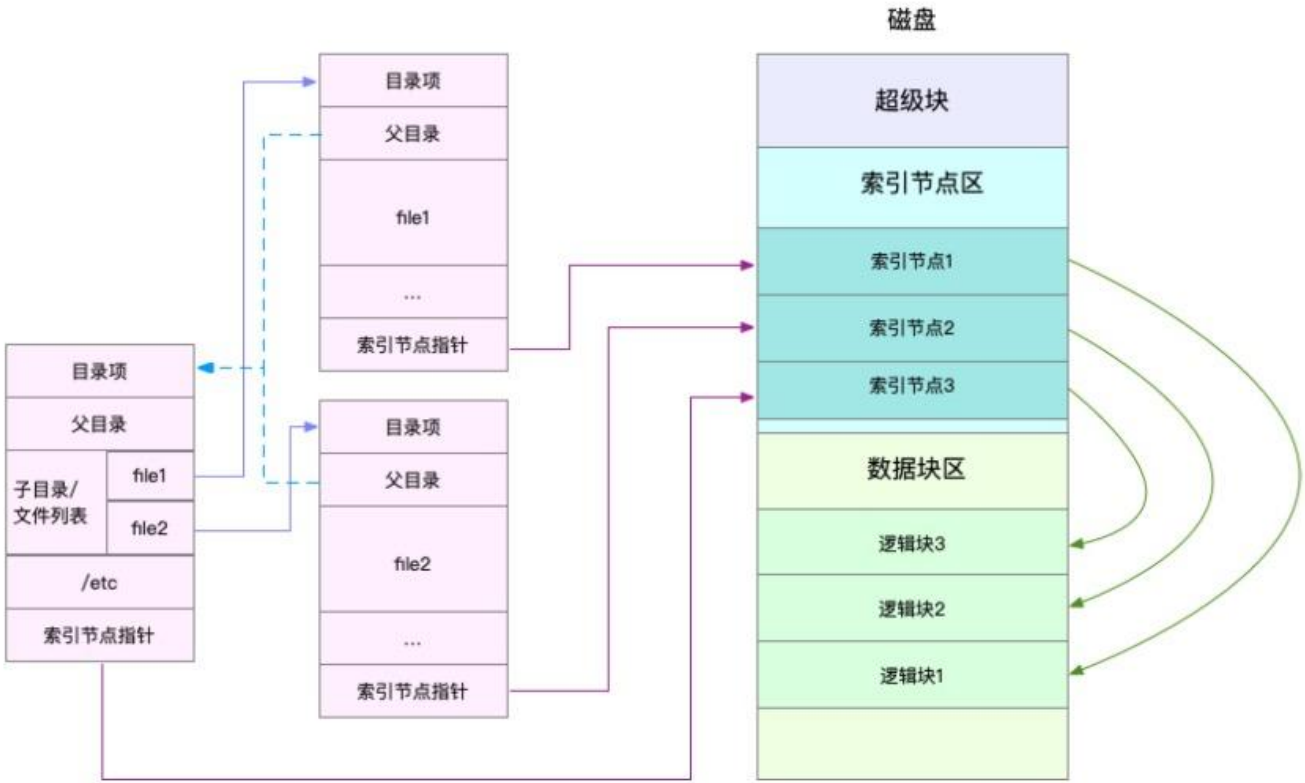
Linux 硬碟 I/O 是怎麼工作（上）

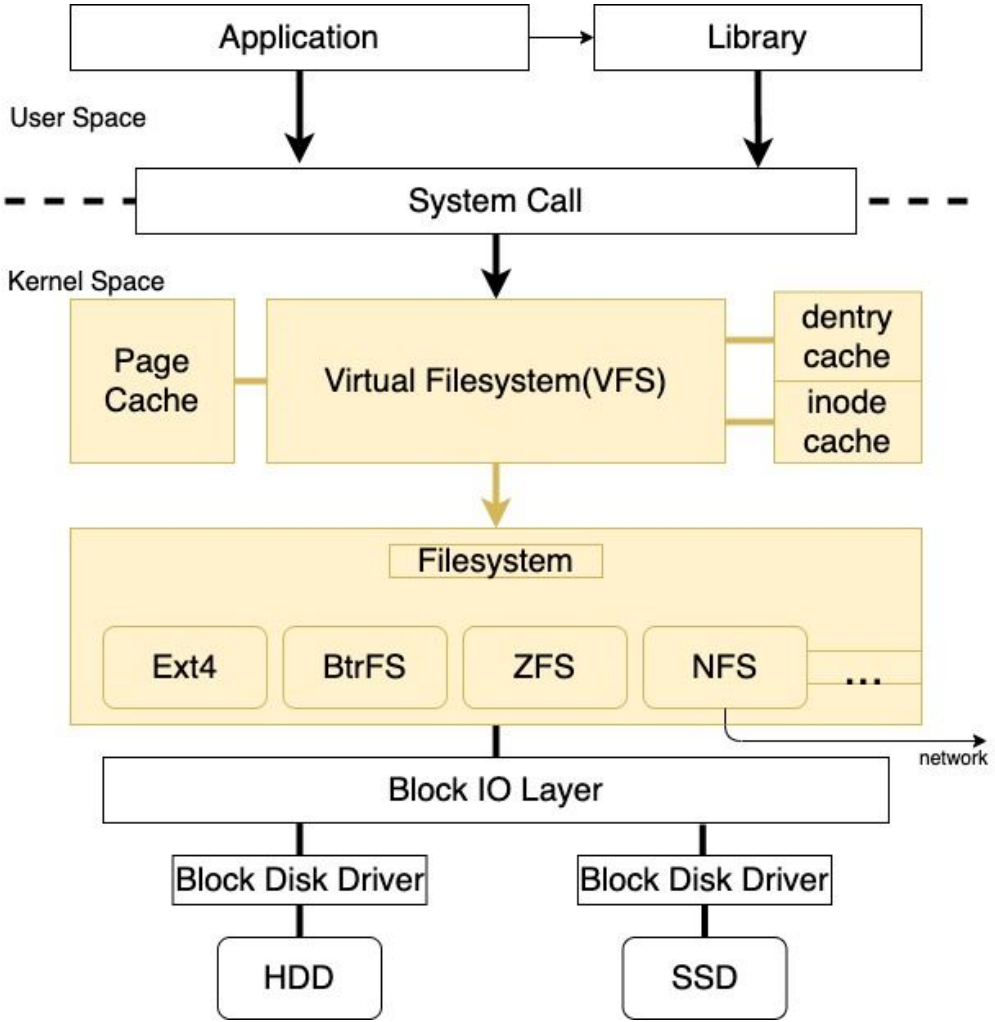
Raix Lai
2020/7/30

Linux 中一切皆為檔案


磁盘
超级块
索引节点区
索引节点1
索引节点2
索引节点3
数据块区
逻辑块3
逻辑块2
逻辑块1
目录项
父目录
file1
...
索引节点指针
目录项
父目录
子目录/
文件列表
file1
file2
/etc
索引节点指针
目录项
父目录
file2
...
索引节点指针

檔案儲存結構

Linux IO Stack



VFS 提供的統一接口

- dentry
- Inode
- file
- block
- superblock

Block Device

- Block devices are characterized by random access to data organized in fixed-size blocks.
- 硬碟是常見的 block device，可以永久化儲存數據。
- 常見的兩大類硬碟
  - 機械硬碟 (Hard Disk Drive, HDD)
  - 固態硬碟 (Solid State Drive, SSD)

機械硬碟

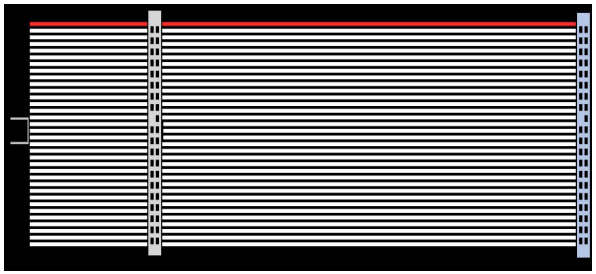
- 硬碟磁盤
- 讀寫磁頭
- 數據儲存在磁盤上的環狀軌道
- 讀寫時需要移動位置
- 需要依順序讀寫
- 隨機讀寫效率差
- 最小讀取單位:
  - 512 byte (sector)

固態硬碟

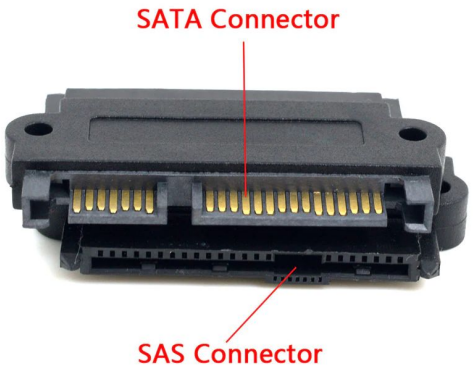
- 固態電子元件組成
- 不需要磁軌尋址
- 速度快
- 價格昂貴
- 損壞時難以拯救
- 有寫入次數壽命
- 最小讀取單位:
  - 4KB (page)

儲存裝置接口

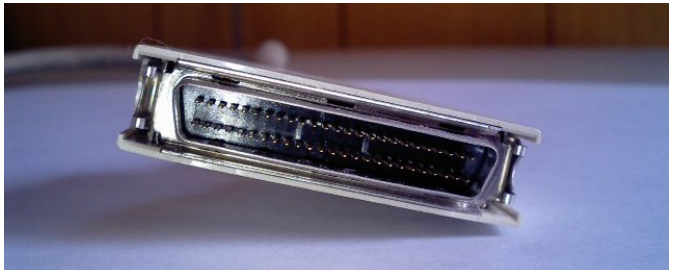
IDE



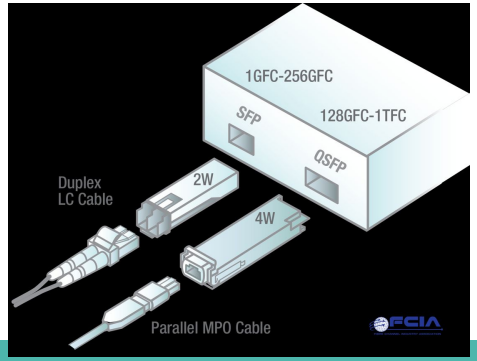
SAS / SATA



SCSI

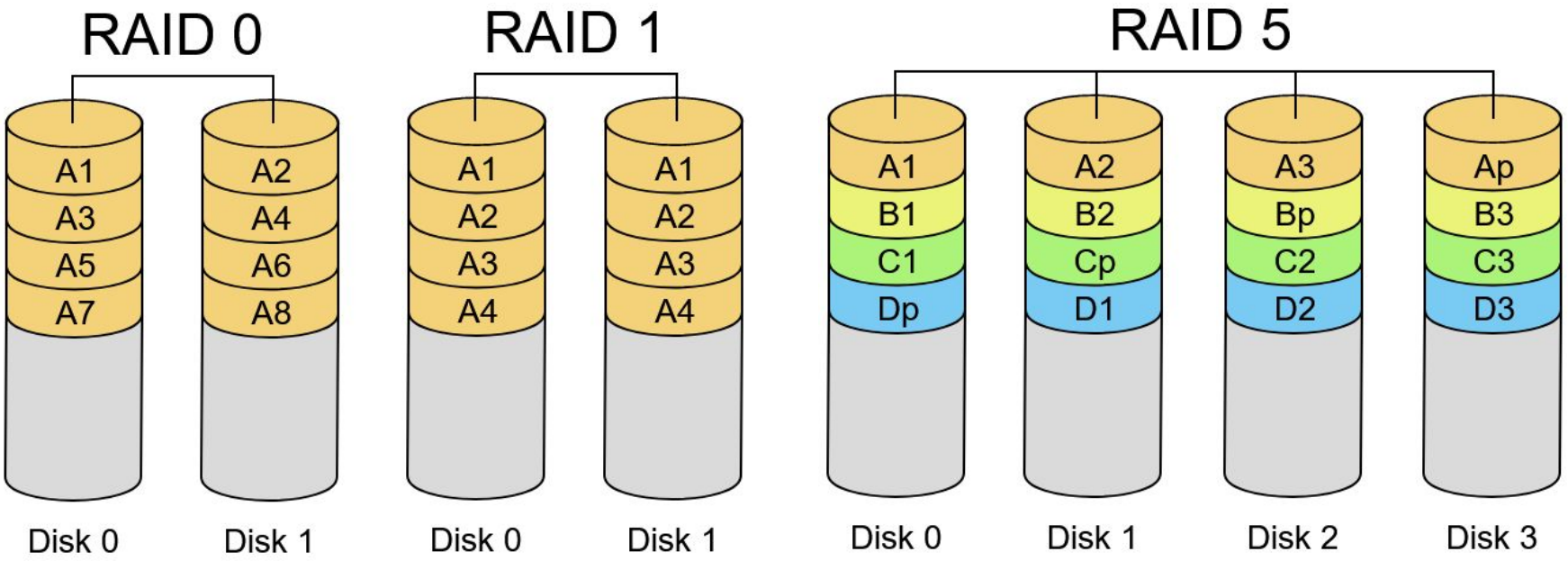


Fibre Channel



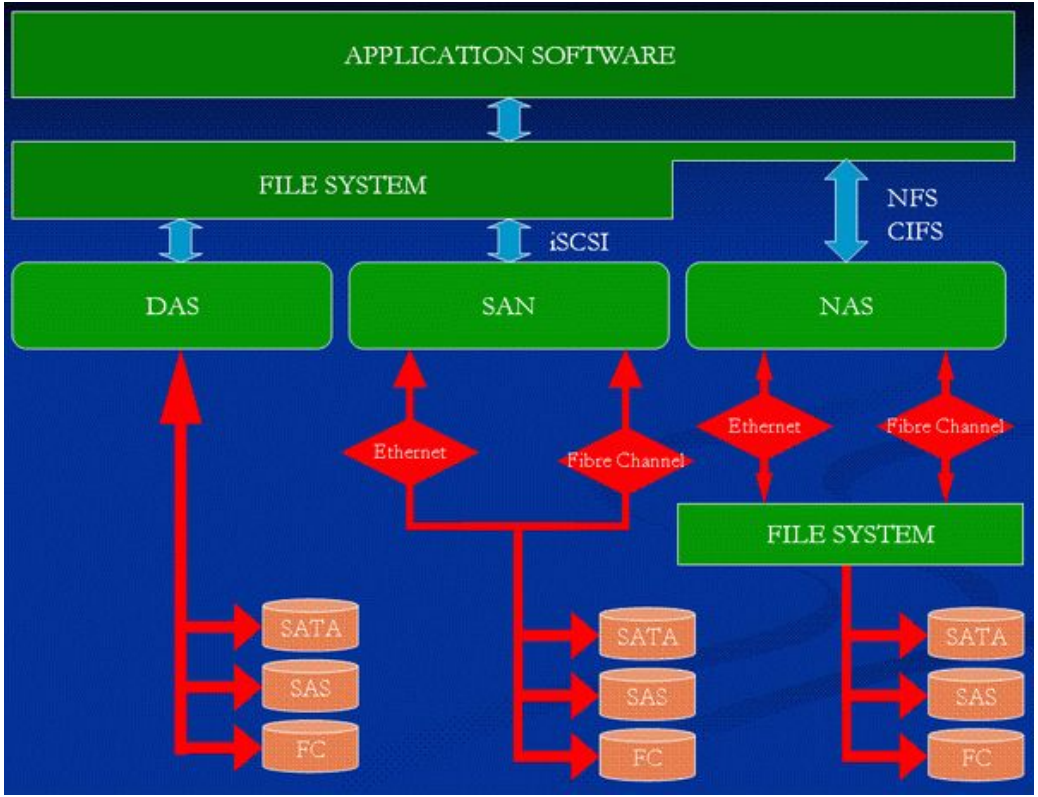
RAID

容錯式磁碟陣列（RAID, Redundant Array of Independent Disks）

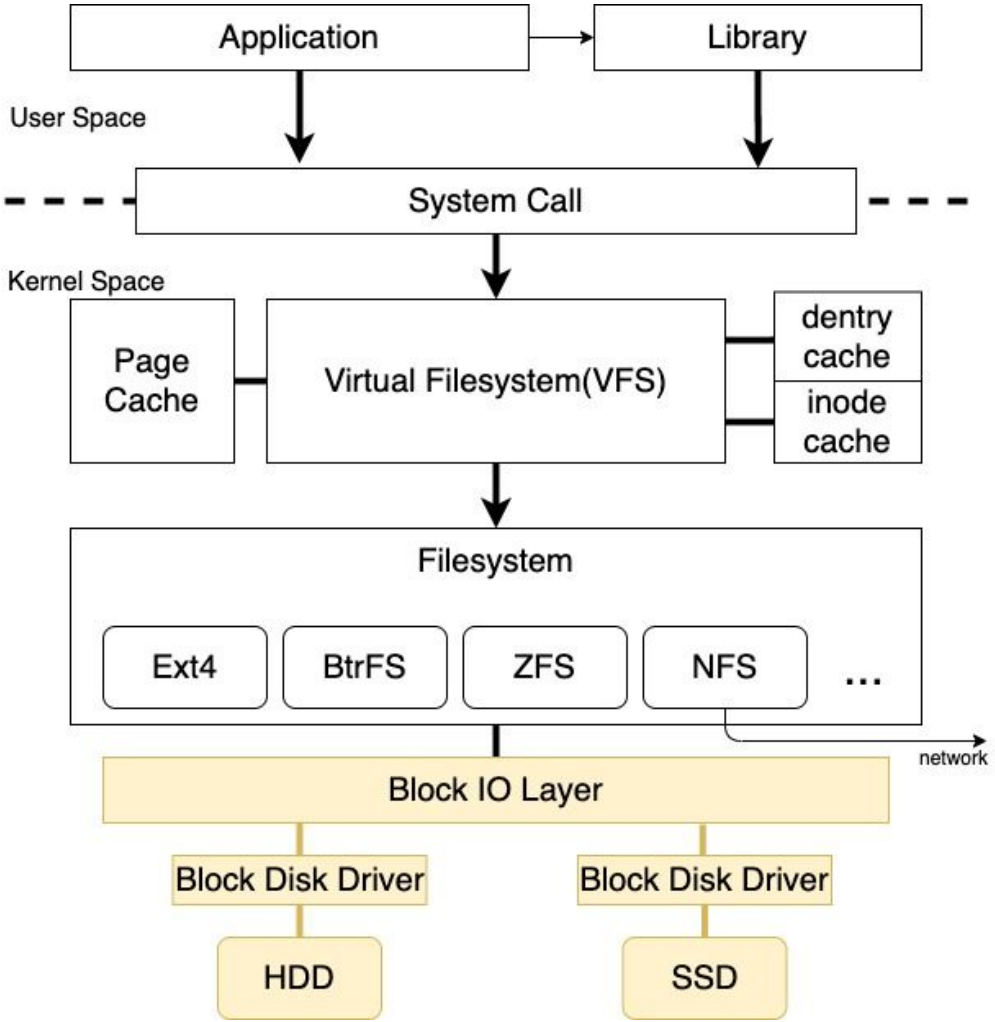


網路硬碟

- NFS
- CIFS/Samba
- iSCSI

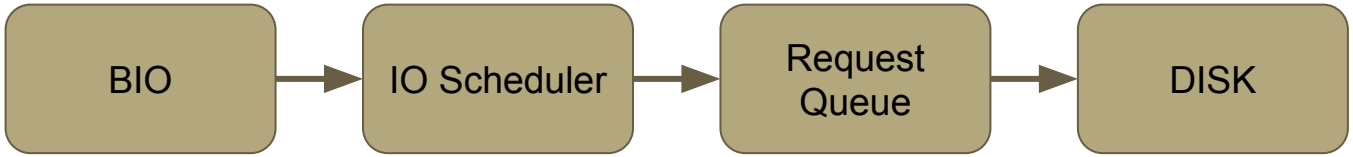


Linux IO Stack



Block IO Layer

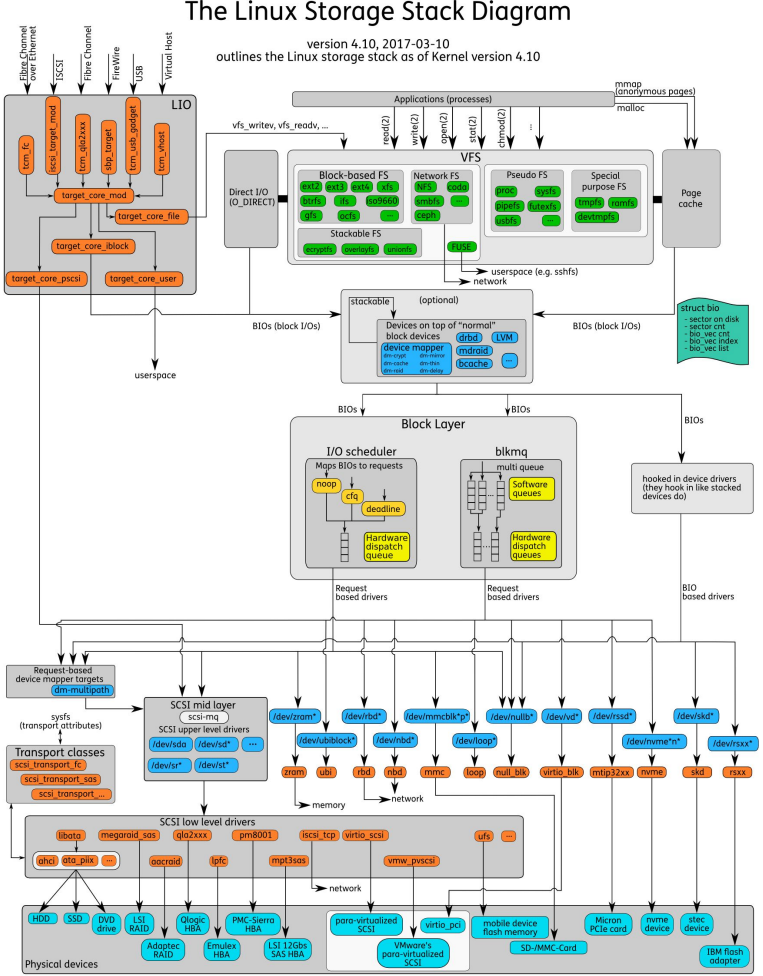
1. 對 block device 提供抽象化的統一介面，提供統一框架管理
2. 對 filesystem 和 application 的 I/O 請求排隊，通過合併、排序等方式，提高硬碟讀寫效率。



IO Scheduler

- None: 完全不做處理，適合 VM
- NOOP: 先入先出對列，只做基本的合併，適合 SSD
- CFQ: 完全公平排程器，許多 Linux 發行版的預設選項。
  - 每個 process 有自己的 queue，按照時間片段來均分每個 process 的 I/O 請求
- Deadline: 分讀、寫兩個 queue，並確保達到 deadline 的請求優先被處理。
  - 多用在 I/O 壓力較大的情境，如 Database

Linux Storage Stack Diagram



https://www.thomas-krenn.com/de/wikiDE/images/e/e0/Linux-storage-stack-diagram_v4.10.png