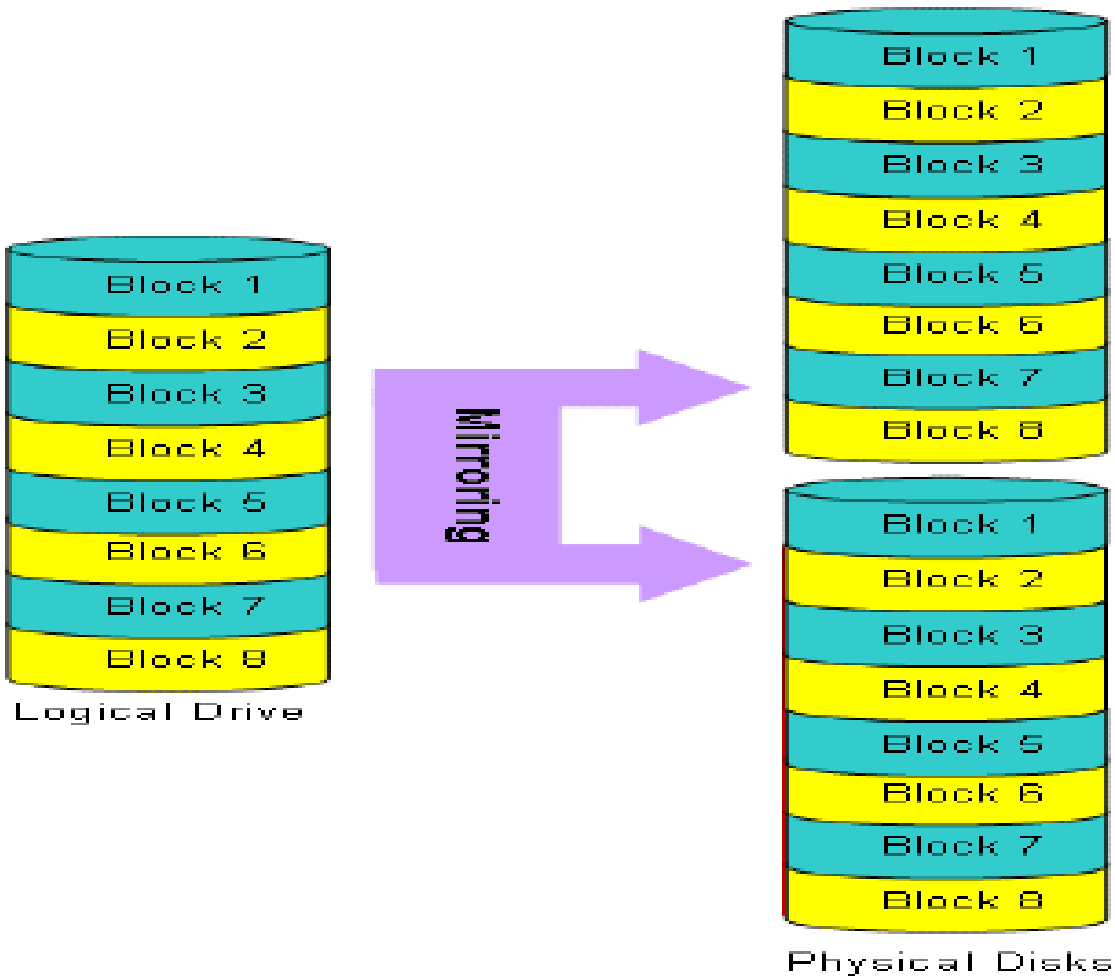
- RAID Structure

- **RAID** – multiple disk drives provides **reliability** via **redundancy**.
- RAID is arranged into six different levels.
- Several improvements in disk-use techniques involve the use of multiple disks working cooperatively.
- **Disk striping** uses a group of disks as one storage unit.
- RAID schemes improve performance and improve the reliability of the storage system by storing redundant data.
  - **Mirroring** or **shadowing** keeps duplicate of each disk.
  - **Block interleaved parity** uses much less redundancy.

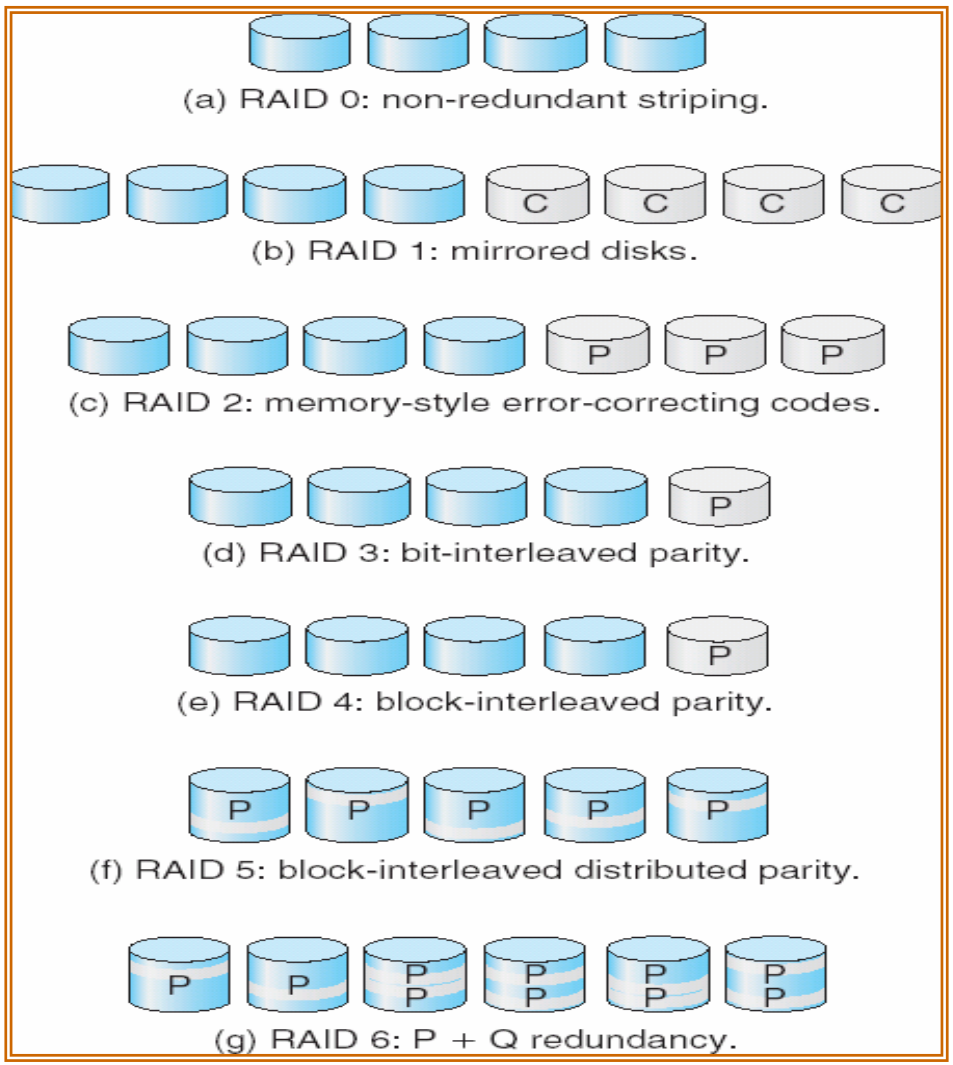
-- Striping


Block 1
Block 2
Block 3
Block 4
Block 5
Block 6
Block 7
Block 8
Logical Drive
striping
Block 1
Block 3
Block 5
Block 7
Block 2
Block 3
Block 5
Block 7
Physical Disks

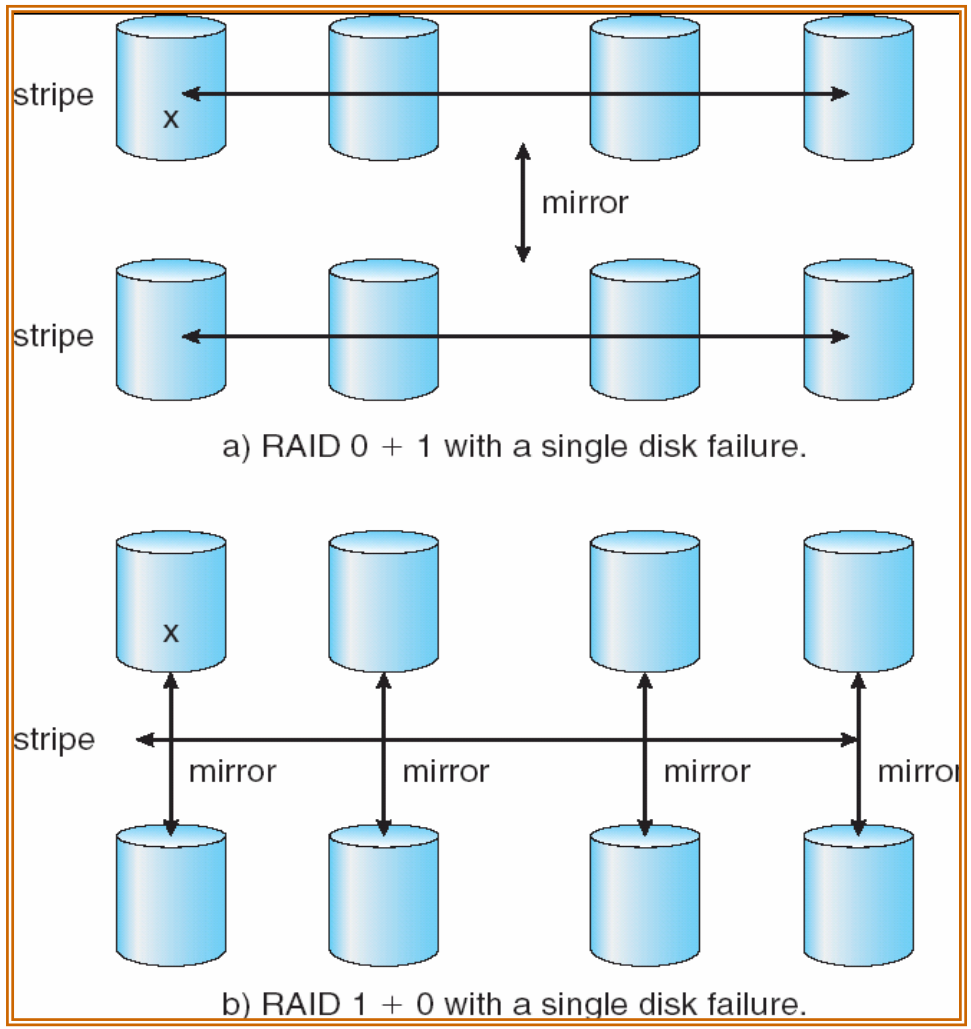
-- Mirroring


Block 1
Block 2
Block 3
Block 4
Block 5
Block 6
Block 7
Block 8
Logical Drive
Mirroring
Block 1
Block 2
Block 3
Block 4
Block 5
Block 6
Block 7
Block 8
Block 1
Block 2
Block 3
Block 4
Block 5
Block 6
Block 7
Block 8
Physical Disks

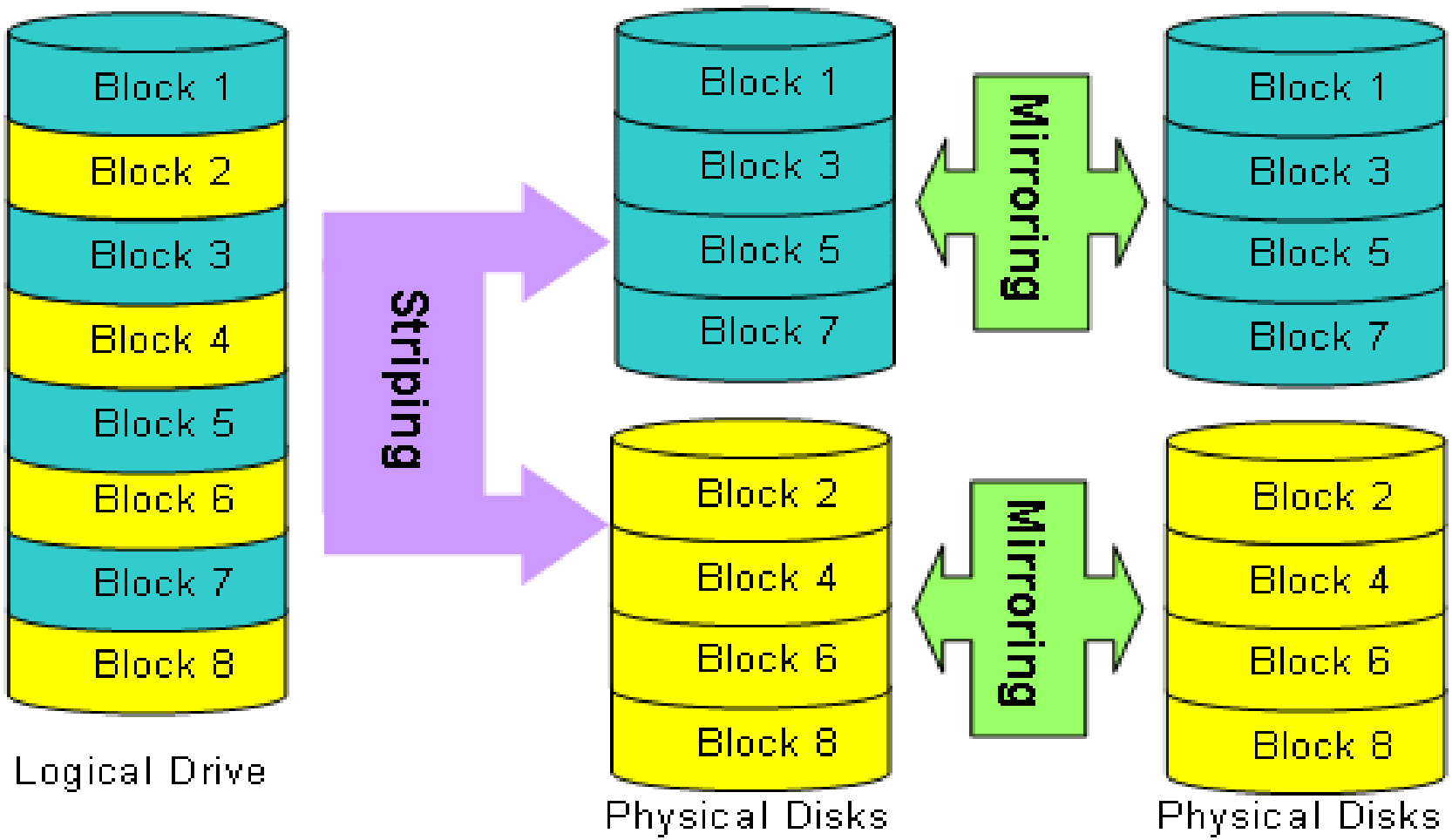
-- RAID Levels ...

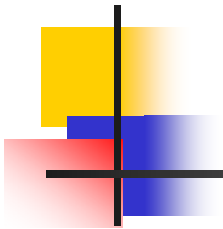

(a) RAID 0: non-redundant striping.
C
(b) RAID 1: mirrored disks.
P
(c) RAID 2: memory-style error-correcting codes.
(d) RAID 3: bit-interleaved parity.
(e) RAID 4: block-interleaved parity.
(f) RAID 5: block-interleaved distributed parity.
(g) RAID 6: P + Q redundancy.

--- RAID (0 + 1) and (1 + 0)



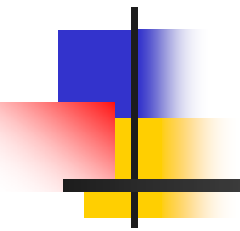
-- Stripping the Mirroring (RAID 0+1)





... -- RAID Levels

- RAID level 0 used in high-performance applications where data loss is not critical.
- RAID 1 used in applications that requires high reliability with fast recovery
- RAID (0+1) and (1+0) are Used where both performance and reliability are important as in small data bases.
- RAID 5 is often preferred to RAID 1 (due to its high overhead) for storing large volumes of data.
- RAID 6 offer better reliability than level 5.
- RAID concepts used also with tapes and data broadcasting over wireless systems.



End of Chapter 12