21 Storage Devices

Disks

- Main components: arm, head, track/cylinder, platter, sector, controller.

- Typically 2-14 surfaces, thousands of tracks per surface, hundreds of sectors per track, 512-4096 bytes per sector.

- In 2015, 3 TByte disk costs less than US$100.

- Sectors can be read and written individually, or in adjacent groups:
  - Seek: move heads to correct track.
  - Select desired read/write head.
  - Wait until desired sector rotates into position under the head.
  - Read or write sector while it spins by.

- Seek time = 5-50 ms, rotational latency is 0–5 ms (drive spins at 7200/10000/15000) RPM).

- Sequential read/write throughput over 100MB/s. Random access much slower.

- Technology advances lately mostly in miniaturization. In 1975, 40 Mbytes took up space the size of a commercial washing machine.

Hardware evolution: capacity grows much faster than speed (it takes longer and longer to read an entire hard disk). As a consequence, disks throughput is likely to become an important bottleneck, and the operating system must work to mitigate this.

A formatted sector includes a preamble to detect beginning and an error correcting code (ECC) that corrects a small number of bit flips and detects up to a certain larger number of bit flips.

The first sectors of different tracks are not aligned. A track differential is required for continuous scanning of consecutive sectors. E.g., suppose a disk with 300 sectors/track, a seek time between adjacent tracks of 0.8 ms and that rotates at 10,000 rpm. How many sectors should the track differential be?
Disk access time = seek time + rotational latency + transfer time.
OS can only try to minimize the first (and to some extent, second) component (see next lecture).

**Solid State Drives (SSDs)**

- Solid state memory, NAND-flash based (other technologies: M-RAM, PCM).
- No mechanically moving parts
- Multiple flash packages + volatile memory for controller
- Fast read access (less than .01ms).
- Still considerably more expensive than disk drives (US$90 for a 250GB SSD, or US$360 versus US$33 per TB in 2015).
- Capacity catching up with disk drives.
- No mechanical seek, no rotational latency.

Internal organization:

- Erasure blocks of 128 or 512KB
- Each erasure block has multiple pages:
  - 4 KB data
  - 128 bytes for metadata (ECC)
- Interface allow reading and writing at page granularity
- Atomically access data + metadata

Access characteristics:

- Fast random-access reads and writes (tens of microseconds per page)
- Sustained write bandwidth a few thousand pages per second
- But small in-place updates are inefficient
  - Need to erase entire block before overwriting
– Erase takes a few milliseconds

Solution: Flash Translation Layer (FTL)

• In-memory remap table
• Map logical pages to physical ones
• Write to new page + update mapping
• What happens upon reboot?
• Need to reconstruct remap table. How?
• Store identity and version number of logical page in metadata part
• Problem: obsolete pages within blocks
• Need GC procedure: Select block, copy valid pages out, erase, add it to free block list

Another limitation: wear

• Reliability degrades after many write-erase cycles (100,000s)
• Can be alleviated using wear-leveling algorithms
• Maximize device lifetime as a whole by shifting blocks around to avoid permanent blocks from never being written while “hot-spots” degrade quickly.

Comparison: Disks vs SSDs

SSDs have:

• Low latency
• Higher read+write throughput (slightly higher in sequential access, much higher for random access)
• Lower energy consumption

Disks have:
• Better (lower) cost / TB
• No write-wear

Different market segments: Disks are still the main choice for enterprise bulk storage, while SSDs are used when performance (enterprise, desktop), energy, and lack of mechanical components matter (portable devices).