Ilia Lebedev (cs162-am)

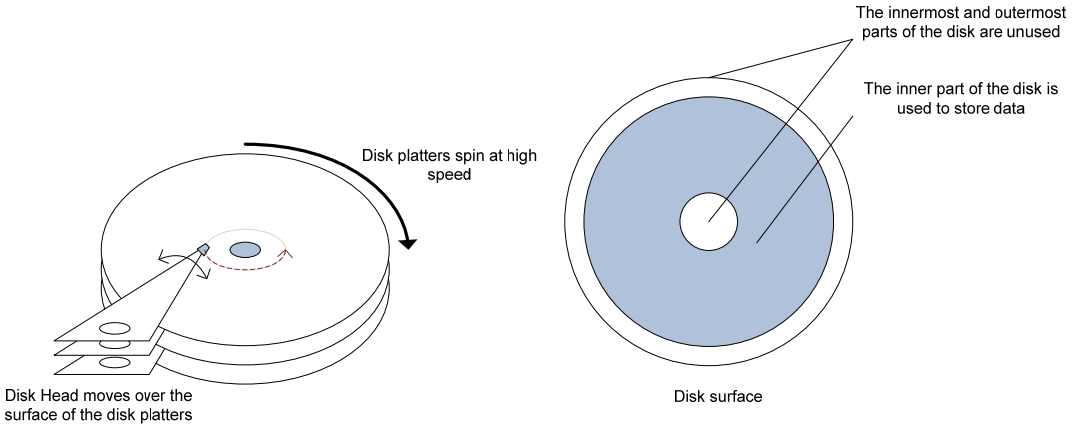
CS162 Lecture Notes: Monday 3/30/2009

Midterm in 2 ½ weeks.

Datasheets for various IO devices will be posted on the course website shortly. There is more reading there than can be easily digested – meant to be used as reference.

An Operating System designer needs to be aware of several key points regarding disk technology:

- A disk is made up of several *disk platters*
- The surface of a disk is broken up into *tracks*
- A track contains sectors.
- A block of data on disk contains a *block header*, followed by some data.



Some technologies tangential to the modern hard drive are removable drivers (such as ZipDisk and JavaDisk). These technologies are no longer around because as bit density grew, the tolerances on moving parts became more restrictive. The disk head and the disk itself became inseparable (very tight tolerance, minimal number of moving parts). Removable disks simply cannot provide the bit density offered by sealed hard drives.

Floppy drives are another example of extinct magnetic disk technology. These devices offered very low access speed and bit density. Currently, floppies are virtually extinct, and have been replaced by flash drives (jump dries, etc).

Disk operation:
The following steps must occur to perform a disk access:

1. A *seek* operation moves the disk heads to line up over a desired set of tracks. Given the incredibly high bit density, the motor precision is low. A feedback control system is used (heads read the track numbers, which are used to adjust the track position until destination is reached). A random seek is an access to a track unrelated to the current one. Random seeks are rare (20%-30%), as data is often arranged on disk with seek

time minimization in mind.
2. Some *rotational latency* is introduced when the disk spins to the correct angular location (beginning of a sector). This step used to require a command separate from the seek. Now this and other functionality is integrated into the on-disk controller.
3. The *read* or *write* is performed. The magnetic head is scans along the track, reading or writing bits.

A hard drive is essentially exploiting an idea similar to magnetic tape. The difference is the geometry of the storage media. Instead of a reel of magnetic tape, the surface of the disk platters is magnetized to record data. Instead of a stationary magnetic head, the head is mounted on a movable arm. The disk head "flies" extremely low over the disk surface on an air cushion created by the spinning platters. The aerodynamics of this system are phenomenal – the disk head must not fly too high (the data would not be read), and must not hit the surface, or *crash* (the disk would be destroyed). In modern drives, the disk head soars only tens of nanometers over the disk surface. This means that the disk head must be retracted when the disk is spinning below a certain RPM to avoid crashing. 3600 RPM was a common rotational speed in older drives (frequency of AC circuit provided a natural clock. Newer drives spin *much* faster). Compared to a magnetic tape device, the hard drive features far fewer moving parts. This allows for tighter manufacturing tolerances and a higher bit density.

The drum storage technology is extremely similar to hard drives. A large number of heads are permanently positioned over the surface of a spinning drum. This approach eliminates seeks, but is highly limited by two design constraints:

Cost: magnetic heads are relatively expensive.

Bit density: The drum cannot be used to achieve a high density of data.

The fast access time of data make drums ideal for paging.

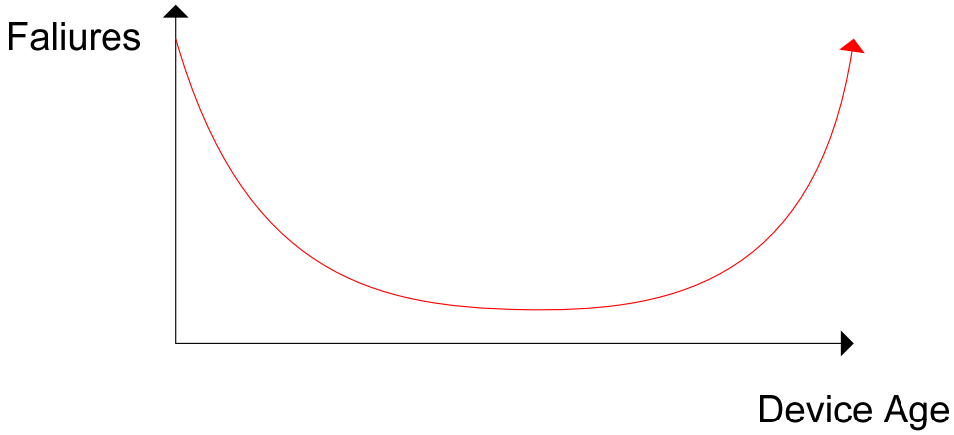
A number of different magnetic disks were presented in class. See datasheets online for more detail.

Tangent: Disk reliability

1.6 M mean time between failures likely a marketing exaggeration (drives *do* fail occasionally).

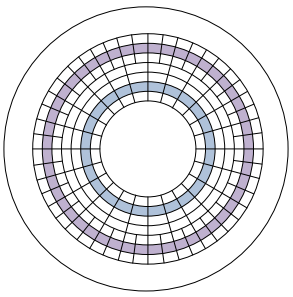
The question of disk reliability is a sensitive subject: manufacturers report error rates in the ballpark of 10^{-14} errors per transaction. Nevertheless, drives fail. Why? Manufacturers likely use optimistic assumptions when testing (such as ideal temperature, humidity, power supply conditions, constant error rate across device lifetime, etc). Due to the low reported error rate, the figure cannot actually be proven easily (would have to test for *months*).

"Bathtub curve of reliability"



New devices exhibit a high failure rate (defects, etc). Old devices succumb to aging. Thus least failures are observed in mature devices that are not yet old.

- The big idea is that all of this disk technology is absolutely incompatible – the interfaces, disk geometry, capacity, and access times vary wildly from manufacturer to manufacturer. Newer drives introduce additional complexity by remapping dead bits to “spare” sectors, and by storing more data on outside tracks (modern drives have ~2000 blocks per inch). Sector size has been historically fixed at 512 bytes.



Outside tracks can store more bits than inside sectors

This places strain on OS developers: an OS must be aware of these factors to be able to use the device. While the marked was dominated by a small number of disk companies, the task was manageable: each manufacturer would have a small number of drives, resulting in a relatively small amount of intelligence required from the operating system in order to use the drive. In the recent years, however, the market has seen an explosion of hard drives of various capacities, speeds, and organization. There is no real standard for drive organization. A hard drive product has a market lifetime of only a few months - new models are released frequently. What implications does this have on an OS?

- Modern disks are far more intelligent than older hard drives. An on-disk controller handles much of the disk details that were previously handled by the OS. In fact, the disk controller exposes the drive as a linear address space, hiding the complexity of seeks, etc. The controller also handles things like remapping of broken sectors. Nowadays, all disks have caches 4, 8, 16 MB

Tangent: sector remapping

To increase yield, manufacturers often dedicate some spare storage on the disk to be used in case of broken sectors. A table of bad sectors is created during drive initialization (each bit is written and read to test its condition), and the spare sectors are used to keep the drive operations. Two approaches are possible:

- Dedicate spare sectors per track: Pros: no seek overhead when using "broken" sectors. Cons: can run out of spares.
- Dedicate special tracks to remap bad sectors. Pros: lots of spares. Cons: if mapping is done on a per-sector basis, seek times can be a problem. If the entire tracks are remapped, small defects can waste a lot of good sectors.

The on-disk controller attempts to optimize the use of the drive via a read/write buffer – disk accesses are queued up and performed in an optimal order. The controller takes care of variable number of blocks per track.

Prior to on-disk controllers, an *RPS Miss* condition could arise: seek and rotation was handled separately, and when the right block was found, the CPU received an interrupt. If the disk used no buffering, a new access could begin before the old one was complete.

It is important to know the trends in technology development, as the changing relative advantages and limitations of technology may shift the problems OS design will face in the near future.

Tangent: Rule of 72
The doubling period of a steadily growing variable can be obtained by taking the percentage growth per year, and dividing it by 72.

A large number of trend graphs were presented in lecture. These graphs are available online. A summary is given below.

An increasing imbalance between capacity increase and read latency decrease is emerging. Drives are becoming enormous, and relatively extremely slow.

Law of diminishing returns applies heavily to all changes in technology – no breakthrough advances, only incremental improvements.

Some vocabulary:

- Media transfer rate – max transfer rate through the head (max on outside track inside a block)
- Max sustained transfer rate – much lower due to seeks, etc.
- Instantaneous max – how fast can a buffer dump data to interface

Interfaces

Each interface has its benefits and tradeoffs. There are quite a few:

PATA (legacy)
SATA (high transfer rate)
SCSI (several devices per bus)
UltraSCSI, etc

In general, the transfer rate of the interface will not correspond to high drive performance (drive data access is slower than the interface).

Detailed progression of each variable below can be found in the supplemental materials posted on the website.

Noise
- Large arrays of hard drives can get rather noisy.
- The noise depends largely on the drive's RPM.

Square feet per TB
- (important for data centers)
- Has been dropping rapidly since the creation of hard drives.

Cost/MB
- Per-MB cost of hard drive storage has been dropping rapidly.

Watt/GB, GB/in^3

A large number of figures for various drives were shown in lecture. These figures are available in supplemental readings on the website.

Modern hard drives range from small, low-power and portable microdrives:
- ~1inch
- ~4.2Mb/s
- Several GB
- 3600RPM
- 1500 G shock survival when off
- 175G shock survival when in use
- 16g in weight

To large, high-power, stationary server drives:
- ~5 inches
- ~150 MB/s
- Several TB

What about NAND flash memory?

This technology provides a constant (predictable) access time slower than fastest disk access time. Due to seek overhead, flash drives are generally faster. Prices are driven low by competition, and due to extremely high initial costs, profit is hard to come by with flash memory.