Operating Systems
Input/Output

Outline

- ❑ I/O devices
- ❑ Disk Arm Scheduling Algorithms
- ❑ RAID

I/O Devices

- Block devices
  - Stores information in fixed-size blocks
  - Transfers are in units of entire blocks
  - Hard disk, USB key
- Character devices
  - Delivers or accepts stream of characters, without regard to block structure
  - Not addressable, does not have any *seek* operation
  - Printer, keyboard

Differences in I/O Devices

- Unit of transfer
  - Data may be transferred as a stream of bytes for a terminal or in larger blocks for a disk

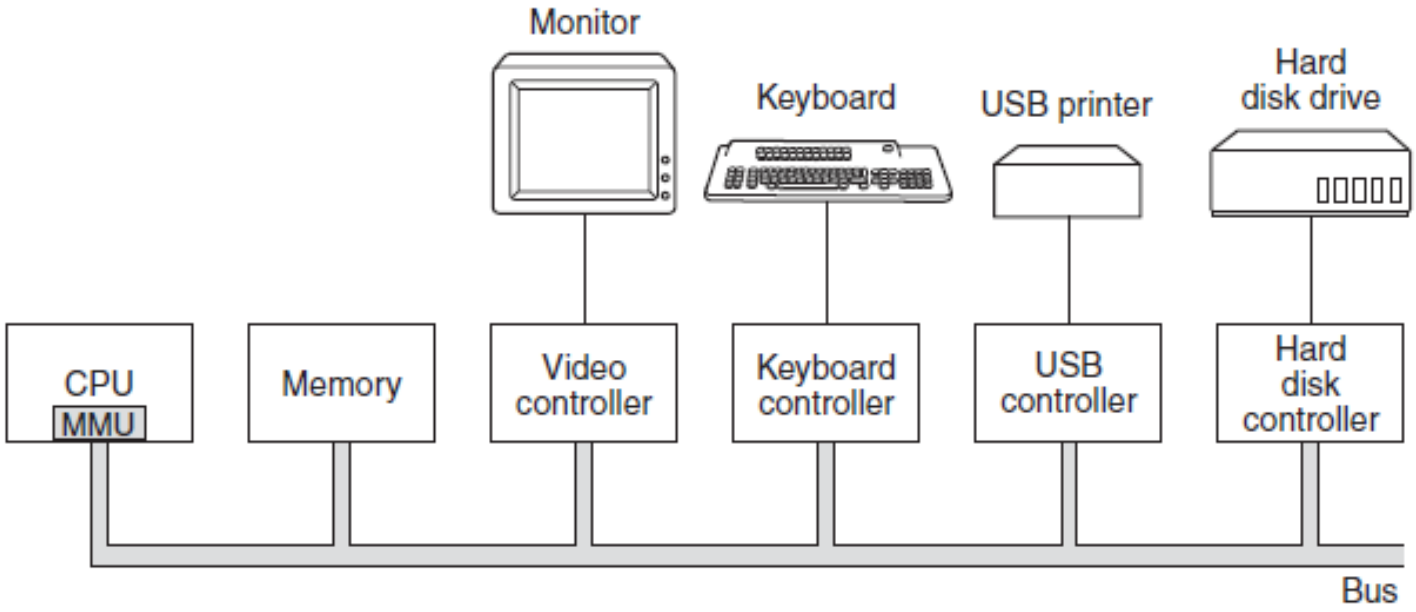
I/O Devices

Device	Data rate
Keyboard	10 bytes/sec
Mouse	100 bytes/sec
56K modem	7 KB/sec
Scanner at 300 dpi	1 MB/sec
Digital camcorder	3.5 MB/sec
4x Blu-ray disc	18 MB/sec
802.11n Wireless	37.5 MB/sec
USB 2.0	60 MB/sec
FireWire 800	100 MB/sec
Gigabit Ethernet	125 MB/sec
SATA 3 disk drive	600 MB/sec
USB 3.0	625 MB/sec
SCSI Ultra 5 bus	640 MB/sec
Single-lane PCIe 3.0 bus	985 MB/sec
Thunderbolt 2 bus	2.5 GB/sec
SONET OC-768 network	5 GB/sec

Figure 5-1. Some typical device, network, and bus data rates.

Tanenbaum & Bo,Modern Operating Systems:4th ed., (c) 2013 Prentice-Hall, Inc. All rights reserved.

Device Controllers



I/O unit often consist of

- device controller (an electronic component)
- device itself (a mechanical component)

Device Controllers

Each controller has

- few registers
  - By writing to these registers, the OS can command to perform some action
- a data buffer
  - OS can read or write

Direct Memory Access

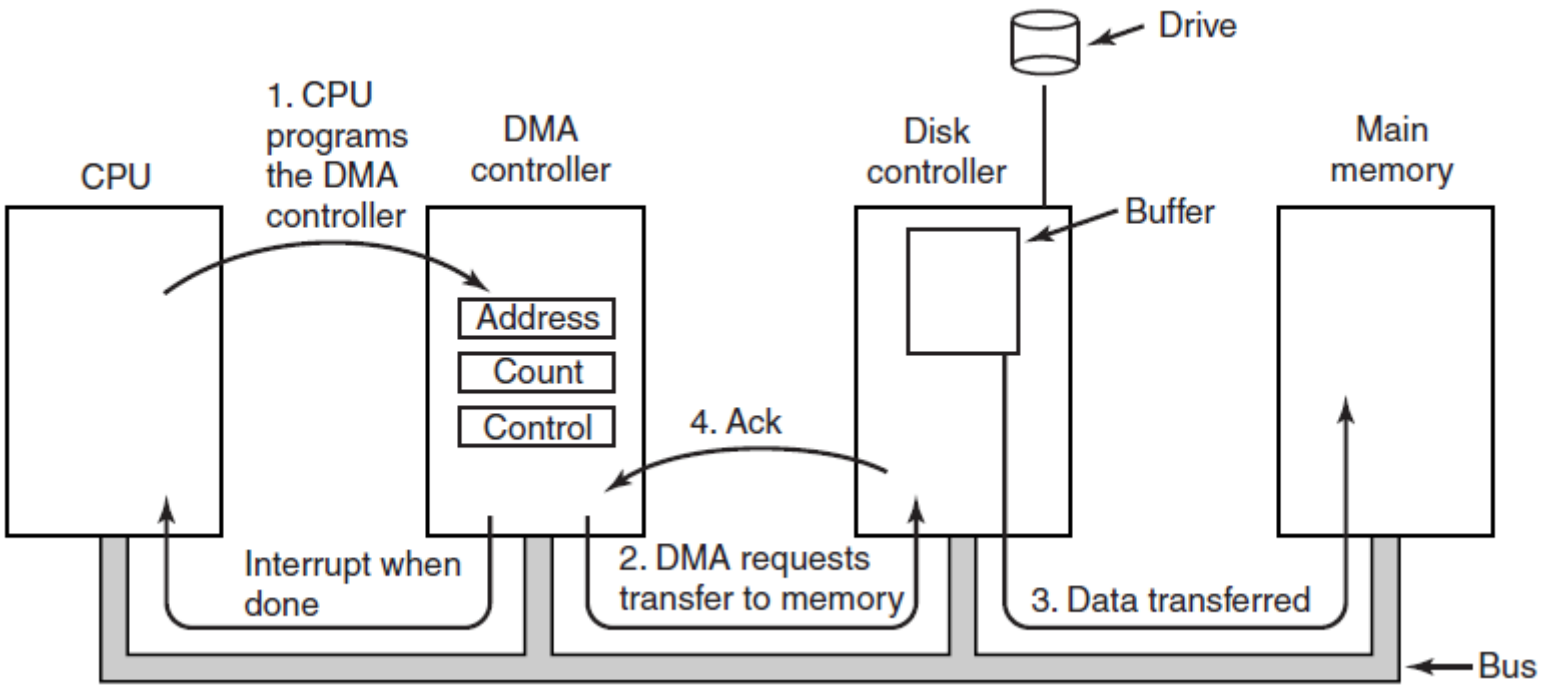
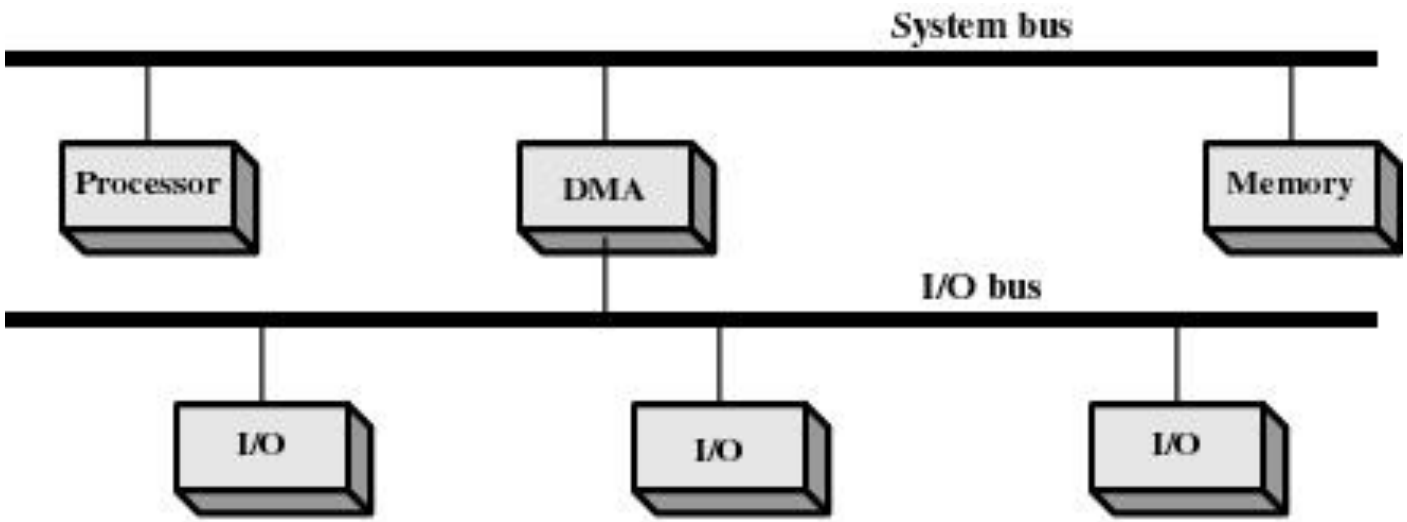


Figure 5-4. Operation of a DMA transfer.

Tanenbaum & Bo,Modern Operating Systems:4th ed., (c) 2013 Prentice-Hall, Inc. All rights reserved.

DMA



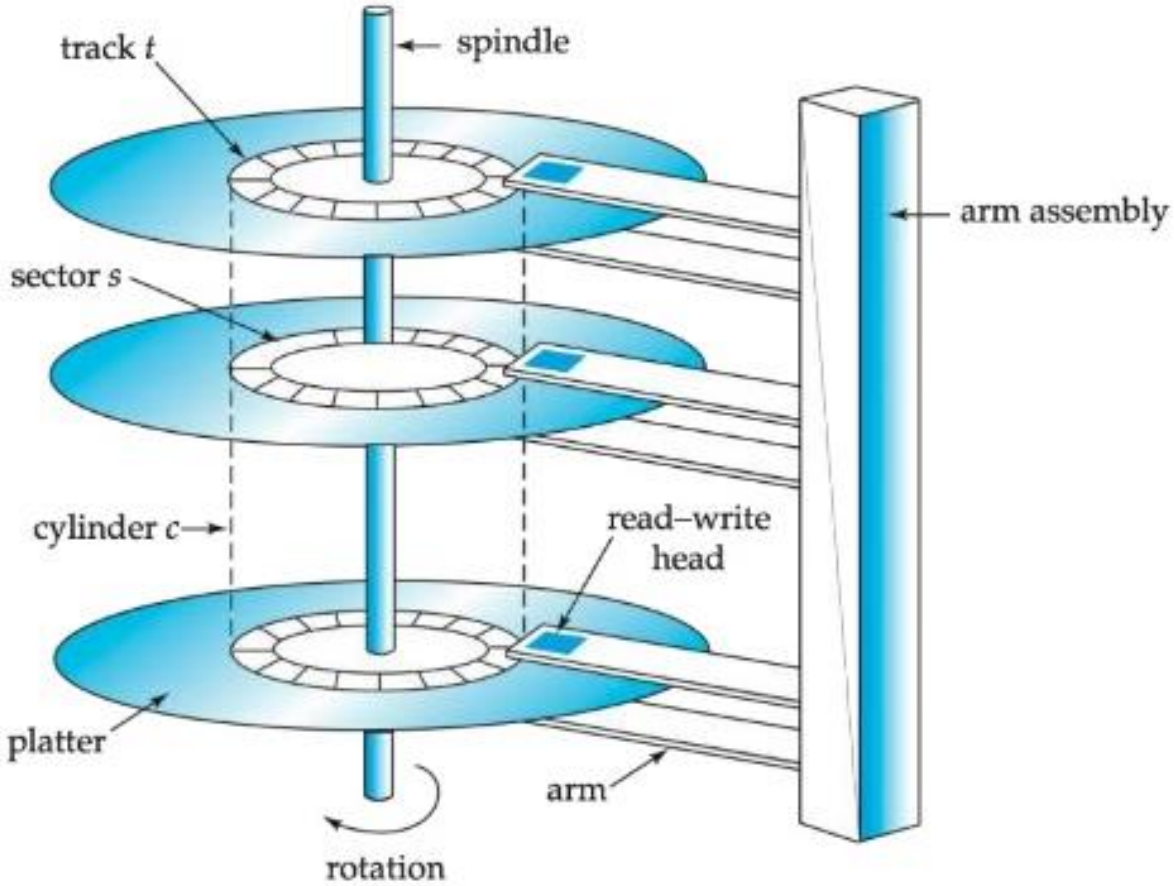
Common way to configure DMA

- I/O bus – to exchange data between I/O and DMA
- System bus – Exchange data with memory and exchange control signals with processor

Outline

- ❑ I/O devices
- ❑ Disk Arm Scheduling Algorithms
- ❑ RAID

Magnetic Disks (2)



Magnetic Disks (1)

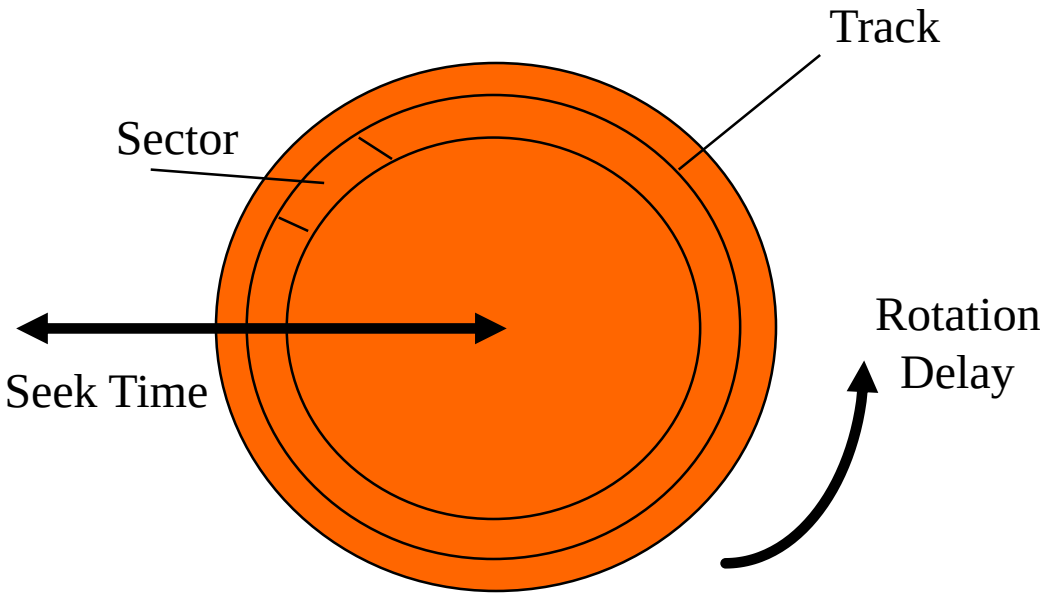
Parameter	IBM 360-KB floppy disk	WD 3000 HLFS hard disk
Number of cylinders	40	36481
Tracks per cylinder	2	255
Sectors per track	9	63 (avg)
Sectors per disk	720	586,072,368
Bytes per sector	512	512
Disk capacity	360 KB	300 GB
Seek time (adjacent cylinders)	6 msec	0.7 msec
Seek time (average case)	77 msec	4.2 msec
Rotation time	200 msec	6 msec
Time to transfer 1 sector	22 msec	1.4 μsec

Figure 5-18. Disk parameters for the original IBM PC 360-KB floppy disk and a Western Digital WD 3000 HLFS (''Velociraptor'') hard disk.

Tanenbaum & Bo,Modern Operating Systems:4th ed., (c) 2013 Prentice-Hall, Inc. All rights reserved.

Disk overheads

- To read from disk, we must specify:
  - cylinder #, surface #, sector #, transfer size, memory address
- Transfer time includes:
  - Seek time: to get to the track
  - Latency time: to get to the sector and
  - Transfer time: get bits off the disk



Disk Arm Scheduling Algorithms

Factors of a disk block read/write:

1. Seek time (the time to move the arm to the proper cylinder).
2. Rotational delay (how long for the proper sector to come under the head).
3. Actual data transfer time.

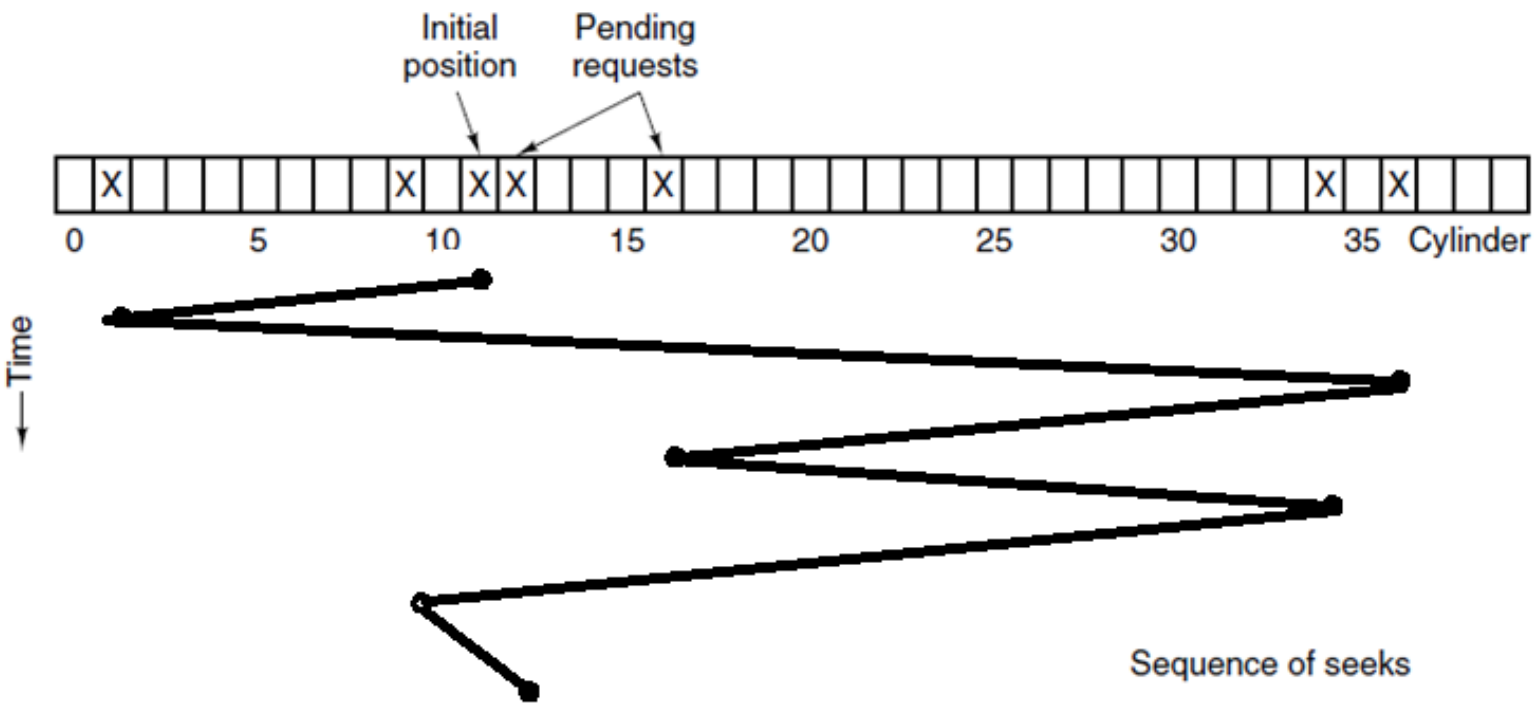
Tanenbaum & Bo,Modern Operating Systems:4th ed., (c) 2013 Prentice-Hall, Inc. All rights reserved

Disk Scheduling Policies - FCFS

- **FCFS (First-Come, First-Served)**
  - Easy to implement
  - Process requests sequentially in the order they arrive
  - Fair to all processes
  - Poor performance if requests are not clustered

Disk Scheduling Policies-FCFS

- Consider imaginary disk with 40 cylinders
- Request to read/write (cylinder number)
  - **11, 1, 36, 16, 34, 9, 12**



Disk Arm Scheduling Algorithms - SSF

- **SSF (Shortest Seek First)**
  - always handle the closest request next, to minimize seek time
  - has risk of starvation

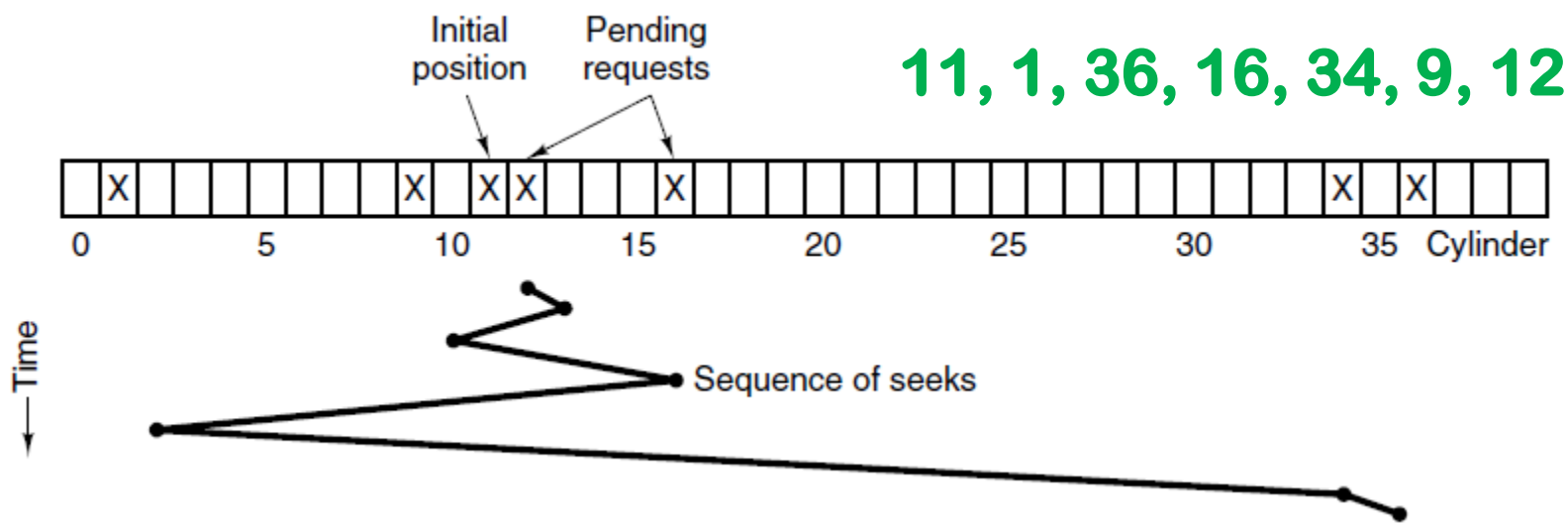
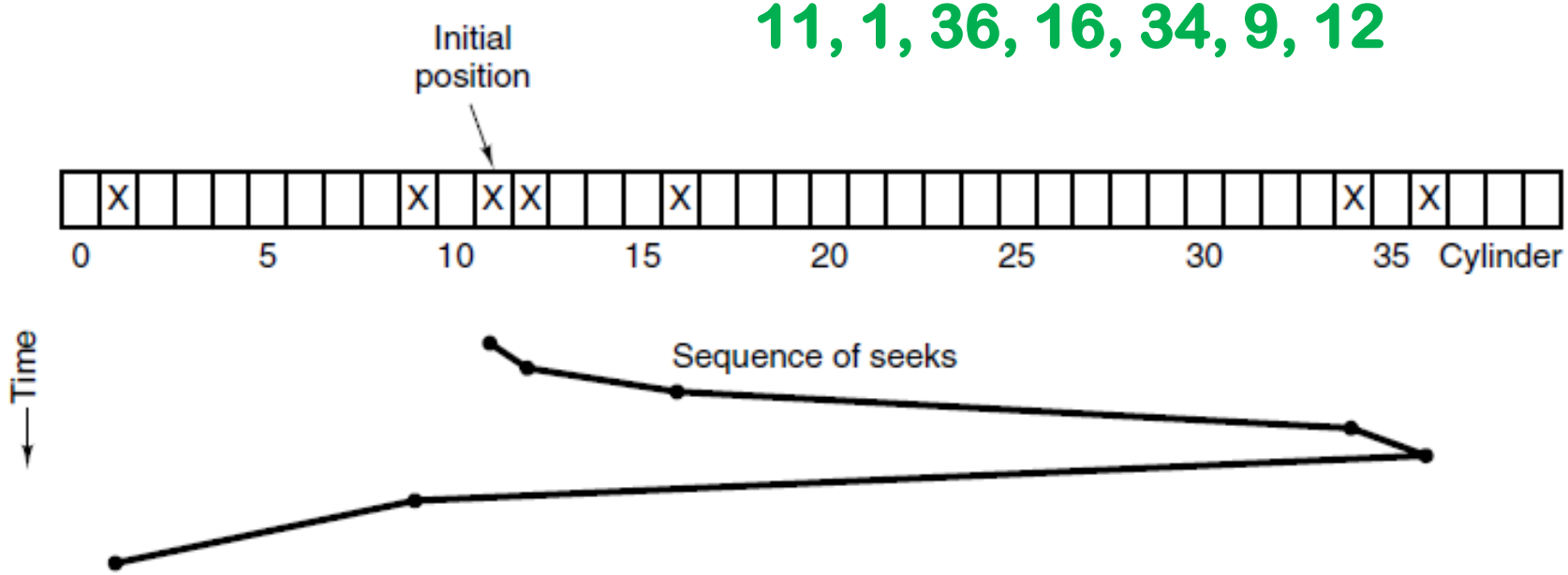


Figure 5-24. Shortest Seek First (SSF) disk scheduling algorithm.

Disk Arm Scheduling Algorithms -SCAN

- **The elevator algorithm/SCAN**
  - maintain 1 bit: the current direction (UP/DOWN)
  - keep moving in the same direction until there are no more outstanding requests in that direction, then switch directions

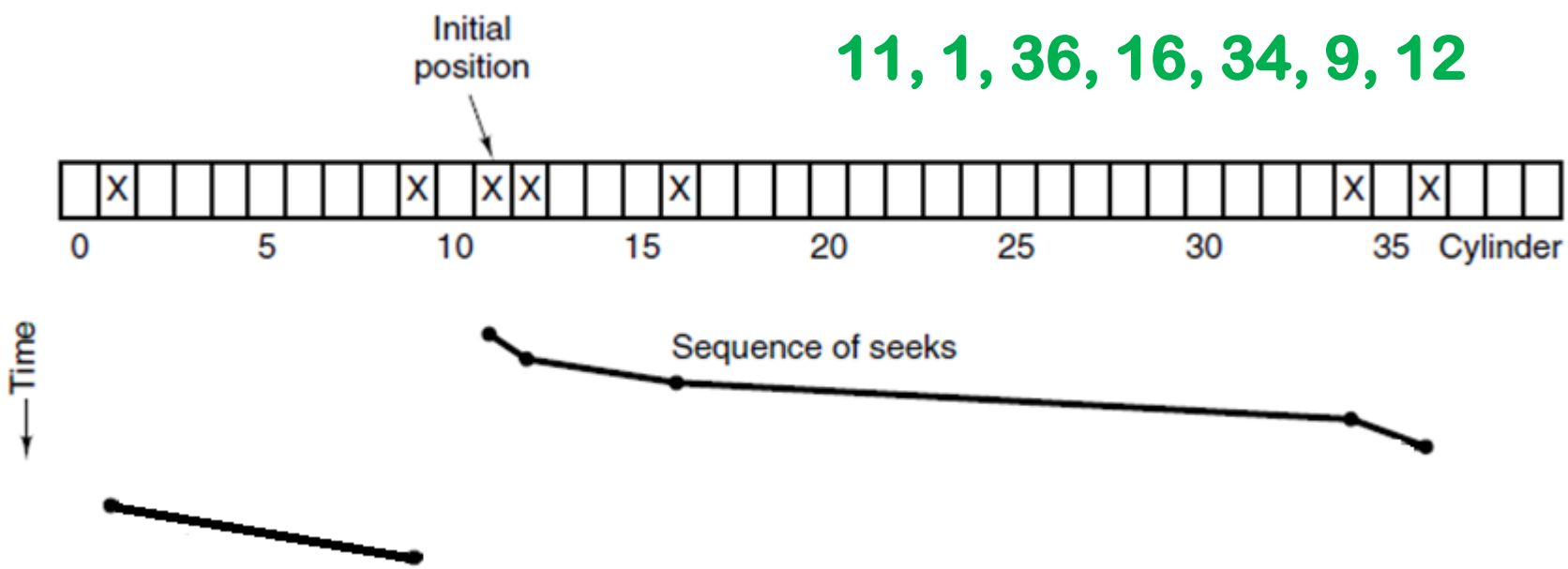
11, 1, 36, 16, 34, 9, 12



Disk Arm Scheduling Algorithms – C- SCAN

- C-SCAN (Circular-SCAN)
  - Restricts scanning to one direction only
  - When the last track has been visited in one direction, the arm is returned to the opposite end of the disk and the scan begins again

11, 1, 36, 16, 34, 9, 12

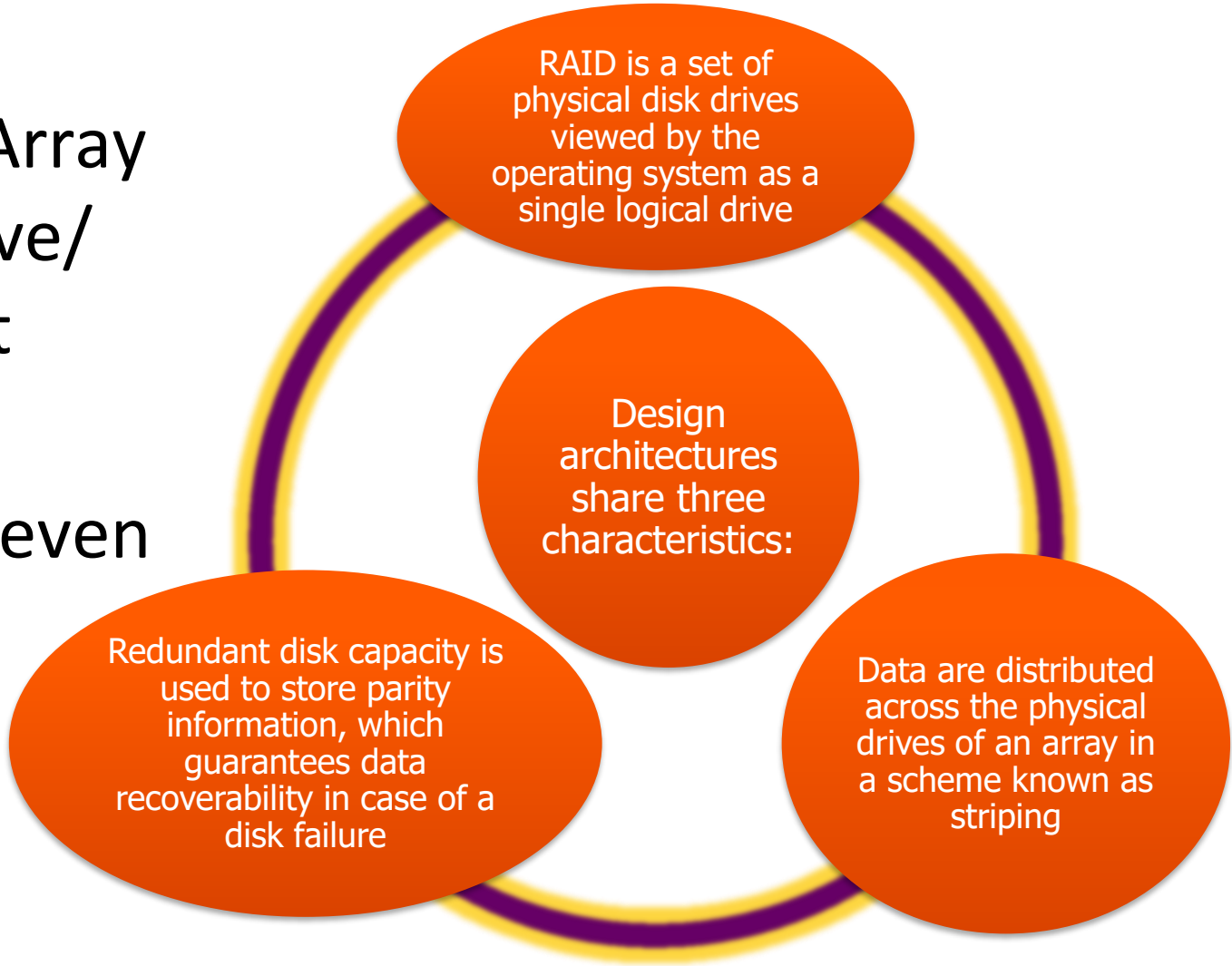


Outline

- ❑ I/O devices
- ❑ Disk Arm Scheduling Algorithms
- ❑ RAID

RAID

- Redundant Array of Inexpensive/ Independent Disks
- Consists of seven levels, zero through six



Copyright © 2018 Pearson Education, Ltd. All Rights Reserved.

REDUNDANT
ARRAY
RAID
ANIMATED
INDEPENDENT
RAID 0
RAID 1
RAID 5
RAID 10
A
D
G
PARITY
Hard Drive
B
E
PARITY
J
Hard Drive
I
K
Hard Drive
PARITY
F
I
L
Hard Drive
DISKS
STRIPING

RAID - Redundant Arrays of Inexpensive Disks

- ראינו שאפשר לחלק דיסק פיזי לדיסקים לוגיים, כעת נראה איך אפשר לחבר כמה דיסקים פיזיים לדיסק לוגי אחד (volume).
- היתרונות בחיבור דיסקים:
  - שיפור ביצועים - אפשר לגשת לדיסקים במקביל.
  - אמינות – אם יש תקלה בדיסק, אפשר לשחזר את הנתונים באמצעות מידע יתיר (redundant) שנשמר ביתר הדיסקים.
- RAID היא שיטה לחיבור כמה דיסקים לדיסק אחד.
- שיטת RAID כוללת כמה אפשרויות לחיבור דיסקים:
  - RAID 0, … , RAID 6
  - להלן נראה את האפשרויות הנפוצות.

RAID

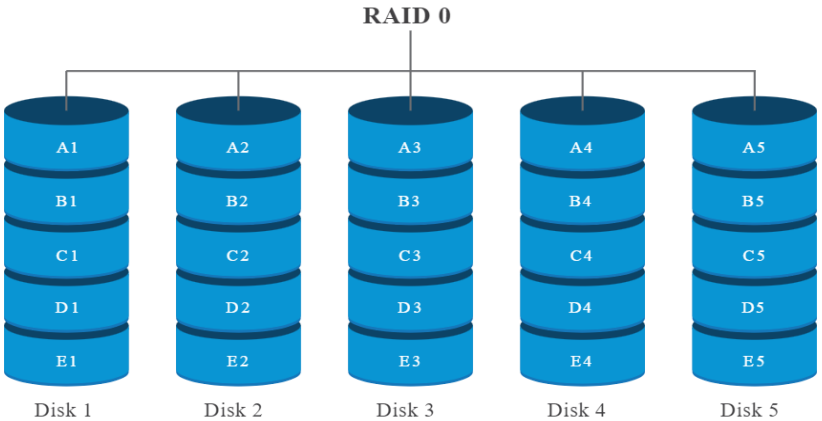
- The term was originally coined in a paper by a group of researchers at the University of California at Berkeley
  - The paper outlined various configurations and applications and introduced the definitions of the RAID levels
- Strategy employs multiple disk drives and distributes data in such a way as to enable simultaneous access to data from multiple drives
  - Improves I/O performance and allows easier incremental increases in capacity
- The unique contribution is to address effectively the need for redundancy
- Makes use of stored parity information that enables the recovery of data lost due to a disk failure

Copyright © 2018 Pearson Education, Ltd. All Rights Reserved.

RAID 0 - Striping (פסים)

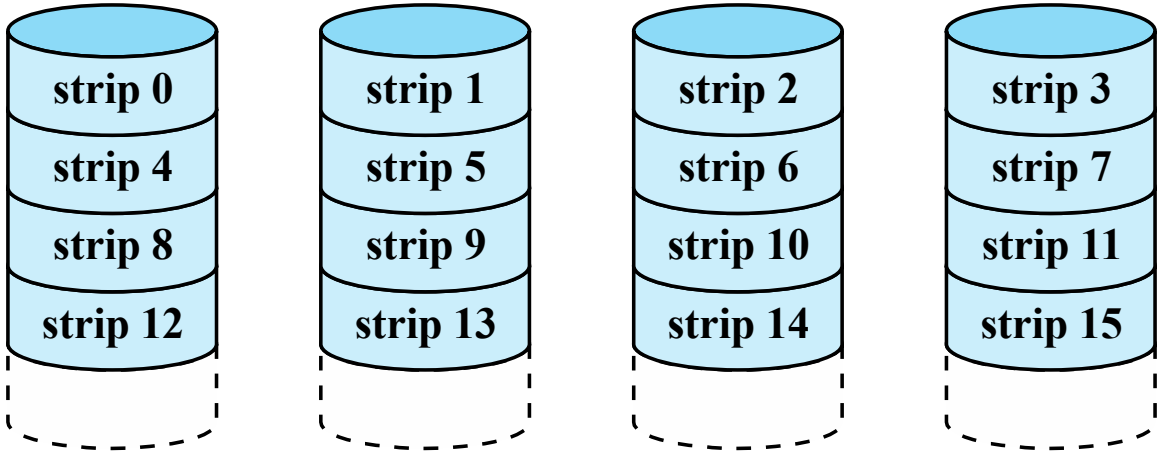
- ראינו שמערכת ההפעלה רואה את הדיסק כמערך של בלוקים.
- בשיטת RAID 0 מחלקים את הדיסקים לפסים, בלוק 0 הוא בדיסק הראשון, בלוק 1 בדיסק הבא וכן הלאה.
  - ביצועים: אפשר לגשת לדיסקים במקביל, נותן את הביצועים הטובים ביותר של RAID.

    לדוגמה, אפשר לקרוא או לכתוב קובץ בגודל שני בלוקים בשני דיסקים בו זמנית.
  - נפח: צרוף הדיסקים לא מפחית מהנפח הכולל של הדיסקים.
  - אמינות: אין יתירות, אם יש תקלה בדיסק אחד, המידע שעל כל הדיסקים משתבש, יותר גרוע מדיסקים ללא RAID.



RAID Level 0

- Not a true RAID because it does not include redundancy to improve performance or provide data protection
- User and system data are distributed across all of the disks in the array
- Logical disk is divided into strips

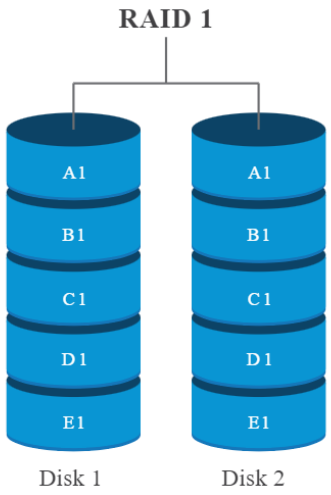


(a) RAID 0 (non-redundant)

Copyright © 2018 Pearson Education, Ltd. All Rights Reserved.

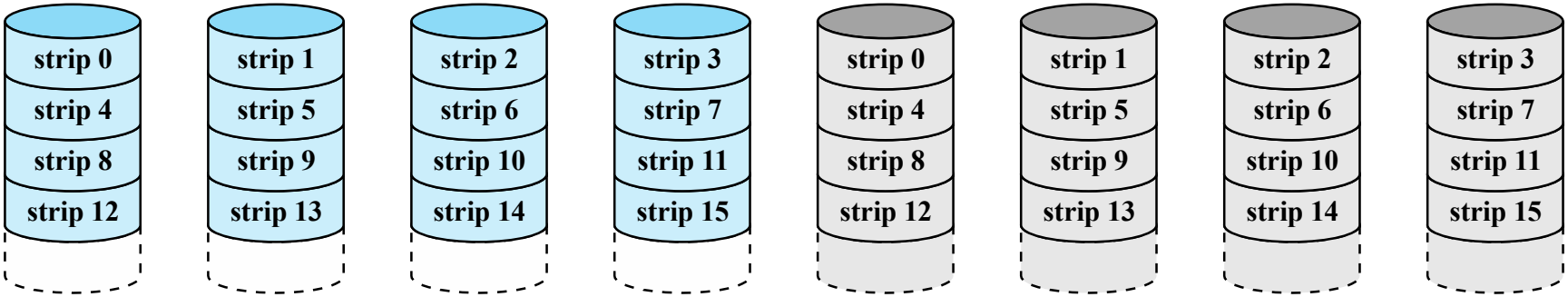
RAID 1 – Mirroring (שכפול)

- בשיטת RAID 1 כל הנתונים נכתבים לשני דיסקים (ראי).
  - ביצועים: קריאה היא יותר מהירה כי אפשר לקרוא משני הדיסקים במקביל.
  - נפח: בשיטה זו אפשר לנצל רק חצי מנפח הדיסקים.
  - אמינות: אם יש תקלה בדיסק, אפשר לקרוא מהדיסק השני.



RAID Level 1

- Redundancy is achieved by the simple expedient of duplicating all the data
- There is no “write penalty”
- When a drive fails the data may still be accessed from the second drive
- Principal disadvantage is the cost

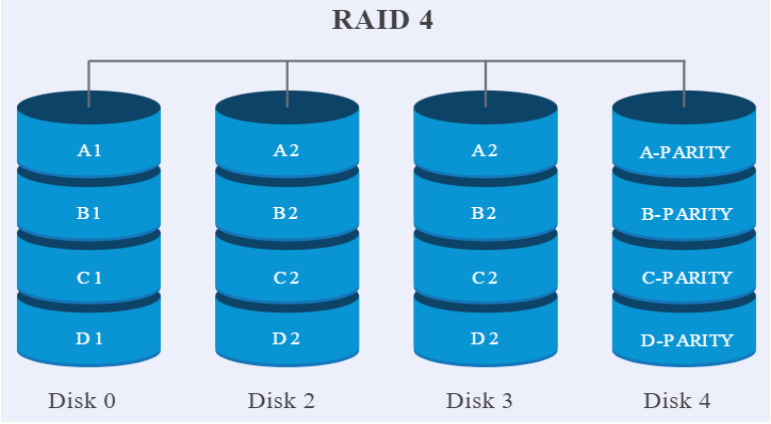


(b) RAID 1 (mirrored)

Copyright © 2018 Pearson Education, Ltd. All Rights Reserved.

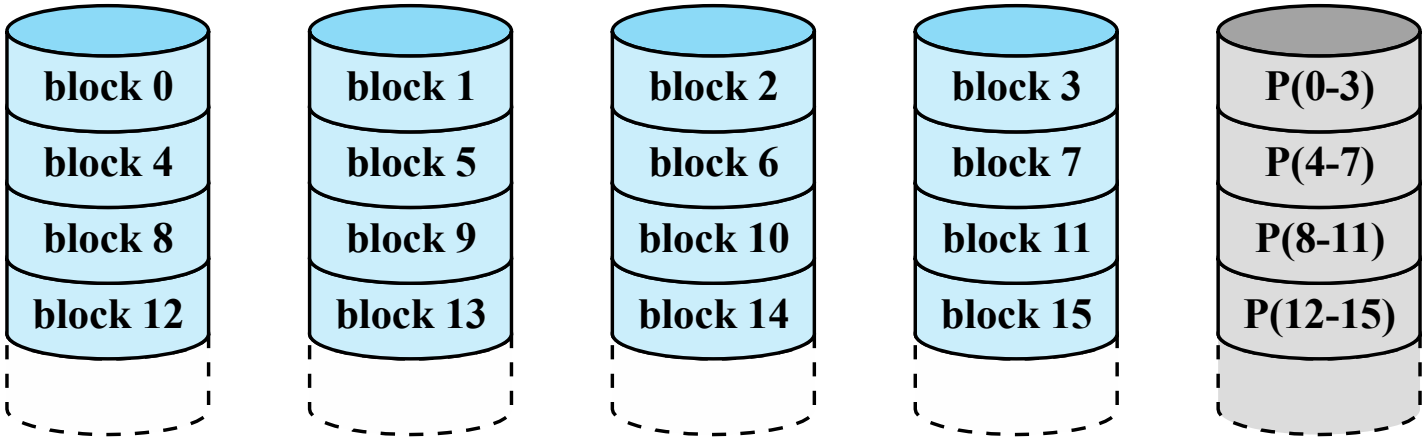
RAID 4 – Parity (זוגיות)

- בשיטת RAID 4 מחלקים את הדיסקים לפסים.
- בדיסק האחרון נשמר מידע (זוגיות) שמאפשר שחזור הנתונים במקרה של תקלה באחד מהדיסקים האחרים.
  - ביצועים: משפר את הביצועים, אפשר לגשת לדיסקים במקביל.
  - נפח: דיסק אחד משמש לזוגיות ולכן הנפח לשמירת נתונים הוא כמספר הדיסקים פחות אחד.
  - אמינות: אם יש תקלה בדיסק **אחד**, אפשר לשחזר את הנתונים מיתר הדיסקים.



RAID Level 4

- Makes use of an independent access technique
- A bit-by-bit parity strip is calculated across corresponding strips on each data disk, and the parity bits are stored in the corresponding strip on the parity disk
- Involves a write penalty when an I/O write request of small size is performed



(e) RAID 4 (block-level parity)

Copyright © 2018 Pearson Education, Ltd. All Rights Reserved.

זוגיות (Parity)

- זוגיות היא פונקציה שמחשבים על הבלוקים של הנתונים של כל פס ושומרים את התוצאה בבלוק שבדיסק האחרון.
  - הפונקציה שמחשבים היא XOR על הביט הראשון של כל הבלוקים של הפס ושמירת התוצאה בביט הראשון של בלוק הזוגיות, XOR על הביט השני ושמירת התוצאה בביט השני של בלוק הזוגיות, וכן הלאה עבור כל ביט בבלוק.
  - התוצאה של XOR היא 0 אם מספר הביטים שהם 1 הוא זוגי, התוצאה היא 1 אם מספר הביטים שהם 1 הוא אי-זוגי.
  - יוצא שמספר הביטים שהם 1 בכל הבלוקים כולל בבלוק הזוגיות הוא זוגי עבור כל ביט.
- בדוגמה הבאה מספר הדיסקים הוא חמש ובכל בלוק ארבעה ביטים:

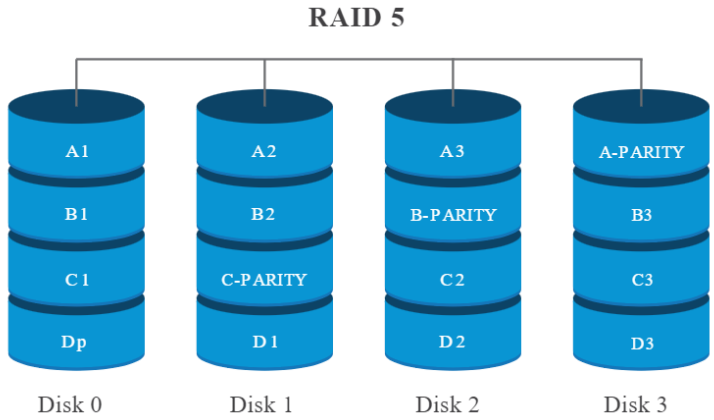
Block0	Block1	Block2	Block3	Parity
00	10	11	10	11
10	01	00	01	10

שחזור נתונים

- אם ארעה תקלה בדיסק, כדי לשחזר את הנתונים בדיסק חדש, נבצע עבור כל בלוק, XOR של כל הדיסקים הנותרים כולל דיסק הזוגיות.
- הסבר:
  - ראינו שבכל הדיסקים כולל דיסק הזוגיות מספר הביטים שהם 1 הוא זוגי.
  - פעולת XOR על הדיסקים הנותרים תיתן 1 אם מספר הביטים הוא אי זוגי, ואז נכתוב בדיסק החדש 1 ונקבל שוב מספר זוגי.
  - פעולת XOR על הדיסקים הנותרים תיתן 0 אם מספר הביטים הוא זוגי, ואז נכתוב 0 בדיסק החדש כי מספר הביטים הוא כבר זוגי.

RAID 5 – Rotating Parity - זוגיות בסבב על כל הדיסקים

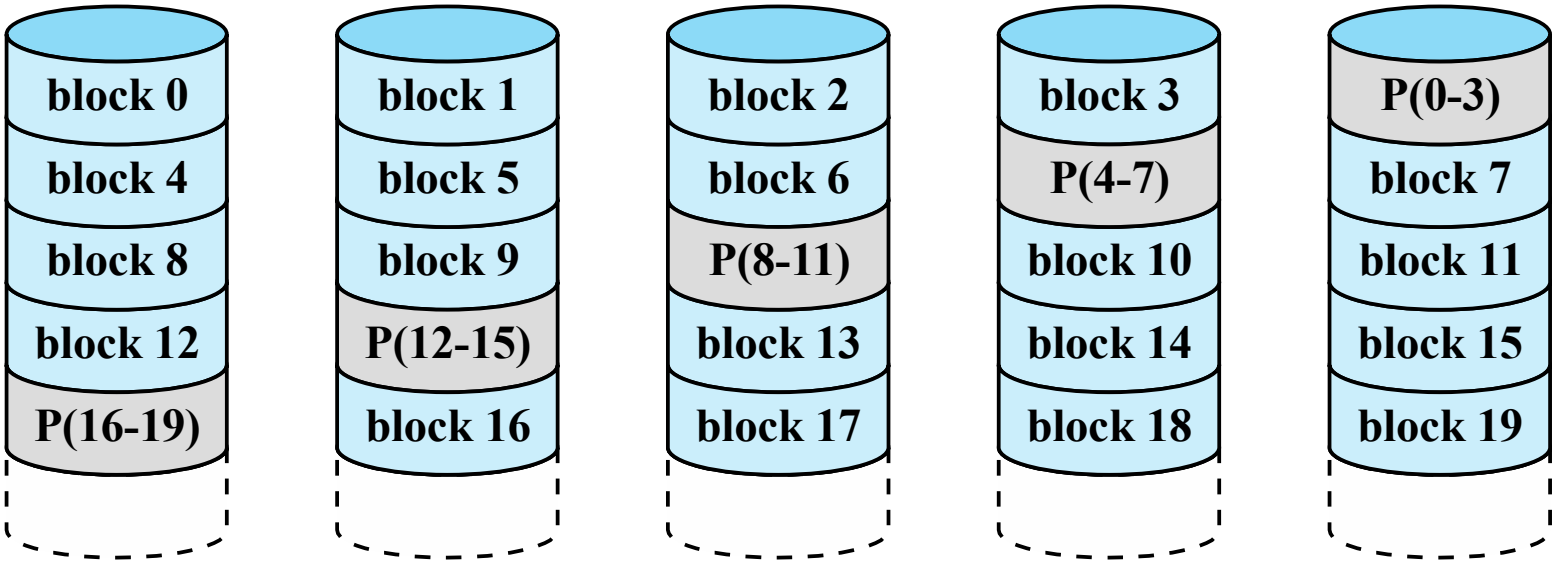
- בשיטת RAID 4, בכל כתיבה של בלוק לדיסק כלשהו צריך לעדכן את הבלוק שבדיסק הזוגיות.
- זה הופך את דיסק הזוגיות לצוואר בקבוק בכתיבות לדיסק.
- בשיטת RAID 5, כדי לשפר את הביצועים, הבלוקים של הזוגיות מפוזרים בסבב על כל הדיסקים.



Disk 0	Disk 1	Disk 2	Disk 3	Disk 4
0	1	2	3	P0
*4	5	6	7	+P1
8	9	10	11	P2
12	*13	14	15	+P3

RAID Level 5

- Similar to RAID-4 but distributes the parity bits across all disks
- Typical allocation is a round-robin scheme
- Has the characteristic that the loss of any one disk does not result in data loss



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

Copyright © 2018 Pearson Education, Ltd. All Rights Reserved.

Category	Level	Description	Disks required	Data availability	Large I/O data transfer capacity	Small I/O request rate
Striping	0	Nonredundant	N	Lower than single disk	Very high	Very high for both read and write
Mirroring	1	Mirrored	$2N$	Higher than RAID 2, 3, 4, or 5; lower than RAID 6	Higher than single disk for read; similar to single disk for write	Up to twice that of a single disk for read; similar to single disk for write
Parallel access	2	Redundant via Hamming code	$N + m$	Much higher than single disk; comparable to RAID 3, 4, or 5	Highest of all listed alternatives	Approximately twice that of a single disk
	3	Bit-interleaved parity	$N + 1$	Much higher than single disk; comparable to RAID 2, 4, or 5	Highest of all listed alternatives	Approximately twice that of a single disk
Independent access	4	Block-interleaved parity	$N + 1$	Much higher than single disk; comparable to RAID 2, 3, or 5	Similar to RAID 0 for read; significantly lower than single disk for write	Similar to RAID 0 for read; significantly lower than single disk for write
	5	Block-interleaved distributed parity	$N + 1$	Much higher than single disk; comparable to RAID 2, 3, or 4	Similar to RAID 0 for read; lower than single disk for write	Similar to RAID 0 for read; generally lower than single disk for write
	6	Block-interleaved dual distributed parity	$N + 2$	Highest of all listed alternatives	Similar to RAID 0 for read; lower than RAID 5 for write	Similar to RAID 0 for read; significantly lower than RAID 5 for write

N = number of data disks; m proportional to log N

(Page 498 in textbook)

Table 11.4 RAID Levels

Copyright © 2018 Pearson Education, Ltd. All Rights Reserved.