# Chapter 12: Secondary-Storage Systems

Prof. Li-Pin Chang

National Chiao Tung University

# Chapter 12: Mass-Storage Systems

- Magnetic tape
- Disk Structure and Attachment
- Disk Scheduling
- RAID Structure
- Solid State Disks

# 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, such as RAID

# Magnetic Tape

- Access time slow, Random access ~1000 times slower than disk
  - Kept in spool and wound or rewound past read-write head 只能進帶或到帶
  - Once data under head, transfer rates comparable to disk
- Relatively permanent 永久的 and holds large quantities of data
  - 1.5~12 TB typical storage
  - Mainly used for backup or cold storage 只能sequential的讀跟寫(很快)

# Magnetic Disks

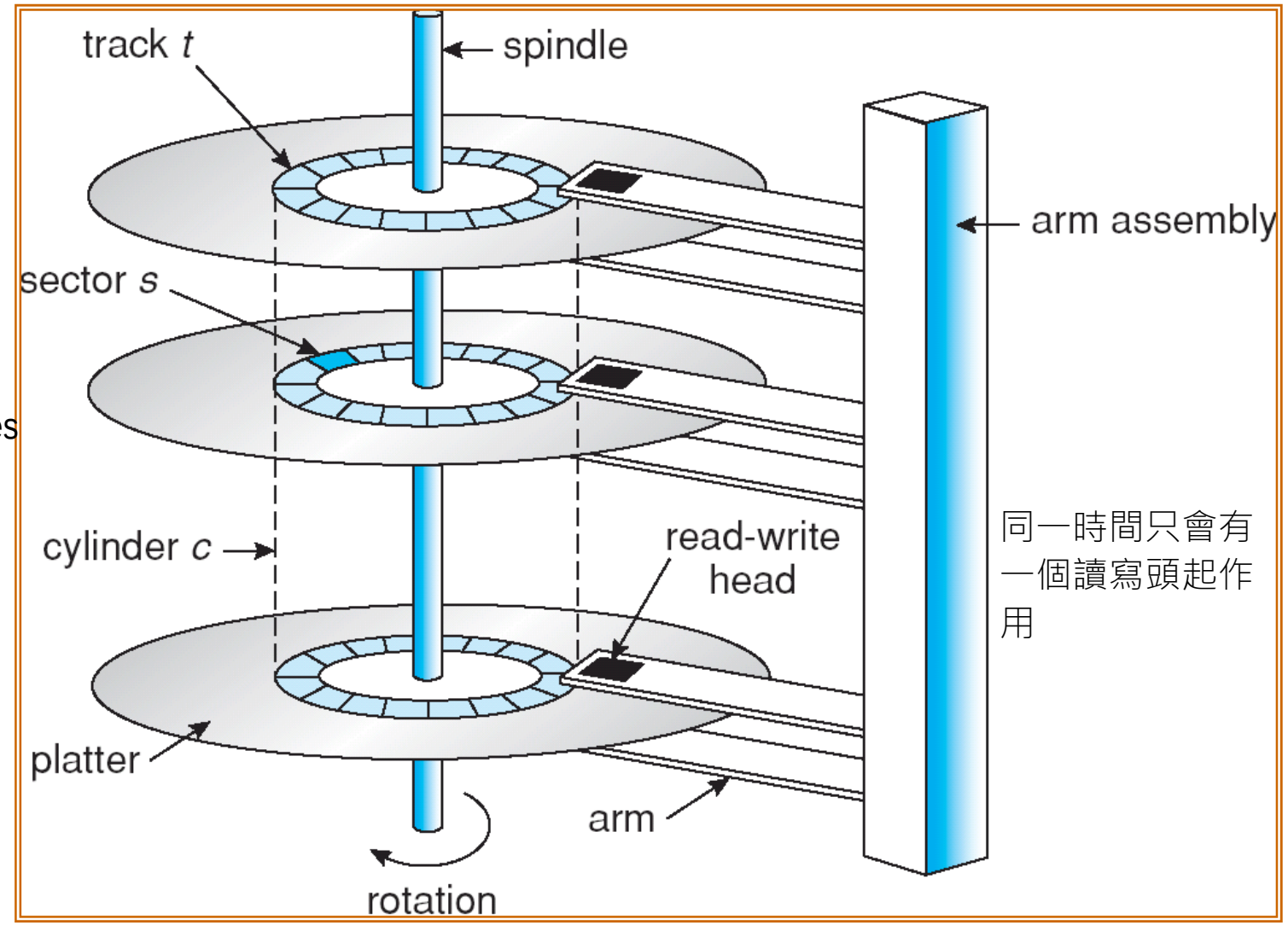
- Provide bulk (大量) of secondary storage of modern computers
- Transfer rate is rate at which data flow between drive and computer
  - SATA3: 600 MB/s
- Positioning time (random-access time) is time to move disk arm to desired cylinder (seek time) and time for desired sector to rotate under the disk head (rotational latency)
  - Typically 5400 rpm (laptop) or 7200 rpm (desktop)
  - Typically 5ms~7ms average seek time
- Head crash results from disk head making contact with the disk surface, causing physical damage

# Moving-Head Disk Mechanism

比較常見的是兩片到四片圓盤，上下兩面都可讀寫

一個sector=512bytes
一個track≒2MB

三個座標能唯一定位資料：
cylinder
head(讀寫頭)
sector



# 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.

cylinder由外圈到內圈

# Disk Attachment

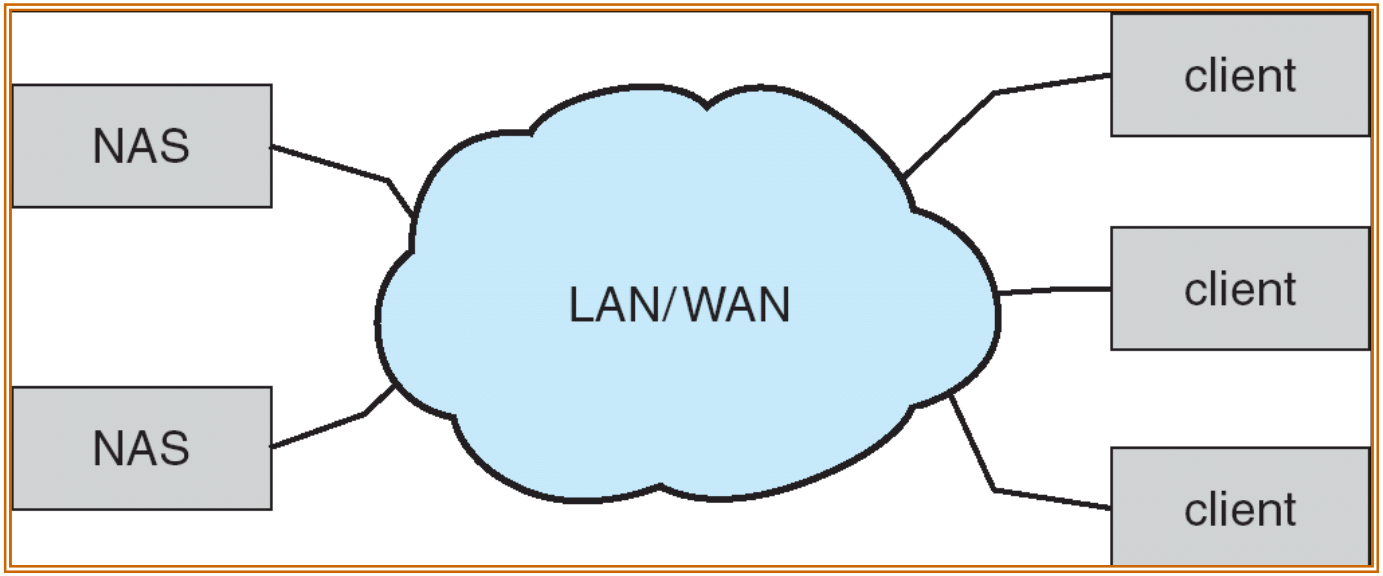
- Host-attached storage accessed through I/O ports talking to I/O busses
- ATA/IDE is the primary disk interface for personal computers
  - Parallel ATA → serial ATA
- SCSI itself is a bus, up to 16 devices on one cable, SCSI initiator requests operation and SCSI targets perform tasks
  - Each target can have up to 8 logical units (disks attached to device controller

最多連接16個裝置

- FC is high-speed serial architecture
  - Can be switched fabric with 24-bit address space – the basis of storage area networks (SANs) in which many hosts attach to many storage units
  - Can be arbitrated loop (FC-AL) of 126 devices

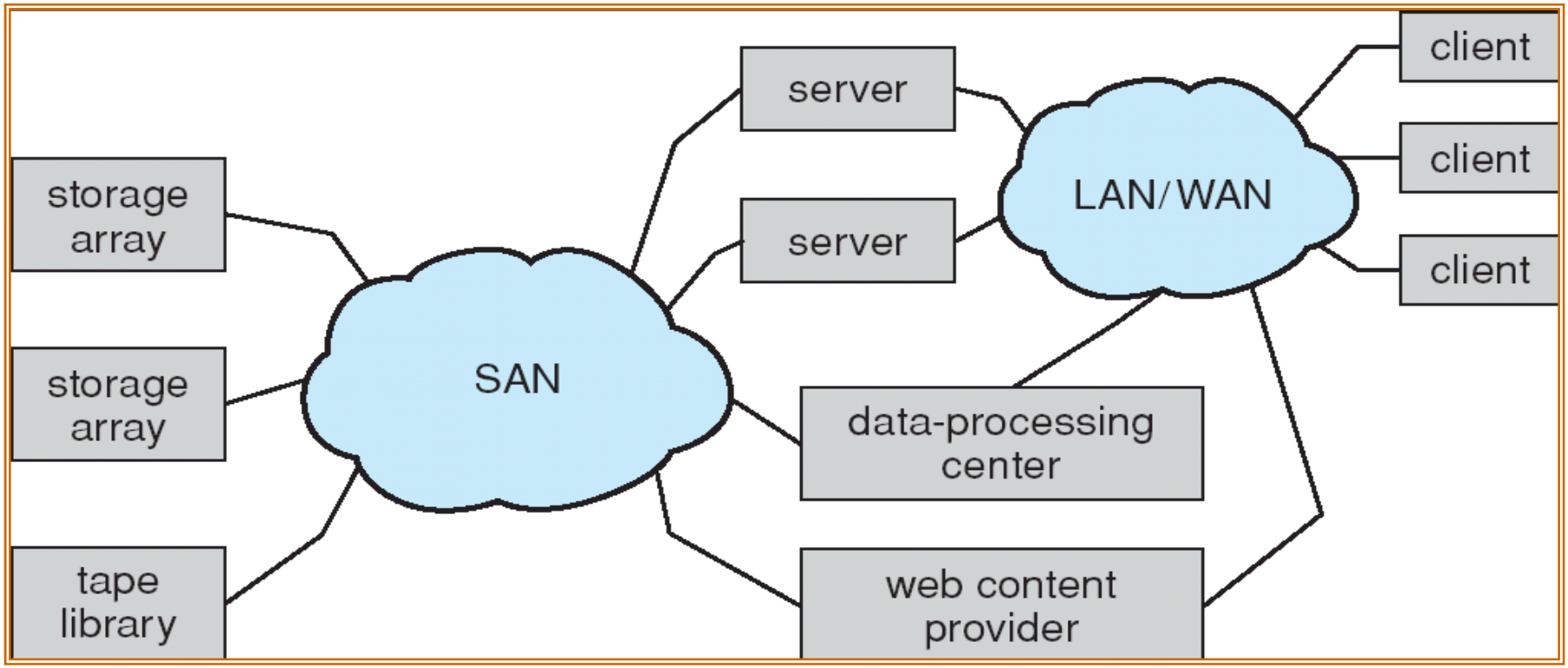
# Network-Attached Storage

- Network-attached storage (NAS) is storage made available over a network rather than over a local connection (such as a bus)
- NFS, CIFS, SAMBA 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



# SAN vs. NAS

- SAN is a network dedicated for storage
  - Performance is the primary concern
  - Topology, bandwidth, cost…
- Storage resource in SAN is hidden from the client of SAN.
  - A volume may sit across many storage devices
- NAS may operate over legacy network
  - Interoperability is much more important

SAN是專為儲存設計的私有網路

傳統網路

互相協同性

# 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 Bandwidth = Total bytes transferred / Total time between first request & completion of last one

# The Needs for Disk Scheduling

- Because of
  - 1) multiprogramming and
  - 2) write buffering,

  there might be a number of pending (待决的) disk requests

- How to select the next request to serve?
  - has impacts on response and throughput

# Disk Scheduling (Cont.)

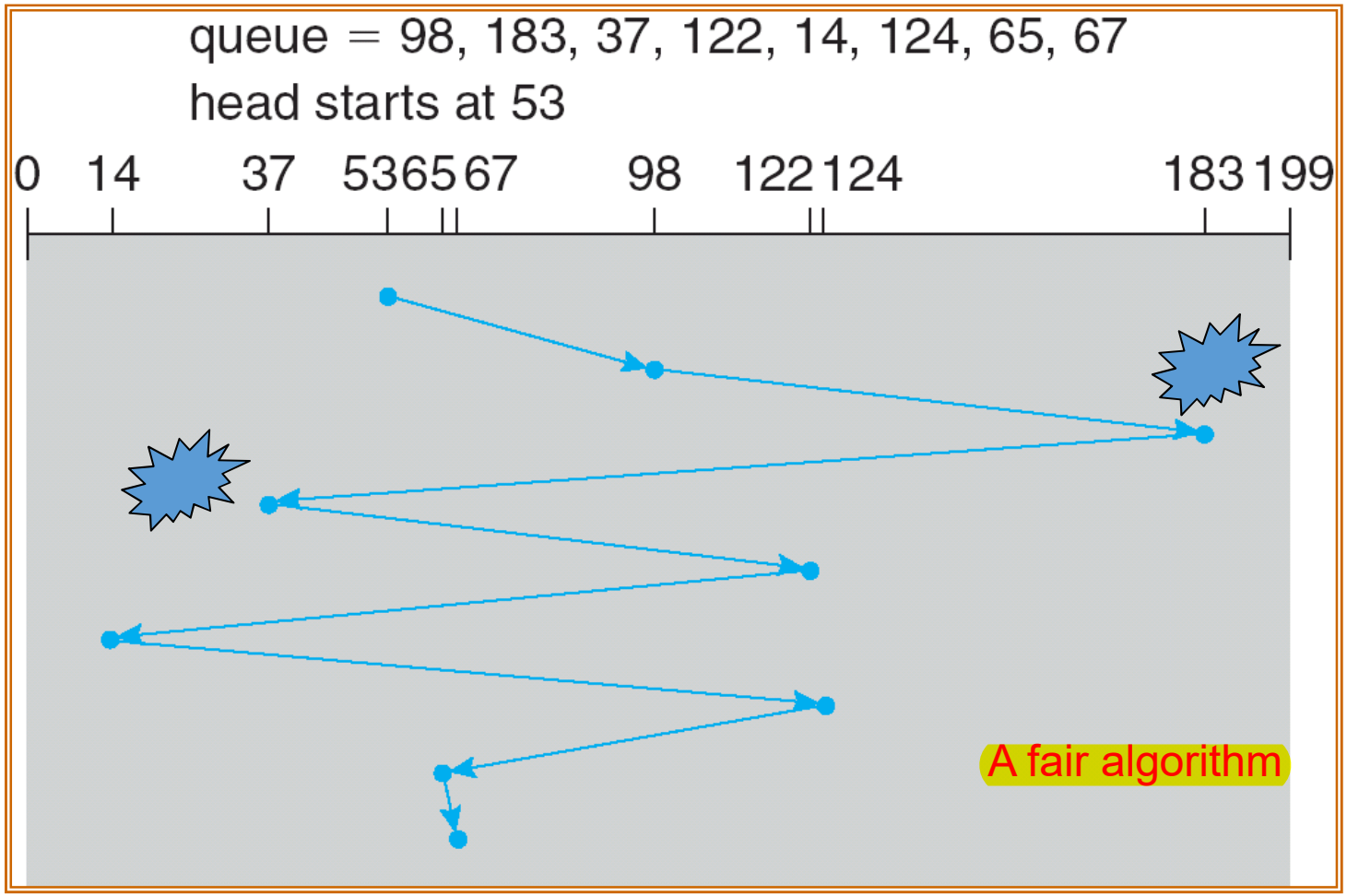
- Several algorithms exist to schedule the servicing of disk I/O requests.
- We illustrate them with a request queue (0-199).

98, 183, 37, 122, 14, 124, 65, 67 這些都是不同的cylinder

- Head pointer 53

# FCFS Disk Scheduling 重點是公平

Illustration shows total head movement of 640 cylinders.

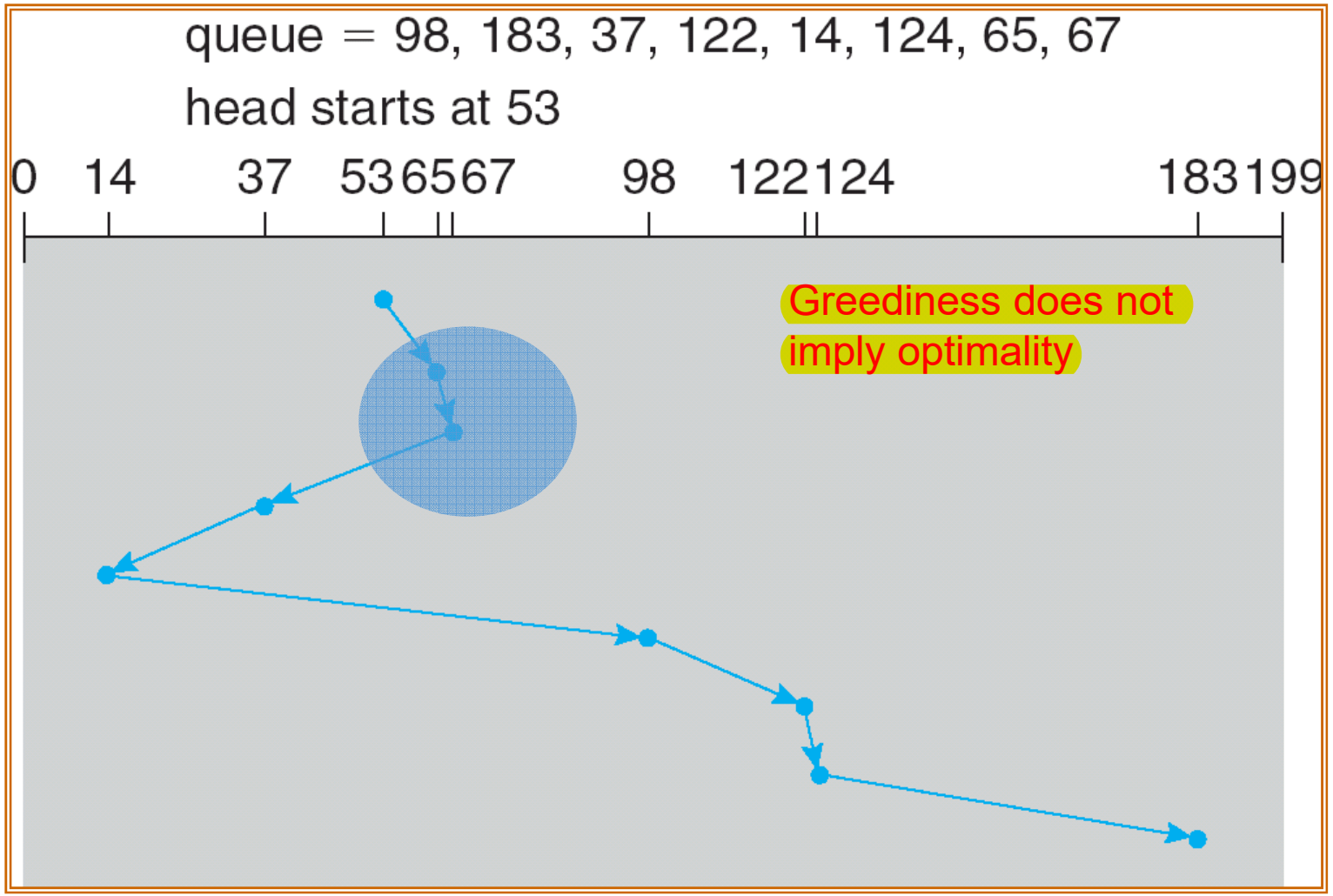


shortest-seek-time-first

# SSTF Scheduling

- Selects the request with the minimum seek time from the current head position
- SSTF scheduling is a form of SJF (shortest-job-first) scheduling; may cause starvation of some requests
  - How to avoid starvation?
- Illustration shows total head movement of 236 cylinders

# SSTF Disk Scheduling

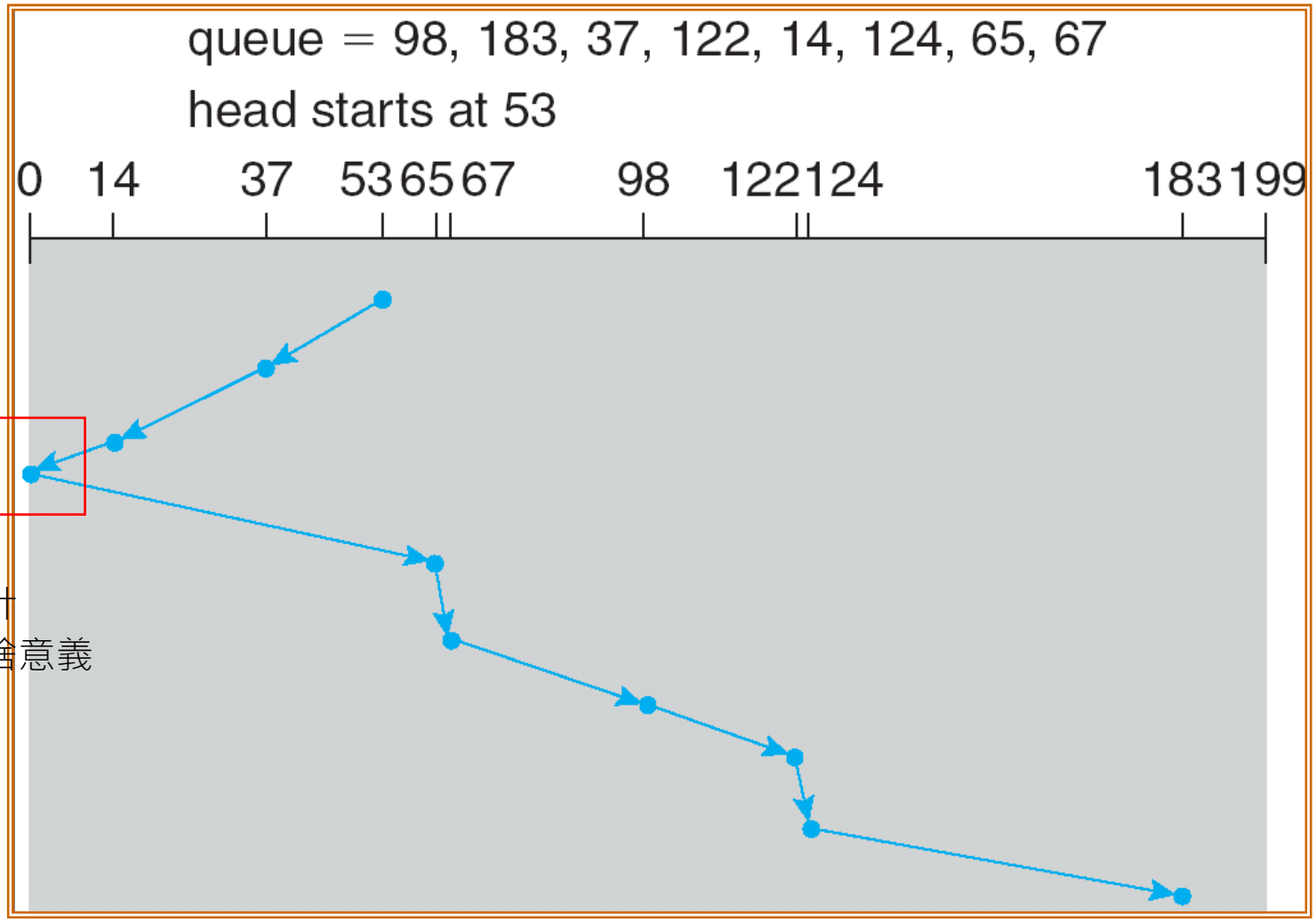


# SCAN Scheduling

- 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 it is called the elevator algorithm
- Illustration shows total head movement of 236 cylinders

磁頭只會左右一維移動
先把一個方向的弄完， 再讀其他的

# SCAN Disk Scheduling




奇怪的設計
走到底沒啥意義

# SCAN

走過去再回來

- A fair algorithm
- In the worst case, a request has to wait for 2 full strokes 最慘就是request在邊邊進來時，磁頭剛好折返回去

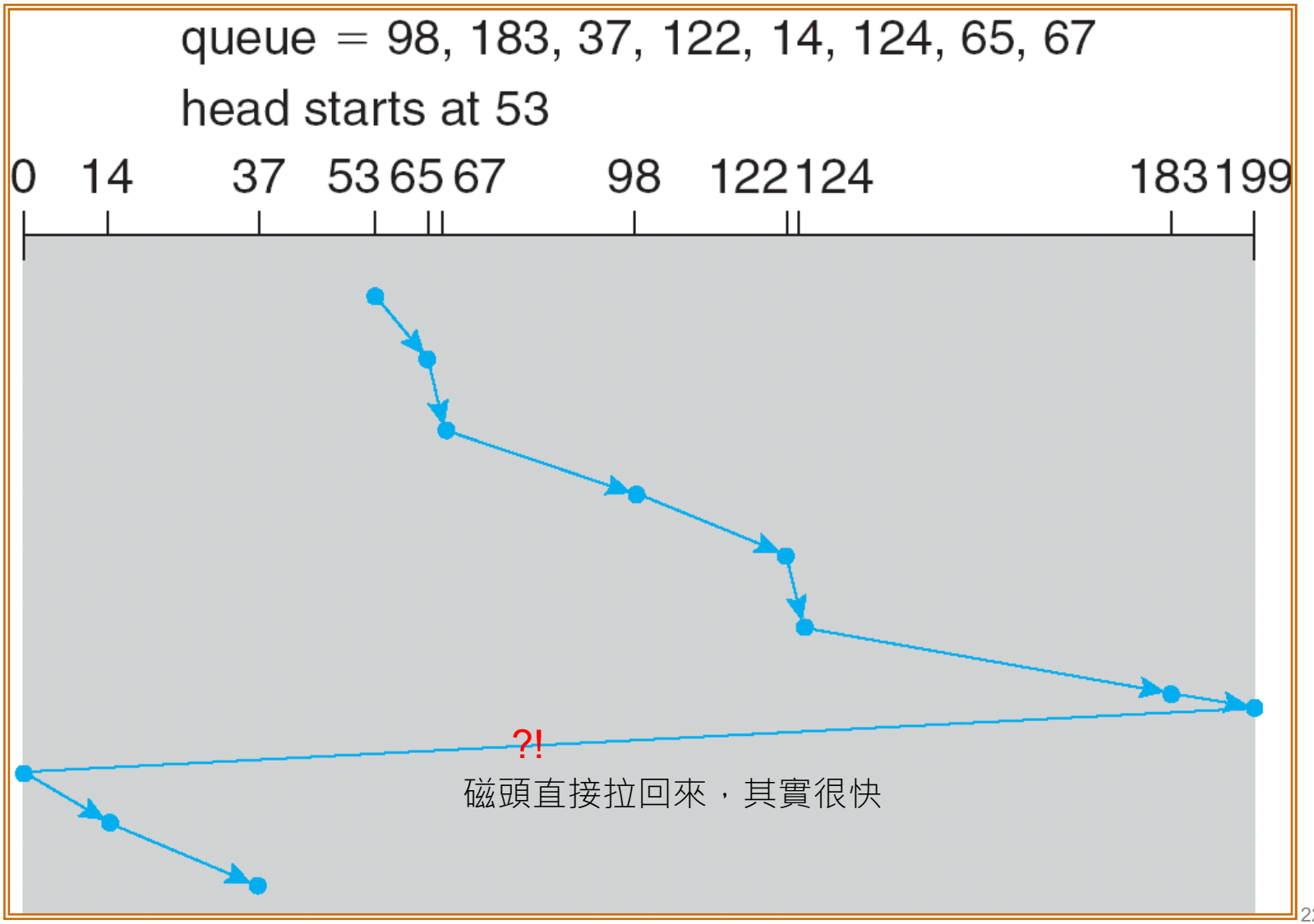
- The waiting time of each cylinder is not uniform
  - At the outermost or the innermost cylinder:
    - Max 2 full strokes and 1 reverse
  - At the middle of the disk:
    - Max: 2 half full strokes and 1 reverse

request的位置會影響wating time
兩邊的request會比中間的request更容易等更久

# C-SCAN Scheduling

- 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.

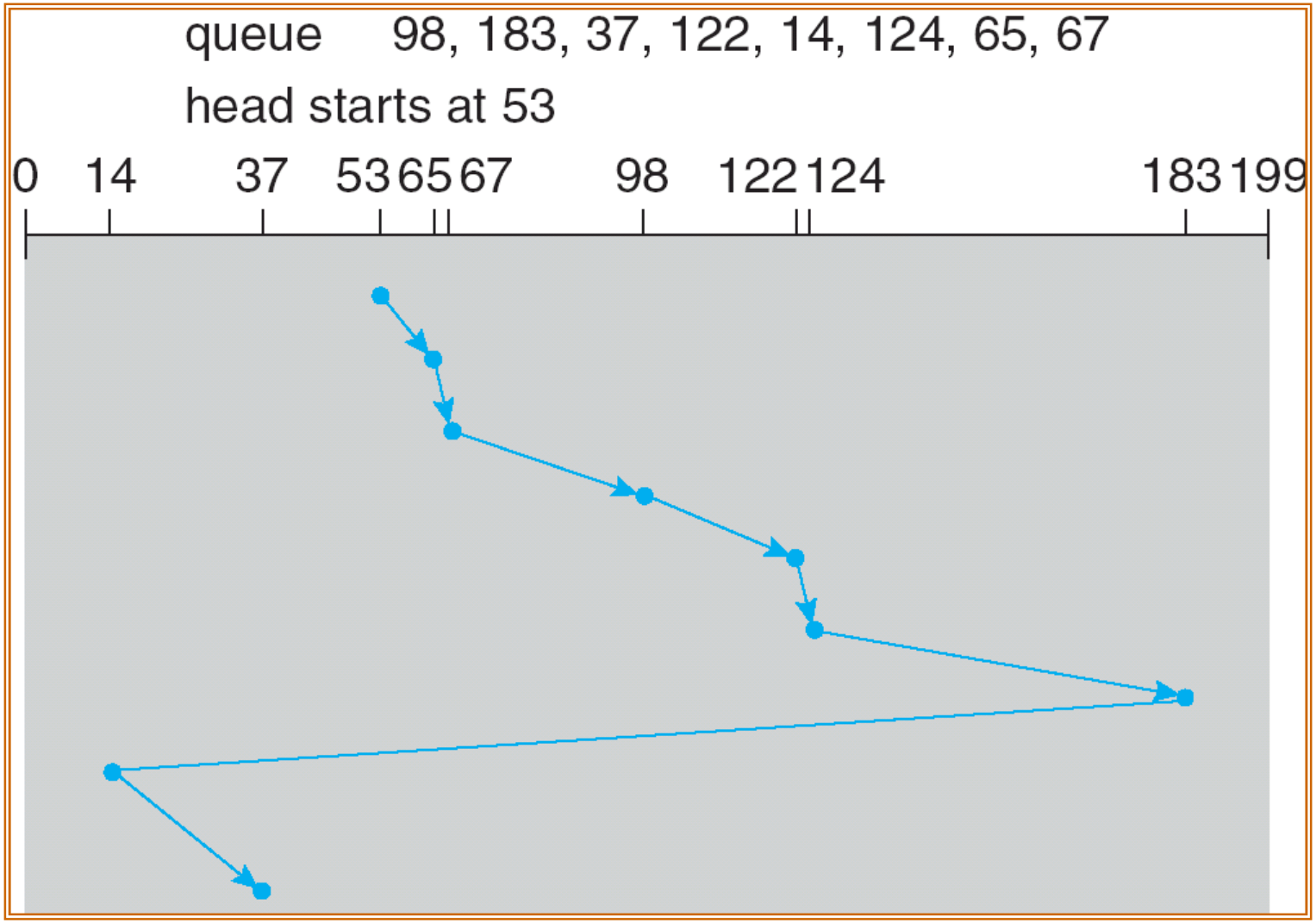
# C-SCAN Disk Scheduling 老師沒看過有實作這種的XD



# 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 DISK Scheduling



# Selection of a Disk-Scheduling Algorithm

大部分用這個

- SSTF is common and has a natural appeal
- SCAN and C-SCAN perform better for systems that place a heavy load on the disk.
- Either SSTF or LOOK is a reasonable choice for the default algorithm
- Performance depends on the number and types of requests.
- Requests for disk service can be influenced by the file-allocation method. 有效率的磁碟讀寫不單是磁碟排程而已
- 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.

實作上會同時維護 FCFS和SCAN，一般狀況用SCAN
如果最早的那個等太久就切成FCFS

# Disk Seek Optimization

lantency = xfor + seek(磁頭) +rotation(碟片旋轉)

- Disk scheduling problem is inherently NP-hard
  - All the methods mentioned above are not optimal (in terms of the total seek distance)
- To reduce seek penalty
  - Highly correlated data can be put in neighbor disk space
- Modern disks impose nearly the same overheads on seek and rotation
  - To optimize rotational delay in OS is hard, and the HDD firmware has better knowledge on the rotation angle
  - New HDDs accepts a number of requests and then reorder them in consideration of rotational delay
  - E,g., SATA NCQ (Native Command Queuing)

OS丟一堆request給磁碟交給他自己排程

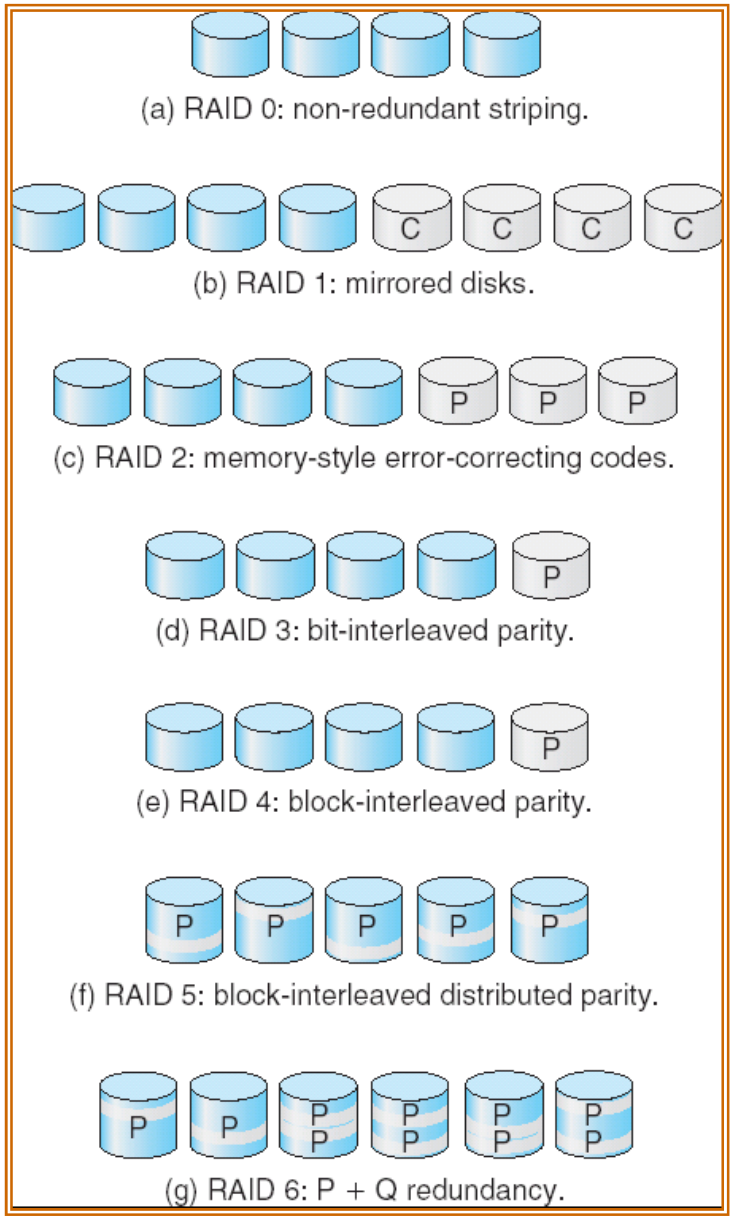
磁碟自己的控制器有比較多資訊(rotation etc.)，可以更好的降低lantency

# RAID Structure

- RAID – Redundant Array of Inexpensive Disks
  - Performance improvement through parallelism
  - Reliability improvement through redundancy

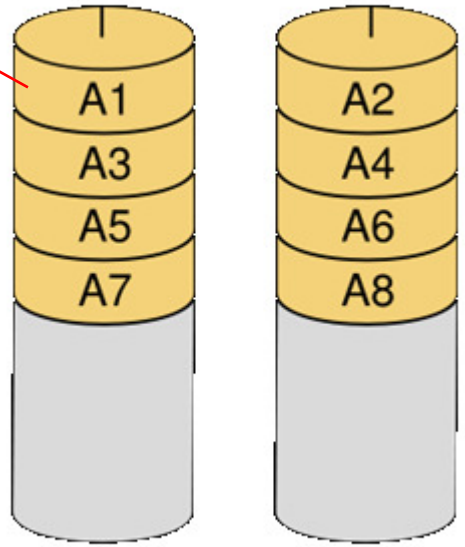
- RAID schemes improve performance and improve the reliability of the storage system by storing redundant data.
  - Mirroring or shadowing keeps duplicate of each disk.
  - Block interleaved parity uses much less redundancy.

# RAID Levels




最常用的應該是0, 1, 5

stripe block
約64K~1M



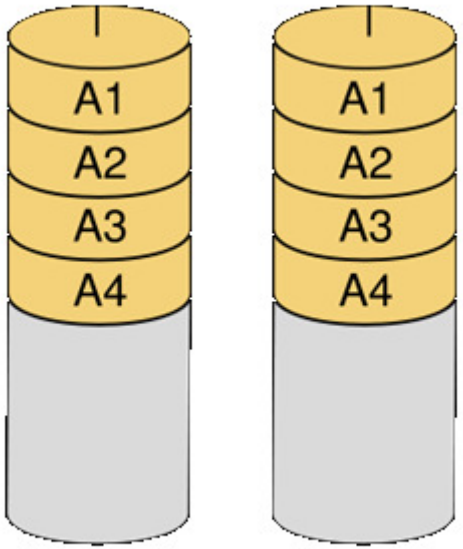

檔案大的話，能提升平行度
提升速度

RAID0

不過現在不太有人直接敢用
因為一旦一顆壞掉就整組壞光XDD

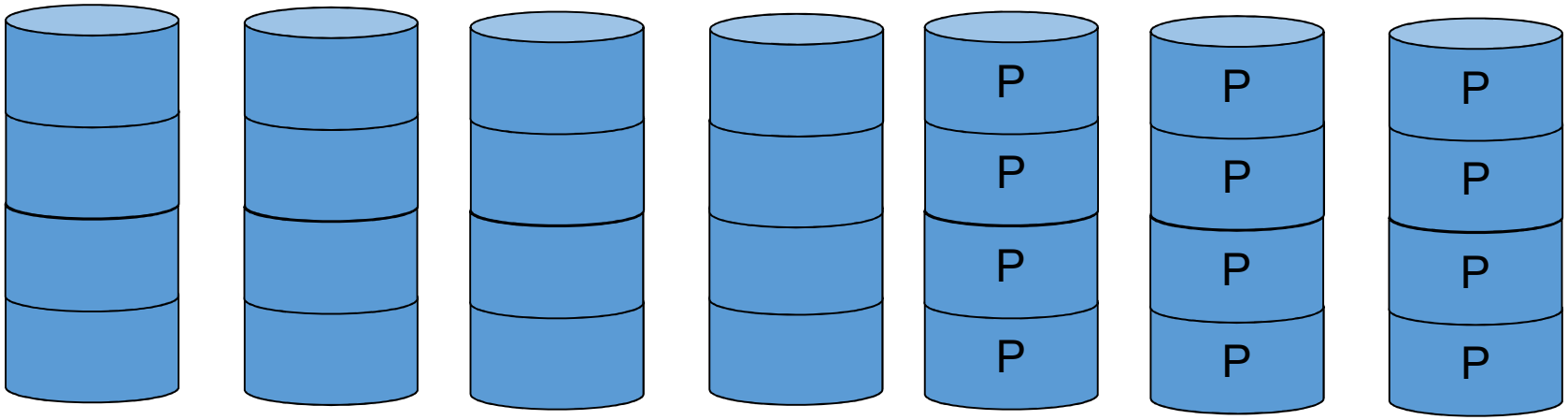
Striping. Aiming at parallelism

Mirror



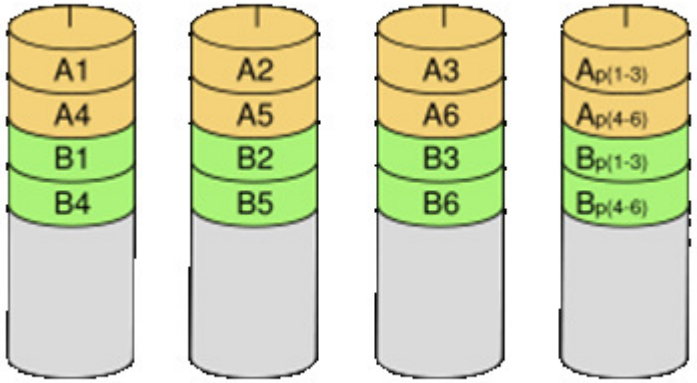

RAID1

Mirroring, 100% redundancy




RAID-2: memory-style ECC

# of parity disks = log2(# of data disks)

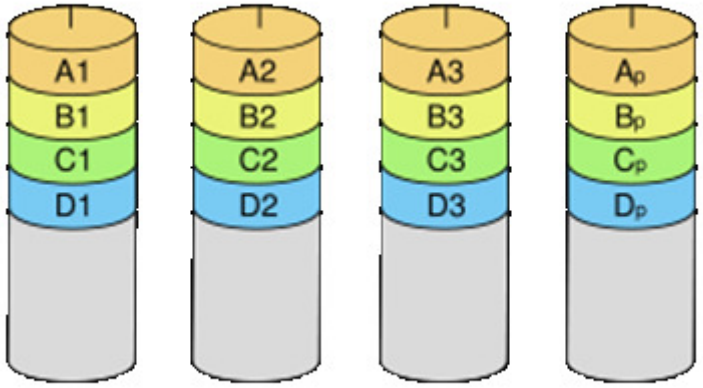



RAID 3

Bit-interleaved
(or sub-block-interleaved)

Fully interleaved, one R/W involves all disks

$A_p = A_1 \oplus A_2 \oplus A_3$
只有一顆壞掉還是能救出來

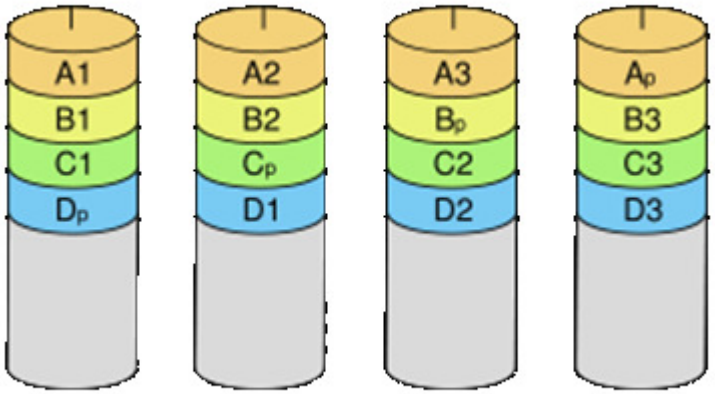



讀寫會很慢，如果A2要更新，$A_p = A_p \oplus A2 \oplus A2'$
P磁碟會變成瓶頸

RAID4

Block-interleaved

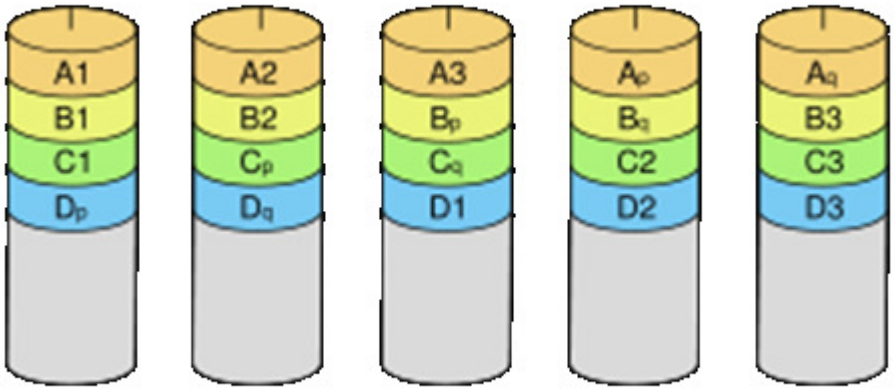
One R involves one disk
One W involves two disks
Parity disk → bottleneck




把P的更新流量分散到每顆磁碟上

RAID 5

One R involves one disk
One W involves two disks
Parity is spread over all disks




編碼方式不是⊕，而是更複雜的
壞掉兩顆依然能救回來

RAID 6

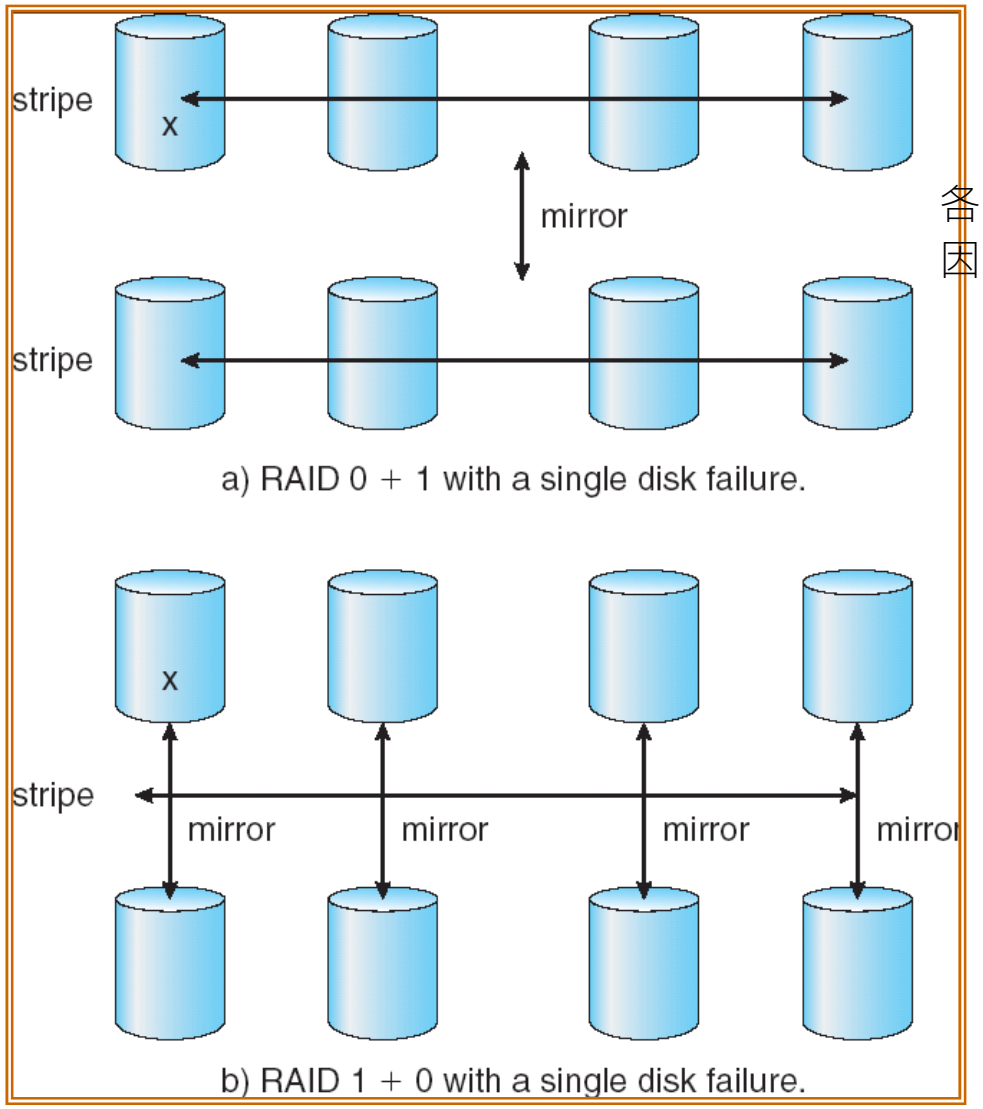
Choosing 4 blocks out of {1,2,3,4,p,q} sufficiently reconstruct {1,2,3,4}

# RAID-5 Relibility

- Let the probability of 1-year up of a disk be *p*
  - The probability of 1-year up of a RAID-0 of 4 disks:
    - $p^4$ (~0.96 if p=0.99)
  - The probability of 1-year up of a RAID-5 of 5 disks:
    - $p^5+(5,1)(1-p)*p^4$ (~0.999 if p=0.99)

壞掉一顆的機率

# RAID 0 + 1 and 1 + 0




各壞一個就全體報銷了
因為不一定會一一對應

← Better survivability

# Solid-State Disks (SSDs)

以block為I/O單位

- Storage devices that emulate standard block devices using non-volatile memory
  - Flash memory or battery-backed RAM
  - The OS use the legacy I/O stack on top of SSDs
- Products
  - Embedded flash cards, SD cards, USB thumb drives, SSDs, PCI-e flash cards
- Performance
  - RAM disk > SSD >> HDD
- Applications
  - Cloud storage: tier storage, cache SSDs
  - Personal computer: HDD replacement, system drive
  - Embedded storage: Smartphones, tablets, laptops, wearables

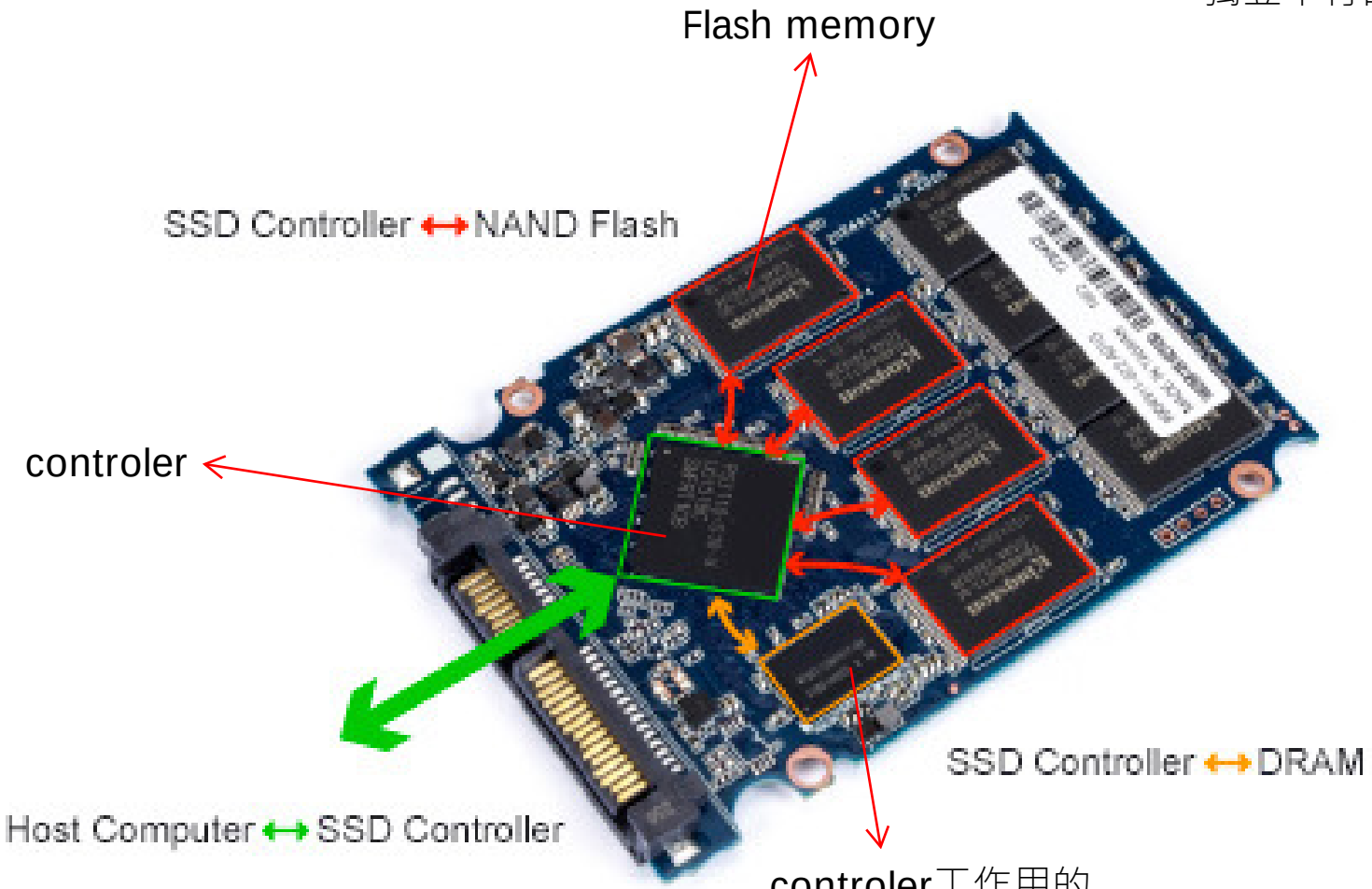
非揮發性記憶體

傳統I/O的流程

在RAID和DRAM中間再加上SSD當cache

# SSD Internal Organization

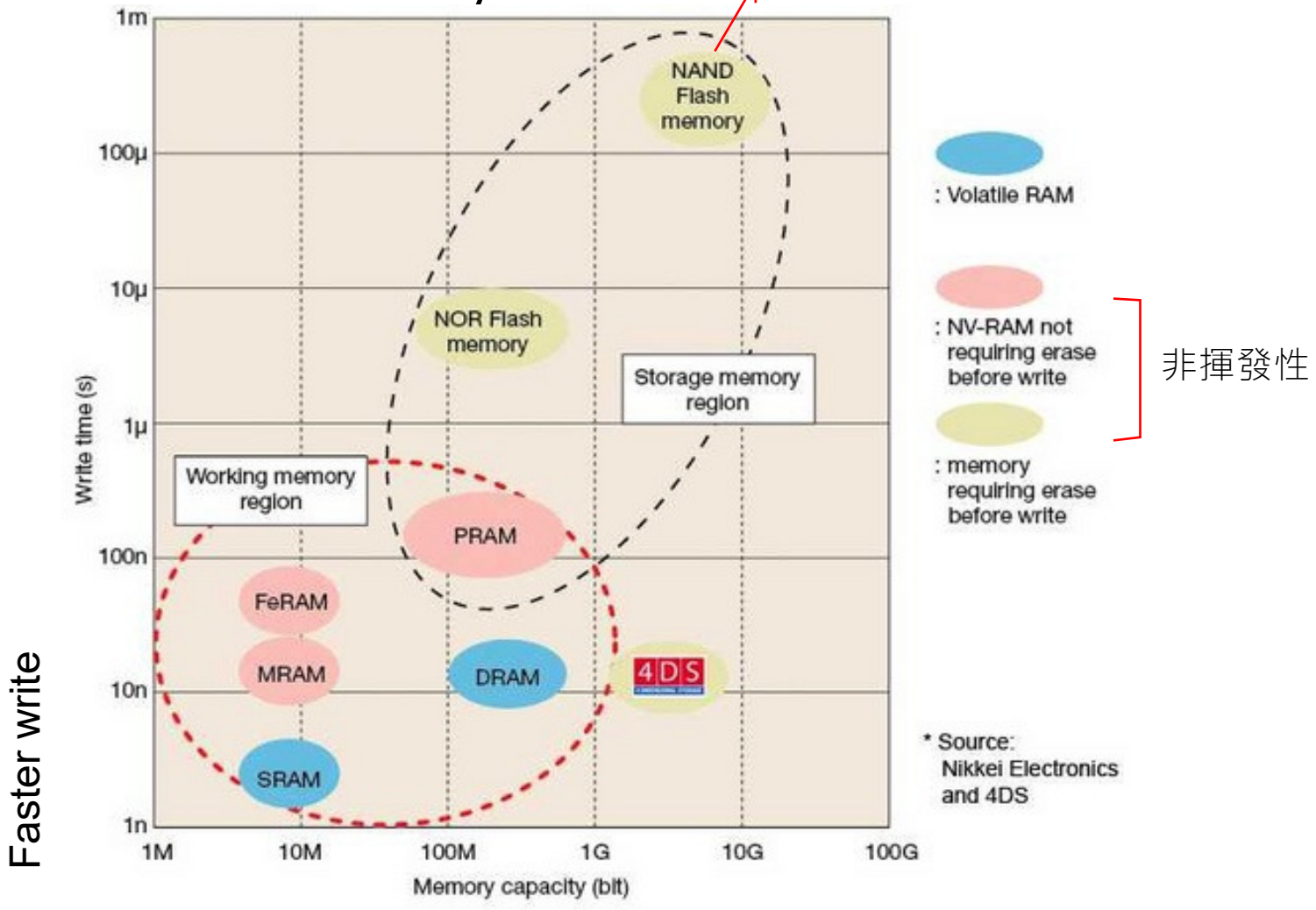
每一顆Flash memory都可以獨立平行讀寫




https://www.kingston.com/tw/ssd/data-protection

# Non-Volatile Memory

有點誇張，大概100u以內




非揮發性

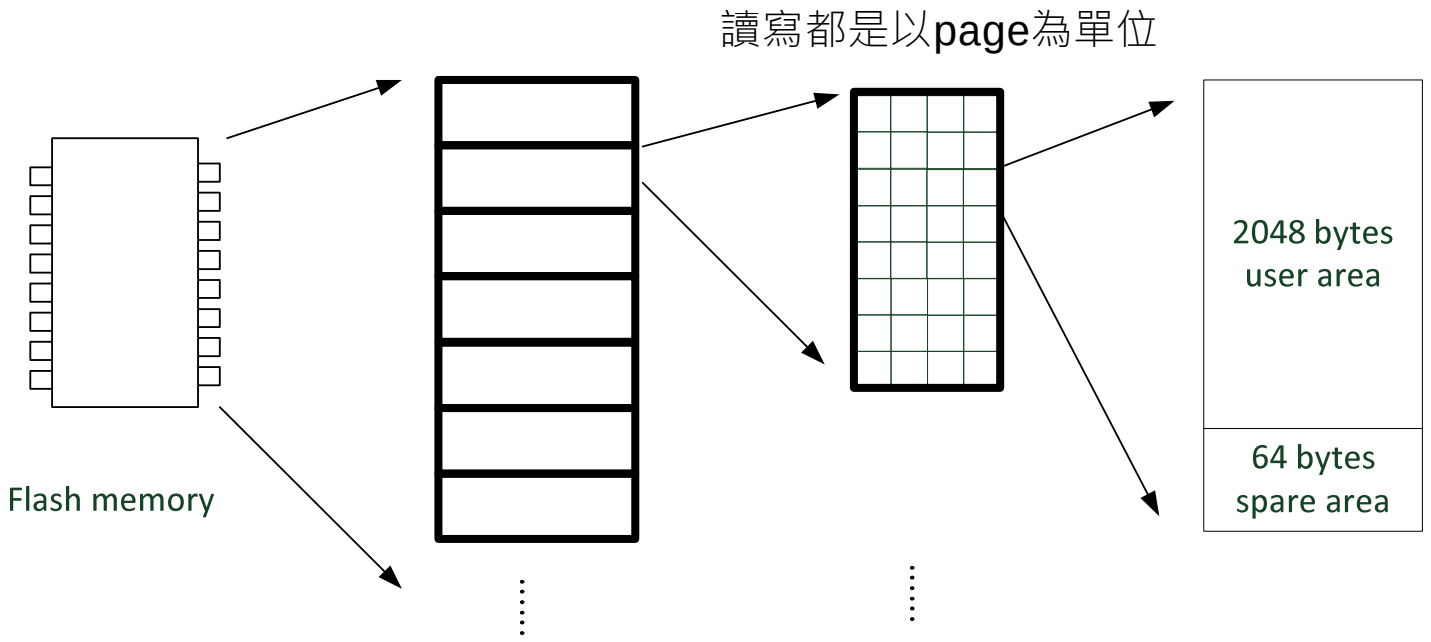
Faster write

Smaller capacity

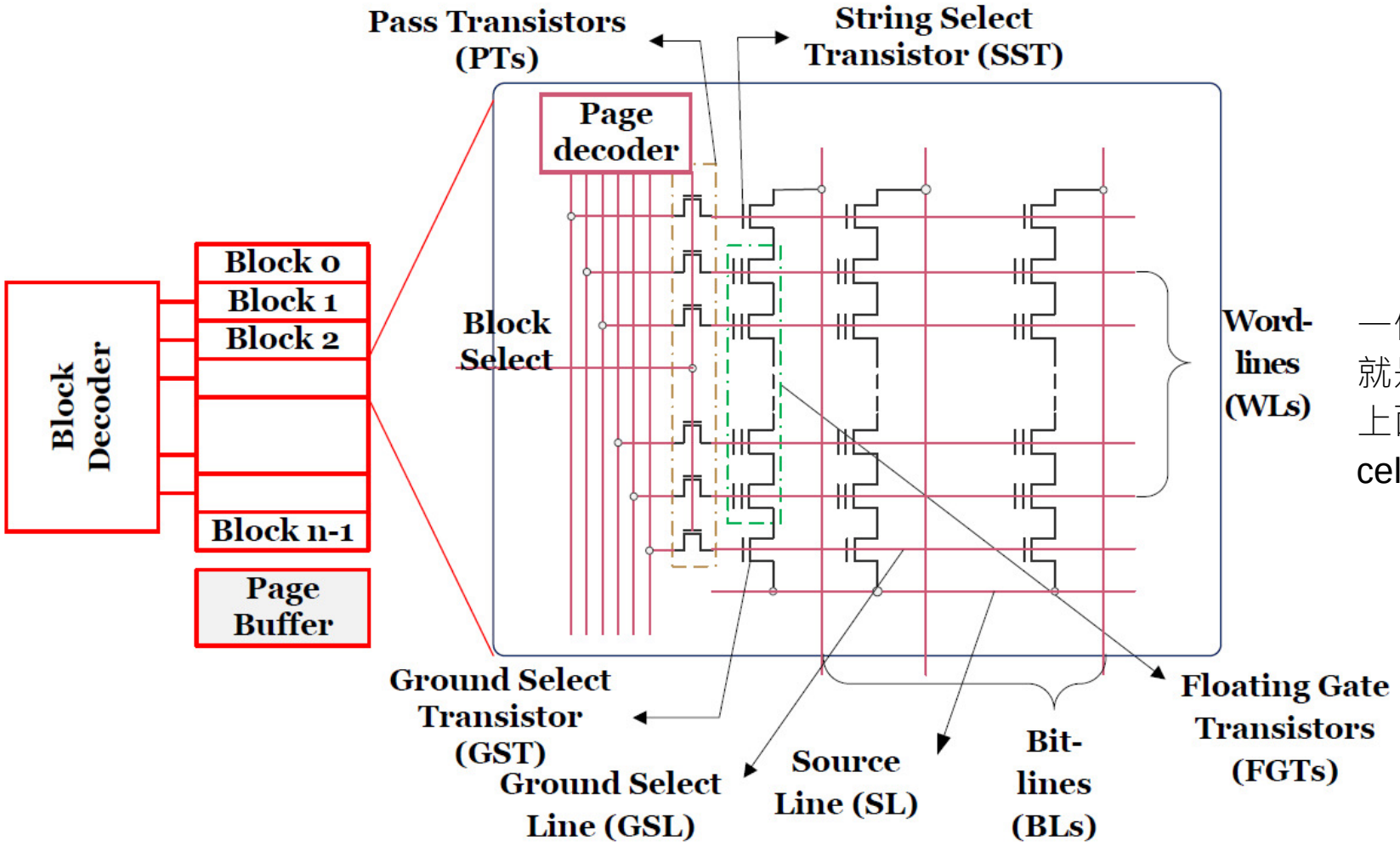
基本上越小的越貴

http://www.storagenewsletter.com/news/flash/4ds-start-up-profile?snc=7835

# NAND Flash Geometry




- Unit size
  - Read/write: page
  - Erase: block

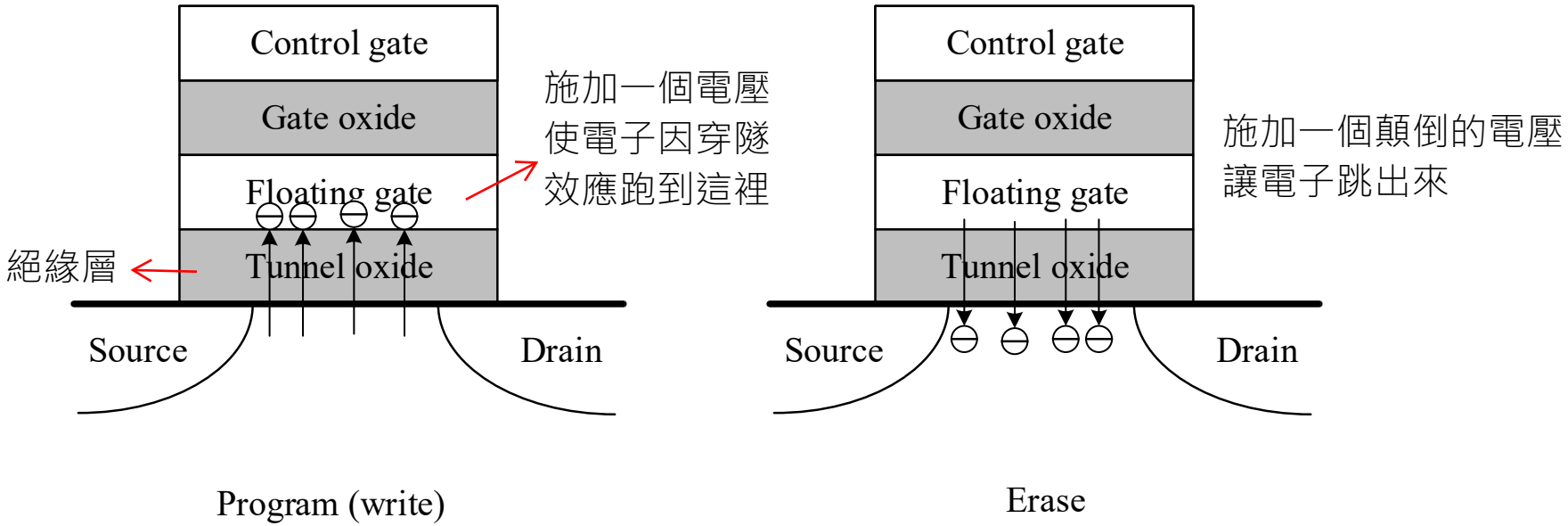



一條word-line就是1 page上面有很多的cells

# Flash Memory

儲存1, 0的單位

- Cell structure, flash program and erase



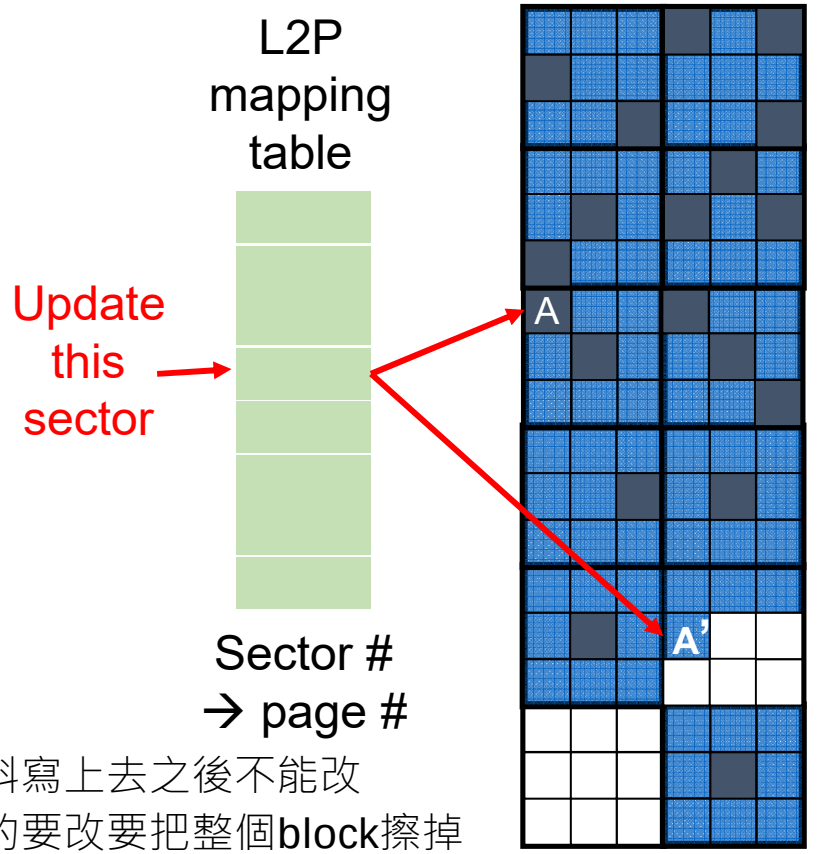
# Flash Translation Layer (FTL)

要把OS對一般硬碟的命令傳換為SSD的方式，由controler來做
把flash memory的物理特性掩藏起來(我是顆HDD XD)

- A firmware layer inside of SSDs
  - Hiding flash memory physics from the host
- Provide block device emulation to the host
- Manage flash memory inside of SSDs
  - Logical-to-physical address translation
  - Garbage collection
  - Wear leveling

# Logical-to-Physical Address Translation

- Pages cannot be overwritten unless being erased
- Erase a block every time a page is overwritten
  - Too inefficient
- Out-of-place update; mark old data invalid
- Need logical-to-physical address translation
  - From logical sector # to physical page #




資料寫上去之後不能改
真的要改要把整個block擦掉
再整塊重寫上去
G8慢→不這樣做
實作→out of place的寫入，直接抓一個空的位置寫入
位置會隨著更新變更→mapping table
將邏輯位置轉為物理的位置

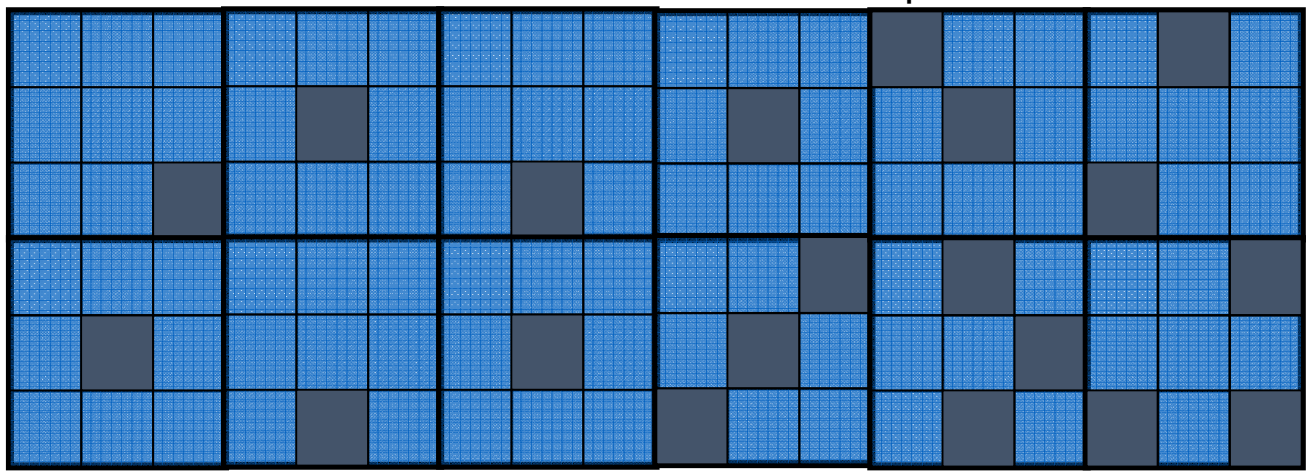
# Garbage Collection

如果滿了就要做Garbage collection

- Recycle memory space occupied by invalid data through block erasure
- Victim selection
  - Minimize the page-copy overhead

把garbage最多block的移到working-space
然後把原來的block整個擦掉→garbage消失

會需要額外的資料結構來區分garbage or clean space



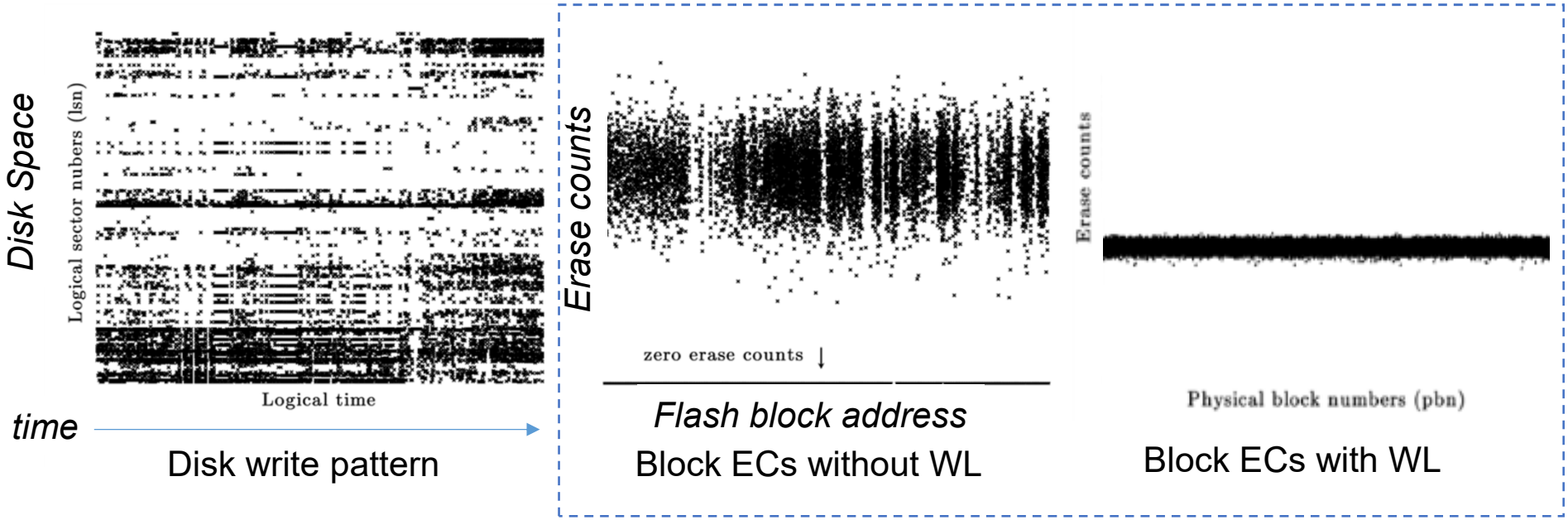
需要一個空白的working-space

Valid page data

Invalid page data

# Wear Leveling

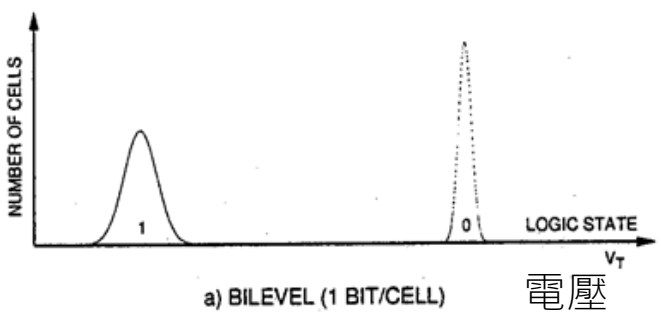
- Typically a (MLC) block endures 3000 cycles of program-erase operations (P/E cycles)
- Locality of write creates frequently written blocks
- Delay the first block retirement by migrating cold data



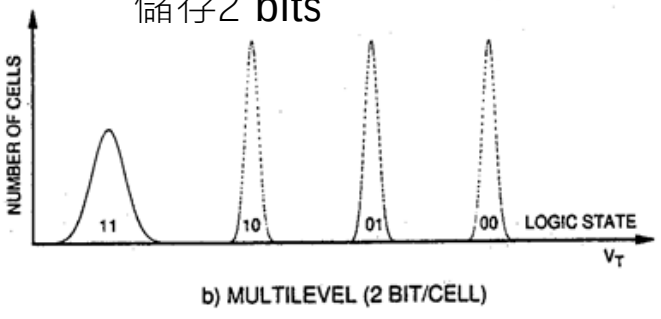

*Li-Pin Chang, Tung-Yang Chou, and Li-Chun Huang, "An Adaptive, Low-Cost Wear-Leveling Algorithm for Multichannel Solid-State Disks," ACM Transactions on Embedded Computing Systems, Volume 13, Issue 3, 2013.*

# Multilevel Cells

切為四段→一個cell可以
儲存2 bits




電壓




Single-level cell

Multi-level cell

快買不到了QQ

- SLC vs. MLC flash
  - Comparable read speed
  - SLC writes about 2x or 3x faster than MLC
  - P/E endurance: 5K cycles (MLC), 100K (SLC)
  - SLC is 2x or 3x more expensive than MLC (and increasing)
  - Hybrid SSDs, dynamic density SSDs
- Now TLC, QLC are in mass production

3bits 4bits

MLC程度越高寫入速度越慢
而且越容易磨損

# End of Chapter 12