Lectures 23: External Memory

I. Magnetic Disk

A. Magnetic Read and Write Mechanisms. Write head is has coils of wire around a gapped rectangular doughnut. Read head uses a partially shielded magnetoresistive sensor made of ferromagnetic materials.

B. Data Organization and Formatting.

1. Tracks = concentric rings on the platter, each the width of a head, and separated by the intertrack gap.
2. Sectors = sections of tracks (512 bytes of data, recently 4K) separated by intersector gaps.
3. Tracks on the outside of the platter have a higher angular velocity than tracks closer to the center.
   a. **Constant angular velocity** layout has the same number of sectors on each track, so outer ones are longer than inner ones. Advantage: easy to find sectors. Disadvantage: Outer sectors have lower density than possible.
   b. **Multiple zoned** layout has the number of sectors based on the circumference of the tracks. Advantage: makes full use of the density possible, and thus more storage capacity.
4. Formatting creates physical sectors (600 bytes) that also include IDs, CRC, and gaps to help with synching.

C. Physical Characteristics

1. Heads: fixed head (one per track, obsolete), or moveable-head (one per surface) mounted on an arm.
2. Non-removable disk (hard disk) or removable disk (floppy).
3. Single platter, or multiple (2 -10) platters on the same spindle.
4. Single sided, or double sided (usually).
5. Head Mechanisms: 1) contact, e.g. floppy; 2) fixed gap (traditional); and 3) aerodynamic gap (Winchester) relies on foil head riding the air current produced by the spinning platters to keep from touching them.

D. Disk Performance Parameters

1. **Seek time** = average time it takes to position a head over the correct track. 4-13ms. Zero on fixed head drives.
2. **Rotational delay** = average time it takes a sector to reach the head. 7200 RPM = 4ms, 10000 RPM = 3 ms
3. **Access time** = seek time + rotational delay.
4. **Transfer time** = time to read or write one sector once it has reached the head, 100 - 300MB/s.
   a. $T = \frac{b}{rN}$, where b = bytes transferred, r = RPMs, N = bytes per track (which varies on multiple zoned disks).
5. Timing comparisons depend on whether data is physically stored sequentially (one access time), or randomly (multiple access times). Hence disk defragmentation is critical.

II. RAID = Redundant Array of Independent Disks

A. Three common characteristics

1. Viewed by the OS as a single logical drive.
2. Data is distributed across the physical drives of an array in a scheme known as striping. A logical drive is divided into *strips* that may be physical blocks, sectors, or some other unit. The way theses strips are mapped determines the type of RAID.
3. Redundant disk capacity is used to store parity information, which guarantees data recoverability in case of disk failure.

B. RAID Level 0 = Non-redundant, with strips distributed round robin among the n disk drives. Transfers can be up to ~n times faster than a single disk if the request is for contiguous data of at least n * sizeof(strip) bytes. If one drive fails, you lose all of your data because it spread across all of the disks!

C. RAID Level 1 = redundant strips on two drives, called *mirroring*. Reads can be twice as fast, but writes are same as for one drive. Costly, but used for accounting storage,

D. RAID Level 2 = data bytes (or bits) spread across n drives, with $\log_2 n$ extra drives holding SECDED Hamming bits. Each I/O operation involves parallel access of all disks. Small transfers are 2 times faster, and large transfers can be up to n times faster than a single drive. Not used because it is costly and data errors are quite rare.

E. RAID Level 3 = data bits spread across n drives, with one extra parity drive that can be used to construct a new drive. Each I/O operation involves parallel access of all disks. Same transfer speed as RAID-2. Used for applications dealing with streaming data, or requiring high throughput.

F. RAID Level 4 = like Level 3, but larger strips are used, and drives can be accessed independently for different requests. Reads are same as RAID-2. Writes always involve the parity drive so it may be a bottleneck. For writes smaller than all of the drives, old updated strip(s) and the parity drive must also be read so the new parity can be calculated. Not used.

G. RAID Level 5 = like RAID-4 except the parity bit is spread across the drives to eliminate the parity drive bottleneck. Highest read rate, but has complex controller. Used for servers. Most versatile RAID level.

H. RAID Level 6 = like RAID-5 except there are now two different data check algorithms create parity bits that are stored on two different drives. This permits the system to recover from the failure of two drives at the same time. Great for mission critical applications.

III. Solid State Drives (SSDs) = semiconductor memory device using flash memory.

A. Flash Memory was developed from EEPROMs (Electrically Eraseable Programable ROMs)

1. Uses insulated *floating gates* between the control gates and the P-substrate in their transistors. The floating gate traps electrons so that the control gate retains its charge without electricity.
2. **NOR flash memory** can retrieve a single byte, and is reliable so it can be used in BIOS chips and to store cell phone OS code. Erasures are by multi-byte blocks.
3. **NAND flash memory** has 16 or 32 bit units, and is accessed by block (512KB). It is denser than NOR flash memory, but may contain faults. Manufacturers try to maximize the amount of usable storage by shrinking the size of the transistor below the size where they can be made reliably, to the size where further reductions would increase the number of faults faster than it would increase the total storage available.

B. SSD Compared to HDD. HDD's have a cost per byte advantage but:

1. SSDs are faster: 45000 I/O operations/second vs 300 in HDD, read 200MB/s vs 80, random access time 0.1ms vs 4-10ms.
2. Also more durable (immune to physical shock and vibration), longer lifespan, lower power consumption, quieter, and cooler.

C. Practical Issues

1. To write a page (4K) involves: 1) Read a 512KB block from SSD into RAM and update the page in this RAM buffer; 2) Erase block on SSD; 3) write entire RAM block back to SSD. As the SSD becomes more fragmented, a file write effects more blocks, and slows the operation. Addressed by providing extra space, or using the TRIM command to free blocks.
2. As flash cells are used they lose their ability to record and retain values. 100,000 writes is a typical limit. Chips use algorithms to spread the writes throughout the chip. SSDs can predict their remaining life.

IV. Optical Memory use high powered lasers to burn bit pits in polycarbonate plastic leaving shiny bit lands. Read with low powered lasers, and sensors that detect the reflective lands and diffusing pits. The track is spiral, and to maintain a **constant linear velocity** the disk increases its RPMs as the laser moves away from the center.

A. For large scale duplication a master die is cut with a laser. Then polycarbonate disks are stamped with the master die, and then coating disk with aluminum and protective acrylic.

B. Recordable optical disks use a dye layer that a modest intensity can "burn" to change the reflectivity of the pits.

C. Rewritable disks rely on a material that is darker in an amorphous phase, and reflective in a crystalline state. A medium laser erases by blending the states, and a high intensity then "burns" the pits in the semi-crystalline material.

D. Compact Disks hold 680MB, and have 2352-byte blocks containing 2048 bytes of data, sync data, 4-byte time/location ID, and 288-byte ECC.

E. Digital Versatile Disk (DVD) hold up to 17GB . This increase over the CD is because of three differences

1. DVD uses a laser with shorter wavelength so distance between bits (.834μm vs 0.4μm) and loops (1.6μm vs 0.74μm) is less. This contributes a seven-fold increase, 4.8GB.
2. DVD has a semi-reflective layer on top of the reflective layer which allows two spirals in opposite directions. The laser changes its focus to read at the different depths. This doubles further the data density, 8.5GB.
3. A DVD can be double sided, which further doubles its capacity to 17GB.

F. High-Definition Optical Disks (Blu-ray) use a laser with even shorter wavelength (blue), and has the data layers closer to the laser. This permits 25GB per layer, so 50GB per side.

V. Magnetic Tape use the same recording and reading techniques as magnetic hard drives, but only has *sequential access*.

A. Tapes vary from 0.15 inch to 0.5 inch wide, and are housed in cartridges.

B. **Parallel recording** had data on the tapes run on 9, 18, or 36 parallel tracks running lengthwise with parity tracks.

C. **Serial recording** records data bits on one track at a time in contiguous blocks called *physical records* separated by *interrecord* gaps. **Serpentine recording** has the tape reverse and the head move to the next position whenever the end of the tape is reached. To speed the system, there are 2 to 8 heads so that 2 to 8 tracks are read/written at a time.

D. The current linear tape-open (LTO-5) holds 3.2TB using 1280 tracks on 0.5 inch x 846 meters, with 16 heads, and 280 MB/s.