TCSS 422: OPERATING SYSTEMS

Hard Disk Drives

Wes J. Lloyd
Institute of Technology
University of Washington - Tacoma

OBJECTIVES

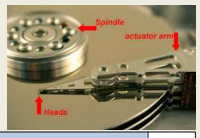
- HDD Internals
- Seek time
- Rotational latency
- Transfer speed
- Capacity
- Scheduling algorithms

HARD DISK DRIVE (HDD)

- Primary means of data storage (persistence) for decades
- Consists of a large number of data **sectors**
- Sector size is 512-bytes
- An n sector HDD
  can be is addressed as an array of 0..n-1 sectors

HDD INTERFACE

- Writing disk sectors is atomic (512 bytes)
- Sector writes are completely successful, or fail
- Many file systems will read/write 4KB at a time
  - Linux ext3/4 default filesystem blocksize - 4096



BLOCK SIZE IN LINUX EXT4

- mkefs.ext4 -i bytes-per-inode

Specify the bytes/inode ratio. mke2fs creates an inode for every bytes-per-inode bytes of space on the disk. The larger the bytes-per-inode ratio, the fewer inodes will be created. This value generally shouldn't be smaller than the blocksize of the filesystem, since in that case more inodes would be made than can ever be used. Be warned that it is not possible to expand the number of inodes on a filesystem after it is created, so be careful deciding the correct value for this parameter.

EXAMPLE: USDA SOIL EROSION MODEL WEB SERVICE (RUSLE2)

- Host ~2,000,000 files totaling 9.5 GB on a ~20GB filesystem on a cloud-based Virtual Machine
- With default inode ratio (4096 block size), only ~488,000 files will fit
- Drive less than half full, but files will not fit !
- HDDs support a minimum block size of 512 bytes
- OS filesystems such as ext3/ext4 can support "finer grained" management at the expense of a larger catalog size

EXAMPLE: USDA SOIL EROSION MODEL WEB SERVICE (RUSLE2) - 2

- Free space in bytes (df)

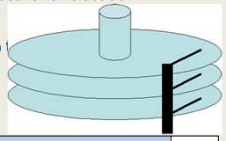
```
Device              total size   bytes-used  bytes-free usage
/dev/vda2           13315844   9556412   3049188  76% /mnt
```

- Free inodes (df –i) @ 512 bytes / node

```
Device            total inodes   used      free     usage
/dev/vda2          3552528 1999823 1552705   57%   /mnt
```

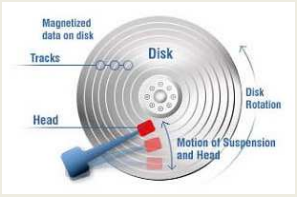

HDD INTERFACE - 2

- Torn write
  - When OS uses larger block size than HDD
  - Upon power failure only a portion of the OS block is written
- HDD access
  - Contiguous reads of sequential sectors is fastest
  - Random sector reads are slow
  - Disk head continuously must jump different tracks



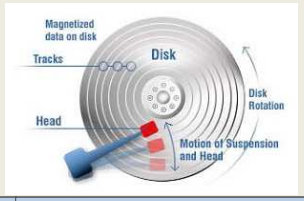
HDD PLATTER

- Made from aluminum coated with thin magnetic layer
- HDD records on both sides of each platter
- Data is stored by inducing magnetic changes



HDD SPINDLE

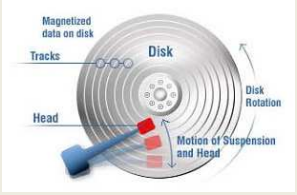
- Connected to motor which spins the disk
- Speed measures in RPM (rotations per minute)
- Typical: 7200-15000 rpm
- 10000 rpm – 1 rotation in 6ms; 15k rpm 1 rotation in 4ms



HDD TRACK

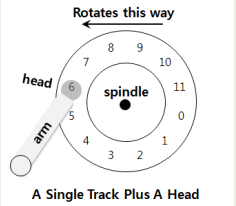
- Concentric circle of sectors
- Single side of platter contains 290 K tracks (2008)
- Zones: groups of tracks with same # of sectors

Outer tracks have More sectors



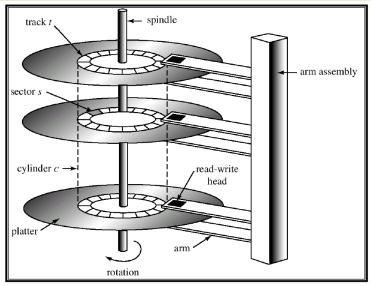
SIMPLE DISK DRIVE

- Single track disk
- Head: one per surface of drive
- Arm: moves heads across surface of platters



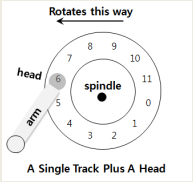
A Single Track Plus A Head

HARD DISK STRUCTURE



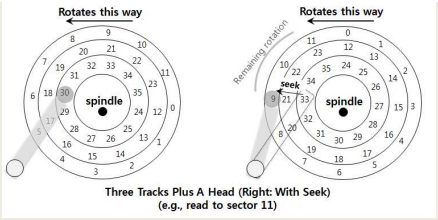
SINGLE-TRACK LATENCY: THE ROTATIONAL DELAY

- Rotational latency ($T_{rotation}$): time to rotate to desired sector
- Average $T_{rotation}$ is ~ half the time of a full rotation
- 7200rpm = 8.33ms per rotation = ~4.166ms
- 10000rpm = 6ms per rotation = ~3ms
- 15000rpm = 4ms per rotation = ~2ms



A Single Track Plus A Head

SEEK TIME



Three Tracks Plus A Head (Right: With Seek)
(e.g., read to sector 11)

- Seek time (T_{seek}): move the disk arm to proper track
- Most time consuming HDD operation

PHASES OF SEEK

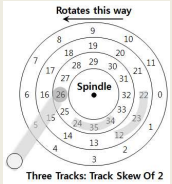
- Acceleration → coasting → deceleration →settling
- Acceleration: the arm gets moving
- Coasting: arm moving at full speed
- Deceleration: arm slow down
- Settling: Head is carefully positioned over track
  - Settling time is often high, from .5 to 2ms

HDD I/O

- Data transfer
  - Final phase of I/O: time to read or write to disk surface
- Complete I/O cycle
  - Seek (accelerate, coast, decelerate, settle)
  - Waiting on rotational latency
  - Data transfer

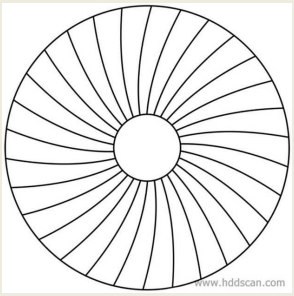
TRACK SKEW

- Sectors are offset across tracks to allow time for head to reposition for sequential reads
- Without track skew, when head is repositioned sector would have already been passed



Three Tracks: Track Skew Of 2

TRACK SKEW - 2



HDD CACHE

- Buffer to support caching reads and writes
- Improves drive response time
- Up to 128 MB, slowly have been growing
- Writeback cache
  - Report write immediately when data transfer to HDD cache
  - Dangerous
- Writethrough cache
  - Reports write only when write is actually completed to disk

TRANSFER SPEED

- I/O Time $T_{I/O} = T_{seek} + T_{rotation} + T_{transfer}$
- The rate of I/O $R_{I/O} = \frac{Size_{Transfer}}{T_{I/O}}$

	Cheetah 15K.5	Barracuda
Capacity	300 GB	1 TB
RPM	15,000	7,200
Average Seek	4 ms	9 ms
Max Transfer	125 MB/s	105 MB/s
Platters	4	4
Cache	16 MB	16/32 MB
Connects Via	SCSI	SATA

Disk Drive Specs: SCSI Versus SATA

I/O SPEED

- Random workload: 4KB random read on HDD
- Sequential workload: read 100MB contiguous sectors

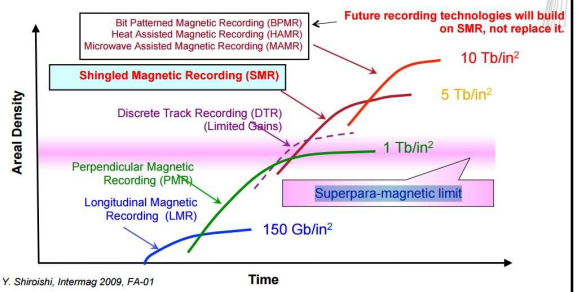
		Cheetah 15K.5	Barracuda
T_{seek}		4 ms	9 ms
$T_{rotation}$		2 ms	4.2 ms
Random	$T_{transfer}$	30 microsecs	38 microsecs
	$T_{I/O}$	6 ms	13.2 ms
	$R_{I/O}$	0.66 MB/s	0.31 MB/s
Sequential	$T_{transfer}$	800 ms	950 ms
	$T_{I/O}$	806 ms	963.2 ms
	$R_{I/O}$	125 MB/s	105 MB/s

Disk Drive Performance: SCSI Versus SATA

There is a huge gap in drive throughput between random and sequential workloads

Magnetic Recording System Technologies

New recording system technologies are needed to keep the HDD industry on its historical track of delivering capacity improvements over time



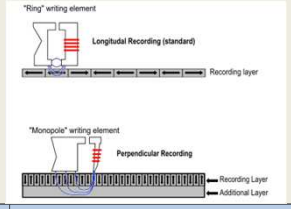
Y. Shiroishi, Intermag 2009, FA-01

HDD CAPACITY

- Superparamagnetism limits HDD capacity
- In sufficiently small nanoparticles, magnetization can randomly flip direction under the influence of temperature.
- HDD capacity is limited by the minimum usable size of particles – the superparamagnetic limit.

HDD CAPACITY - 2

- Longitudinal recording: 100-200GB/in
- Perpendicular recording: 667 GB/in
- Future technologies under development



MODERN HDD SPECS

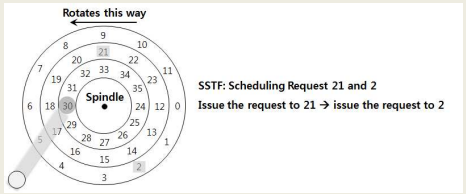
- See sample HDD configurations here:
- https://www.hgst.com/products/hard-drives

DISK SCHEDULING

- Disk scheduler: determine how to order I/O requests
- Multiple levels - OS and HW
- OS: provides ordering
- HW: further optimizes using intricate details of physical HDD implementation and state

SSTF – SHORTEST SEEK TIME FIRST

- Disk scheduling – which I/O request to schedule next
- Shortest Seek Time First (SSTF)
- Order queue of I/O requests by nearest track



SSTF ISSUES

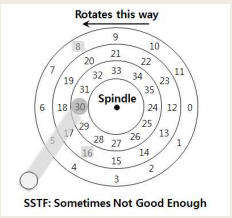
- Problem 1: HDD abstraction
- Drive geometry not available to OS. Nearest-block-first is a comparable alternate algorithm.
- Problem 2: Starvation
- Steady stream of requests for local tracks may prevent arm from traversing to other side of platter

DISK SCHEDULING ALGORITHMS

- **SWEEP**
- Single repeated passes across disk
- Issue: if request arrives for a recently visited track it will not be revisited until a full cycle completes
- **F-SCAN**
- Freeze request queue during sweep
- Cache arriving requests until later
- **Elevator (C-SCAN)** – circular scan
- Sweep from outer to inner track and reverse, inner to outer track, etc.

SHORTEST TIME POSITIONING FIRST

- Determine next sector to read?
- On which track?
- On which sector?



SSTF: Sometimes Not Good Enough

On modern drives, both seek and rotation are roughly equivalent:
Thus, SPTF (Shortest Positioning Time First) is useful.

I/O MERGING

- Group temporary adjacent requests
- Reduce overhead
- Read (memory blocks): 33 8 34
- How long we should wait for I/O ?
- When do we know we have waited too long?

QUESTIONS