Adattárolás és fájlrendszerek

Sulyok András Attila

2018. 10. 18.

Bevezetés a számítástechnikába

Háttértárak

Perzisztens tár (nonvolatile): kikapcsolás után megmarad az adat

Jellemzők:

- sávszélesség
- hozzáférési idő
- tárolókapacitás
- élettartam
- ár

(lásd: memóriahierarchia)

Optikai tárolók

Egy lemez egy hosszú reflektív spirál mentén tárolja az adatokat.
Az adatok az anyag felületén (pit (0) és land (1)) tárolódnak, kiolvasáskor a felpörgetik a lemezt, és egy lézersugárral megvilágítják.
A land visszatükrözi a sugarat, a pit nem (ön-interferál).

Típusai:

- Technológiától függően: CD, DVD, BluRay
- Írhatóságtól függően: olvasható (ROM), írható (R), újraírhaó (RW)
- Alkalmazástól függően: Audio-CD

Külön fájlrendszer (ISO-9660 $\rightarrow$ ISO-13490, UDF)
Jellemző sávszélesség: 6 MB/s (CD 40$\times$), 21 MB/s (DVD 16$\times$), 54 MB/s (BD 12$\times$)
(busz sávszélessége korlátozhatja)

Merevlemez

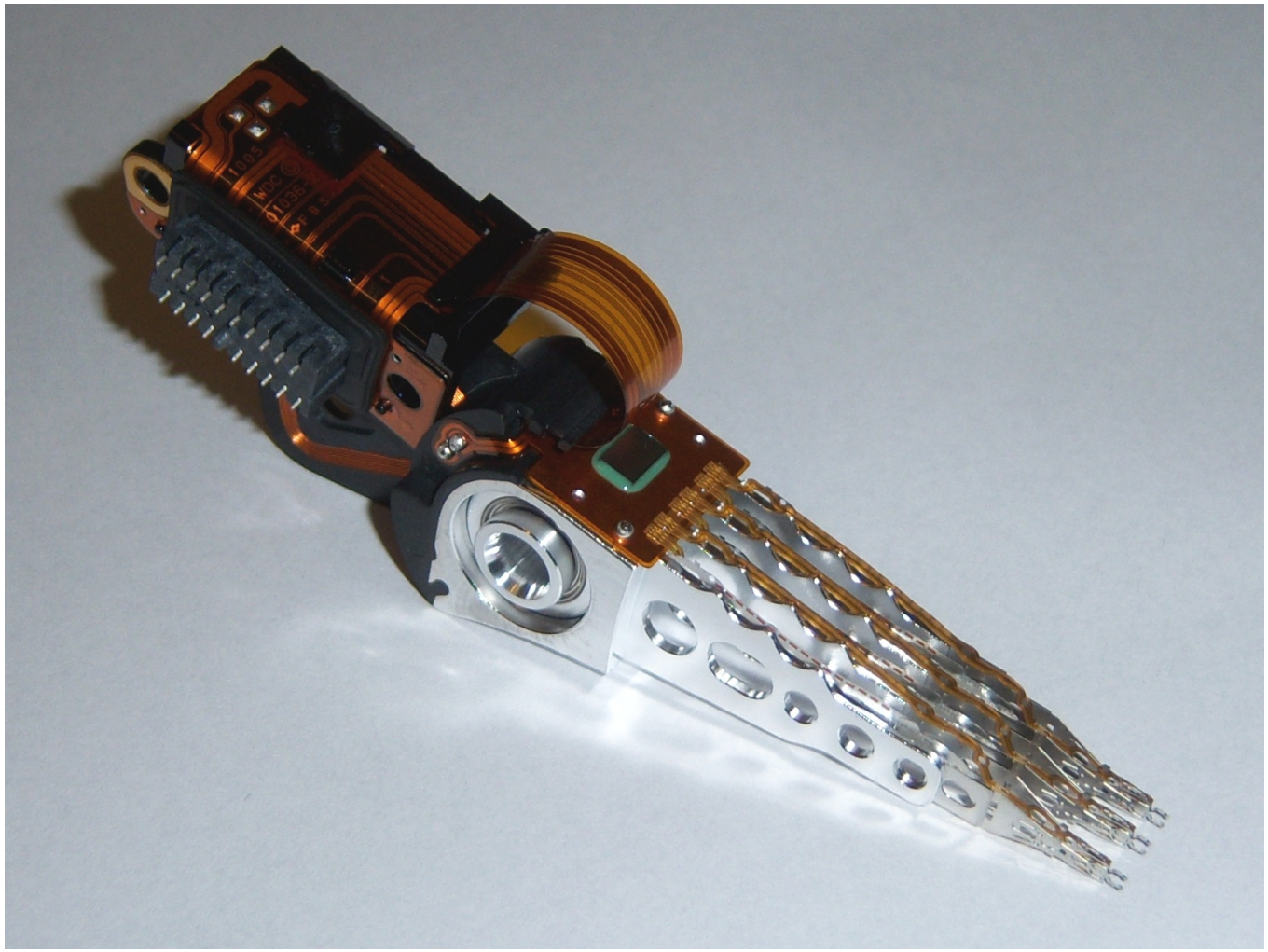
Hard Disk Drive (HDD)

Mágnesezhető, forgó (ca. 7200 RPM) lemezekre írja az adatot.
Ellentétben az optikai lemezekkel, koncentrikus körökre.
(Egyenletes sebességgel lehet forgatni.) Az író/olvasó fej a lemezek között mozog néhány nanométerre
légpárnát generál maga alá.

Régebben a cilinder-sáv-szektor (cylinder-head-sector) alapján címezték
(megfelelően az írási procedúrának)
manapság inkább logikailag, folytonosan (Logical Block Addressing)

Merevlemez

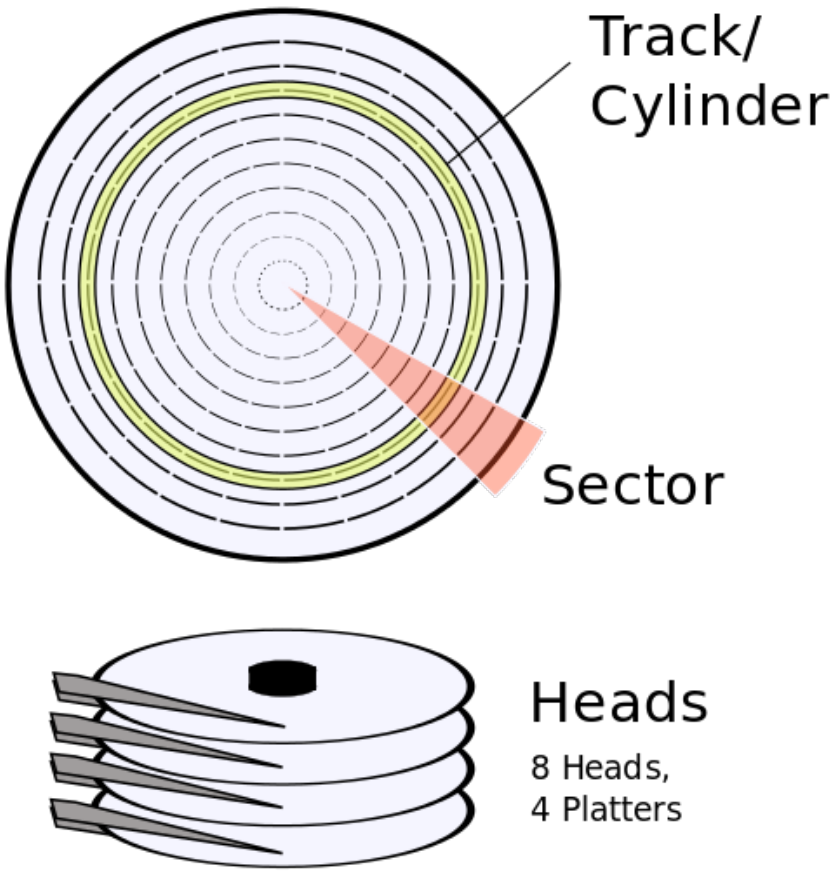
Merevlemez író/olvasó fej



Kép forrása: Wikipedia

Merevlemez

Merevlemez címzése



Kép forrása: Wikipedia

Merevlemez

Elérési idő áll:

- Keresési időből ($\sim$9 ms)
- Forgatási időből: a lemez forgási sebességétől függ (7200 RPM mellett $\sim$4 ms)

Jellemző sávszélesség: 220 MB/s

Flash alapú háttértárak

Floating-gate transzisztor technológián alapuló perzisztens tár
Random-access: bármely cím elérése ugyanannyi idő

Módosítás előtt törölni kell az adatot, ezt viszont csak blokkonként lehet:
a módosítás helye lemásolódik (write amplification).

A HDD-vel ellentétben, az operációs rendszer explicit törlésre jelöli a nem használt blokkokat (TRIM művelet).
A törlésre jelölt blokkok szemétgyűjtésre kerülnek (garbage collection), ekkor törlődik ténylegesen az adat.

Az újraírás fizikailag terheli az eszközt
Erre megoldás, hogy az írásokat egyenletesen elosztjuk az eszközön (wear levelling)
Pl.: pendrive, Solid State Drive (SSD) Jellemző sávszélesség:

Egyéb technológiák

- mágnesszalag: hosszú távú (évtizedek) adattárolásra, amit ritkán olvasnak
- memrisztor alapú technológiák (ReRAM):
  a memrisztor egy nemlineáris passzív elem, amelynek ellenállása függ a múltbeli töltésáramlástól
  az ellenállás az állapot
- 3D XPoint: SSD és DRAM között árban és sebességben
  automatikus fájlrendszer-cache-nek szánják

Partícionálás

Egy háttértáron belül lehetőség van több területet: partíciót kialakítani

- több operációs rendszernek
- többféle fájlrendszert alkalmazni
- külön partíció az adatoknak és rendszernek
- swap

Két nagy formátum van a partíciók kialakítására: MBR és GPT

Partícionáláshoz használható parancsok Linuxon:

```
fdisk -l
gparted /dev/sda
```

Master Boot Record (MBR)

Az MBR jelenti egyúttal a partíciók előtti boot rekordot is, amely az első szektorban helyezkedik el, tartalmazza:

- bootoláskor lefuttatandó kódot (bootstrap code)
- partíciók helyét, hosszát és típusát

Alapvetően négy (primary) partíció,
de az egyik lehet extended, amelyben több logikai partíció foglal helyet

GUID Partition Table (GPT)

MBR utódja, annak hiányosságai javítására
(pl. 2 TiB helyett 2 ZiB max méret, 4 helyett 128 partíció)

Kompatibilitás: az MBR helyén olyan kód, amely képes kezelni az GPT-t (protective MBR)

A partícióknak globális azonosítójuk (UUID vagy GUID)
és nevük van.

Boot folyamat (BIOS)

1. Számítógép bekapcsol, ellenőrzi magát (power-on self test)
2. Basic Input-Output System (BIOS) inicializálja a hardware-t alapvető hardware-t kezelő kódok
3. Végrehajtja az MBR-ben talált boot loader első fázisát
4. Az megkeresi a megfelelő partíció elejéről a boot loader második fázisát
5. Az betölti a tényleges boot loadert
6. Amely elindítja az operációs rendszert

Boot folyamat (UEFI)

1. Számítógép bekapcsol, ellenőrzi magát (power-on self test)
2. Unified Extended Firmware Interface (UEFI) inicializálja a hardware-t
   BIOS utódja
   UEFI szabvány része a GPT
3. Saját NVRAMból kikeresi, hogy melyik EFI partíción és mit kell elindítani
   Secure Boot esetén ellenőrzi ennek aláírását
4. Ez lehet maga az operációs rendszer, vagy egy boot loader

Logical Volume Management (LVM)

A normális partícionálás helyett egy flexibilisebb módja a háttértárak kezelésének

A háttértárak (Physical Volumes) fel vannak osztva kis részekre (extents), ezekből állnak a logikai partíciók (logical volumes).

- egymás utáni extentek lehetnek különböző fizikai eszközökön: szekvenciális olvasást gyorsítja
- könnyen lehet a partíciók méretét változtatni
- fizikai eszközöket ki lehet cserélni futó rendszer alatt (hot swap)
  az LVM átpakolja az extenteket
- képes biztonsági mentésekre
- hibrid partíciók: különböző sebességű eszközökön, a többször elért adatok automatikusan a gyorsabb eszközökre kerülnek
- külső töredezettség

Redundant Array of Independent Disks (RAID)

Redundánsan tárol több háttértáron, hogy néhány elromlása esetén se történjen adatveszteség

- RAID0: csíkozás (striping)
  nem redundáns, csak a sávszélességet növeli
- RAID1: tükrözés:
  két eszköz ugyanazt az adatot tartalmazza
- RAID2: bitenkénti csíkozás paritásbittel
  nem igazán használják
- RAID3: byte-onkénti csíkozás paritással
  nem igazán használják
- RAID4: blokkonkénti csíkozás paritással
- RAID5: blokkonkénti csíkozás elosztott paritással
  ellentétben az előzőekkel, a paritás nem egy háttértáron van, hanem elosztva az összesen
- RAID6: blokkonkénti csíkozás elosztott dupla paritással
  két háttértár is kieshet egyszerre

Fájlrendszerek (file systems)

A fájlrendszerek feladata, hogy a rendelkezésre álló helyen elhelyezze az adatokat:
hogyan reprezentálja, milyen stratégiával hozza létre és törölje, stb., és adminisztrálja a fájlokat (metaadatok).

A felhasználó felé ez fájlok egy (Linux, Mac) vagy több (Windows) fa struktúráját jelenti.
A belső csomópontok a könyvtárak vagy mappák.

Metaadatok (adatok az adatokról) lehetnek:

- A fájl tulajdonosa
- Készítés, hozzáférés, módosítás dátuma
- Hozzáférési jogosultságok
- Kiterjesztett attribútumok, pl.: előkép-ikon, cím, összefoglaló ezekket nem értelmezi a fájlrendszer, csak tárolja

Töredezettség (file fragmentation)

Általában sok fájlt tárolunk, s ezek mérete folyamatosan változik. Probléma: nehéz megoldani, hogy fizikailag egymás után következő blokkokban tároljuk őket (külső töredezettség vagy single file fragmentation).

Ha a szabad hely töredezik, például a fájlok törlésének következtében (vagy mert a blokkméret egységes), az a belső töredezettség vagy free space fragmentation.

Ennek elkerülésére rendszeresen érdemes töredezettség-mentesíteni. Vagy: extentek (lásd ext4).

File Allocation Table (FAT)

Egyszerű, régóta használatos fájlrendszer,
főleg hordozható eszközökön, illetve az EFI partíción.

Alapvető foglalási egysége a cluster, a különböző verziókban ez különböző nagyságú.

A fájlok clustereit a File Allocation Table tartalmazza, láncolt lista formában.

Maximális fájlméret: 4 GiB, fájlrendszer-méret: 2 GiB (FAT16) vagy 2 TiB (FAT32)

ext4

Linux kernelhez fejlesztették ki

Alapegység: blokk

Több technológia a fragmentáció elkerülésére:

- extents: több blokkot foglal egyszerre
  kevesebb fragmentáció és egyszerűbb adminisztráció
  alapvetően fájlonként max. 4 extent, a többit egy fában tárolja
- allocate-on-flush: késleltetett allokáció (lehet többet egyben)

Maximális fájlméret: 16 TiB, fájlrendszer-méret: 1 EiB

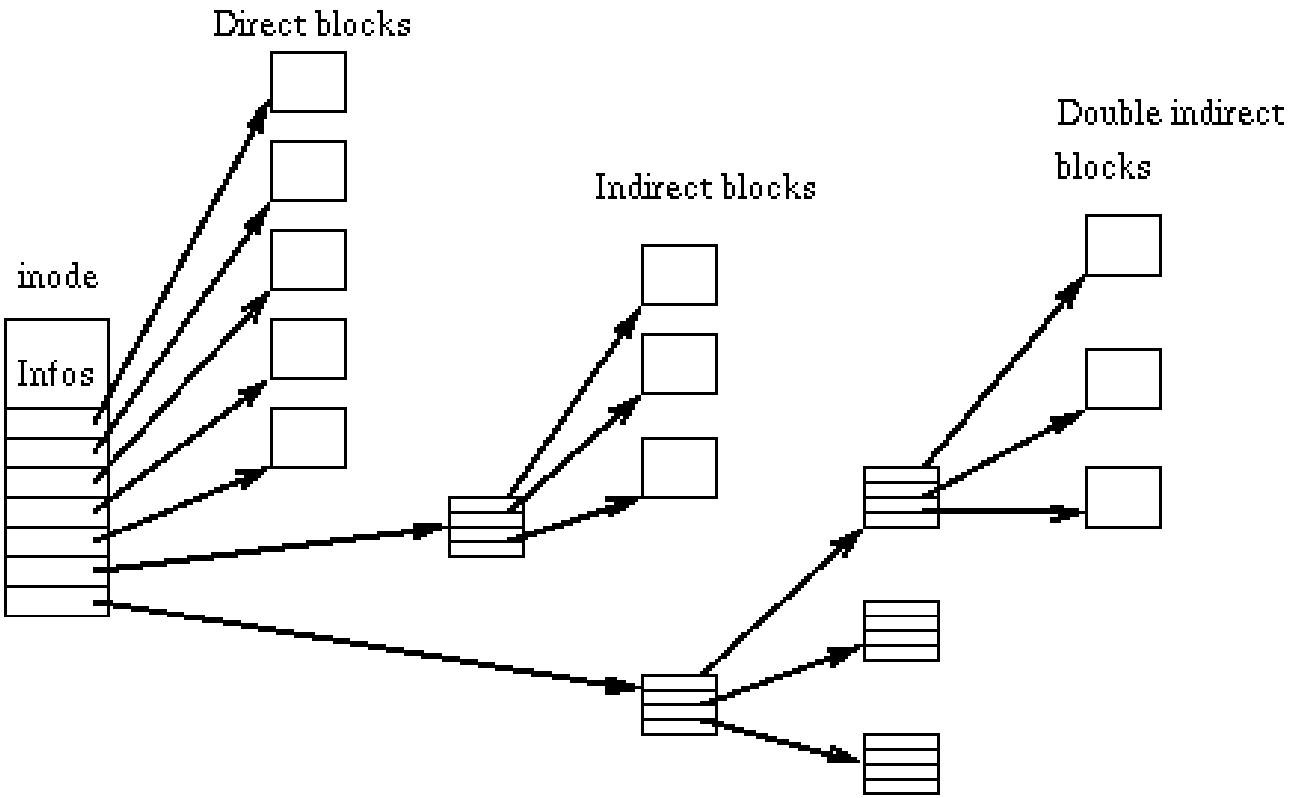
Inode

Adatokat tárolnak a fájlokról: név, létrehozás/módosítás dátuma, tulajdonos, jogosultságok (metaadatok);
illetve tárolják a tényleges adatokat tartalmazó blokkokat.
(Kis fájl esetén az inode tartalmazhatja magát a fájlt (inlining).)

Az inode-ok adott helyen egy táblázatban vannak felsorolva.
inode-index: a fájl e táblázatbeli indexe

Inode

ext2 fájlrendszer esetén



Kép forrása: Wikipedia

Inode

Adatokat tárolnak a fájlokról: név, létrehozás/módosítás dátuma, tulajdonos, jogosultságok (metaadatok);
illetve tárolják a tényleges adatokat tartalmazó blokkokat.
(Kis fájl esetén az inode tartalmazhatja magát a fájlt (inlining).)

Az inode-ok adott helyen egy táblázatban vannak felsorolva.
inode-index: a fájl e táblázatbeli indexe

Speciális inode: link: egy másik bejegyzésre hivatkozik

- hard link: a fájlra mutat
  tulajdonképpen minden fájl egy hard link, amely a névtől az inode-indexre mutat
  ha nem mutat egy hard link sem a fájlra, akkor az törlődik
- soft link: a fájl elérési helyét mutatja (pl.: `/home/sulan`)

A könyvtárak egyszerű hard link listák (`ls -i`).

Egyéb fájlrendszerek

- Nagy megbízhatóságú fájlrendszerek: Btrfs, ZFS
- Párhuzamos fájlrendszer: Lustre
- Hálózati fájlrendszerek: NFS, SMB, SSHFS
- Memória-alapú fájlrendszerek: tmpfs, RAM disk
- Konfigurációs fájlrendszerek: sysfs, procfs, devfs
- Swap:
  ide menti ki az operációs rendszer a folyamatok memóriáját, ha a fizikai memória megtelt

További tanulmányozásra

- Az Operációs rendszerek tárgy
- Az Arch Linux disztribúció boot folyamata:
  `https://wiki.archlinux.org/index.php/Arch_boot_process`
- Előadás az ext4 töredezettségkezelő módszeréről
  `https://events.static.linuxfound.org/slides/2010/linuxcon_japan/linuxcon_jp2010_fujita.pdf`