מערכות אחסון משני

אמצעי אחסון משניים בעבר

כרטיסים מנוקבים

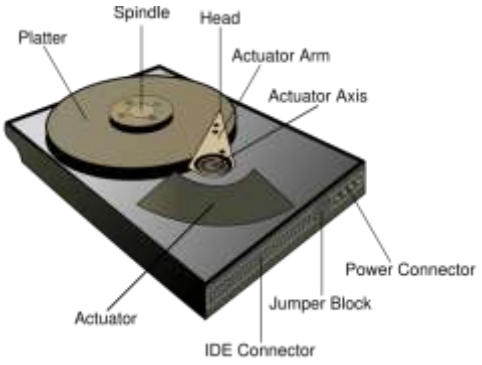
Floppy disks (דיסקטים)

סרטים מגנטיים

CD & DVD

סרטי נייר מנוקבים

דיסק קשיח (Hard disk)



Solid State Drive

מחירי זיכרון (2012)

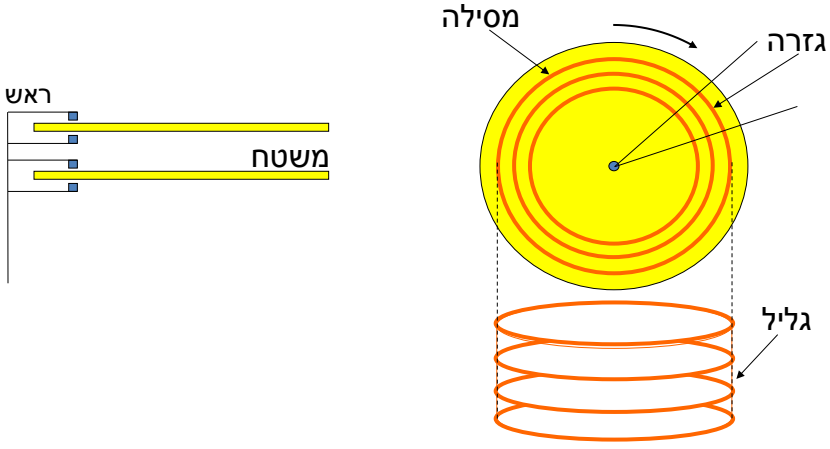
סוג זיכרון	נפח אופייני	מחיר אופייני	מחיר ל-1GB	מהירות העברת מידע
ראשי (SDRAM) 800MHz	4 GB	120$	30$	10 GB/Sec
Disk on key	16 GB	100$	6$	20 MB/Sec
Solid State Drive	100 GB	300$	3$	Read 200, Write 100 MB/Sec
משני (hard disk) 7200 סל"ד	500 GB	100$	20 cent	100-120 MB/Sec
דיסק קשיח חיצוני לגיבוי	1 TB	100$	10 cent	100 MB/Sec
CD-ROM (52x)	700 MB	0.2$	30 cent	7.8 MB/Sec
DVD (8x)	4.7 GB	0.25$	5 cent	10 MB/Sec

הדיסק

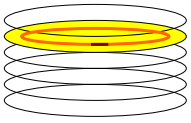
6 משטחים

הראש הקורא

תיאור סכמתי של מבנה הדיסק

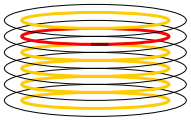


כתובת בדיסק



ארגון פיזי: משטח (Plate), מסילה (Track), גזרה (Sector).

החיסרון: תמיד ישנה הזזת זרוע במעבר ממסילה למסילה.



ארגון לוגי: גליל (Cylinder), מסילה, גזרה.

גליל: אוסף כל המסילות בעלות רדיוס זהה.

היתרון: במעבר על כל הדיסק מספר הזזות הזרוע קטן (פי מספר המשטחים).

נתונים לדיסק אופייני (אבל ישן)

IBM Ultrastar 36z15 scsi http://www.hgst.com/hdd/ultra/ul36z15.htm

גודל: 4 פלטות, 8 משטחים, 10500 מסילות למשטח,
450 גזרות למסילה, 512 בתים לגזרה

קיבולת:

גזרה:	512 B		
מסילה:	225 KB	= 230,400 B	= 512 B x 450
גליל:	1.8 MB	= 1800 KB	= 225 KB x 8
מארז:	18.9 GB	= 18900 MB	= 1.8MB x 10,500

מהירות:

סיבוב: 15,000 סל"ד = 250 סיבובים לשניה ⇐ 4 ms לסיבוב

זמן תזוזה:	מינימלי	0.65 ms,
	מקסימלי	6.7 ms,
	ממוצע משוקלל (יצרן)	3.4 ms

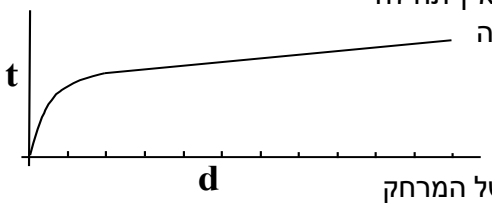
זמן קריאת רשומה

- סכום של שלושה גורמים:
  - זמן תזוזת זרוע (seek time)
  - השהיית סיבוב (rotational delay, latency)
  - זמן קריאת גזרה - העברה (transfer time)

חישוב זמן התזוזה I

אם d המרחק (במסילות) בין המקום הנוכחי של הזרוע למסילת היעד.

- את רוב המרחק הזרוע עוברת במהירות קבועה (מהירות שיוט) v
- אך כדי להגיע למהירות השיוט, יש להאיץ תחילה
- גם כדי לקרוא, על הזרוע להאט לעצירה



לכן זמן התזוזה אינו פונקציה לינארית של המרחק

נניח שזמן ההאצה למהירות שיוט הוא 1 (וכך גם זמן ההאטה)

$t(0) = 0$

$t(1) = T_{min}$

$t(2) = T_{base}$

$t(d) = (d-2) / v + T_{base}$ $(d>2)$

חישוב זמן התזוזה II

דוגמה: בדיסק Ultrastar 36z15

V	= 1800 tracks / msec	
t(1)	= 0.65 msec	
t(2)	= 0.87 msec	
t(10)	= 8 / 1800 + 0.87	≈ 0.874 msec
t(100)	= 98 / 1800 + 0.87	≈ 0.924 msec
t(1000)	≈ 1000 / 1800 + 0.87	≈ 1.43 msec
t(3500)	≈ 3500 / 1800 + 0.87	≈ 2.8 msec
		זמן תזוזה מקסימלי
t(10500)	≈ 10,500 / 1800 + 0.87	≈ 6.7 msec

מרחק תזוזה צפוי

משפט: נתון דיסק עם n גלילים.

אם מיקום הראש הקורא ויעד התזוזה נבחרים באופן בלתי-תלוי ובהתפלגות אחידה על פני כל הגלילים,

אז בממוצע בכל תזוזה, עוברים $E(d) = \frac{n}{3} - O\left(\frac{1}{n}\right)$ מסילות

לדיסק בדוגמה יש 10,500 מסילות

⇦ מרחק תזוזה ממוצע = 3,500 מסילות

⇦ זמן תזוזה ממוצע = 2.8 msec

☝ היצרן חישב זמן תזוזה ממוצע אחרת (כנראה בעזרת ניסויים) וקיבל **3.4 msec**

מרחק תזוזה צפוי: הוכחה

ישנם n גלילים

- גליל המוצא נבחר בהתפלגות אחידה: P(src=i) = 1/n
- גליל היעד נבחר בהתפלגות אחידה: P(dest=i) = 1/n
- גליל היעד אינו תלוי בגליל המוצא

$$P(src=i \wedge dest=j) = P(src=i) \cdot P(dest=i) = 1/n^2$$

$$\text{Expected distance} = \sum_d d \cdot P(d) = \sum_{1 \le i,j \le n} |i-j| \cdot \frac{1}{n^2} = \frac{1}{n^2} \cdot 2 \cdot \sum_{i=1}^{n} \sum_{j<i} j =$$

$$\frac{1}{n^2} \cdot 2 \cdot \sum_{i=1}^{n} \sum_{j=1}^{i-1} j = \frac{1}{n^2} \cdot 2 \cdot \sum_{i=1}^{n} \frac{i \cdot (i-1)}{2} = \frac{1}{n^2} \cdot \left(\sum_{i=1}^{n} i^2 - \sum_{i=1}^{n} i \right) =$$

$$\frac{1}{n^2} \cdot \left(\frac{n(n+1)(2n+1)}{6} - \frac{n(n+1)}{2} \right) = \frac{1}{n^2} \cdot \left(\frac{2n^3 - 2n}{6} \right) = \frac{n}{3} - O\left(\frac{1}{n}\right)$$

השהיית סיבוב

15,000 סיבובים לדקה (סל"ד) = 250 סיבובים לשניה

⇐ זמן סיבוב = 4 msec

ממוצע השהיית הסיבוב = זמן חצי סיבוב = 2 msec

זמן העברה

- חישוב בבתים:

$$transfer = \frac{\#\ bytes\ transferred}{\#\ bytes\ per\ track} \times \left(rotation\ time\right)$$

- חישוב לפי גזרות (סקטורים):

אם יש 450 גזרות למסילה,

זמן קריאת גזרה = 1 / 450 זמן סיבוב.

עבור 15,000 סל"ד (סיבוב = 4 ms):

זמן העברת גזרה 0.0089 msec = 4 / 450 .

תרגיל כיתה

קוראים באופן סידרתי קובץ של 900MB אשר מאוחסן ברציפות על גלילים סמוכים בדיסק עם:

- 10,500 גלילים
- 8 מסילות בגליל
- 450 גזרות למסילה
- 512 B לגזרה
- זמן תזוזת הזרוע:
  - מינימלית 0.65 msec
  - ממוצעת 3.4 msec
  - מהירות הסיבוב: 15,000 סל"ד

קיבולת גליל:

$512\ B \times 450 \times 8 = 1800KB$

מספר הגלילים לאחסון הקובץ:

$900\ MB / 1800\ KB = 512$

זמן קריאת מסילה = זמן סיבוב:

$4\ msec$

זמן קריאת גליל:

$8 \times 4\ msec = 32\ msec$

זמן קריאת הקובץ:

$512 \times 32\ msec = sec\ 16.4$

תרגיל כיתה (המשך)

- בתחילת הקריאה, הגעה לגליל הראשון של הקובץ
  - תזוזת זרוע ממוצעת 3.4 msec
  - הגעה לתחילת המסילה (חצי סיבוב): 2 msec בממוצע
- בסיום קריאת גליל, במעבר לגליל הבא
  - תזוזת זרוע מינימלית 0.65 msec
  - הגעה לתחילת המסילה (חצי סיבוב): 2 msec בממוצע
  - עבור 512 גלילים יש 511 מעברים שידרשו: 1354 msec = (2+0.65)*511
  - סה"כ ל- 16.4 שחושבו בשקף הקודם יש להוסיף 1.354 שניות עבור מעברים בין גלילים.

ואם קוראים מסילות אקראיות?

לכל מסילה יש לבצע:

- תזוזה אקראית (512/3-2)/1800+0.87=0.96 msec
  (בתוך 512/10500 מהדיסק)
- השהיית סיבוב אקראית 2 msec
- קריאה 4 msec
- סה"כ 6.96 msec

קריאת כל הקובץ:

- בקובץ 4096 = 8 × 512 מסילות.
- סה"כ: 4096 × 6.96 msec = 28.5 sec

הכי גרוע: קריאה של גזרות אקראיות

בכל מסילה יש 450 גזרות, ועבור כל-אחת צריך לבצע:

- תזוזה אקראית 0.96 msec
  בתוך 512/10500 מהדיסק
- השהיית סיבוב אקראית 2 msec
- קריאה 0.0089 msec
- סה"כ 2.9689 msec

קריאת כל הקובץ:

- בקובץ 450 × 4,096 = 1,843,200 גזרות.
- סה"כ:

1,843,200 × 2.9689 msec = 5,472 sec = 91 min = 1:31 hours

כתיבת רשומה בגזרה לא ריקה

- אפשר לכתוב רק גזרה שלמה.

- הפעולות שיש לבצע כאשר תוכן הגזרה לא נמצא בזיכרון:
  - חישוב כתובת היעד
  - קריאת גזרה לחוצץ
  - עדכון החוצץ
  - המתנה עד שהגזרה שוב מול הראש הקורא/כותב
  - כתיבה (העברה)

- סה"כ: כמו קריאה + סיבוב שלם נוסף.

נתונים לדיסק אופייני (2012)

Seagate Savvio 10K.5 SAS http://www.seagate.com/internal-hard-drives/enterprise-hard-drives/

גודל: ~~4~~ 3 פלטות, ~~8~~ 6 משטחים, ~~10500~~ 184,799 מסילות למשטח,
~~450~~ גזרות למסילה, 512 בתים לגזרה
1643 בממוצע

קיבולת:

גזרה:	512 B		
~~מסילה:~~	~~225 KB~~	~~= 230,400 B~~	~~= 512 B x 450~~
~~גליל:~~	~~1.8 MB~~	~~= 1800 KB~~	~~= 225 KB x 8~~
~~מארז:~~	~~18.9~~ GB	~~= 18900 MB~~	~~= 1.8MB x 10,500~~

900 מפורמט

מהירות:

סיבוב: ~~15,000~~ **10,000** סל"ד = 250 סיבובים לשניה ⇐ ~~4~~ **6** ms לסיבוב

זמן תזוזה:	מינימלי	~~0.65~~ **0.2** ms
	מקסימלי	~~6.7~~ **7.7** ms
	ממוצע משוקלל (יצרן)	~~3.4~~ **3.7** ms

Old Wisdom

- זיכרון ראשי קטן ומהיר ⇔ זיכרון משני ענק ואיטי
- קריאה של בלוקים באופן רציף מהירה בהרבה
- זמן כתיבה ≈ זמן קריאה
- שיפור מועט בביצועים עם השנים

(Solid State Drives) SSD

- אמצעי זיכרון משני בלא חלקים נעים

עדיין לא "בשלים" להחליף את הדיסק הקשיח

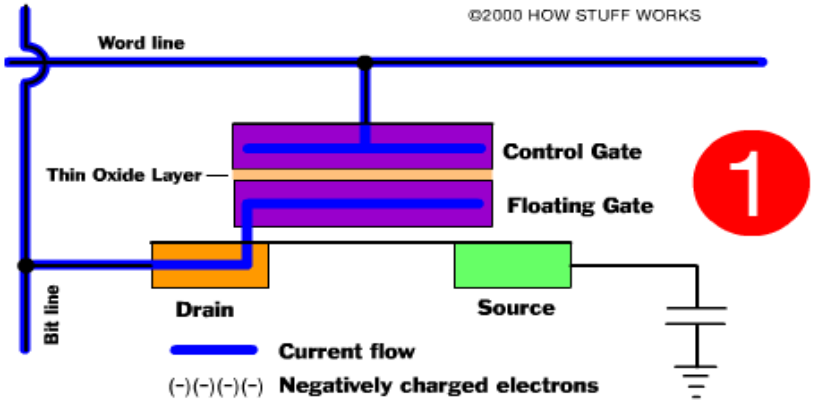
- DRAM דורשים מתח מתמיד ולכן לא מתאימים לגיבוי או הפצה
- flash drive, כמו disk on key, יקרים ואיטיים.

יתרונות וחסרונות של SSD

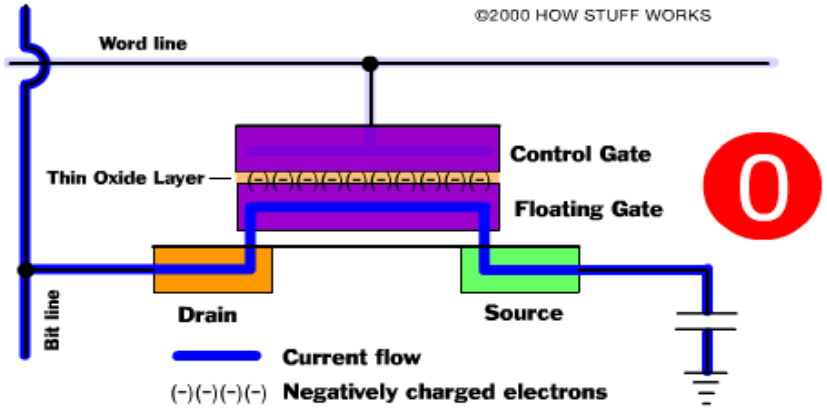
- ✓אתחול מהיר
- ✓גישה מהירה לרשומה
- ✓לא מתחמם
- ✓צריכת אנרגיה נמוכה (ב-flash)
- ✓עמידות בפני זעזועים פיסיים
- ✓זמן הקריאה של קובץ השמור באופן רציף (סדרתי) כמעט זהה לקריאת הקובץ כאשר רשומותיו מפוזרות על פני הדיסק
- ✓פחות שגיאות בעת קריאת הנתונים

- ×עדכון דורש "מחיקה" של ערכים ולכן כתיבה איטית מהקריאה
- ×מספר כתיבות מוגבל לכל תא זיכרון
  - צריך לפזר את השימוש על פני הדיסק
- ×העברת נתונים איטית
- ×מחיר יקר ליחידת אחסון
- ×עדיין לא מותאם לנפחי מידע גדולים
- ×קריסת דיסק היא לרוב טוטאלית (יותר קשה לשחזר מידע)

תא זיכרון Flash מורכב משני טרנזיסטורים

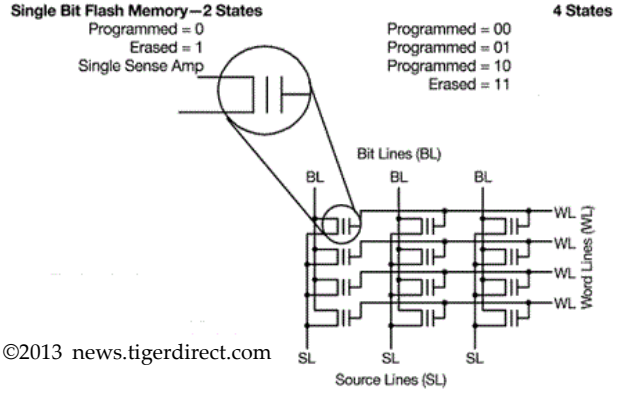


כתיבה היא שינוי מ-1 ל-0, ושינוי הפוך דורש "מחיקה"

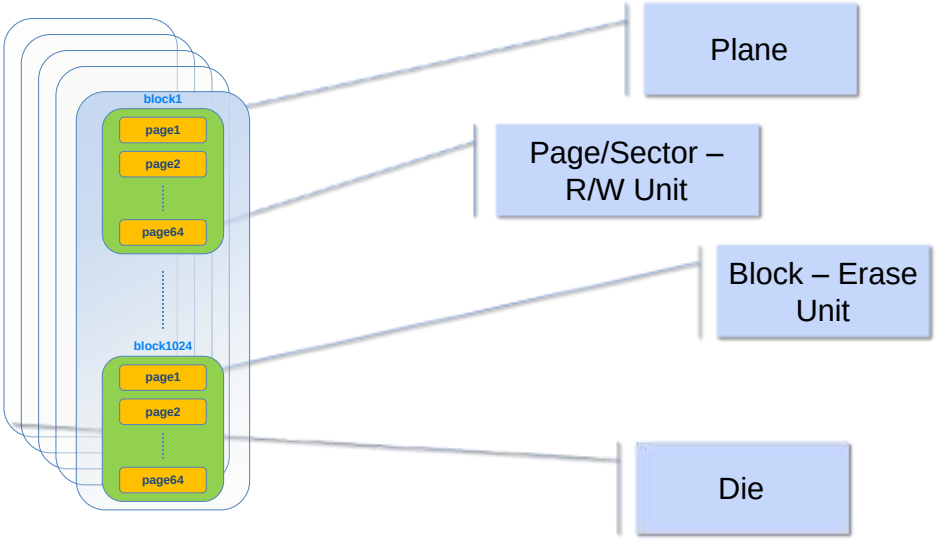


גישה לתא

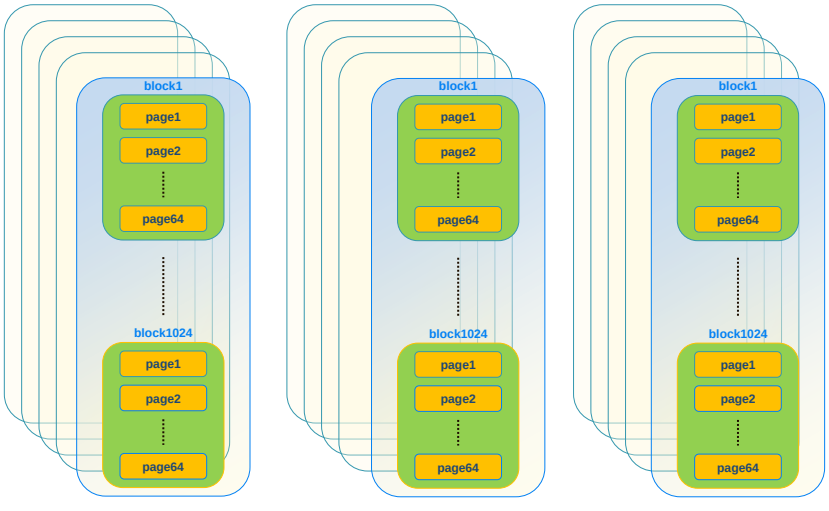
- קומבינציית המתח בין ה-word line ל-bit line מגדירה את הפעולה על התא (כתיבה, קריאה, מחיקה)



מבנה הדיסק



מבנה הדיסק



פרמטרים לדוגמה

יחידת מחיקה

- דף (page): 4KB
- בלוק (block): 128 דפים
- משטח (plane): 2048 בלוקים
- תבנית (die): 4 משטחים
- תבנית מכילה: 4 GB

- מארז (package): 2 תבניות
- דיסק: 10 מארזים

ניתן לקרוא ממשטחים שונים במקביל.
- איך זה משפיע על הביצועים?
- איך זה משפיע על ארגון קבצים?

פרמטרים ל- Intel SSD 320

- השהיית קריאה: 75 מיקרו-שניות (0.075 מילי-שניה)
- קריאה רציפה: 270 MB/Sec
- השהיית כתיבה: 90 מיקרו-שניות
- כתיבה רציפה: 90 MB/Sec

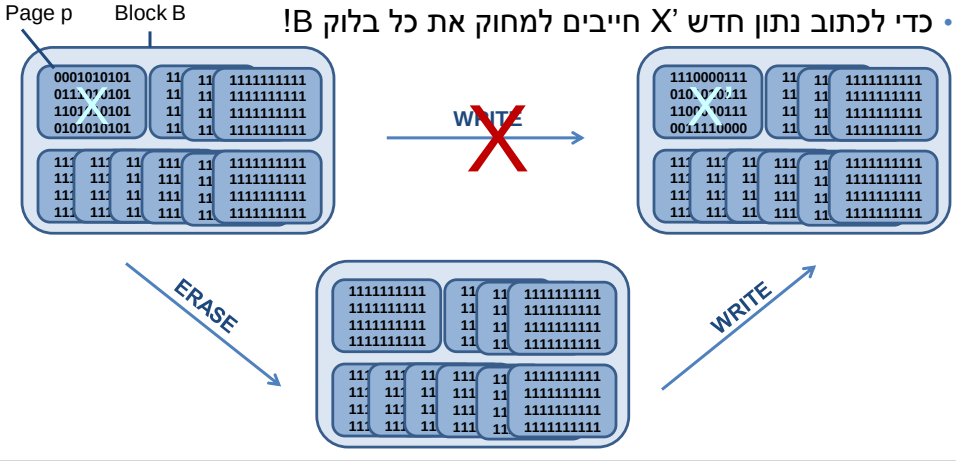
- (זמנים טיפוסיים, על פי מדידות יצרן)

דוגמה של קריאת/כתיבת קובץ מ/ל SSD

- קריאת 900 MB:
  - זמן קריאה: 900/270 = 3.33 Sec
  - אין הבדל בין קובץ רציף לקובץ לא רציף
- כתיבת 900 MB:
  - זמן כתיבה: 900/90 = 10 Sec

מימוש כתיבות: הבעיה עם עדכון

- בתוך בלוק נתונים B נמצא דף p ועליו כתוב נתון X
  - נתעלם כרגע משאר הדפים, ונניח שהם ריקים
- כדי לכתוב נתון חדש 'X חייבים למחוק את כל בלוק B!



מימוש כתיבות: נסיון 1 (לא יעיל)

- $T=t_0$: מצב התחלתי
- $T=t_1$: כתיבה של W,X,Y
- $T=t_2$: כתיבה של Z,A,B',C',R'

Read/Erase/Modify/Write

$T=t_0$

Erase Block I			
b	c	-	-
j	-	k	l
m	-	-	-
-	-	q	r

$T=t_1$

Erase Block I			
b	c	W	X
j	Y	k	l
m	-	-	-
-	-	q	r

$T=t_2$

Erase Block I			
B'	C'	w	x
j	y	k	l
m	Z	A	-
-	-	q	R'

- שמור נתונים בזיכרון
- מחק את Block 1

- שמור נתונים בזיכרון
- מחק את Block 1

Read/Erase/Modify/Write

- לא יעיל
  - כתיבה אטית
    - יש לחכות למחיקת בלוק שלם בכל שינוי/כתיבה (לדף משומש)
  - "מקצר" את חיי הדיסק
    - כתיבות חוזרות לאותו בלוק

הדוגמא מבוססת על
NAND Flash Solid State Storage: Performance and Capability – an In-depth Look
Jonathan Thatcher © 2009 SNIA

מימוש כתיבות: נסיון 2 (לא יעיל)

- $T=t_0$: מצב התחלתי
- $T=t_1$: כתיבה של W,X,Y
- $T=t_2$: כתיבה של Z,A,B',C',R'

Read/Modify/Write

$T=t_0$

Erase Block I			
b	c	-	-
j	-	k	l
m	-	-	-
-	-	q	r

$T=t_1$

Erase Block 2			
b	c	W	X
j	Y	k	l
m	-	-	-
-	-	q	r

$T=t_2$

Erase Block 3			
B'	C'	w	x
j	y	k	l
m	Z	A	-
-	-	q	R'

- שמור נתונים בזיכרון
- מחק את Block 1

- שמור נתונים בזיכרון
- מחק את Block 2

Read/Modify/Write

- הנחה: Block 2 ו-Block 3 נמחקו מבעוד מועד
- יעיל יותר מהדוגמא הקודמת
  - אין צורך לחכות למחיקה לשם השלמת הכתיבה
- פיזור אחיד (wear leveling)
  - נתונים שמעודכנים בתדירות גבוהה מפוזרים על פני הדיסק
- כתיבות מוגברות (write amplification)
  - 8 כתיבות ביוזמת המשתמש (25%)
  - 24 כתיבות תקורה בשל העברת הנתונים (75%)

מימוש כתיבות: נסיון 3

- $T=t_0$: מצב התחלתי
- $T=t_1$: כתיבה של W,X,Y
- $T=t_2$: כתיבה של Z,A,B',C',R'

Garbage Collection

$T=t_0$

Erase Block I			
b	c	-	-
j	-	k	l
m	-	-	-
-	-	q	r

$T=t_1$

Erase Block 2			
W	b	c	X
Y	j	k	l
m	q	r	

$T=t_2$

Erase Block 2			
w	-	-	x
y	j	k	l
m	q	-	B'
C'	Z	A	R'

כתוב למקום פנוי ב- Block 2

אסוף נתונים ומחק את Block 1

Garbage Collection

- מתבצע ברקע
  - התדירות נקבעת על פי האלגוריתם
  - תדירות גבוהה ← כתיבות מהירות, חיי הדיסק מתקצרים
  - תדירות נמוכה ← כתיבות עלולות להתעכב, חסכון בכתיבות
  - ניתן לכוון כתיבות לבלוקים שהיו פחות בשימוש
- שיפור בתקורה
  - 8 כתיבות ביוזמת המשתמש (50%)
  - 8 כתיבות תקורה בשל העברת הנתונים (50%)