Solid State Drives (SSDs)

Daniel Mosse

(slides are modified from **Dr. Ahmed Amer's CS 1550 Slides and Sherif Khattab**)

Historical Disk drive specifics

	IBM 360KB floppy	WD 18GB HD	Seagate ST9250410AS 250GB HD (16MB cache) https://www.seagate.com/staticfiles/support/disc/manuals/notebook/momentus/7200.4%20(Holliday)/100534376a.pdf https://www.manualslib.com/manual/440126/Seagate-St9250410as-Momentus-7200-4-250-Gb-Hard-Drive.html
Cylinders	40	10601	16K (2 read/write heads)
Tracks per cylinder	2	12	2
Sectors per track	9	281 (average)	63
Sectors per disk	720	~36M	~500M (LBA)
Bytes per sector	512	512	512-4K (diff sizes per partition)
Capacity	360 KB	18.3 GB	256GB-2TB
Seek time (min)	6 ms	0.8 ms	1.5ms
Seek time (average)	77 ms	6.9 ms	11 ms
Rotation time	200 ms	8.33 ms	4.17ms
Spinup time	250 ms	20 sec	From off 4.5, from standby 3
Sector transfer time	22 ms	17 µsec	1.7 µsec (300MBps)

Solid State Drive (SSD)

Also known as

solid-state disk
or
electronic disk

- SSDs, unlike HDDs, have **no moving mechanical components**
  - Uses electronic interfaces compatible with traditional HDDs
  - Faster start up (no spin up)
- More resistant to physical shock, run more quietly, have lower access time and less latency
- But, SSDs are **more expensive** per unit of storage than HDDs.
- SSDs are more reliable than HDDs, BUT
  - SSD failures are often catastrophic/immediate
  - SSDs have 10-100K write cycles
  - HDDs give warning to save/recover data
- HDDs need to seek+spin for random IOPS (not sequential)

SSD organization (example)

- 1 page = 4KB
- 1 block = 64 pages
- 1 plane = 2048 blocks
- 1 die = 4 planes
- Reading and programming is performed on a page basis
- Erasure can only be performed on a block basis
  - NAND SSDs need to write whole block to write 1s (“erase” before writing), but 0s can be set individually
  - The erase state: 0xFF or 0x00
  - 1.5ms (25µs for reading a page)
  - Finite number of erase-write cycles

SSD vs. HDD

property	SSD	HDD
Spin up time	--	Seconds
Data transfer rate	100-600 MBps	300MBps
Noise	--	”lots” (?)
Cost/GB, capacity	12-20c, 2TB (2018)	2c, 8TB (2018)
Performance	Random = seq	Seek, rotational
Power consumption	5-20W	2W

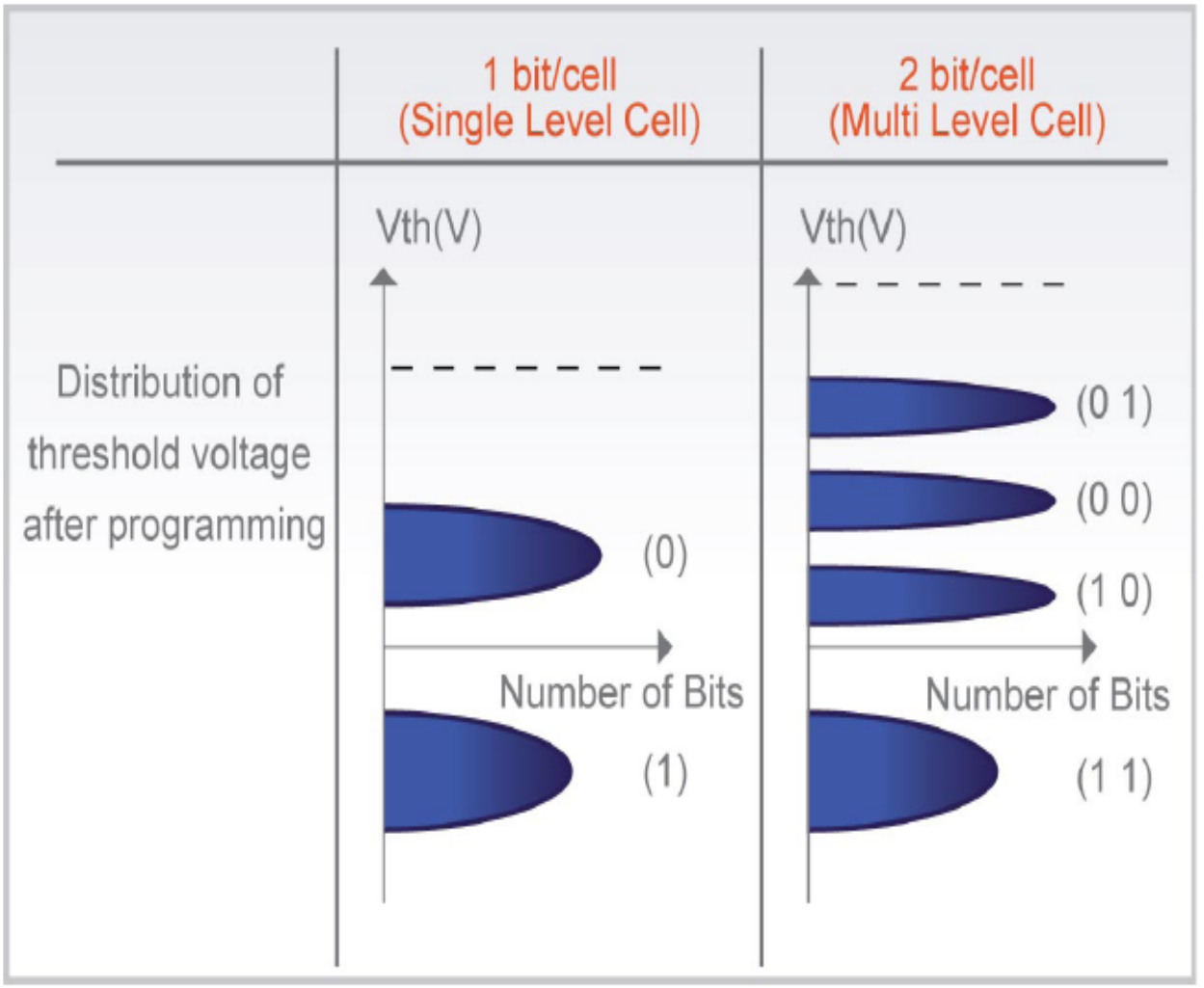
	Drive Model	Description	Seek Time			Latency	Read XFR Rate		Write XFR Rate	
			Track to Track	Average	Full Stroke		Outer Tracks	Inner Tracks	Outer Tracks	Inner Tracks
Hard Drives	Western Digital WD7500AYYS	7200 RPM 3.5” SATA	0.6 ms	8.9 ms	12.0 ms	4.2 ms	85 MB/sec	60 MB/sec*	85 MB/sec	60 MB/sec*
	Seagate ST936751SS	15K RPM 2.5” SAS	0.2 ms	2.9 ms	5.0 ms*	2.0 ms	112 MB/sec	79 MB/sec	112 MB/sec	79 MB/sec
Flash SSDs	Transcend TS8GCF266	8GB 266x CF Card	0.09ms				40 MB/sec		32 MB/sec	
	Samsung MCAQE32G5APP	32G 2.5” PATA	0.14ms				51 MB/sec		28 MB/sec	
	Sandisk SATA5000	32G 2.5” SATA	0.125ms				68 MB/sec		40 MB/sec	

MLC SSD vs. HDD

Disk type	IOPS read	IOPS write
HDD 15K rpm	500	133
Consumer 1	60K	34K
Consumer 2	170K	6K
Enterprise 1	750K	83K
Enterprise 2	585K	113K

% Writes	Total IOPS	Performance vs 15K SAS Hard Drive
0%	5400	20x better
5%	252	1.25x better
10%	130	1.5x worse
20%	65	3x worse
50%	26	8x worse
100%	13	16x worse

NAND SLC vs. MLC Technology



Source: Toshiba. 2008

(Source Toshiba)

HDD vs. SDD

Random access

	Read	Write	Erase
NAND (SLC)	25us	300us	1ms
NAND (MLC)	50us	800us	1ms
HDD	3ms	3ms	N.A.

Erase are hidden by operating the erase during the idle period.

Sequential access

	NAND : Single chip operation		NAND : 4 chip interleaving	
	Read	Write	Read	Write
NAND (SLC)	25MB/sec	20MB/sec	100MB/sec	80MB/sec
NAND (MLC)	20MB/sec	10MB/sec	80MB/sec	40MB/sec
HDD	80MB/sec	80MB/sec	-	-

(Source - Ken Takeuchi INRET, 08)

Wear-leveling

Remember: # write cycles of NAND is ~100K for SLC and ~10K for MLC

Reducing Wear Level:

- Write data to be evenly distributed over the entire storage
- Count # of Write/Erase cycles of each NAND block
- Based on the Write/Erase count, NAND controller re-map the logical address to the different physical address (flash translation table, which is similar to what?)
- Wear-leveling is done by the NAND controller (FTL), not by the OS
- What happens if the OS does it? In addition or instead of the FTL

(Source - Ken Takeuchi INRET, 08)

Static Vs. Dynamic wear-leveling

Static data: Data that does not change such as system data (OS, application SW).

Dynamic data: Data that are rewritten often such as user data.

Dynamic wear-leveling: Wear-level only over empty and dynamic data.

Static wear-leveling: Wear-level over all data including static data.

(Source - Ken Takeuchi INRET, 08)

OS changes for SSDs

- Wear Leveling can be done at the device
- LBA is useful for SSDs also
- OS needs to minimize the number of writes
  - Use more caching, smarter caching
  - TRIM operations: inform devices which pages are no longer used. How often? Who does it?
  - Device can use buffers also
  - Massive use of file system:
    - Should hibernation be allowed?
    - Should OSs rethink swapping?
    - Prefetching and caching (mainly flushing)
  - Can the memory manager and file systems be merged?