作業系統：Storage Management

**目錄**

[作業系統：Storage Management 1](#_Toc203917709)

[1. 大量儲存結構概觀(Mass-Storage Structure Overview) 1](#_Toc203917710)

[1.1 硬碟(Hard Disk Drive, HDD) 1](#_Toc203917711)

[1.2 非揮發性記憶體裝置(NVM Devices) 2](#_Toc203917712)

[1.3 使用揮發性記憶體作為儲存媒介(Volatile Memory as Mass Storage) 3](#_Toc203917713)

[2. 位址對應(Address Mapping) 3](#_Toc203917714)

[2.1 邏輯區塊位址(Logical Block Address, LBA) 3](#_Toc203917715)

[2.2 實體位址(Physical Address) 3](#_Toc203917716)

[2.3 硬碟位址映射問題：為什麼不能直接對應 3](#_Toc203917717)

[3. 硬碟排程(HDD Scheduling) 4](#_Toc203917718)

[3.1 FCFS(First Come First Serve)演算法 4](#_Toc203917719)

[3.2 SCAN(電梯演算法) 4](#_Toc203917720)

[3.3 Circular SCAN(C-SCAN) 5](#_Toc203917721)

[3.4 NVM Scheduling(非揮發性儲存裝置排程) 5](#_Toc203917722)

[3.5 Write Amplification(寫入放大) 6](#_Toc203917723)

[3.6 錯誤偵測與更正(Error Detection and Correction) 6](#_Toc203917724)

[4. 儲存裝置管理(Storage Device Management) 6](#_Toc203917725)

[4.1 格式化：低階格式化(Low-level Formatting) 6](#_Toc203917726)

[4.2 格式化：邏輯格式化(Logical Formatting) 6](#_Toc203917727)

[4.3 分割(Partitioning) 6](#_Toc203917728)

[4.4 檔案系統掛載(Mount) 6](#_Toc203917729)

[5. 儲存裝置連接方式(Storage Attachment) 7](#_Toc203917730)

[5.1 主機直連儲存(Host-Attached Storage) 7](#_Toc203917731)

[5.2 網路附加儲存(Network-Attached Storage, NAS) 7](#_Toc203917732)

[5.3 Cloud Storage 7](#_Toc203917733)

[6. RAID 結構(Redundant Arrays of Independent Disks) 8](#_Toc203917734)

[6.1 RAID Levels(RAID 等級) 8](#_Toc203917735)

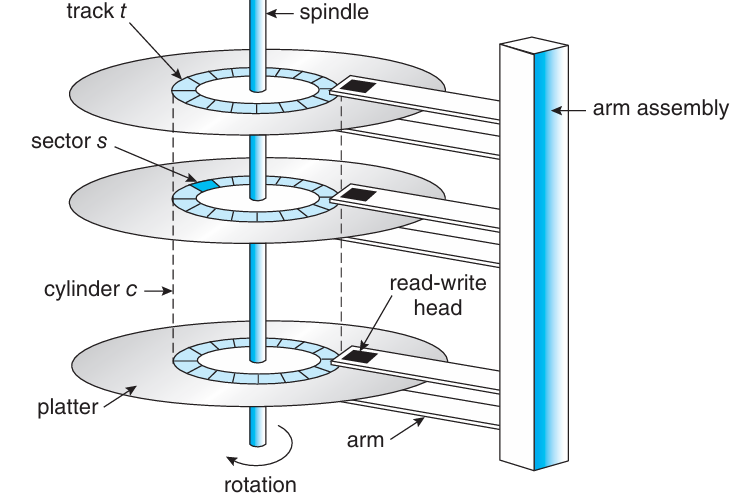
# 基本概念

## 硬碟(Hard Disk Drive, HDD)

基本組成如下：

|  |  |
| --- | --- |
| Platter(碟片) | 每個碟片為扁平的圓形面，表面塗有磁性物質(用來儲存資料) |
| 磁頭(Read-Write Head)： | 浮在碟片表面上方，用來讀寫資料 |
| Arm(手臂) | 將磁頭固定在一起，移動 arm 就可同時移動所有磁頭。 |
| Track | 圓形軌道，磁碟表面被邏輯劃分為一圈圈的 track |
| Sector | 每個 track 再細分為 sector(最小單位)，大小通常為 512 bytes 或 4 KB。 |
| Cylinder | 所有磁頭在同一 arm 位置下的 track 組成的集合 |
| Spindle(軸心) | 碟片會高速旋轉，轉速以 RPM(轉/分)表示，例如 5400 RPM、7200 RPM |

Figure 1：HDD構成



對於HDD存取時間與效能指標：

|  |  |
| --- | --- |
| Seek Time | 移動磁頭到對應 cylinder 的時間 |
| Rotational Latency | 等待 sector 旋轉到磁頭下方的時間(平均為半圈時間) |
| Transfer Rate | 資料從磁碟到記憶體的傳輸速度 |
| DRAM Buffer | 現代 HDD 有內建快取記憶體(快取讀寫資料) |

HDD常見的風險為Head Crash，即磁頭接觸到碟面，導致資料損壞，需更換整顆硬碟。

## 非揮發性記憶體裝置(NVM Devices)

NVM(Non-Volatile Memory)即使斷電資料仍保留。常見於SSD、USB 隨身碟與嵌在主機板上的 eMMC。

Table 1：SSD(NVM) vs. HDD

|  |  |  |
| --- | --- | --- |
| 項目 | SSD(NVM) | HDD |
| 結構 | 無移動部件(電子元件) | 有磁碟片、磁頭與馬達 |
| 速度 | 快(無 seek time、低 latency) | 慢(有 seek time、旋轉延遲) |
| 耐用度 | 擦寫次數有限(需配合 wear-leveling) | 易受物理損壞(如 head crash) |
| 價格/容量 | 較貴，單位容量較小 | 較便宜，單位容量大 |

雖然 NVM 是一個統稱，但絕大多數市面上的 NVM 裝置(SSD、USB、eMMC)內部都是Flash memory，尤其是 NAND Flash。其中Flash memory有以下獨特性：

Table 2：Flash的限制

|  |  |
| --- | --- |
| 無法直接覆寫 | 修改資料時需整個 block 擦除後重寫 |
| 擦除單位大(block) | 雖然寫入單位是 page(例如 4KB)，但擦除要整個 block(例如 256KB) |
| 有限寫入壽命 | 每個 cell 有次數上限(如 10 萬次) |
| 需做 wear-leveling | 為避免熱區損壞，須平均寫入分布 |

## 使用揮發性記憶體作為儲存媒介(Volatile Memory as Mass Storage)

雖然 DRAM 是揮發性記憶體(Volatile Memory)，但它可以被用來模擬成一種儲存裝置，稱為 RAM disk 或 RAM drive。

Table 3：RAM Disk / RAM Drive

|  |  |
| --- | --- |
| **特性** | **說明** |
| 本質 | 用驅動程式將部分 DRAM 模擬成儲存裝置，像是一顆「虛擬硬碟」 |
| 操作方式 | 可以格式化成檔案系統、像普通磁碟一樣存取 |
| 使用情境 | 暫存檔案、高速臨時資料存取(比 SSD 快) |
| 儲存風險 | 斷電資料即失(因為是揮發性記憶體) |

# 位址對應(Address Mapping)

現代作業系統中，檔案資料的儲存通常會以兩種位址方式來描述：

## 邏輯區塊位址(Logical Block Address, LBA)

是一種抽象化的位址編號方式。儲存空間視為一個連續的一維陣列(array)，每個元素稱為一個邏輯區塊(Logical Block)。每個 LBA 對應到底層儲存裝置上的一個實體儲存單位，像是HDD的sector與SSD的page。

## 實體位址(Physical Address)

是裝置實際讀寫時所對應的物理位置，例如磁碟上的磁道、柱面、扇區，或是 NAND Flash 上的 page/block。

## 硬碟位址映射問題：為什麼不能直接對應

* 壞軌替換(Bad Sector Remapping)：硬碟內部會偷偷把壞的 sector 替換掉，LBA 沒變但實體 sector 改變了。
* 不同 track 的 sector 數不同(Zoned Bit Recording)：外圍的 track 可以放比較多 sector。
* 廠商不公開 mapping 機制：現代硬碟內部的 sector 排列方式是製造商決定的，不再單純由 OS 控制。

Table 4：RAM Drive、HDD、NVMe 比較

|  |  |  |  |
| --- | --- | --- | --- |
| **項目** | **RAM Disk** | **SSD / NVMe** | **HDD** |
| 儲存類型 | DRAM | Flash NAND | 磁碟片 |
| 是否揮發 | ✅ 是 | ❌ 否 | ❌ 否 |
| 存取速度 | 🚀 非常快 | 🔼 快 | 🐢 慢 |
| 資料保存 | ❌ 斷電即失 | ✅ 可長期保存 | ✅ |
| 使用場合 | 暫存檔案、初始化、安裝程序 | 作業系統、應用程式、快取 | 大容量備份、冷資料 |

# 硬碟排程(HDD Scheduling)

在 HDD 上，作業系統的目標是：

* 最小化 Access Time(存取時間)
* 最大化 Bandwidth(資料傳輸頻寬)

其中，對於HDD存取時間組成可以分成：

Table 5：HDD存取時間組成

|  |  |
| --- | --- |
| **項目** | **說明** |
| Seek Time | 磁頭移動至目標磁軌(Cylinder)的時間 |
| Rotational Latency | 等待磁碟旋轉至目標區塊(Sector)的位置 |
| Bandwidth | 總資料量 ÷ 第一次請求到最後一筆資料完成的時間 |

## FCFS(First Come First Serve)演算法

操作方式是照請求到達的順序處理。

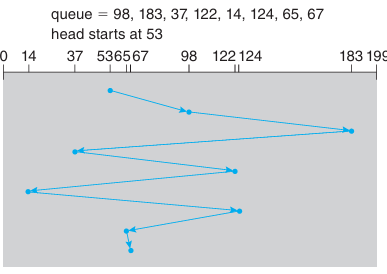
* 優點：公平簡單。
* 缺點：效率差(磁頭來回跳動嚴重)。

舉例而言：

53 → 98 → 183 → 37 → 122 → 14 → 124 → 65 → 67

總移動距離即為640 cylinders。

Figure 2：FCFS disk scheduling



## SCAN(電梯演算法)

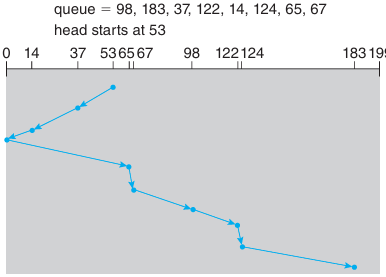
磁頭向一個方向掃描(如從內往外)，服務沿途的請求，碰到邊界後反轉方向。

* 優點：效率提升，避免極端跳動，
* 缺點：兩端的請求可能等比較久(遠端先被服務)

舉例而言：

53 → 37 → 14 → (到達 0，反轉) → 65 → 67 → 98 → 122 → 124 → 183

Figure 3：SCAN disk scheduling



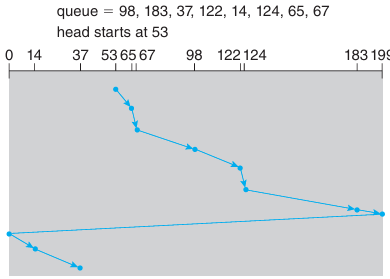
## Circular SCAN(C-SCAN)

只往一個方向掃描(如從 0 到最大磁軌)，到邊界後**瞬間回到開頭，**不服務回程請求

* 優點：提供一致的等待時間(避免方向切換造成的延遲)

舉例而言：

53 → 65 → 67 → 98 → 122 → 124 → 183 → (到達199，回到0) → 14 → 37



# NVM Scheduling(非揮發性儲存裝置排程)

NVM(Non-Volatile Memory)特性：

* 沒有機械構造：沒有磁頭移動 → 不需考慮「Seek Time」與「Rotational Latency」
* 資料存取方式：快速、支援隨機存取(IOPS 非常高)
* 排程策略：多為簡單的 FCFS(先來先服務)為主

然而，NVM也有些寫入效能限制：

* 讀取比寫入快
* 寫入效能會隨裝置使用壽命下降
* 寫入需考慮：擦除(Erase)限制、耗損壽命(Wear Leveling)與垃圾回收(Garbage Collection, GC)

## Write Amplification(寫入放大)

一次寫入請求可能導致多次額外的讀寫動作。以實際流程來看：

1. 寫入新資料 → 若該區塊已寫滿
2. GC 啟動：
   1. 讀出「有效資料」
   2. 寫入到空白區(Over-Provisioning 空間)
   3. 清除整個舊區塊
3. 寫入新資料

## 錯誤偵測與更正(Error Detection and Correction)

錯誤可以分為：

* Transient(軟性錯誤)：一次性的位元翻轉，例如：電磁干擾
* Permanent(硬性錯誤)：裝置受損或老化導致的資料錯誤

# 儲存裝置管理(Storage Device Management)

作業系統除了排程之外，還負責儲存裝置的其他管理，包括：磁碟初始化、從磁碟開機、損壞區塊(Bad Blocks)復原

## 格式化：低階格式化(Low-level Formatting)

又稱物理格式化，作業系統可將裝置劃分為可讀寫的「區段(Sector)」或「頁面(Page)」。

每個儲存單位會有：

* 標頭(Header)
* 資料區(Data Area)
* 結尾區(Trailer)

## 格式化：邏輯格式化(Logical Formatting)

建立檔案系統的資料結構(如檔案表、目錄、空間配置表等)

## 分割(Partitioning)

作業系統將磁碟劃分為多個區段(Partitions)，每個區段可以被當作獨立裝置處理，例如系統執行檔、Swap 空間、使用者檔案系統。一個磁區可對應一個 partition，也可跨多個分割區(如 RAID)。

## 檔案系統掛載(Mount)

挂載代表將檔案系統設為可用狀態。Windows 使用「磁碟代號」(C:, D:)。Linux 則掛載於一個統一的樹狀結構中。

# 儲存裝置連接方式(Storage Attachment)

作業系統可透過以下三種方式存取次級儲存設備：主機直連儲存(Host-Attached Storage)、網路附加儲存(Network-Attached Storage, NAS)、雲端儲存(Cloud Storage)。

## 主機直連儲存(Host-Attached Storage)

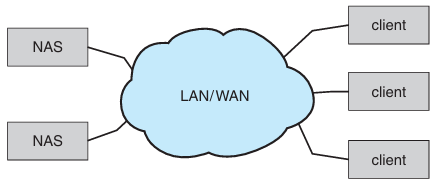
通過本機 I/O 介面連接，如：SATA、USB、FireWire、Thunderbolt。可連接裝置：HDD、SSD、CD/DVD、藍光、磁帶、SAN。儲存裝置之間以「邏輯資料區塊」進行讀寫(以 bus ID 或邏輯單元識別)。

## 網路附加儲存(Network-Attached Storage, NAS)

透過網路(通常為 LAN)提供儲存資源。NAS 可以是專用儲存設備，也可以是分享儲存空間的普通電腦。

* 優點
  + 可讓多台電腦共享檔案
  + 使用者可在多台 NAS 客戶端中，同步存取個人目錄
* 缺點：
  + 效能比不上直連儲存(會消耗網路頻寬)
  + 用於大型用戶端－伺服器架構時，可能導致延遲增加

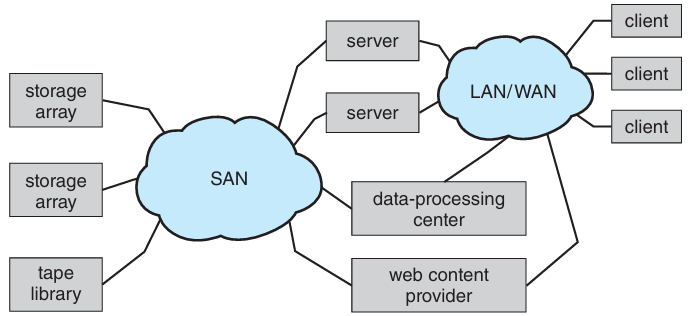
Figure 4：Network-attached storage



## Cloud Storage

和 NAS 類似，但存取的是遠端資料中心，透過 WAN(如 Internet) 連線。像是Amazon S3、Dropbox、OneDrive、iCloud。

Figure 5：Storage-area network.



# RAID 結構(Redundant Arrays of Independent Disks)

RAID 是一種磁碟組織技術，主要目標是：

* 提高效能(透過平行存取)
* 增強可靠性(透過冗餘資料)

## RAID Levels(RAID 等級)

待整理