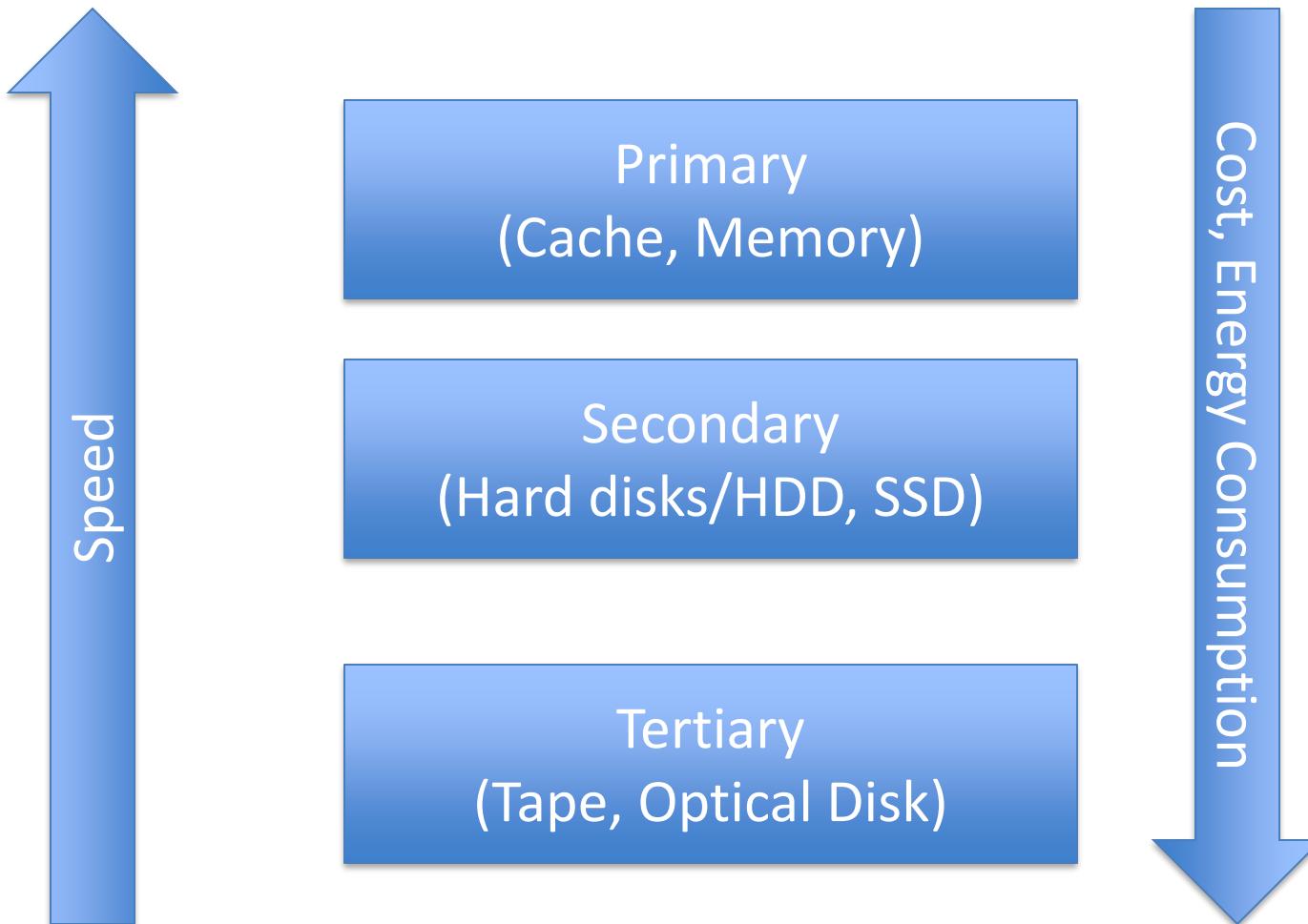


# Storage Systems

DSCI 551

Wensheng Wu

# Storage hierarchy



# Notes

- Hdfs block size = 128MB
- store a file (data set) with 200MB
- storage block (hdd) = 4KB
  - $128\text{MB} / 4\text{KB} = 32\text{K}$

1s = 1000ms

1ms = 1000microsecon

1 micro = 1000 ns

# Access times



Time taken before drive is ready to transfer data

LEVEL	ACCESS TIME	TYPICAL SIZE
Registers	"instantaneous"	under 1KB
Level 1 Cache	1-3 ns	64KB per core
Level 2 Cache	3-10 ns	256KB per core
Level 3 Cache	10-20 ns	2-20 MB per chip
Main Memory	30-60 ns	4-32 GB per system
Hard Disk	3,000,000-10,000,000 ns	over 1TB

SSD: 25,000 ns

Resource: <https://arstechnica.com/information-technology/2012/06/inside-the-ssd-revolution-how-solid-state-disks-really-work/>

# CRUD

- Basic functions of a storage device
- CRUD:
  - (C)reate/write
  - (R)ead
  - (U)pdate/overwrite
  - (D)elete

# Characterizing a storage device

- Capacity (bytes)
    - How much data it can hold
  - Cost (\$\$\$)
    - Price per byte of storage
  - Bandwidth (bytes/sec)
    - Number of bytes that can be transferred per second
    - Note that read and write bandwidth may be different
  - Latency (seconds)
    - Time elapsed, waiting for response/delivery of data
- $4KB = 4*2^{10}$   
 $= 2^{12}B$
- Hdfs block size  
 $128MB = 2^7MB = 2^{27}B$   
 $128MB/4KB = \dots$

# Time to complete an operation

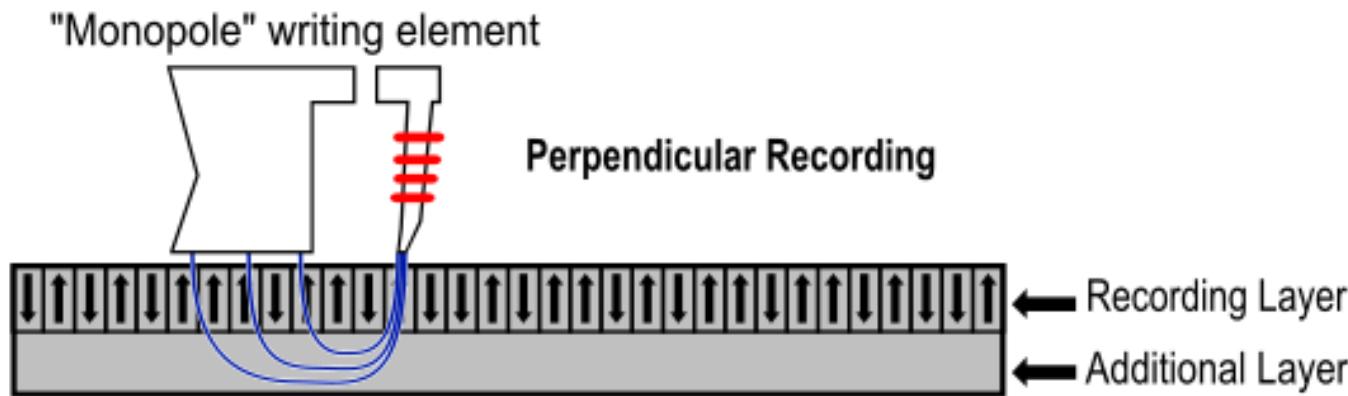
- Time to complete an operation depends on both bandwidth and latency
  - $\text{CompletionTime} = \text{Latency} + \frac{\text{Size}}{\text{Bandwidth}}$
- The time for a workload may depend on
  - Technology, e.g., hard drive/SSD
  - Operation type, e.g., read/write
  - Number of operations in the workload
  - Access pattern (random vs. sequential)

# Access pattern

- Sequential
  - Data to be accessed are located next to each other or sequentially on the device
- Random
  - Access data located randomly on storage device

# Magnetic recording

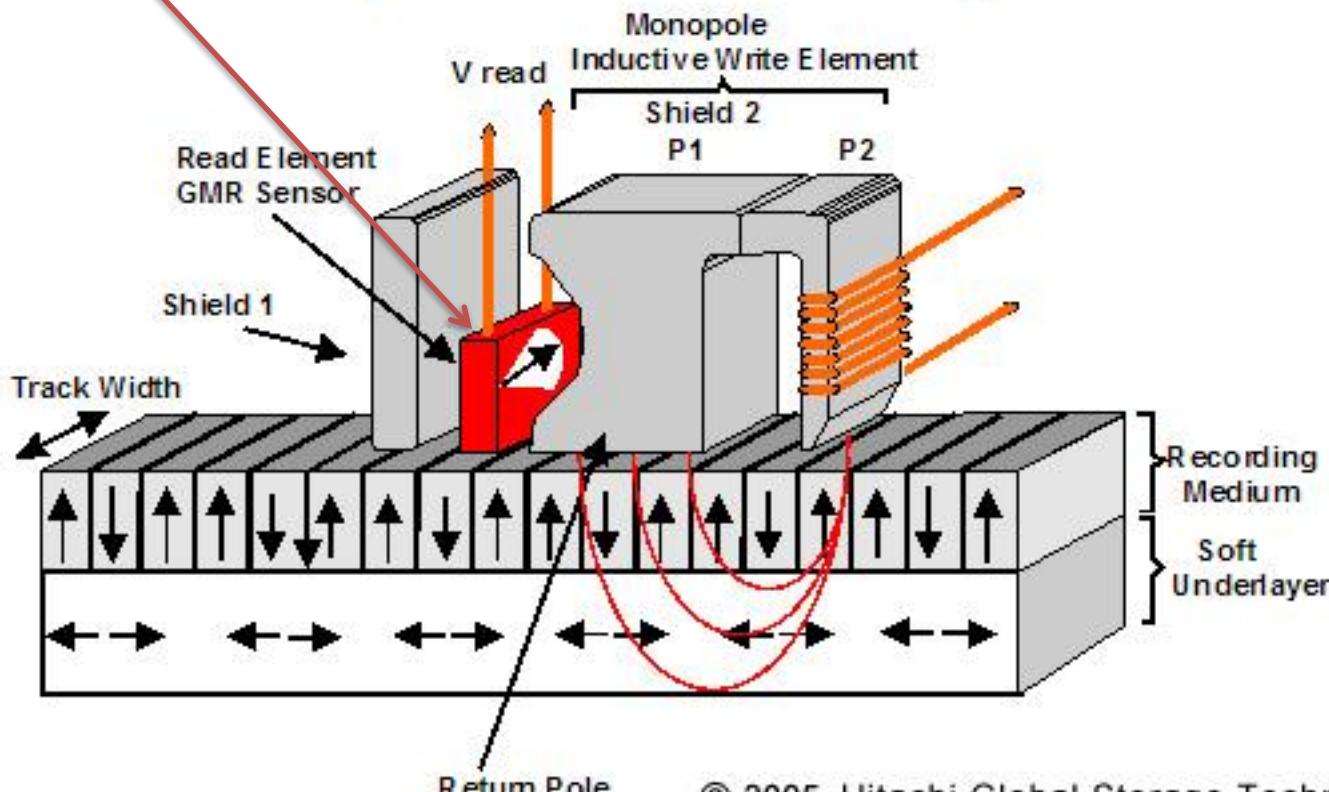
- Write head
  - Applies electrical current to write head
  - Changes direction of magnetic field under head



# Reading

- Read head senses direction of magnetic field

## Perpendicular Recording

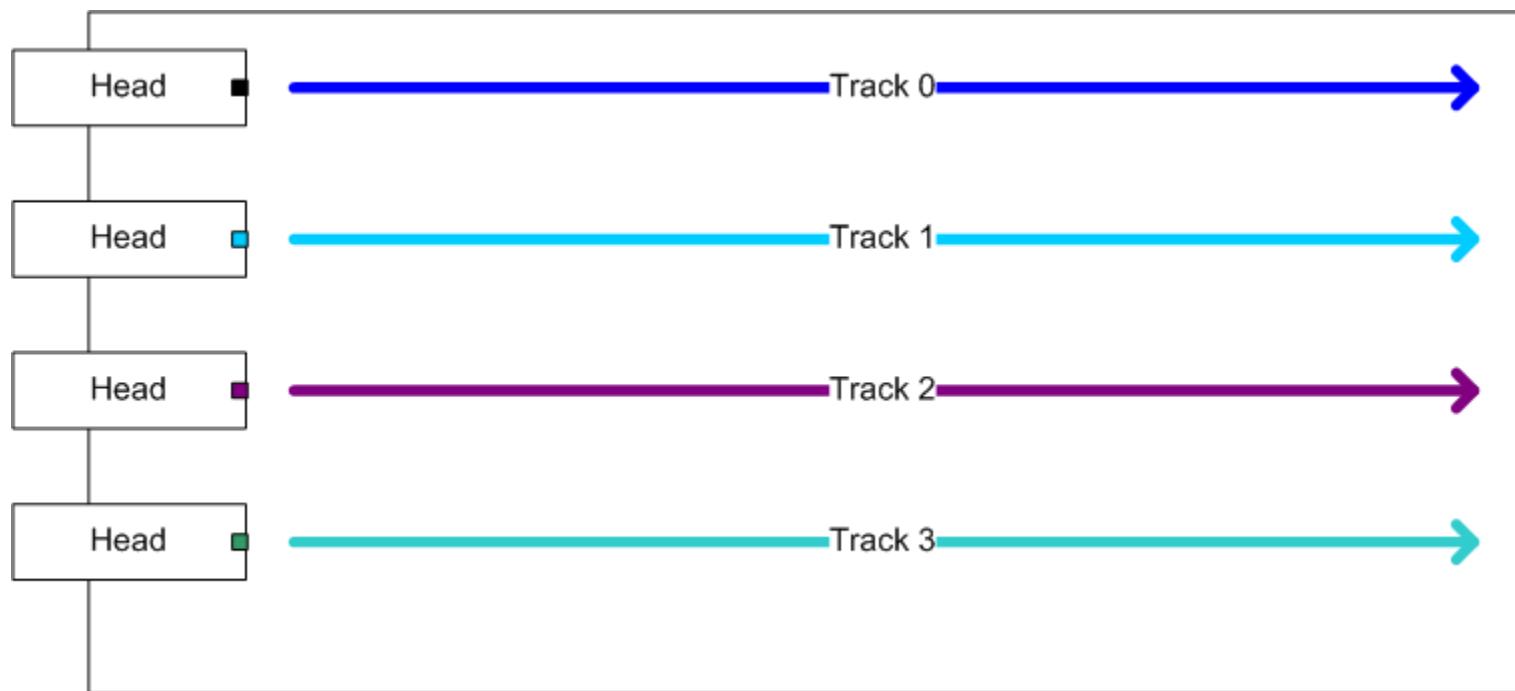


# Road map

- Tapes 
- Hard disk
- Solid state drive

# Linear tape

- Data recorded on parallel tracks that span the length of the tape



# Tapes

- Current technology is LTO
  - Linear Tape-Open (an open standard)
- Characteristics
  - Capacity up to 6.25TB per tape (LTO-7)
  - Drive cost ~ \$2000
  - Tape cost ~ \$60 for 6TB tape
- Tape access time (~ minute)
  - Time to mount the tape
  - Time to wind the tape to correct position
- Data transmission rates ~ 250MB/sec



\$60
300MB/s
6TB
LTO-7

# Performance characteristics

- High latency/low cost makes tape most appropriate for "archival" storage
  - Low frequency of (mostly sequential) reads
  - Very large data objects
- Random access will be slow due to latency
  - Sequential reads will be fast

# Linear tape file system

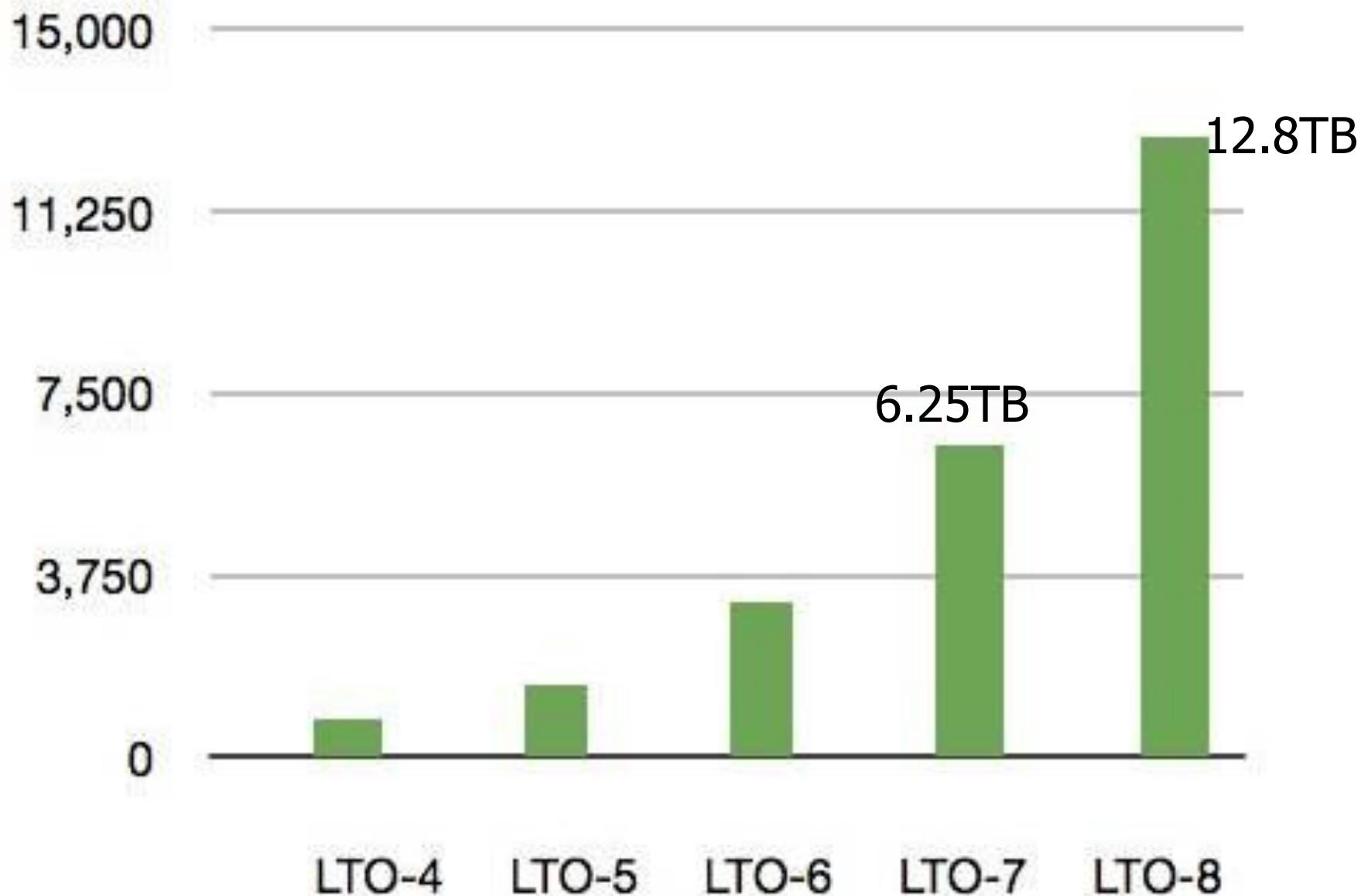
- Two partitions on tape
  - First contains metadata and directories. Tape reader can find and load this very quickly
  - Second contains blocks for data
- Directory structure coded in XML
  - Self describing file format...



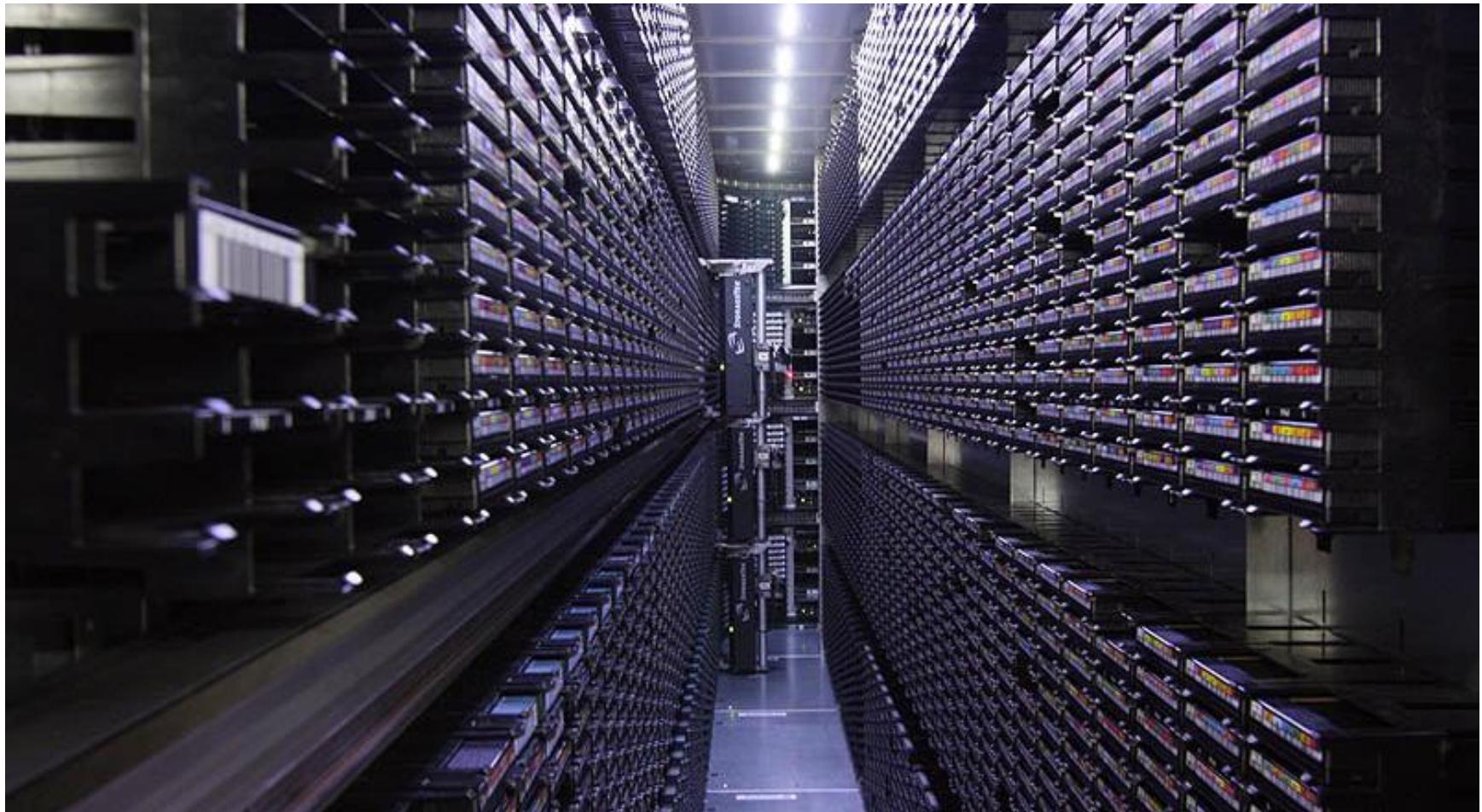
# Tape Cartridge



## **Raw Capacity in GB**



# A tape library



# Inside a robotic tape library

- <https://www.youtube.com/watch?v=nYfTtvpQ778>



# Road map

- Tapes
- Hard disk 
- Solid state drive

# Hard disk drives

- Perhaps the most pervasive form of storage
- Basic Idea:
  - One or more spinning magnetic platters
    - Typically two surfaces per platter
  - Disk arm positions over the radial position (tracks) where data are stored
    - It swings across tracks (but do not extend/shrink)
  - Data is read/written by a read/write head as platter spins



Sponsored ⓘ

WD 4TB 3.5 Inch SATA III, 7200 RPM, 64 MB Cache Enterprise Hard Drive (WD4000FYYZ)

★★★★★ 321

\$123<sup>00</sup>

✓prime FREE Delivery Thu, Jan 23

Amazon's Choice



WD Blue 1TB PC Hard Drive - 7200 RPM Class, SATA 6 Gb/s, 64 MB Cache, 3.5" - WD10EZEX

★★★★★ 24,141

\$44<sup>09</sup> \$109.99

✓prime FREE Delivery Thu, Jan 23

More Buying Choices

\$34.98 (76 used & new offers)

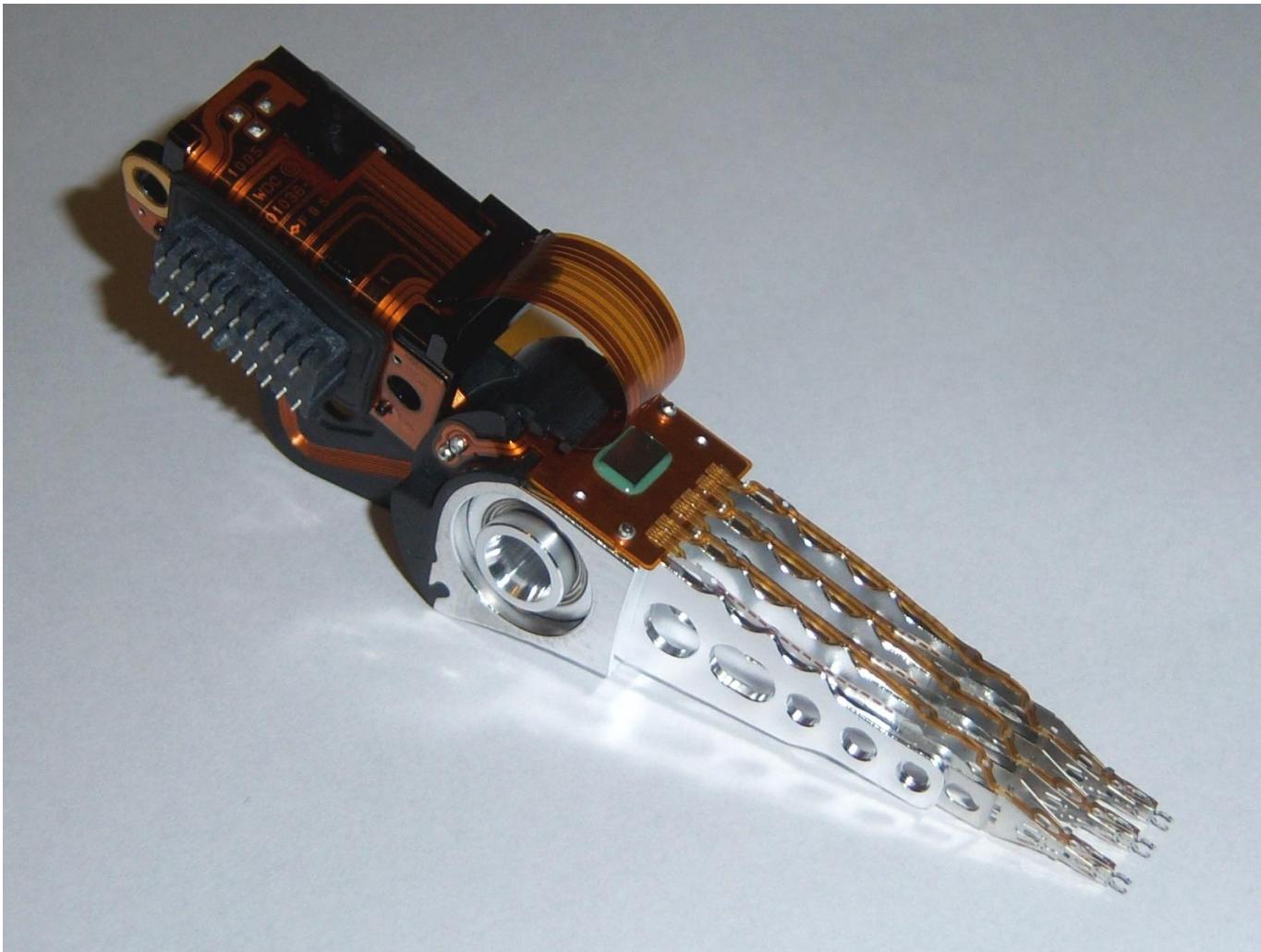
# Internal of hard disk



# Disk arm and platter

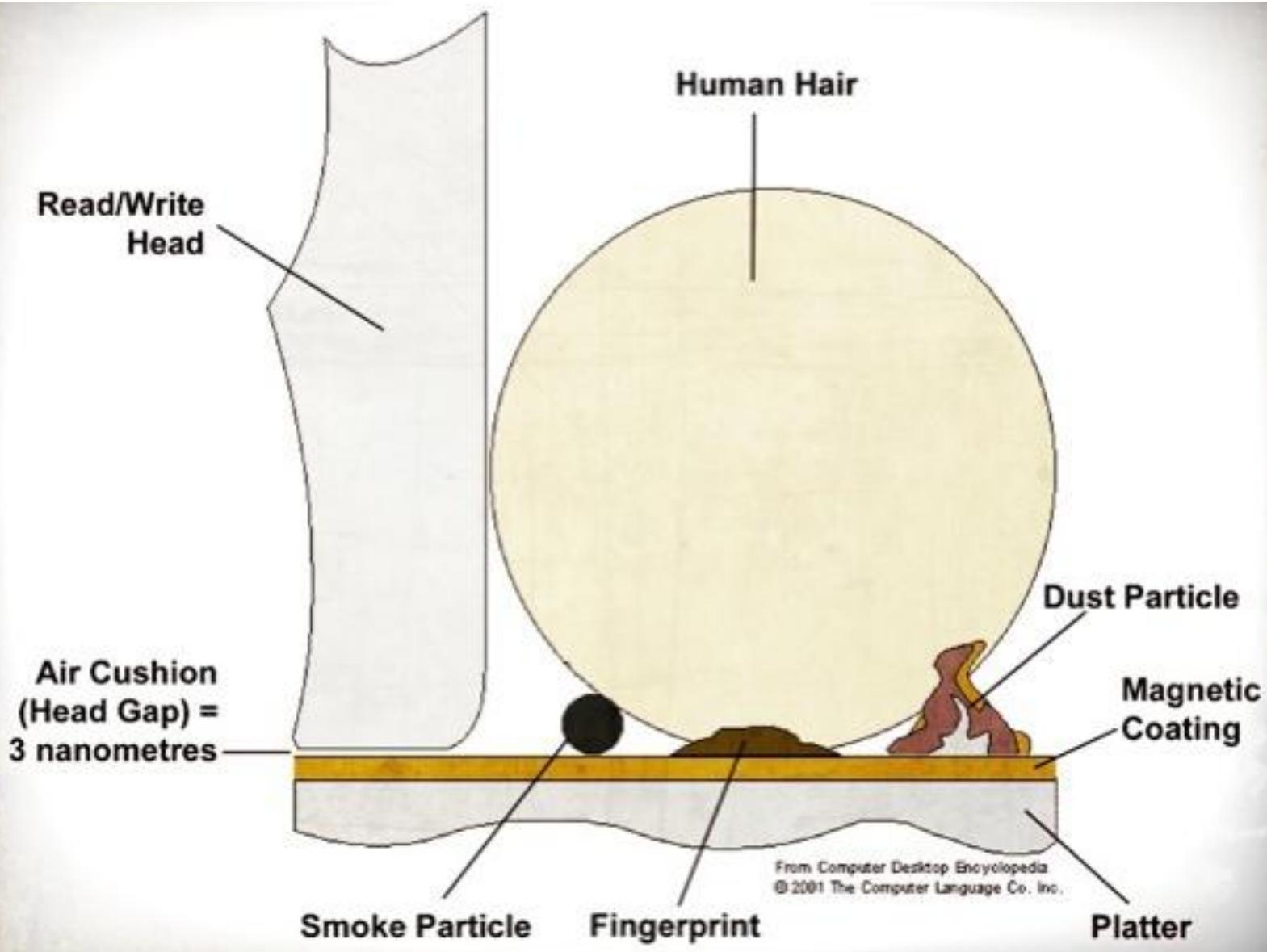


# Disk arm & head close-up



# Disk head close-ups





# Disk head movement

- Hard disk head movement while copying files between two folders (e.g., partition c to d)
  - <https://www.youtube.com/watch?v=BIB49F6ExkQ>



# 2GB Storage in 1980s (\$250,000!)



# Physical characteristics

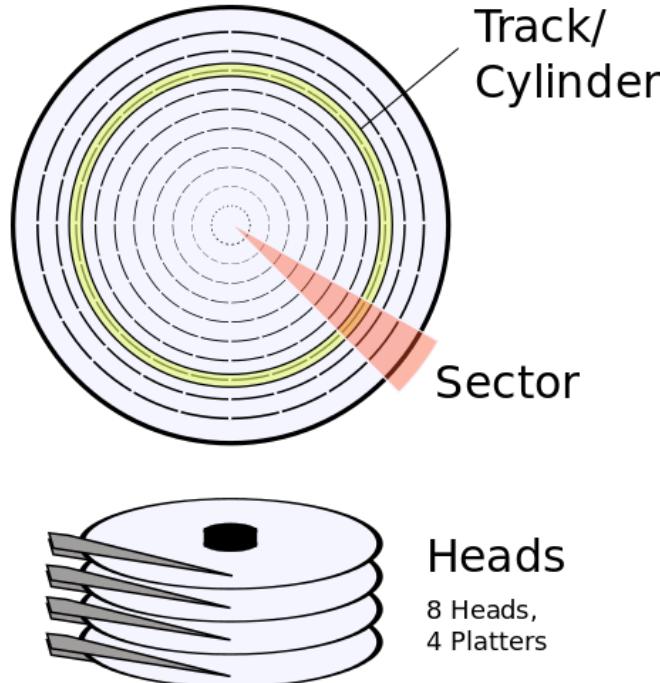
- 3.5" (diameter, common in desktops)
- 2.5" (common in laptops)
- Rotational speed
  - 4,800 RPM
  - 5,400 RPM
  - 7,200 RPM
  - 10,000 RPM => how long for one rotation?
    - $60000 \text{ ms} / 10000 \text{ rotations} = 6\text{ms/rot.}$
- Between 5-7 platters
- Current capacity up to 10TB (Western Digital)

# Disk organization

- Each platter consists of a number of tracks
- Each track is divided into N fixed size sectors
  - Typical sector size is 512 bytes (old) or **4KB (new)**
    - **$4 * 2^{10}$  bytes**
  - Sectors can be numbered from 0 to N-1
  - Entire sector is written "atomically"
    - All or nothing

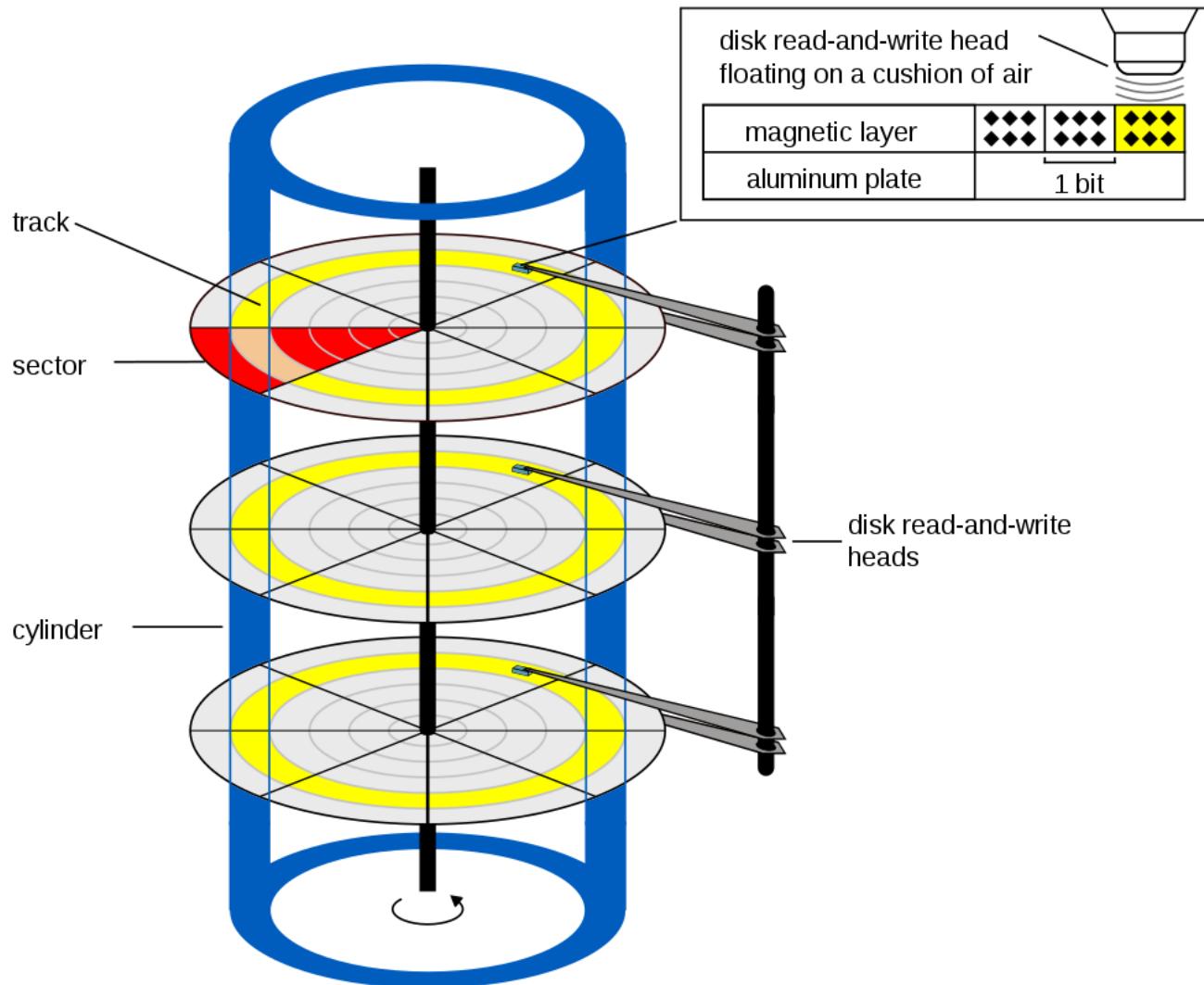
# CHS (cylinder-head-sector) (5-0-3)

- Early way to address a sector
  - Now LBA ([Logical Block Addressing](#)) more common



<https://en.wikipedia.org/wiki/Cylinder-head-sector>

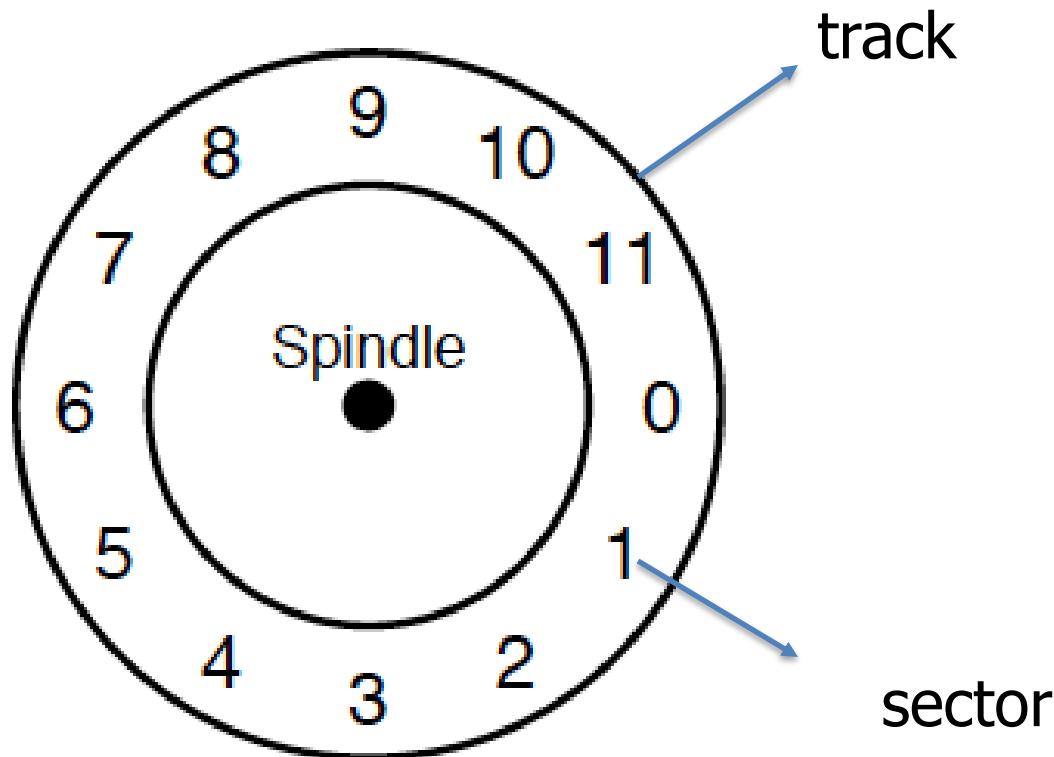
# CHS (Wikipedia)



# Example

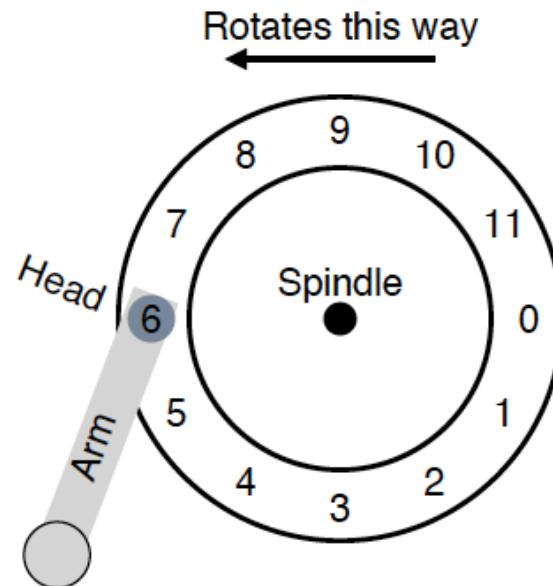
- # cylinders:  $256 = 2^8$
- # heads:  $16 = 2^4$  (i.e., 8 platters, 2 heads/platter)
- # sectors per track:  $64 = 2^6$
- Sector size = 4KB =  $2^{12}$ B
  - What is the capacity of the drive?  
 $2^{(8+4+6+12)} = 2^{30}$  (bytes) = 1GB

# A simple disk drive (one track only)



# Rotational latency

- Waiting for the right sector to rotate under the head
  - On average: about  $\frac{1}{2}$  of time of a full rotation
  - Worst case?
  - Best case?

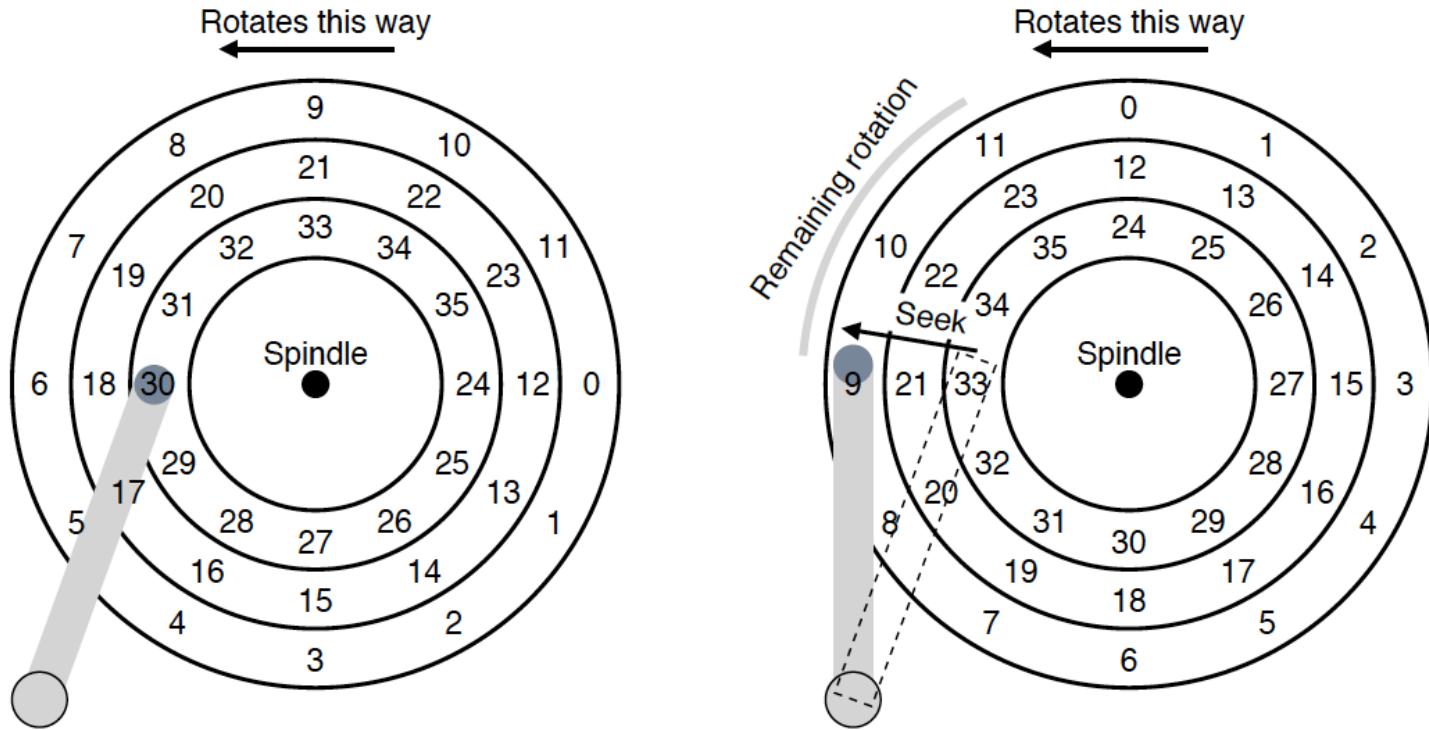


# Rotation time

- Assume 10,000 RPM (rotations per minute)
  - SCSI
  - 1 s = 1000 millisecond (ms)

$$\frac{60000 \text{ ms}}{10000 \text{ rotations}} = \frac{6 \text{ ms}}{\text{rotation}}$$

# Multiple tracks: add seek times



Average seek time is about 1/3 max seek time  
(see reading: Chapter 37, page 9 for more details)

# Transfer time

- Assume that we transfer 512KB
- Assume 128 MB/sec transmission bandwidth
- Transfer time:  
$$512\text{KB}/128\text{MB} * 1000\text{ms} = 4\text{ms}$$
$$2^9/2^7 = 2^2 = 4$$
$$\text{ASCII} = 2^7 = 128$$

# Completion time

- $T = T_{\text{seek}} + T_{\text{rotation}} + T_{\text{transfer}}$ 
  - $T_{\text{seek}}$ : Time to get the disk head on right track
  - $T_{\text{rotation}}$ : Time to wait for the right sector to rotate under the head
  - $T_{\text{transfer}}$ : Time to actually transfer the data

# Example

- Capacity 4TB
- # platters: 4
- # heads: 8
- Bytes per sector: 4096 (4KB)
- Transmission bandwidth: 100MB/sec
- Maximum seek time: 12ms
- RPM: 10,000 => 6ms/rotation

# Time to transfer a file

- The file occupies 100 sectors (sequentially)
- Avg. seek time = 4ms
- Avg. rotational latency = 3ms
  - 10,000 RPM => 6ms for a full rotation
- Transfer time =  $4\text{ms} = 100 * 4\text{KB} / (100\text{MB/sec})$   
 $= 400 \text{ KB*100 ms} / 100\text{MB} = (400\text{MB}/100\text{MB})\text{ms}$
- Actual bandwidth =  $100 * 4\text{KB} / 11\text{ms}$ 
  - Is actual bandwidth < 100MB/sec

# Sector vs. block

- Block has 1 or more sectors
- Disk typically transfers one block at a time
- We will assume one block = one sector
  - Unless stated otherwise

# Sequential operations

- May assume all sectors involved are on same track
  - We may need to seek to the right track or rotate to the first sector
- But no rotation/seeking needed afterward

# Actual bandwidth (OS)

- Consider a workload  $w$ 
  - E.g.,  $w$  = sequential access of 100 blocks of data
  - Denote size (# of bytes) of data in  $w$  as  $|w|$
  - E.g.,  $w$  = 400KB (100 blocks, 4KB/block)
- Suppose completion time for  $w = t$
- Actual bandwidth (with respect to  $w$ ) =  $|w|/t$

# Sequential vs. random

- Consider disk with 7ms avg. seek, 10,000 RPM platter speed and 50 MB/sec transfer rate, 4KB/block
- Sequential access of 10 MB
  - Completion time =  $7 + 3 + 10/50 * 1000 = 210\text{ms}$
  - Actual bandwidth =  $10\text{MB}/210\text{ms} = 47.62 \text{ MB/s}$
- Random access of 10 MB ( $=10\text{MB}/4\text{KB} = 2,560$  blocks)
  - Completion time =  $2560 * (7 + 3 + 4/50) = 25.8\text{s}$
  - Actual bandwidth =  $10\text{MB} / 25.8\text{s} = .387 \text{ MB/s}$

# notes (sequential)

- 10MB
  - $10\text{MB}/4\text{KB} = 2.5 * 1024 \sim 2500$
- completion time
  - latency: 7ms + 3ms
  - transmission:  $10\text{MB} * 1000\text{ms} / 50\text{MB} = .2 * 1000$  (ms) = 200ms
  - total = 10 + 200 = 210ms
- actual bandwidth =  $10\text{MB} / 210\text{ms}$ 
  - b ? 50MB/sec

# notes (random)

- 10MB (2560 blocks/sectors)
- $2560 * (7\text{ms} + 3\text{ms} + 4\text{KB} * 1000\text{ms}/50\text{MB})$ 
  - $2560 * (10\text{ms} + 4/50\text{ms})$
  - $2560 * (10\text{ms} + .08\text{ms})$
  - $t = 25.8\text{s}$
- actual bandwidth =  $10\text{MB}/25.2\text{s} = b'$ 
  - $b' \ll 50\text{MB/s}$

# Road map

- Tapes
- Hard disk
- Solid state drive

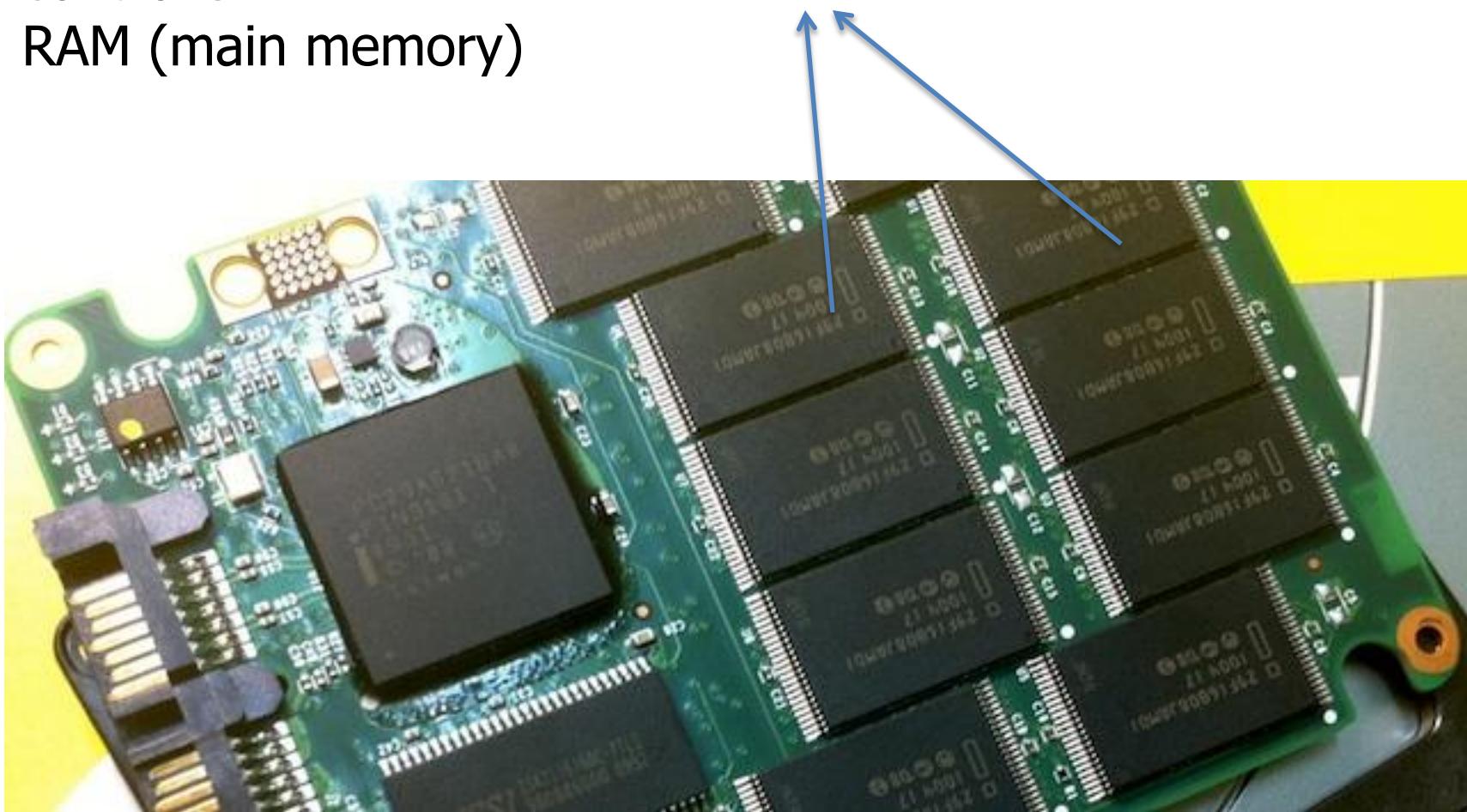


# Solid State Drive



wear leveling  
controller  
RAM (main memory)

# Memory chips

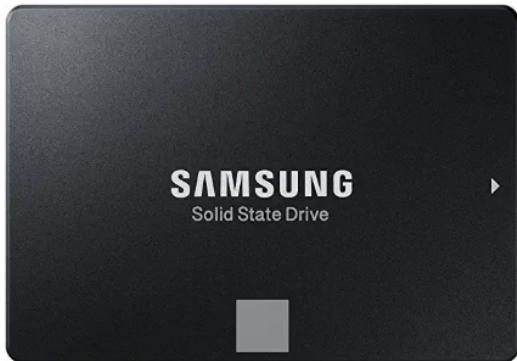


# Solid State Drives

- All electronic, made from flash memory
- Lower energy consumption than hard drive
- More expensive, less capacity
  - 3 times or more expensive
- Limited lifetime, can only write a limited number of times.
  - E.g., 100K program/erase cycles for SLC (single-level cell) memory

# SSD vs. Hard Drive (price)

Amazon's Choice



Samsung SSD 860 EVO 1TB 2.5 Inch SATA III Internal SSD (MZ-76E1T0B/AM)

★★★★★ 13,615

\$146<sup>43</sup> \$199.99

✓prime FREE One-Day

Get it Tomorrow, Jan 20

More Buying Choices

\$123.09 (37 used & new offers)

Amazon's Choice



WD Blue 1TB PC Hard Drive - 7200 RPM Class, SATA 6 Gb/s, 64 MB Cache, 3.5" - WD10EZEX

★★★★★ 24,141

Personal Computers

\$44<sup>09</sup> \$109.99

✓prime FREE Delivery Thu, Jan 23

More Buying Choices

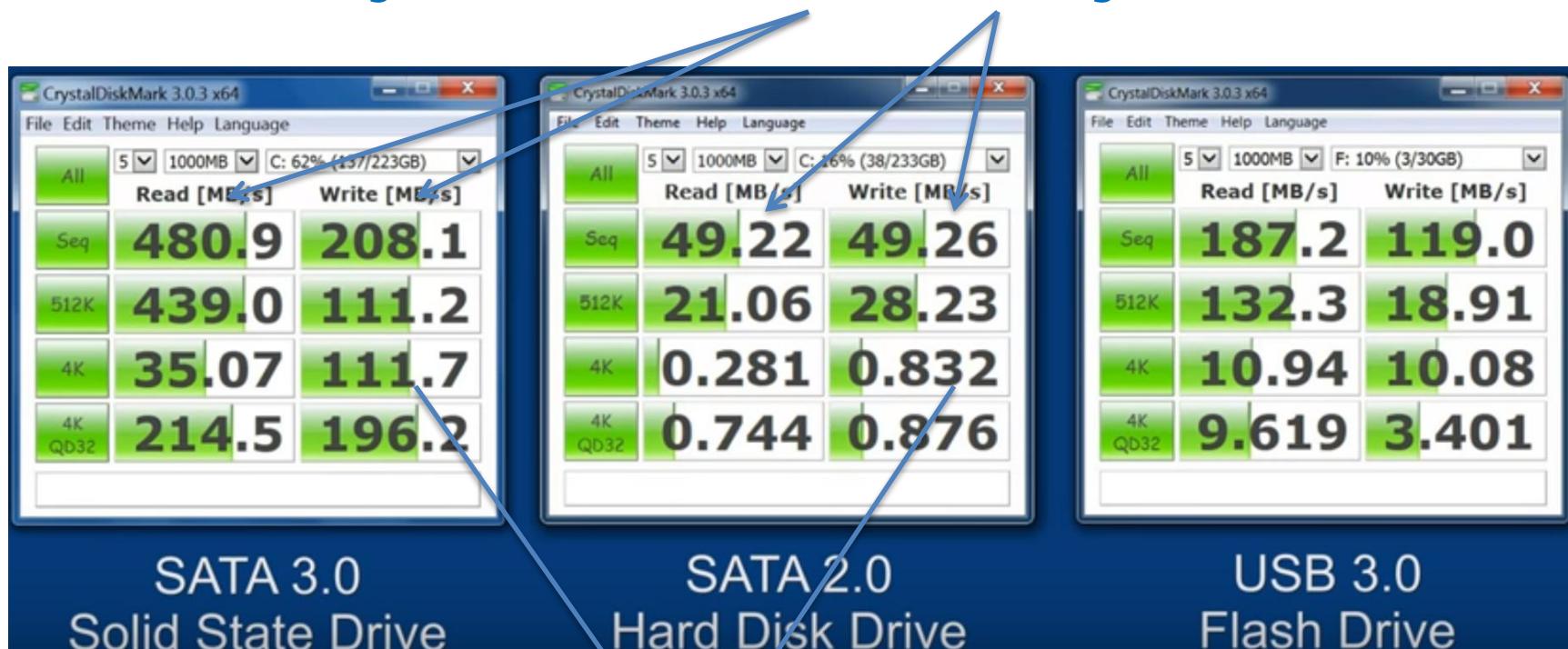
\$34.98 (76 used & new offers)

# Solid State Drives

- Same form-factor and control interface as magnetic disks
- Significantly better latency
  - No seek or rotational delay
- Much better performance on random workload:
  - Benefits from improved latency
  - However, write has much higher latency (but see next slide)

# Speed comparison ([YouTube](#))

Read vs write: significant difference in SSD vs marginal in HDD



SATA 3.0  
Solid State Drive

SATA 2.0  
Hard Disk Drive

USB 3.0  
Flash Drive

Due to buffered writes

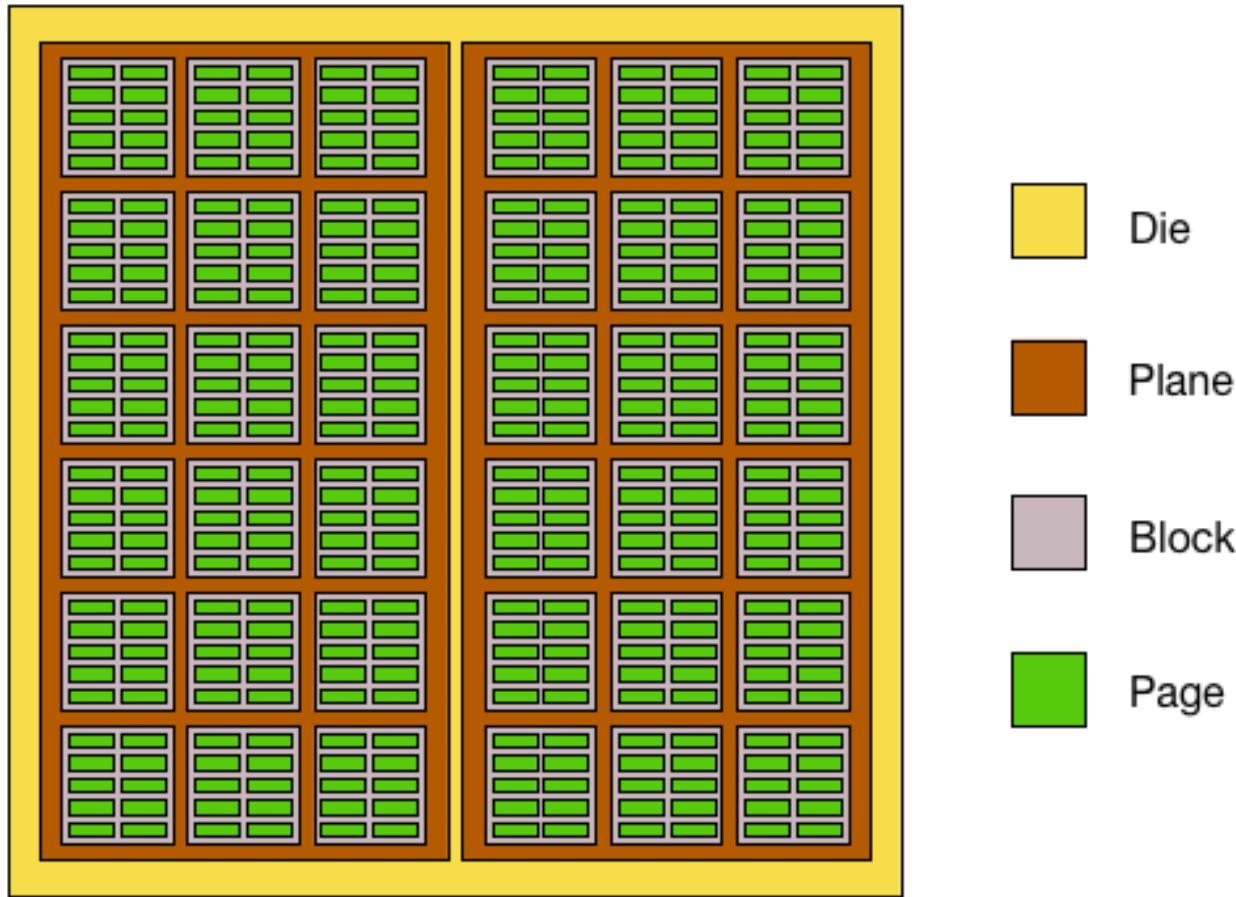
# Writing to SSD is complicated

- Can not overwrite a page
  - Need to erase its block (at a certain point) instead
- SSD controllers take care of all these details

# SSD

- Contains a number of flash memory chips
  - Chip -> dies -> planes -> blocks -> pages (rows) -> cells
  - Cells are made of floating-gate transistors
- Page is the smallest unit of data transfer between SSD and main memory
  - Much like a block in hard disk

# Die Layout



# Dies, planes, block, and pages

- Typically, a chip may have 1, 2, or 4 dies
- A die may have 1 or 2 planes
- A plane has a number of blocks
  - Block is the smallest unit that can be erased
- A block has a number of pages
  - Page is the smallest unit that can be read, programmed/written

# Typical page and block sizes

- Common page sizes: 2K, 4K, 8K, and 16K
- A block typically has 128 to 256 pages

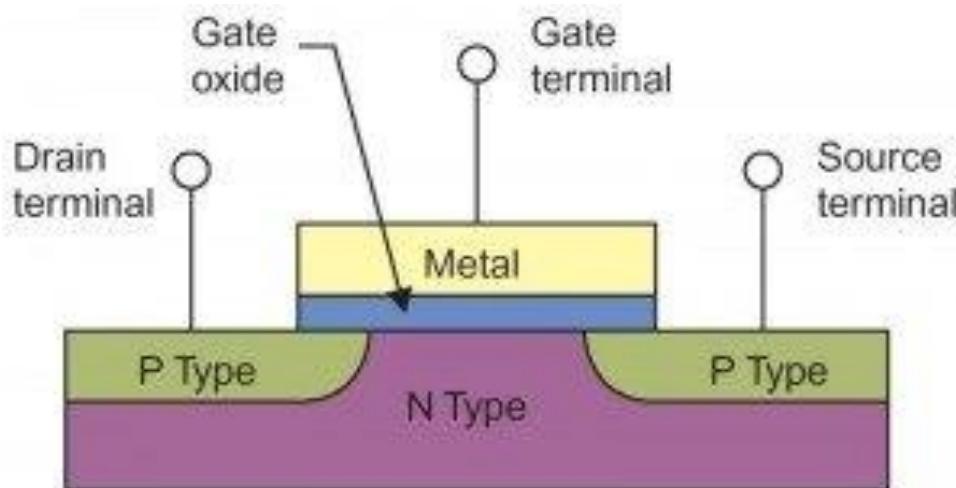
⇒ Block size: 256KB to 4MB

$$11 \text{ (binary)} = 1 * 2^1 + 1 * 2^0 = 3 \text{ (decimal)}$$

$$111 \text{ (binary)} = 7 \text{ (decimal)}$$

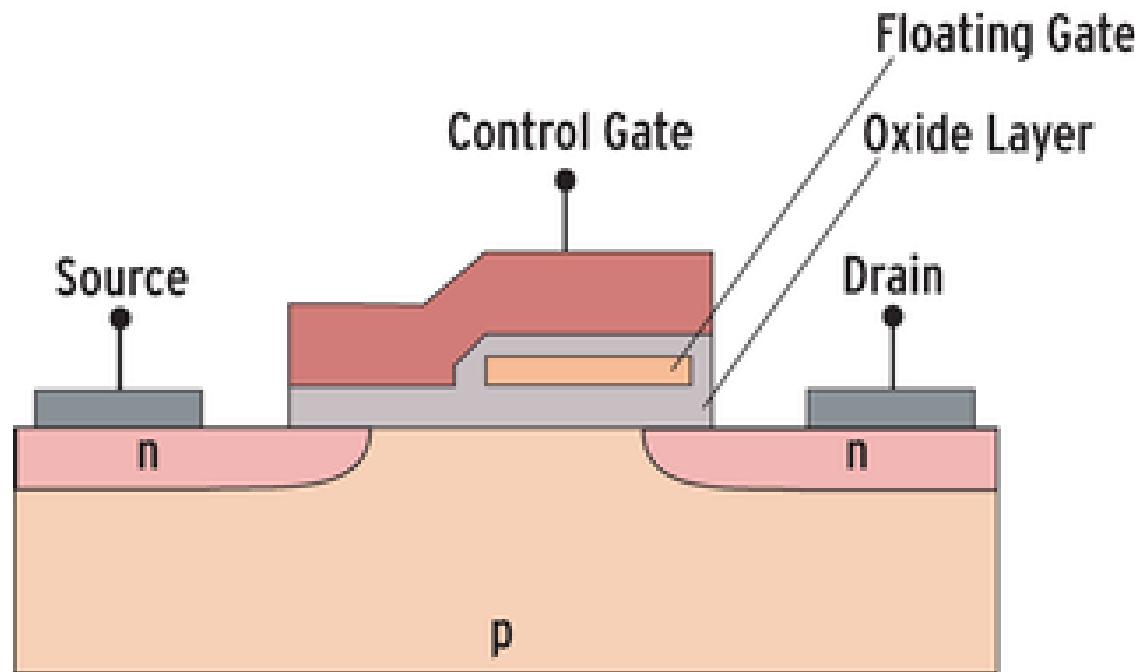
# Normal transistor (MOSFET)

- When current applied to gate terminal
  - semi-conducting region (purple) becomes conductive



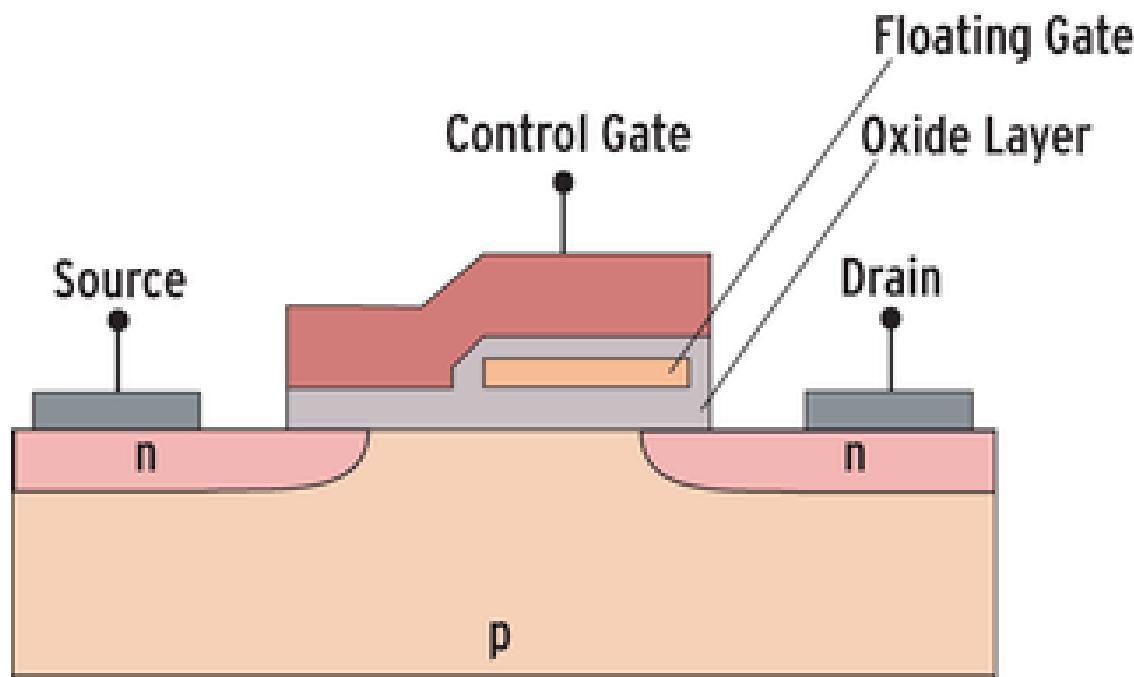
# Floating gate transistor

- Contain an additional gate: floating gate
  - Floating, since isolated by oxide layer
  - (thus not connected to other components)



# Floating gate transistor

- By applying high positive/negative voltage to control gate, electrons can be attracted to or repelled from floating gate



# notes

- 1 1 1 1
- 1 0 0 1 (program)
  - 0 1 1 1 (new data)
- 1 1 1 1 (erase)
- 0 1 1 1 (program)

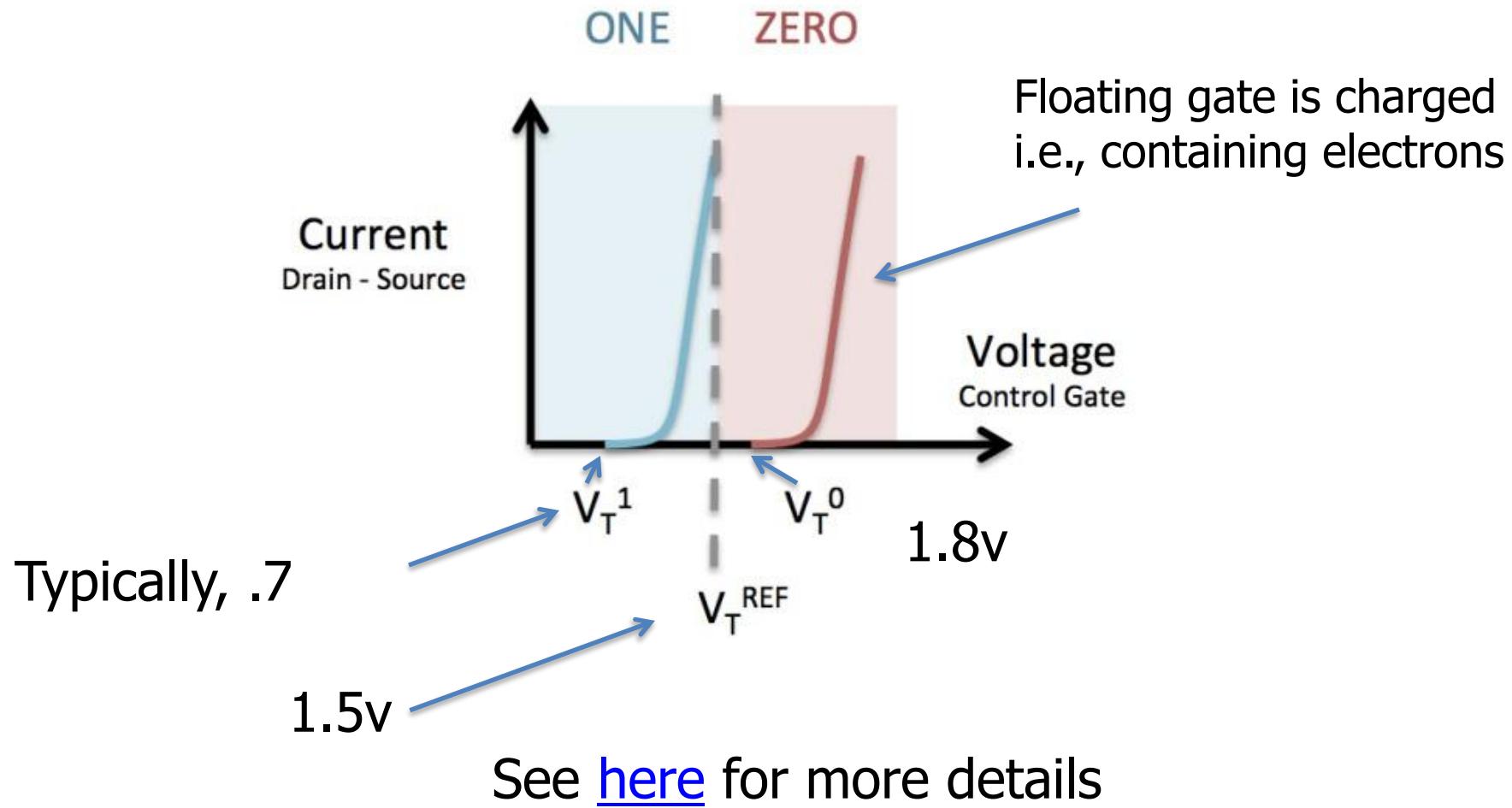
# Floating gate transistor

- State = 1, if no electrons in the floating gate
- State = 0, if there are electrons (negative charges)
  - Electrons stuck there even when power is off
  - So state is retained

# Read operations

- Electrons on the floating gate affect the threshold voltage for the floating gate transistor to conduct
- Higher voltage needed when gate has electrons

# Read operations

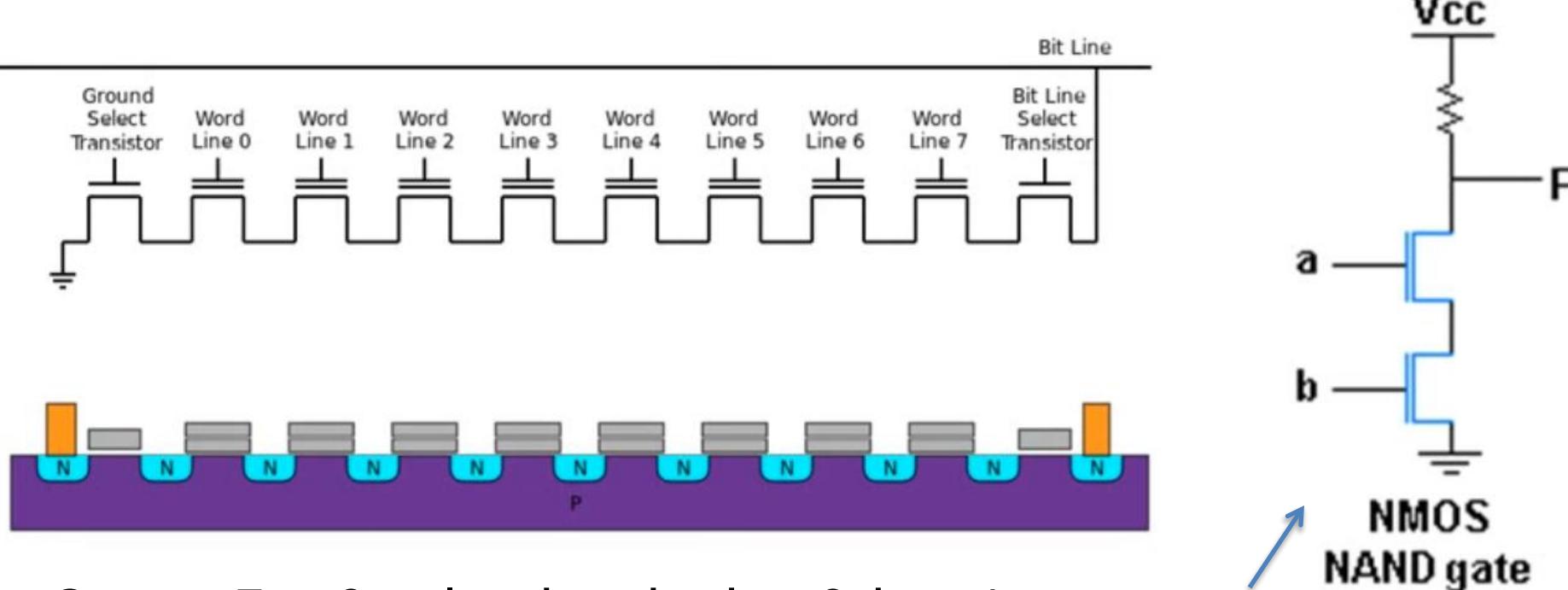


# Read operations

- Apply  $V_{int}$  (intermediate voltage)
- If the current is detected, gate has no electrons  
=> bit = 1
- If no current, gate must have electrons  
=> bit = 0

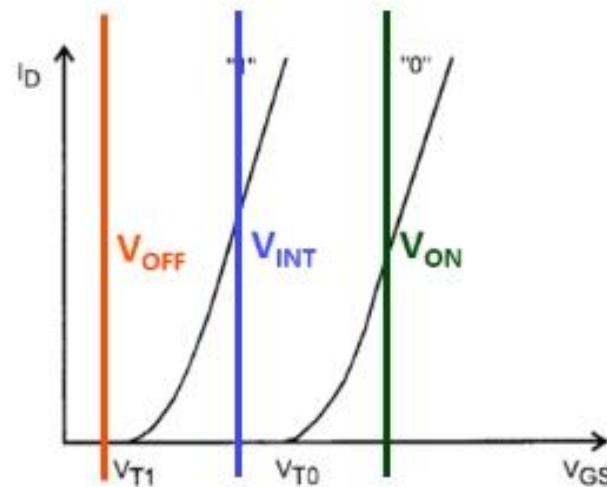
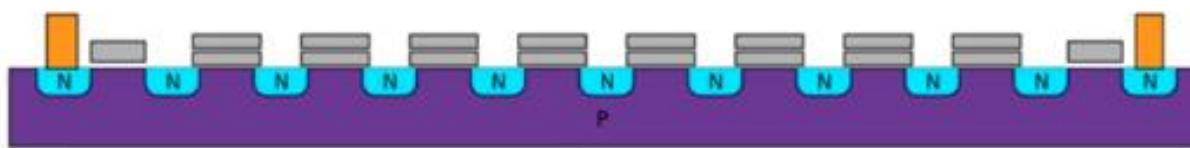
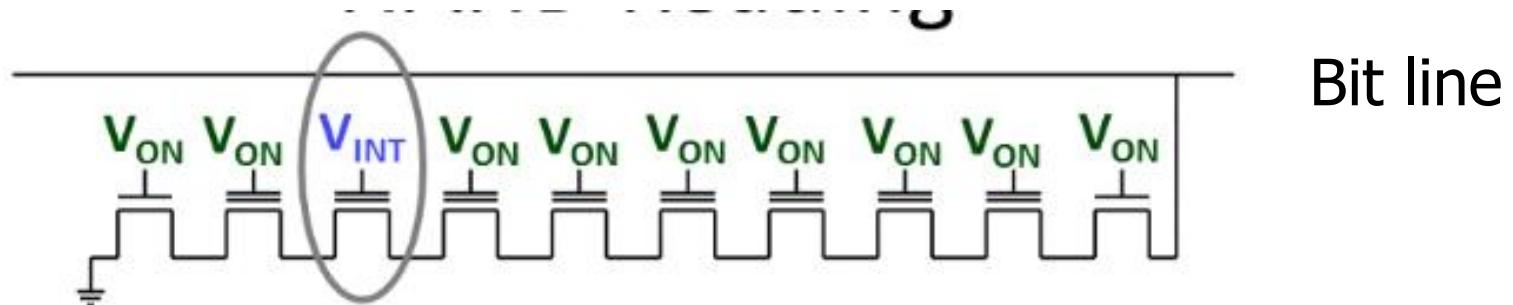
# NAND flash layout

- Transistors are strung together in a series
  - Similar to the transistors in an NAND gate



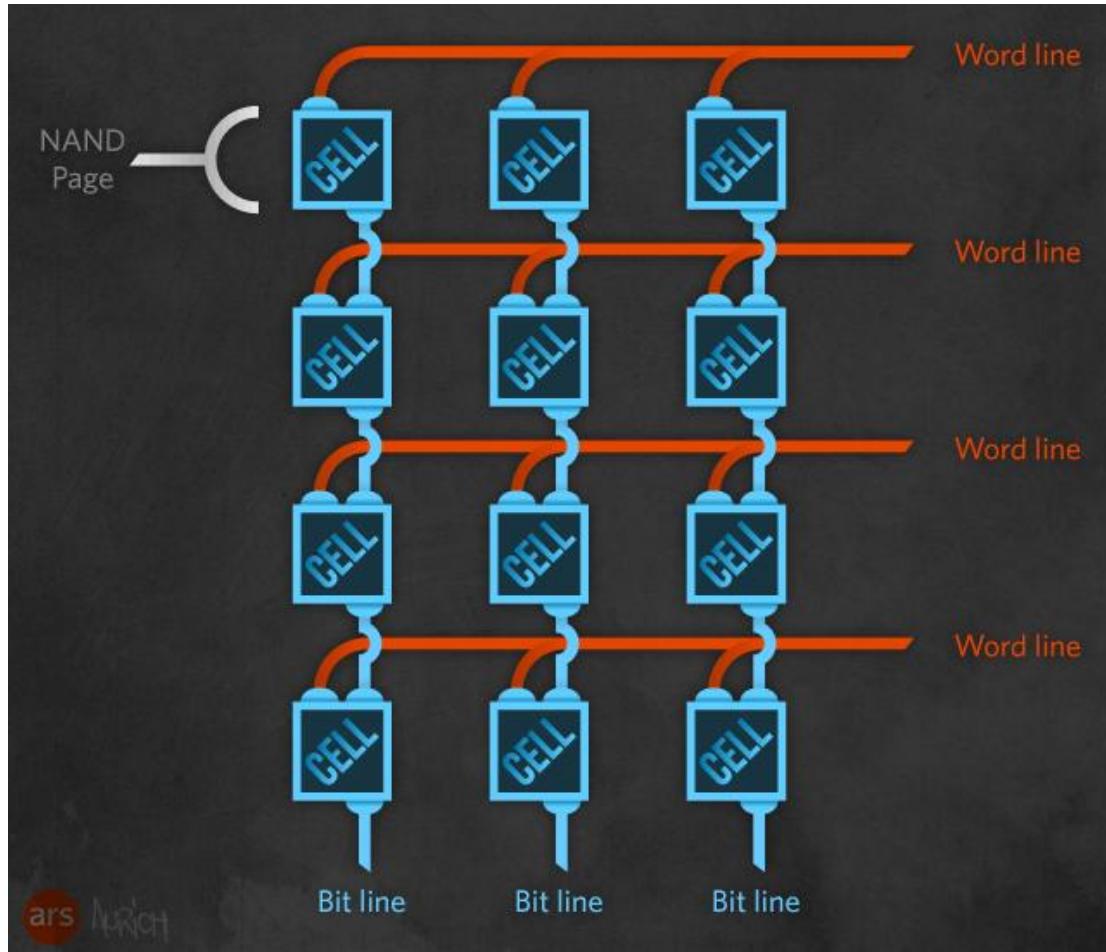
Output  $F = 0$  only when both  $a \& b = 1$   
I.e.,  $F$  will be grounded only when both  $a$  and  $b$  conduct

# NAND reading

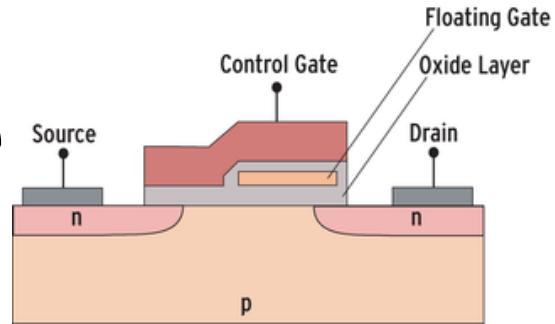


Apply  $V_{on}$  to all others so that they all conduct, no matter they are charged or not.

# NAND flash



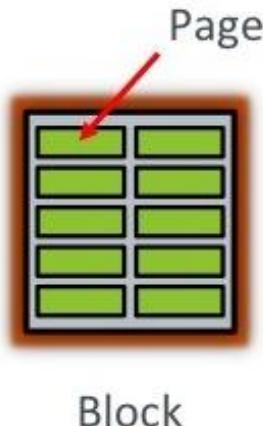
# Write and erase



- Write:  $1 \Rightarrow 0$ 
  - Apply high POSITIVE voltage ( $\gg$  voltage for read) to the control gate
  - Attract electrons from channel to floating gate (through quantum tunneling)
- Erase:  $0 \Rightarrow 1$ 
  - Need to apply much higher NEGATIVE voltage
  - Get rid of electrons from floating gate
  - May stress surrounding cells
  - So dangerous to do on individual pages

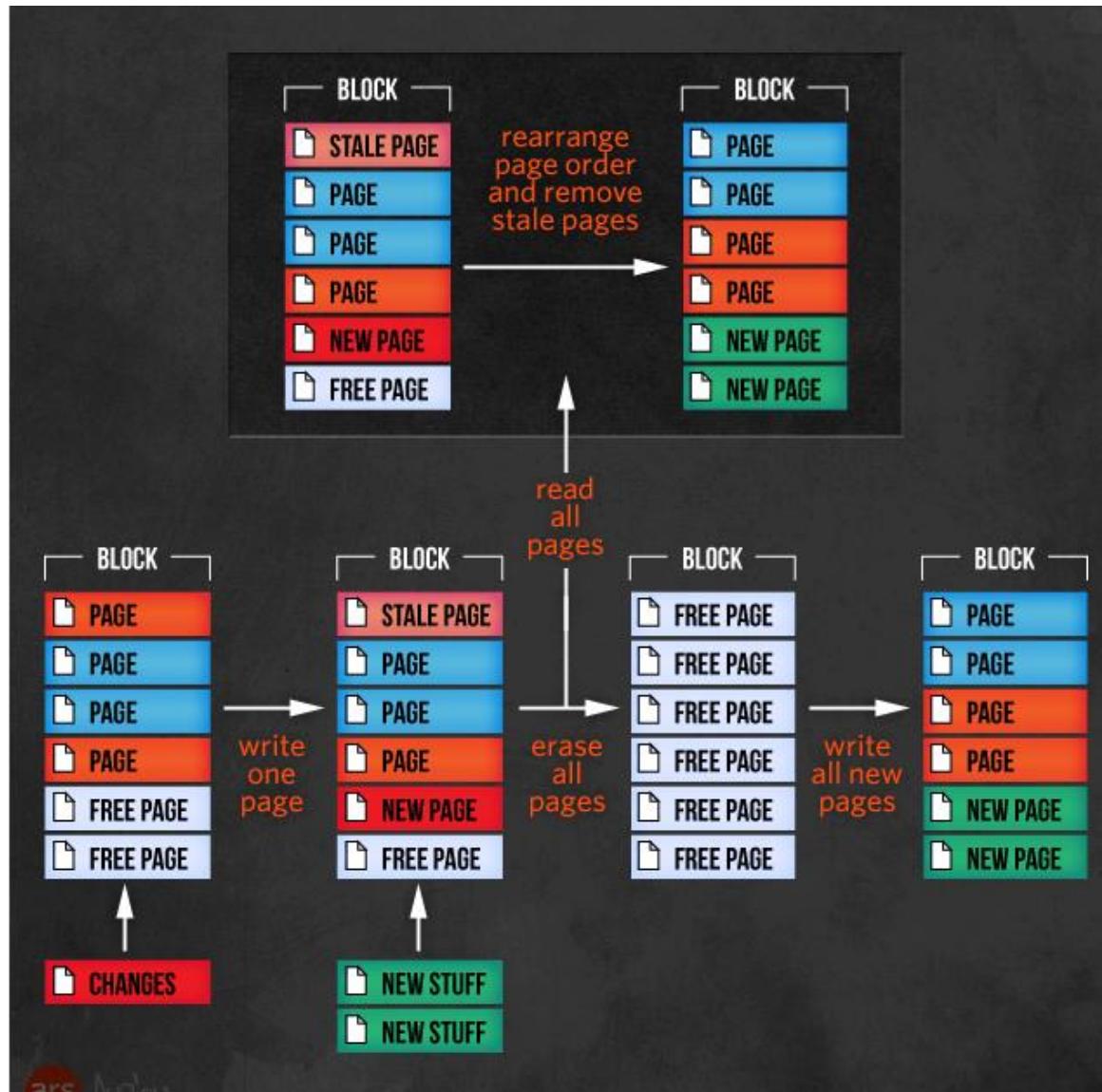
# Read/write units

- **Page** is the smallest unit for read and write (write is also called program, 1->0)
- **Block** is the smallest unit for erase (0->1)
  - i.e., make cells "empty" (i.e., no electrons)



Operation	Area
Read	Page
Program (Write)	Page
Erase	<b>Block</b>

# Example



# notes

- one write (page)
- 4/6 reads (pages)
- one erase (block)
- 6 writes (pages)
- total time = 4 reads + 7 writes + 1 erase
  - $4 * 25 + 7 * 250 + 1500$  (microseconds)

# Latencies: read, write, and erase

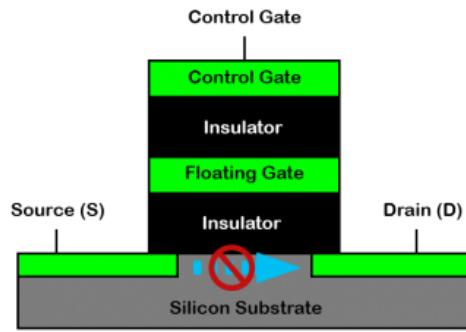
	SLC	MLC	TLC	HDD	RAM
P/E cycles	100k	10k	5k	*	*
Bits per cell	1	2	3	*	*
Seek latency (μs)	*	*	*	9000	*
Read latency (μs)	25	50	100	2000-7000	0.04-0.1
Write latency (μs)	250	900	1500	2000-7000	0.04-0.1
Erase latency (μs)	1500	3000	5000	*	*
Notes	* metric is not applicable for that type of memory				
Sources	<p>P/E cycles [20] SLC/MLC latencies [1] TLC latencies [23] Hard disk drive latencies [18, 19, 25] RAM latencies [30, 52] L1 and L2 cache latencies [52]</p>				

1111 (blank)  
1010 (program)  
0101

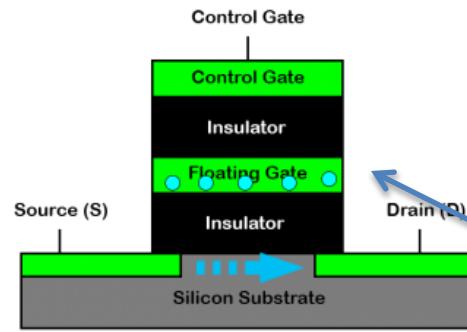
# P/E cycle

- P: program/write; E: erase
- Every write & erase damages oxide layer surrounding the floating-gate to some extent
- P/E cycle:
  - Data are written to cells (P): cell value from 1  $\rightarrow$  0
  - Then erased (E): 0  $\rightarrow$  1

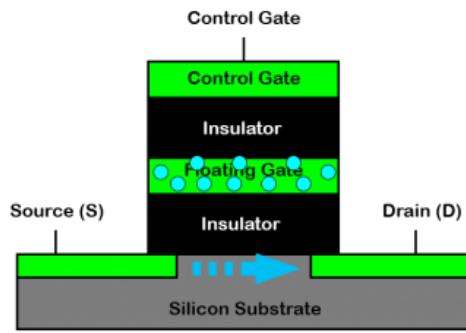
# Multi-level cell (MLC)



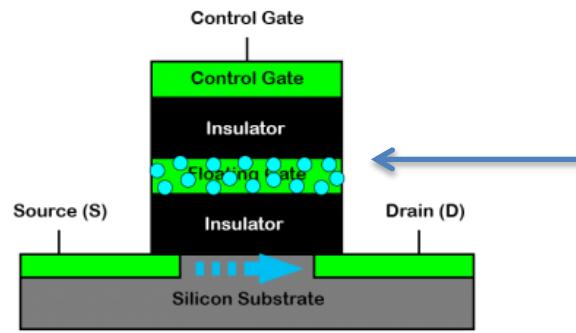
State 1 - No Charge



State 2 - Lightly Charged



State 3 - Medium Charge



State 4 - Highly Charged

Note different levels of electrons

1 => 0 (program)  
0 => 1 (overwritten)

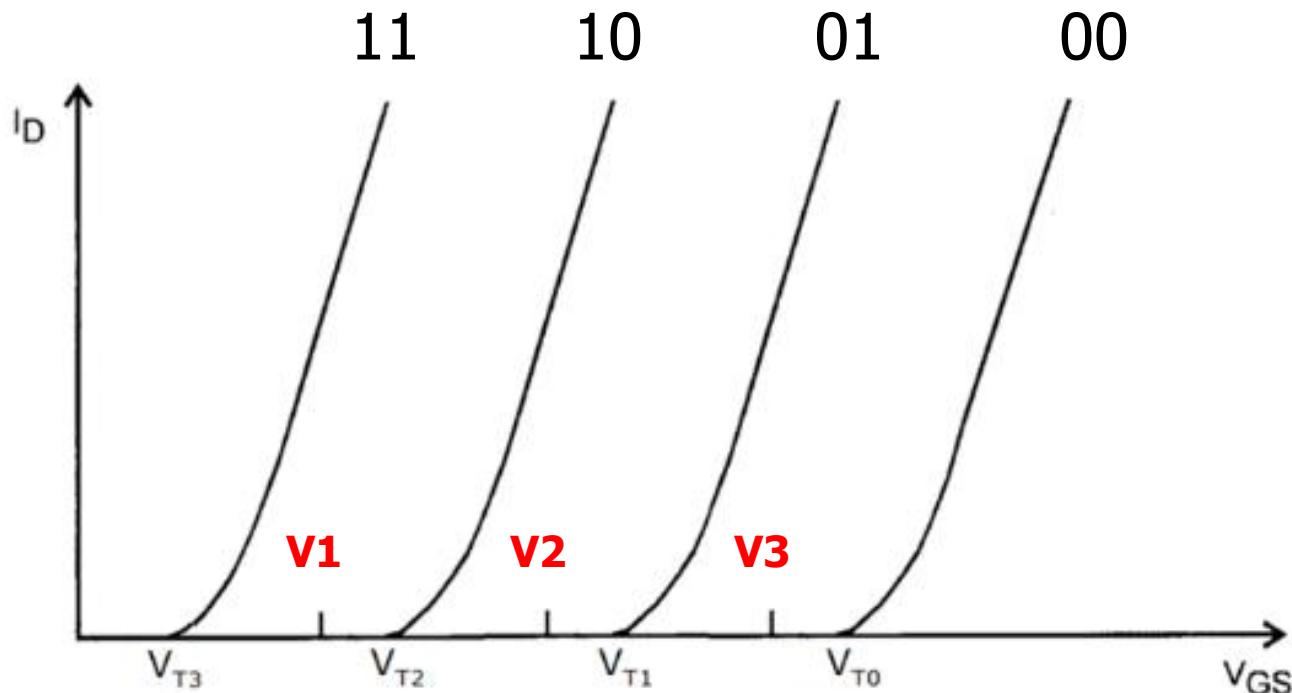
3-5 volts

<http://www.cactus-tech.com/resources/blog/details/solid-state-drive-primer-2-slc-mlc-and-tlc-nand-flash>

# MLC reading

3	5	10	15	16	20	25	32	40	58
---	---	----	----	----	----	----	----	----	----

- 2 bits, 3 intermediate voltages



Current vs voltage

# SLC compared with MLC

- SLC:
  - Less complex
  - Faster
  - More reliable
  - Less storage
  - More costly

# Read more

- Solid-state revolution: in-depth on how SSDs really work
- How do SSDs work?
  - <http://www.extremetech.com/extreme/210492-extremetech-explains-how-do-ssds-work>

# References

- How Flash Memory Works
  - <https://www.youtube.com/watch?v=msi5GDz9Jlw>
- Floating Gate Basics
  - <http://www.cse.scu.edu/~tschwarz/coen180/LN/flash.html>
- Friend of Flash
  - [http://www.nnc3.com/mags/LM10/Magazine/Archive/2008/86/040-041\\_logfs/article.html](http://www.nnc3.com/mags/LM10/Magazine/Archive/2008/86/040-041_logfs/article.html)

# References

- Understanding Flash: Floating Gates and Wear
  - <https://flashdba.com/2015/01/09/understanding-flash-floating-gates-and-wear/>
- From Transistors to Functions
  - <http://www.cs.bu.edu/~best/courses/modules/Transistors2Gates/>

# References

- Solid State Drive Primer
  - <https://www.cactus-tech.com/resources/blog/details/solid-state-drive-primer-1-the-basic-nand-flash-cell>
- How Does a Transistor Work?
  - <https://www.youtube.com/watch?v=lcrBqCFLHIY&feature=youtu.be>

# References

- How Flash Memory Works
  - <https://www.youtube.com/watch?v=s7JLXs5es7I>
- HDD, SSD