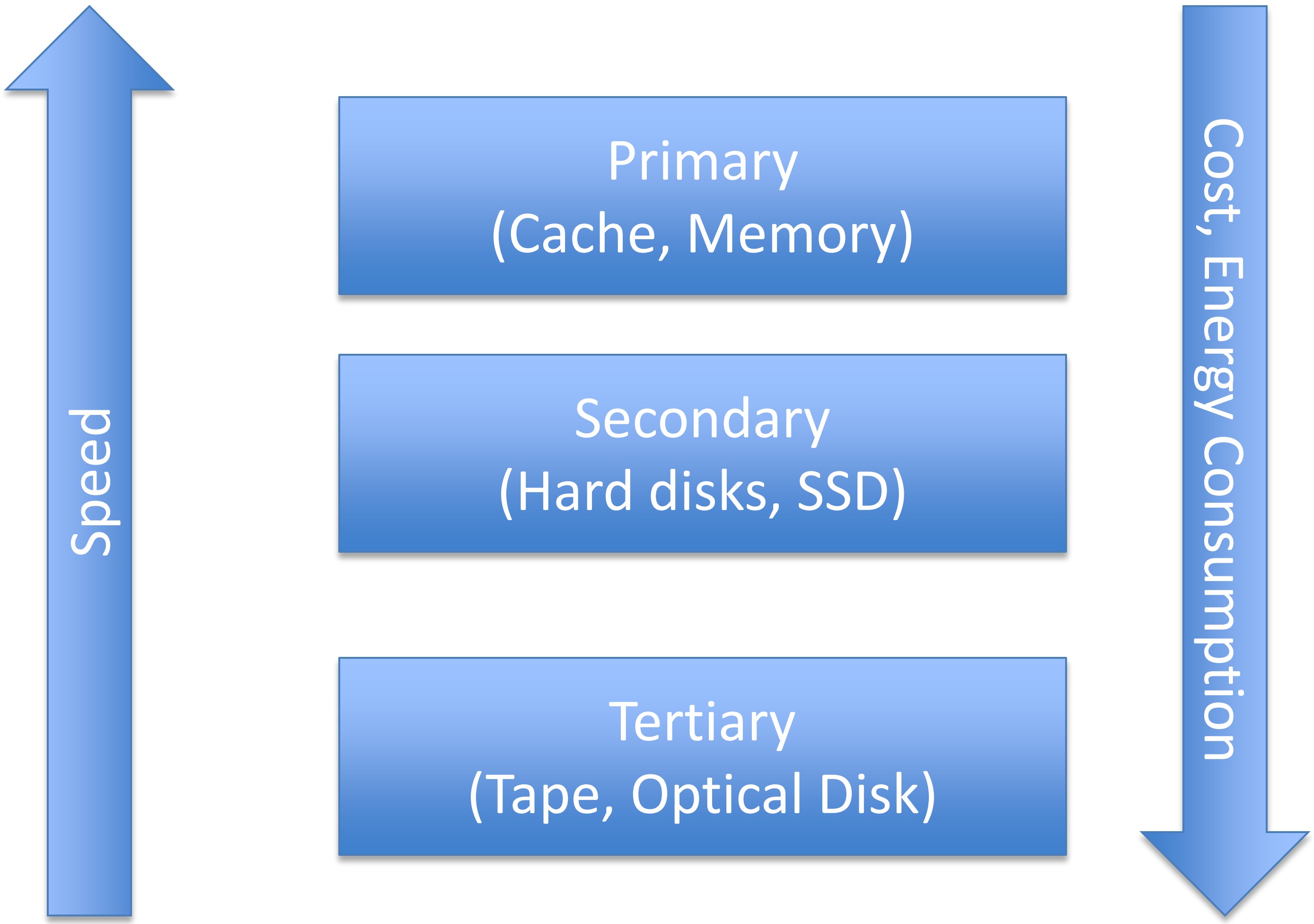
Storage Systems

INF 551

Wensheng Wu

Storage hierarchy



Access times

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 (secs)
  - Time between initiating a request and an action
  - In the case of storage, action is to deliver 1^{st} Byte

Time to complete an operation

- Time to complete an operation depends on both bandwidth and latency
  - CompletionTime = Latency + Size/Bandwidth
- The time of a workload will 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

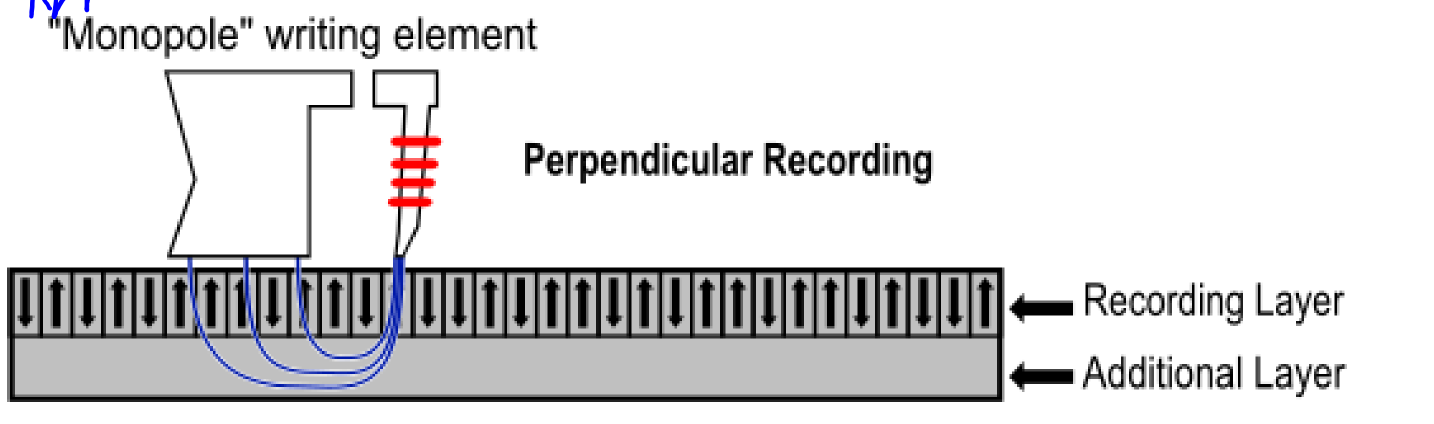
- Sequential
  - Data to be accessed are located next to each other or sequentially on the device

- Random
  - Access data located randomly on storage device

Magnetic recording

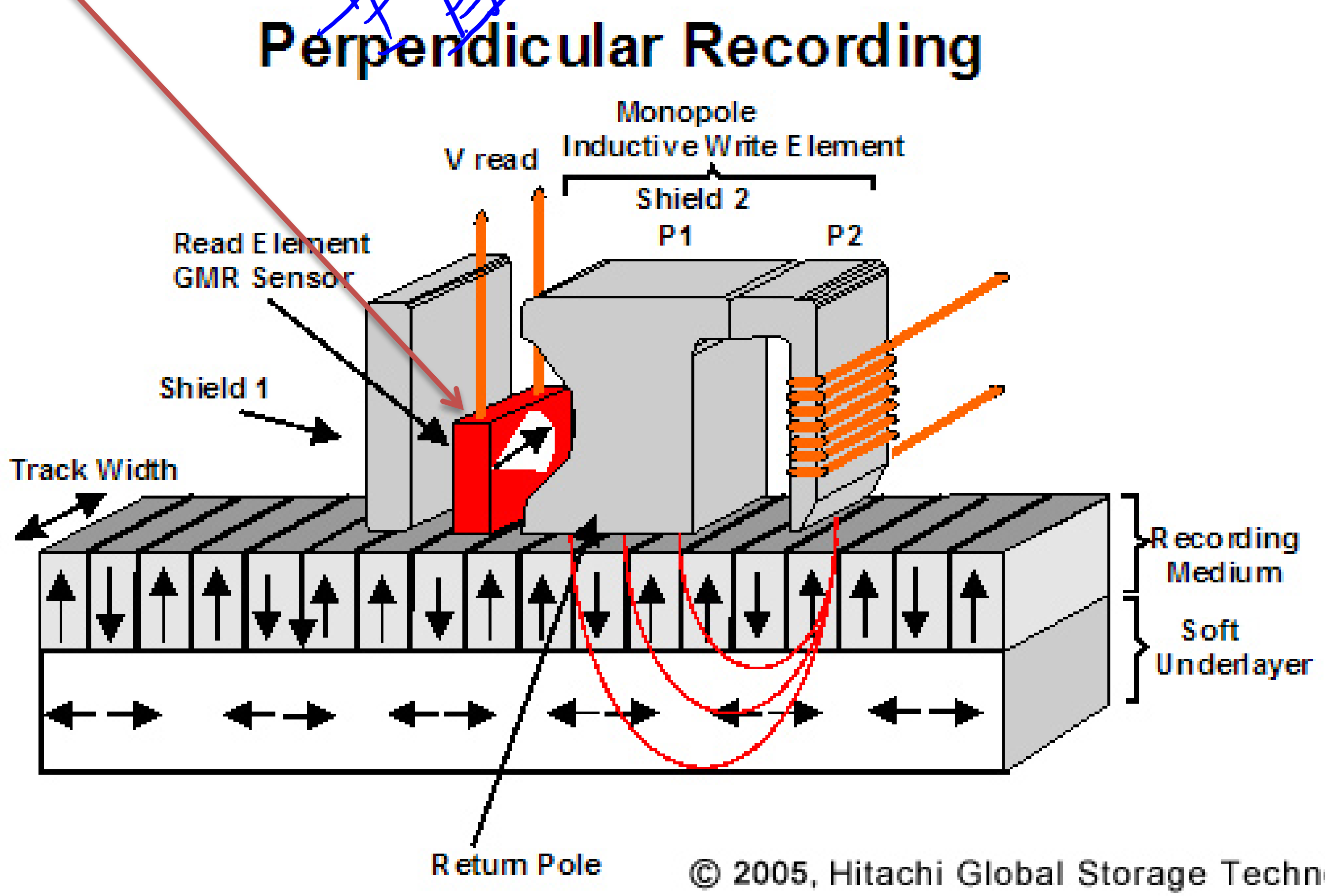
- Write head
  - Applies electrical current to write head
  - Changes direction of magnetic field under head

磁极



Reading

- Read head senses direction of magnetic field

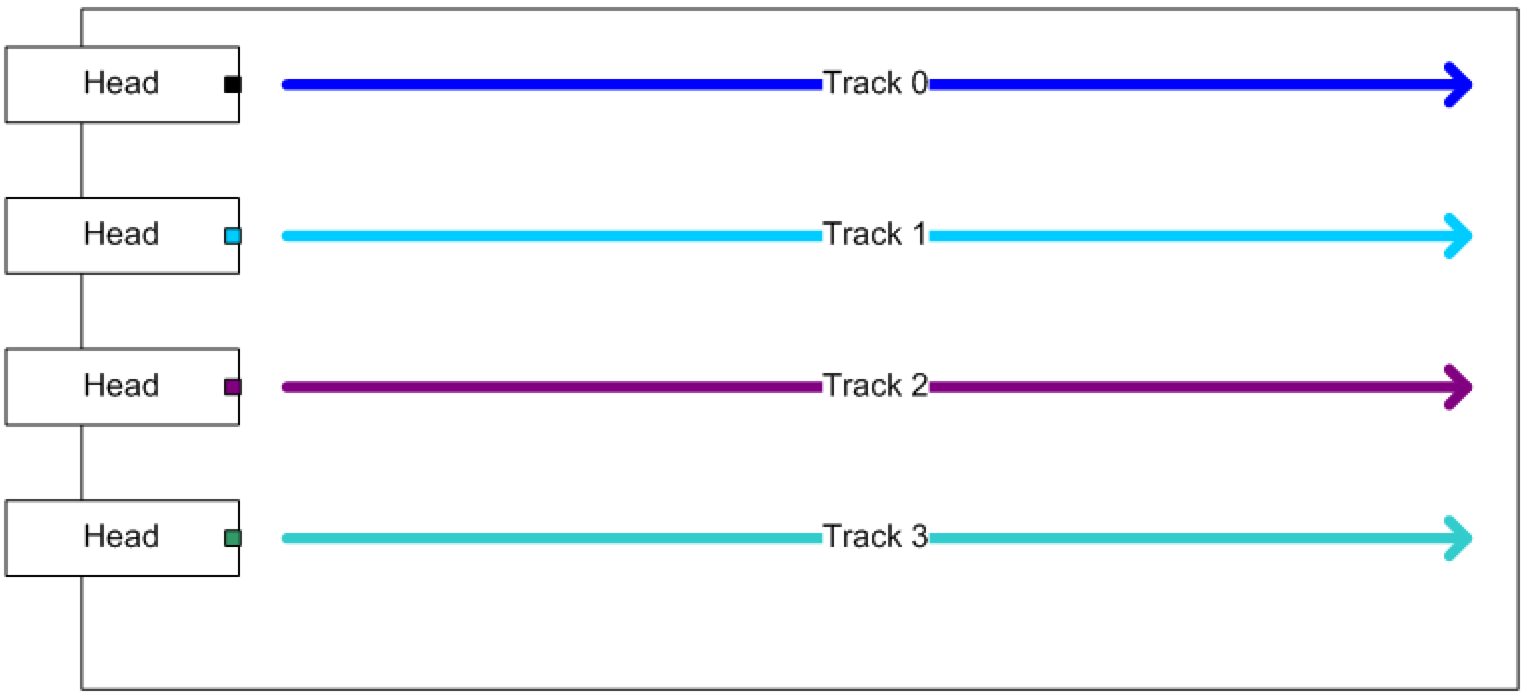


Road map

- 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 6.25 TB per tape (LTO-7)
  - Drive cost ~ $2500
  - Tape cost ~ $45 for 2.5TB tape
- Tape access time (~ minute)
  - Time to mount the tape
  - Time to wind the tape to correct position
- Data transmission rates ~ 250MB/sec

$200
300MB/s
LTO-7

Performance characteristics

- High latency/low cost makes tape most appropriate for "archival" storage
  - Low frequency of reads
  - Very large data objects

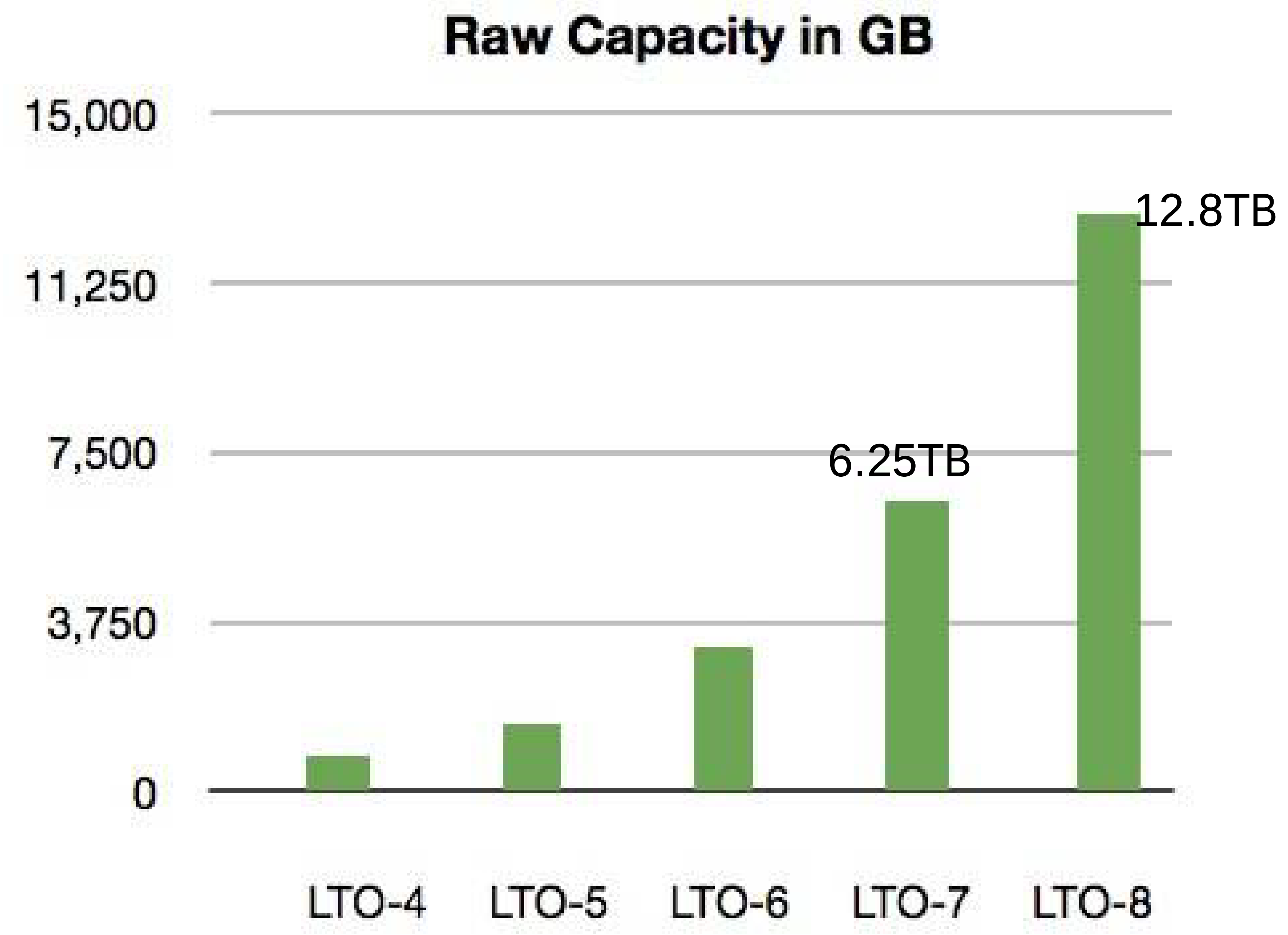
- Random access will be slow due to latency
  - Sequential reads will be fast

Linear tape file system

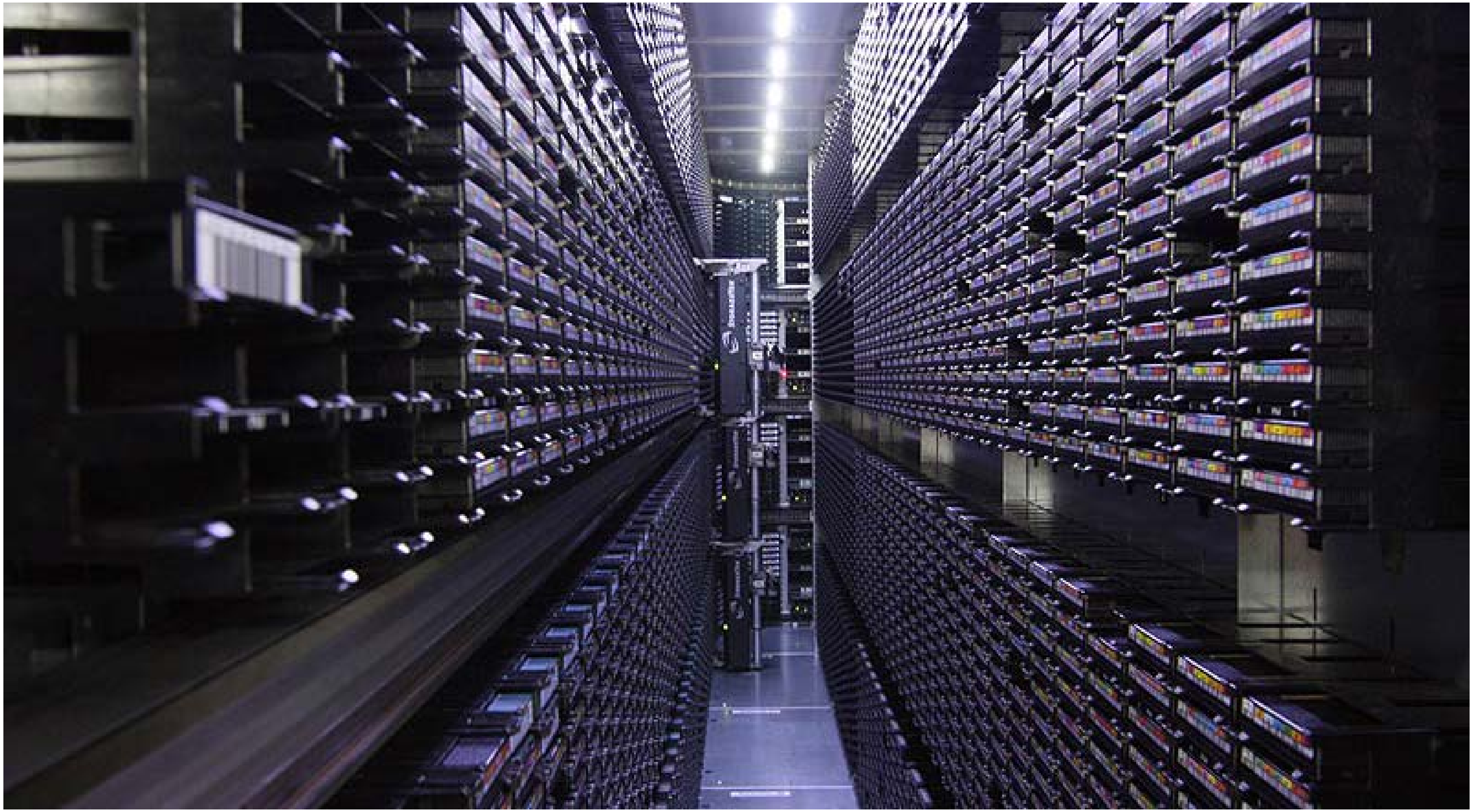
- Two partitions on tape
  - First contains metadata and directories. Tape reader can find and load this very quickly
  - Second contains blocks for data

- Directory structure coded in XML
  - Self describing file format…

Tape Cartridge 磁带盒.


Raw Capacity in GB
15,000
11,250
7,500
3,750
0
12.8TB
6.25TB
LTO-4
LTO-5
LTO-6
LTO-7
LTO-8

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

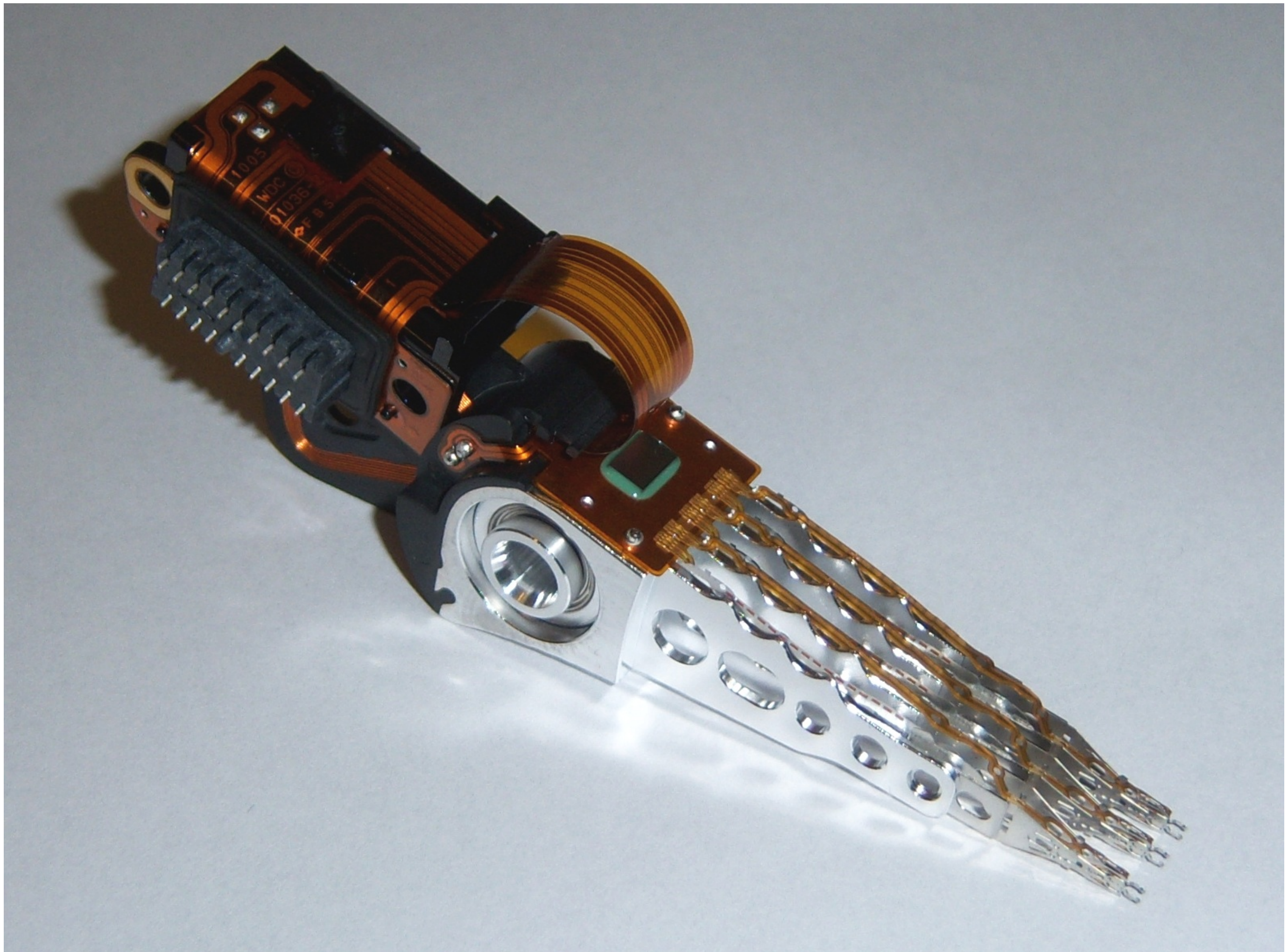
WESTERN DIGITAL WD30EZRS Caviar Green 3TB 64MB cache SATA 3.0Gb/s 3.5 internal hard drive (Bare Drive)

About $120 at Amazon

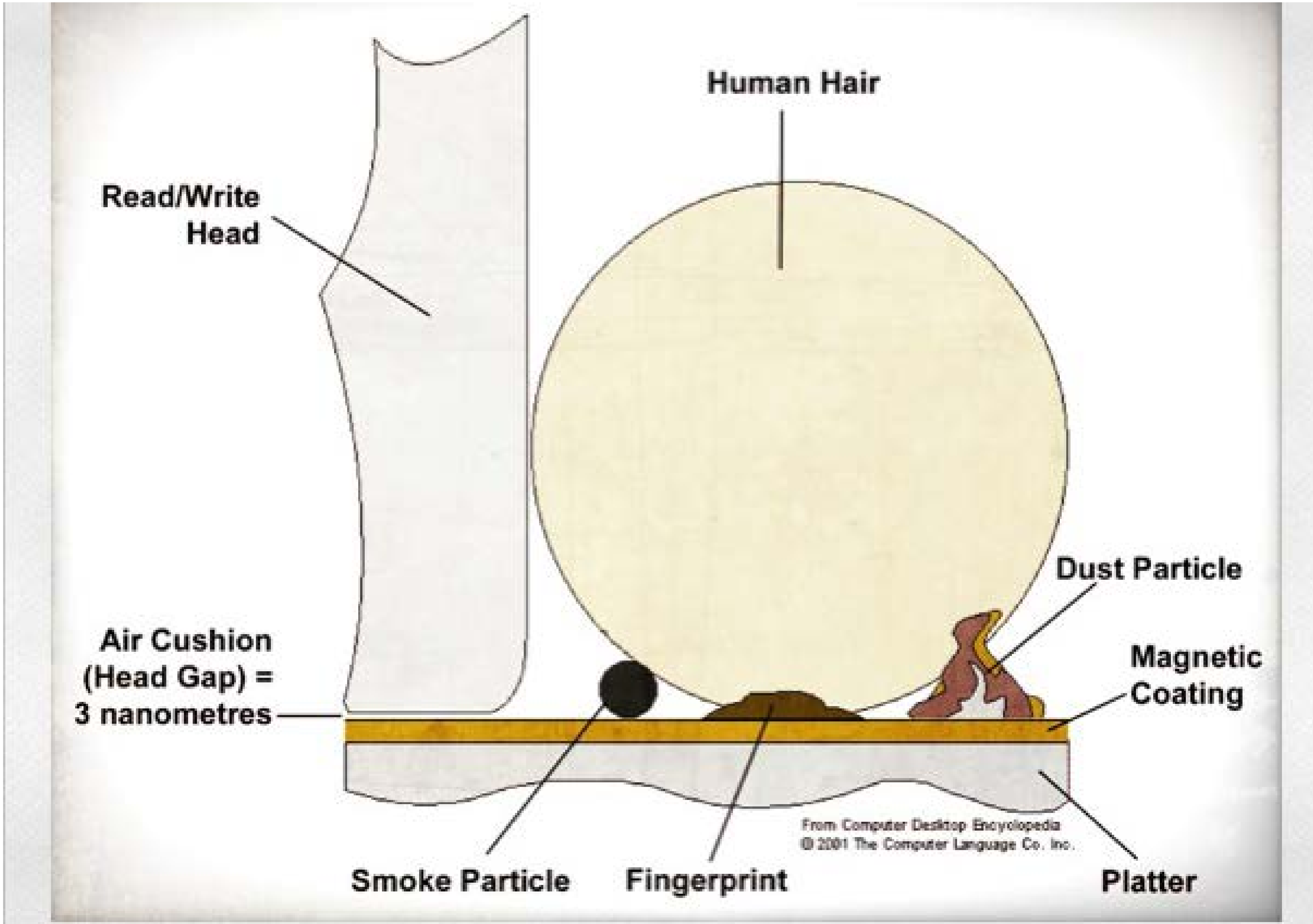
Internal of hard disk

Disk arm and platter

Disk arm & head close-up



Disk head close-ups


Human Hair
Read/Write
Head
Dust Particle
Air Cushion
(Head Gap) =
3 nanometres
Magnetic
Coating
From Computer Desktop Encyclopedia
© 2001 The Computer Language Co. Inc.
Smoke Particle
Fingerprint
Platter

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=BlB49F6ExkQ

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

Physical characteristics

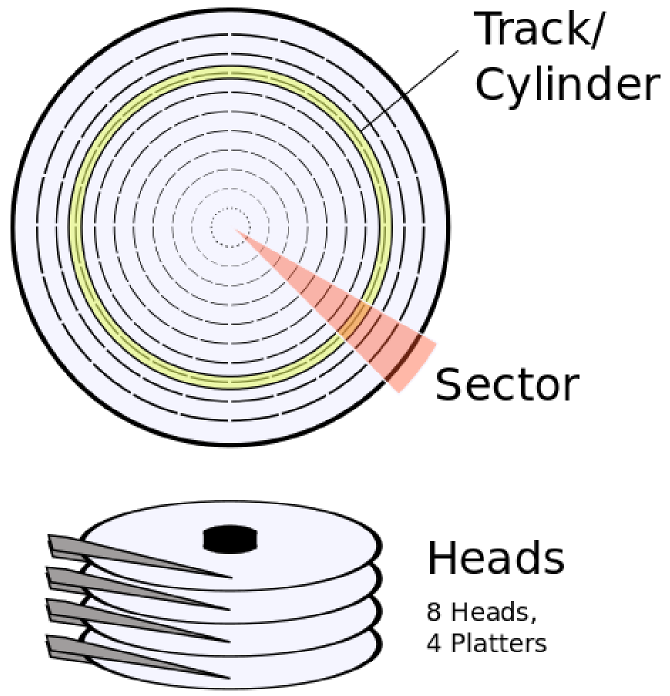
- 3.5" (diameter, common in desktops)
- 2.5" (common in laptops)
- Rotational speed
  - 4,800 RPM
  - 5,400 RPM
  - 7,200 RPM
  - 10,000 RPM
- Between 5-7 platters
- Current capacity up to 10TB (Western Digital)

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)
  – Sectors can be numbered from 0 to N-1
  – Entire sector is written "atomically"
    - All or nothing

CHS (cylinder-head-sector)

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



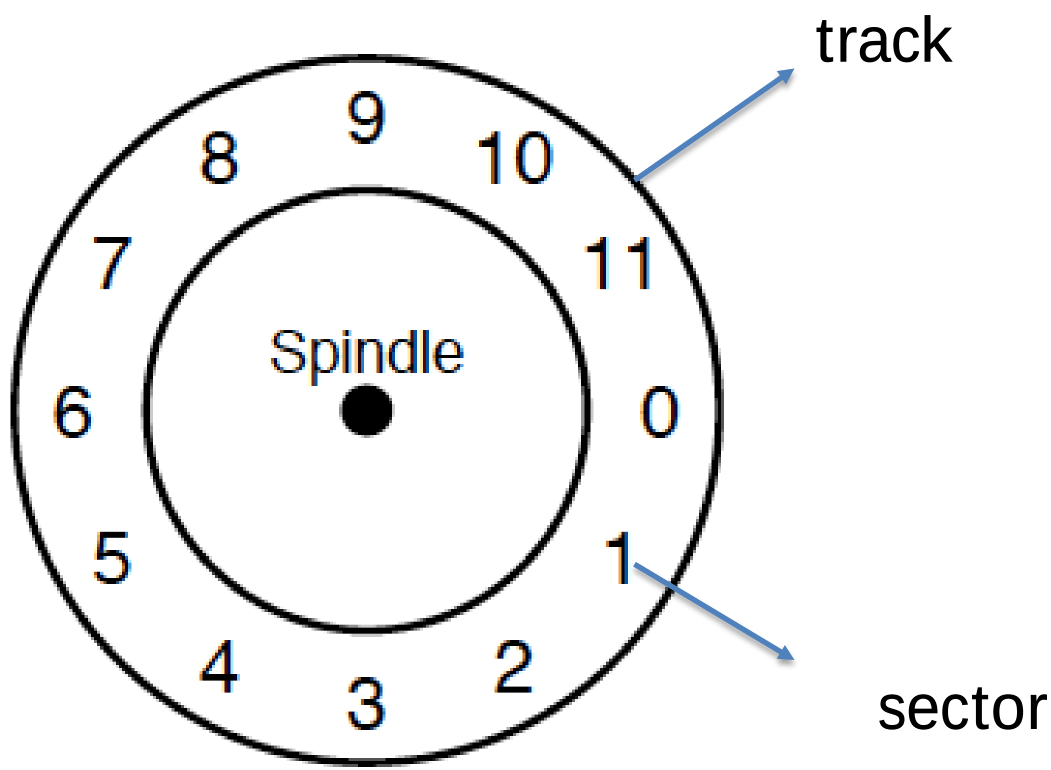
https://en.wikipedia.org/wiki/Cylinder-head-sector

Example

- # cylinders: 256
- # heads: 16 (i.e., 8 platters, 2 heads/platter)
- # sectors/track: 64
- Sector size = 4KB

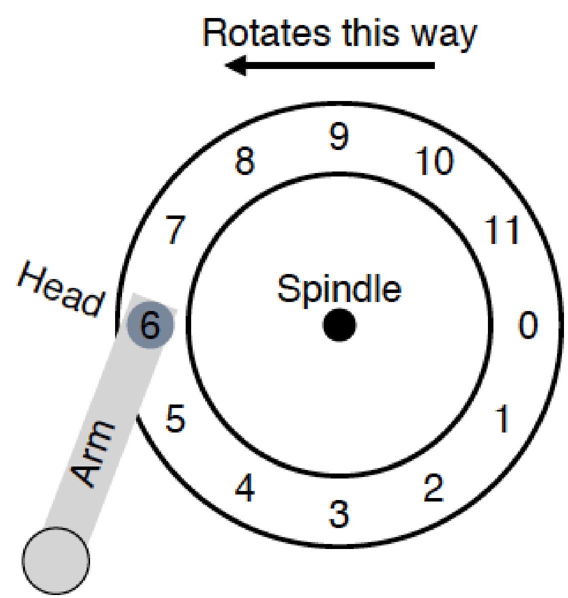
What is the capacity of the drive?

A simple disk drive (one track only)



Rotational latency

- Waiting for the right sector to rotate under the head
  - On average: about ½ of time of a full rotation
  - Worst case?
  - Best case?

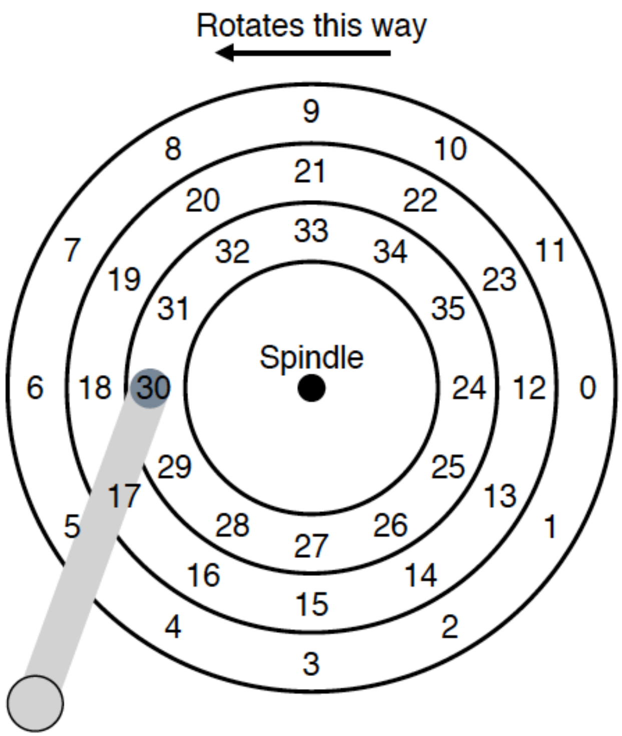


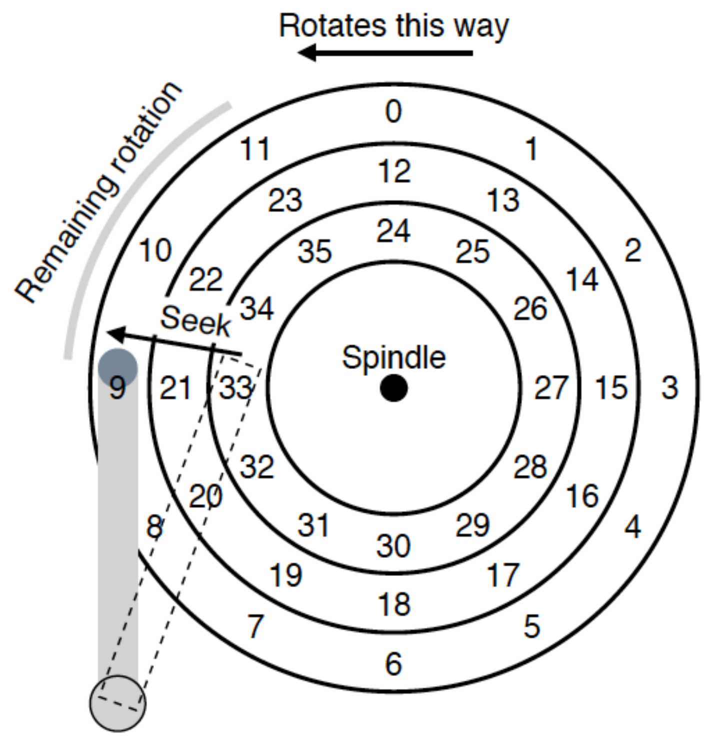
Rotation time

- Assume 10,000 RPM (rotations per minute)

$$\frac{60000\,ms}{1000\,rotations} = \frac{6\,ms}{rotation}$$

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:

  512KB/128MB * 1000ms = 4ms

Completion time

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

Example

- Capacity 4TB
- # platters: 4
- # heads: 8
- Bytes per sector: 4096

- Transmission bandwidth: 100MB/sec
- Maximum seek time: 12ms

- RPM: 10,000

Time to transfer a file

- The file occupies 100 sectors (sequentially)
- Avg. seek time =?
- Avg. rotational latency =?
- Transfer time = ?

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
  - And rotate to the first sector

- But no rotation/seeking needed afterward

Actual bandwidth

- 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 = 210ms
  - Actual bandwidth = 10MB/210ms = 47.62 MB/s
- Random access of 10 MB (2,500 blocks)
  - Completion time = 2500 * (7 + 3 + 4/50) = 25.2s
  - Actual bandwidth = 10MB / 25.2s = .397 MB/s

Road map

- Tapes
- Hard disk
- Solid state drive

Solid State Drive

Memory chips

Solid State Drives

- All electronic, made from flash memory
- Lower energy consumption than hard drive
- Significantly more expensive, less capacity
  - About a factor of 5-10 more expensive
- Limited lifetime, can only write a limited number of times.
  - E.g., 100K program/write cycles for SLC (single-level cell) memory

SSD vs. Hard Drive (price)

MLC (multi-level cell)

Samsung 850 EVO 1TB 2.5-Inch SATA III Internal SSD (MZ-75E1T0B/AM)
by Samsung

$309.99 Prime
Get it by **Thursday, Jan 19**

More Buying Choices
$309.99 new (80 offers)
$273.37 used (6 offers)

2,566

- Hard Disk Size: **1 TB**
- Hardware Interface: **sata 3 0 gb**

Seagate 1TB Barracuda SATA 6GB/s 128MB Cache 2.5-Inch 7mm Internal Bare/OEM Hard Drive (ST1000LM048)
by Seagate

$59.22 ~~$59.99~~ Prime | FREE Same-Day
Get it by **TODAY, Jan 17**

More Buying Choices
$58.24 new (27 offers)

526

Product Features
The thinnest and highest-capacity *2.5-Inch Hard Drive* available

Electronics: See all 7,236 items

Solid State Drives

- Same form-factor and control interface as magnetic disks
- Significantly better latency
  - No seek or rotational delay
- Consistent bandwidth for sequential & random:
  - Benefits from improved latency
  - However, writes take significantly longer then reads

Writing to SSD is complicated

- Can not overwrite a page
  - Need to erase its block (at a certain point) instead

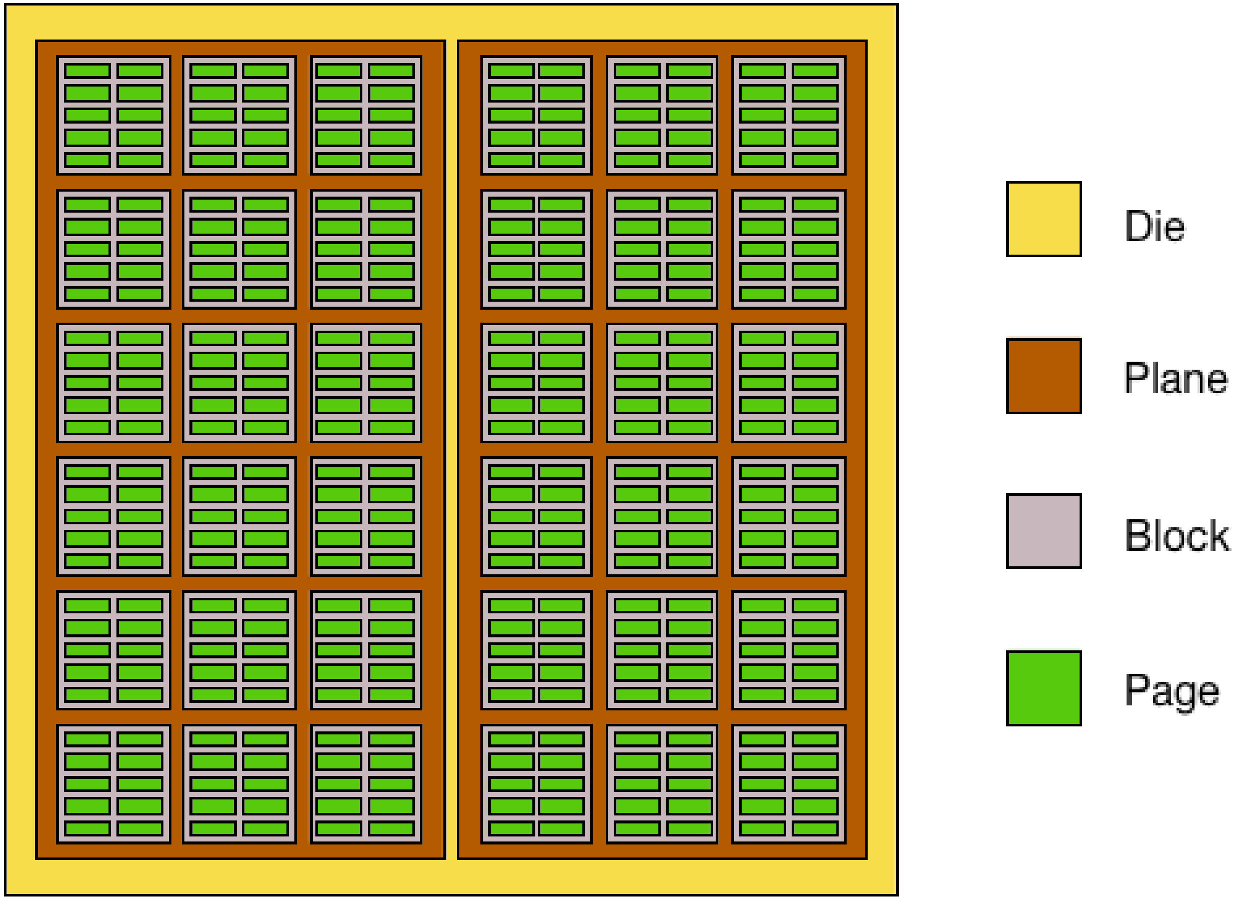
- SSD controllers take care of all these details

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



[Understanding Flash: Blocks, Pages and Program / Erases]

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

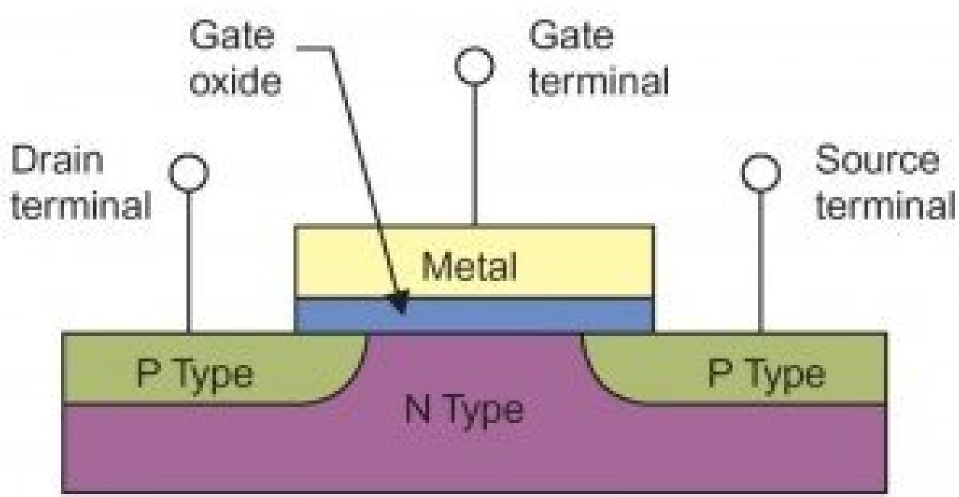
- Common page sizes: 2K, 4K, 8K, and 16K

- A block typically has 128 to 256 pages

=> Block size: 256KB to 4MB

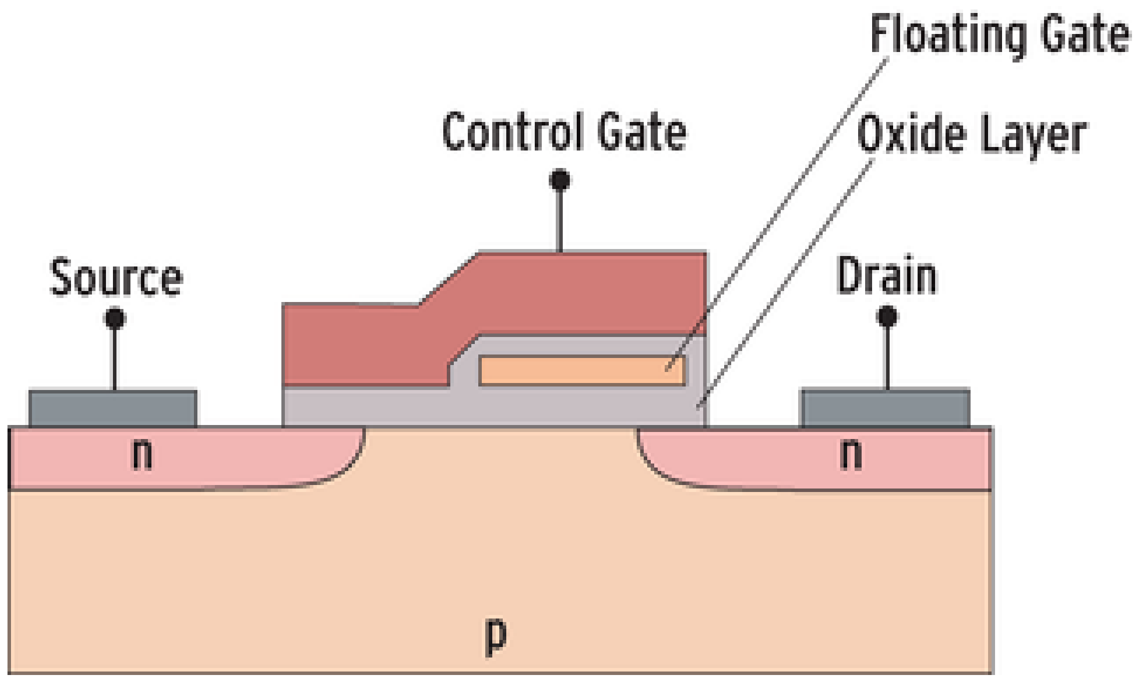
Normal transistor (MOSFET)

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



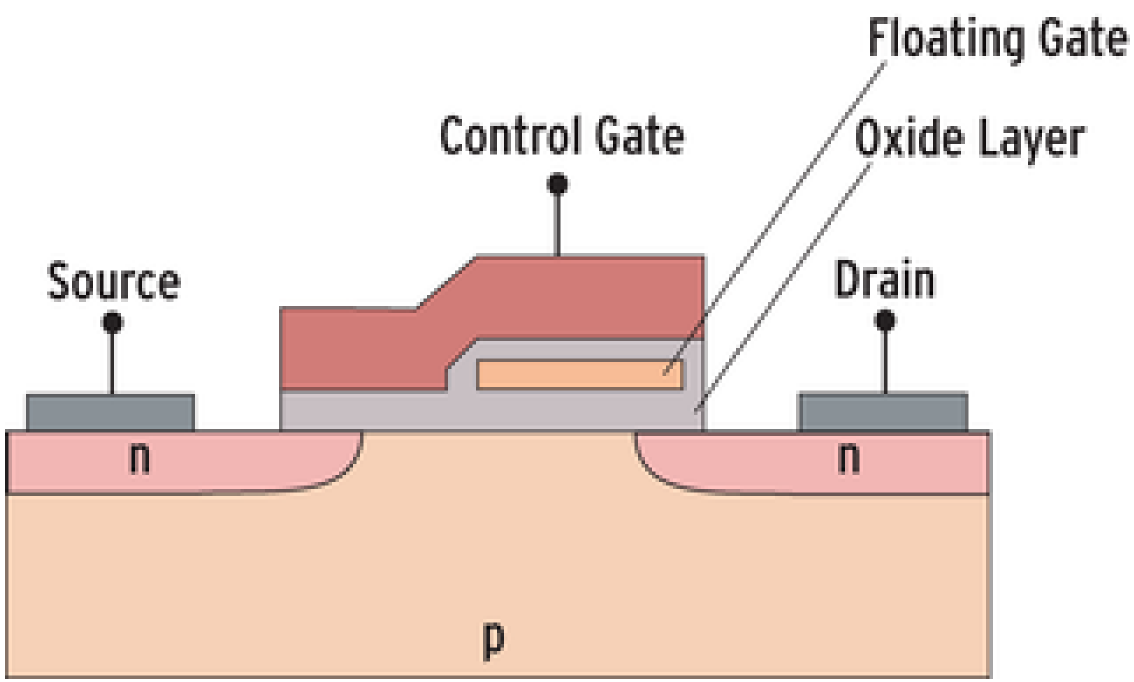
Floating gate transistor

- 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



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

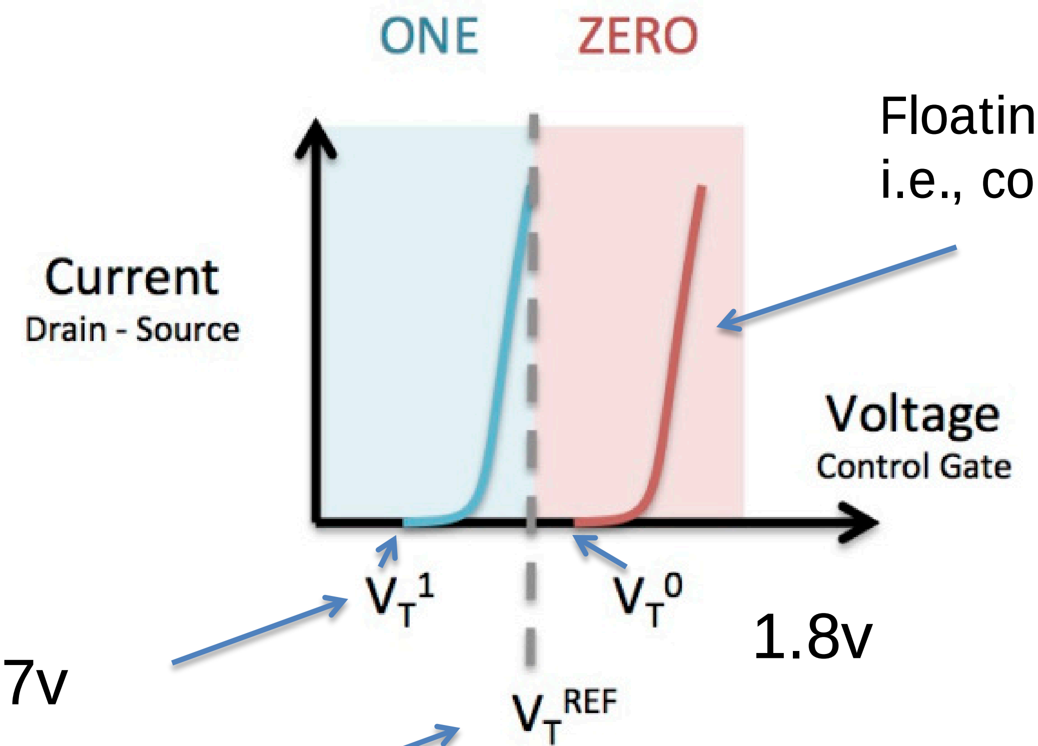
Read operations

- 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

Floating gate is charged
i.e., containing electrons



1.8v

Typically, .7v

1.5v

See here for more details

Read operations

- Apply V_{int} (intermediate voltage)

- If the current is detected, gate has no electrons

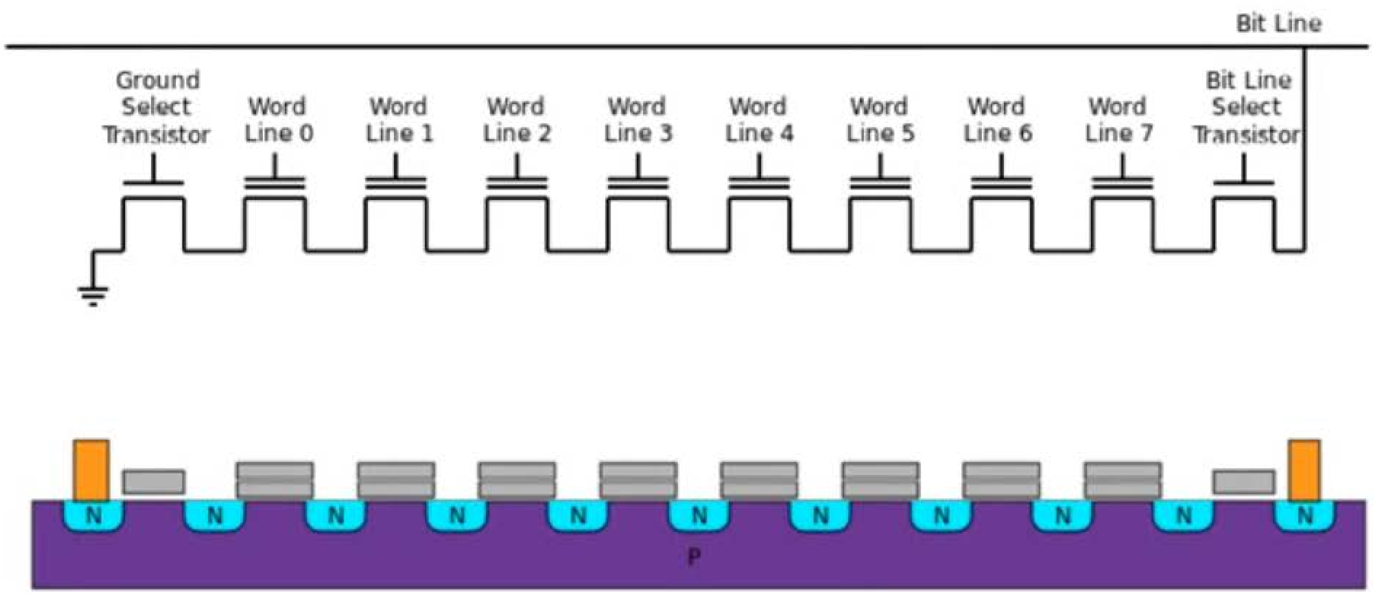
  => bit = 1

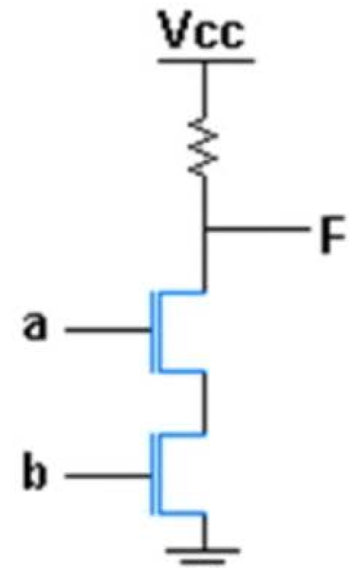
- If no current, gate must have electrons

  => bit = 0

NAND flash layout

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

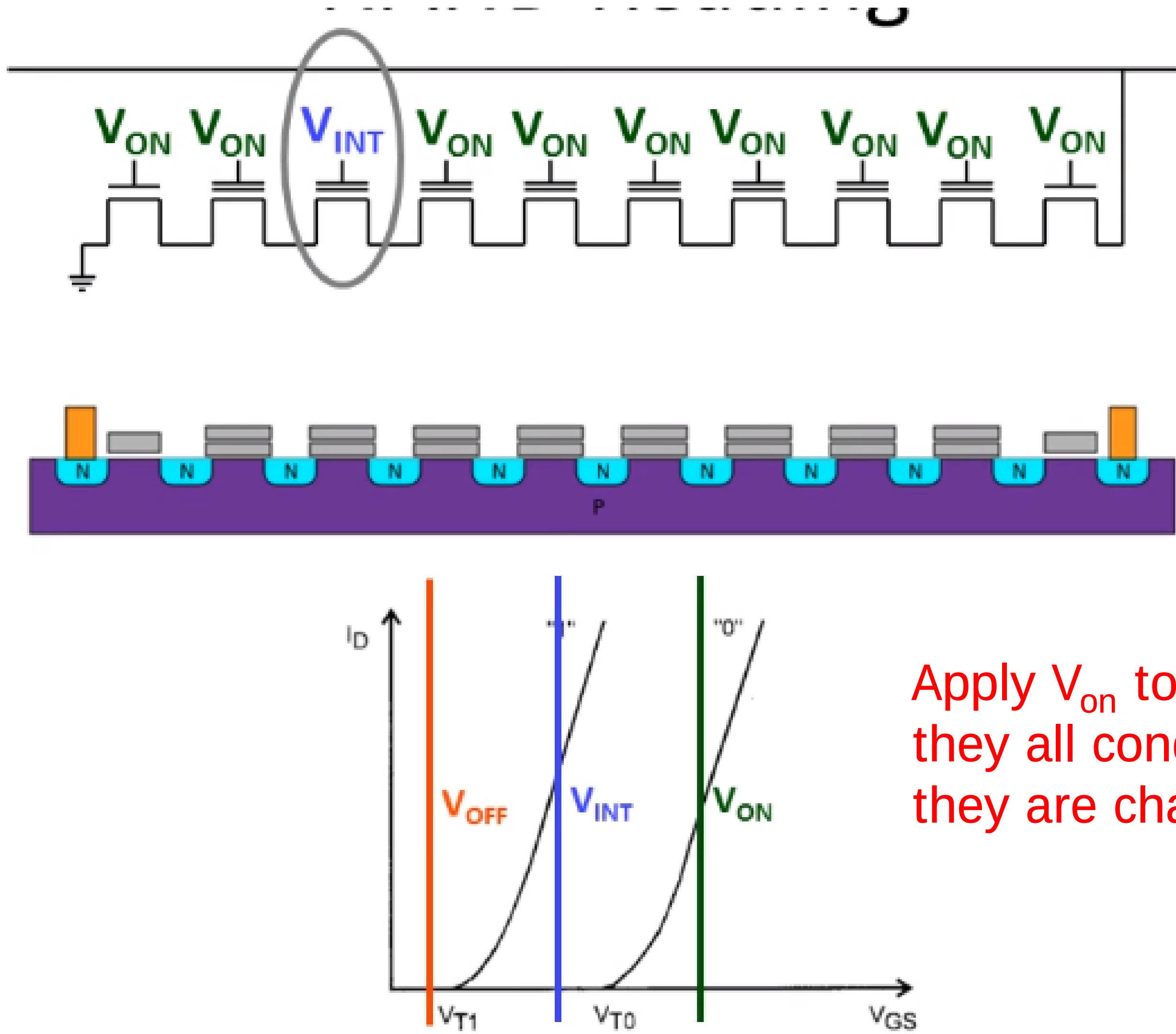




NMOS NAND gate

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

NAND reading

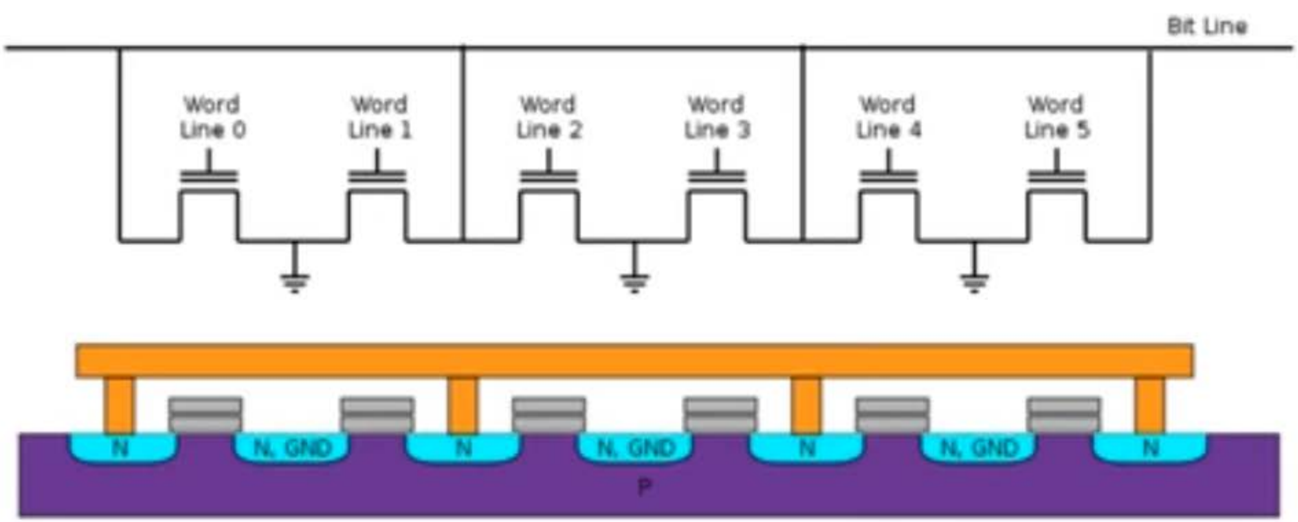


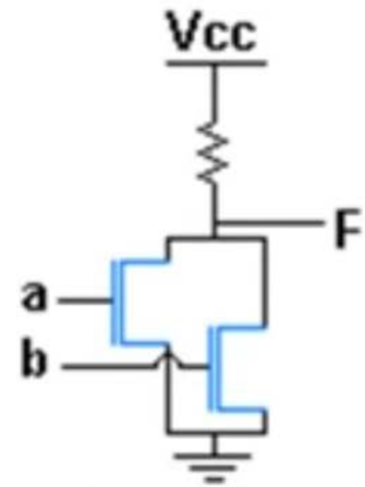
Bit line

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

NOR flash layout

- Each transistor is directly connected to bit line (also ground)
  - Similar to transistors in a NOR gate (to output F)

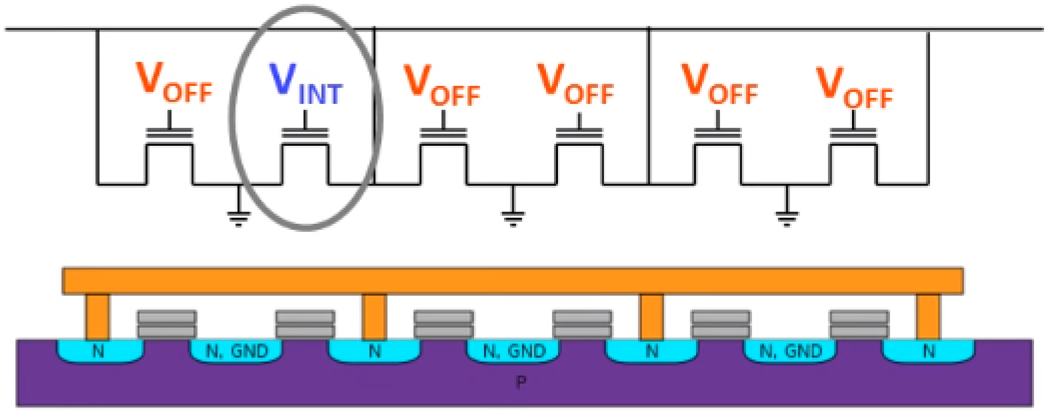


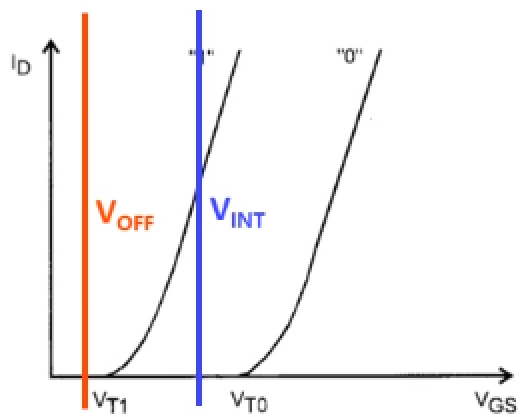


NMOS NOR gate

F will be grounded as long as one of a & b conducts

NOR reading



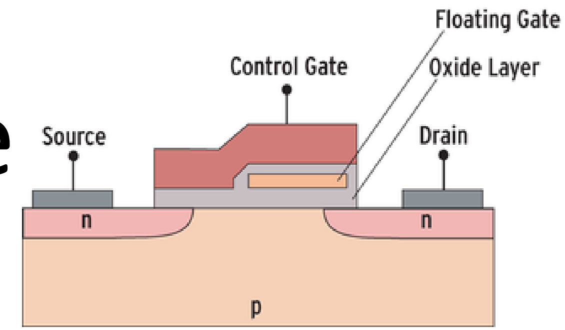


Apply V_{OFF} to all others so that they do NOT conduct, charged or not.

NAND vs NOR

- NOR has individual bit lines and ground connections
  - ➔ circuit more complex
- NAND does not have individual bit lines, nor separate ground connections
  - ➔ save space, larger capacity

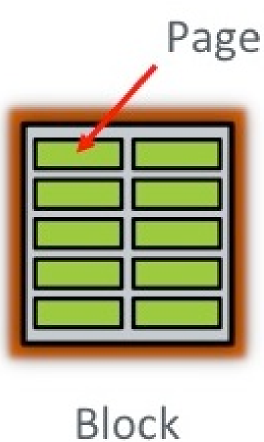
Write and erase



- Write: 1 => 0
  - Apply high POSITVE voltage (>> voltage for read) to the control gate
  - Attract electrons from channel to floating gate (through quantum tunneling)
- Erase: 0 => 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

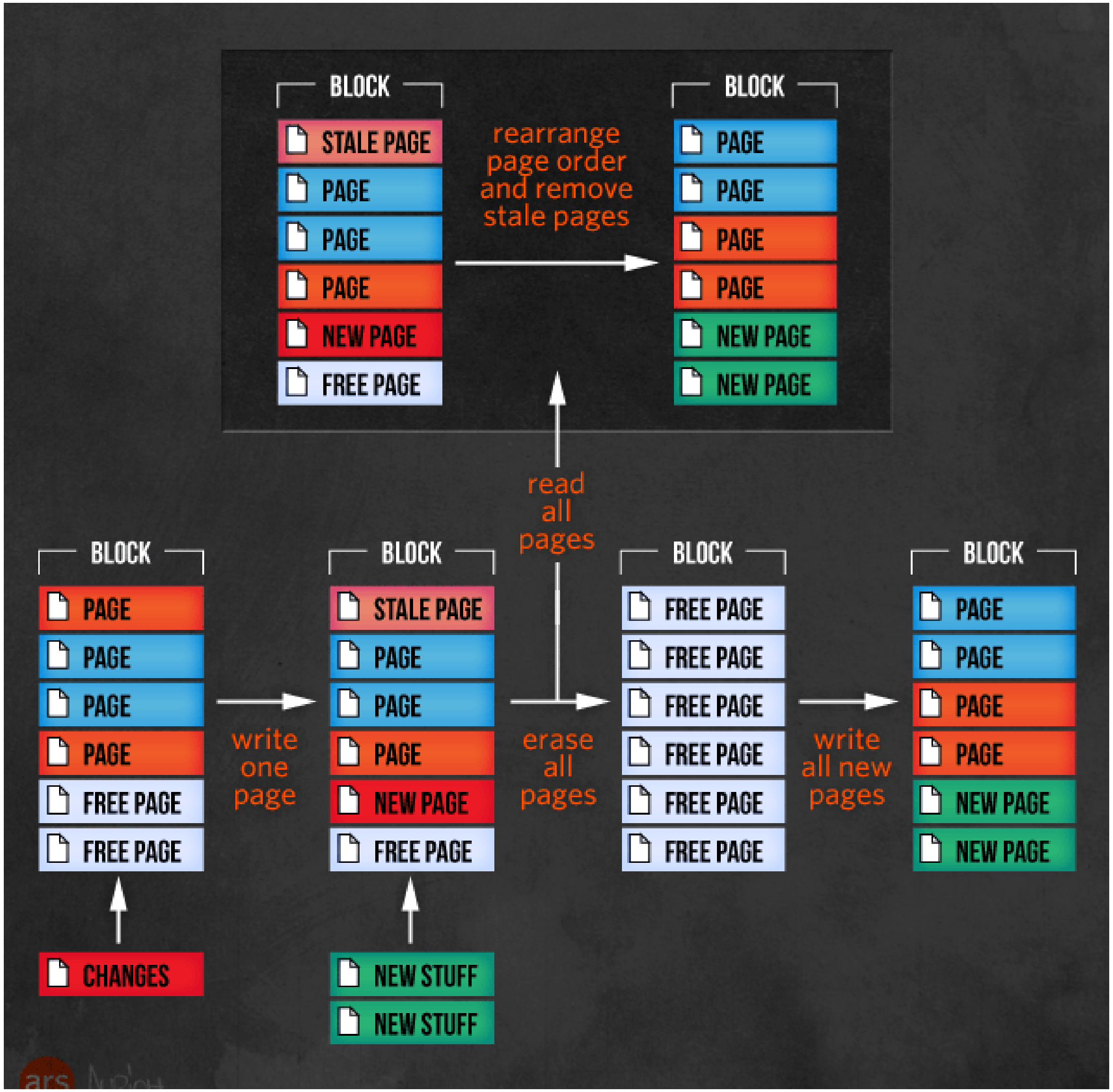
Read/write units

- 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	Block

Example



BLOCK
STALE PAGE
PAGE
PAGE
PAGE
NEW PAGE
FREE PAGE
rearrange page order and remove stale pages
BLOCK
PAGE
PAGE
PAGE
PAGE
NEW PAGE
NEW PAGE
read all pages
BLOCK
PAGE
PAGE
PAGE
PAGE
FREE PAGE
FREE PAGE
CHANGES
write one page
BLOCK
STALE PAGE
PAGE
PAGE
PAGE
NEW PAGE
FREE PAGE
NEW STUFF
NEW STUFF
erase all pages
BLOCK
FREE PAGE
FREE PAGE
FREE PAGE
FREE PAGE
FREE PAGE
FREE PAGE
write all new pages
BLOCK
PAGE
PAGE
PAGE
PAGE
NEW PAGE
NEW PAGE

Latencies: read, write, and erase

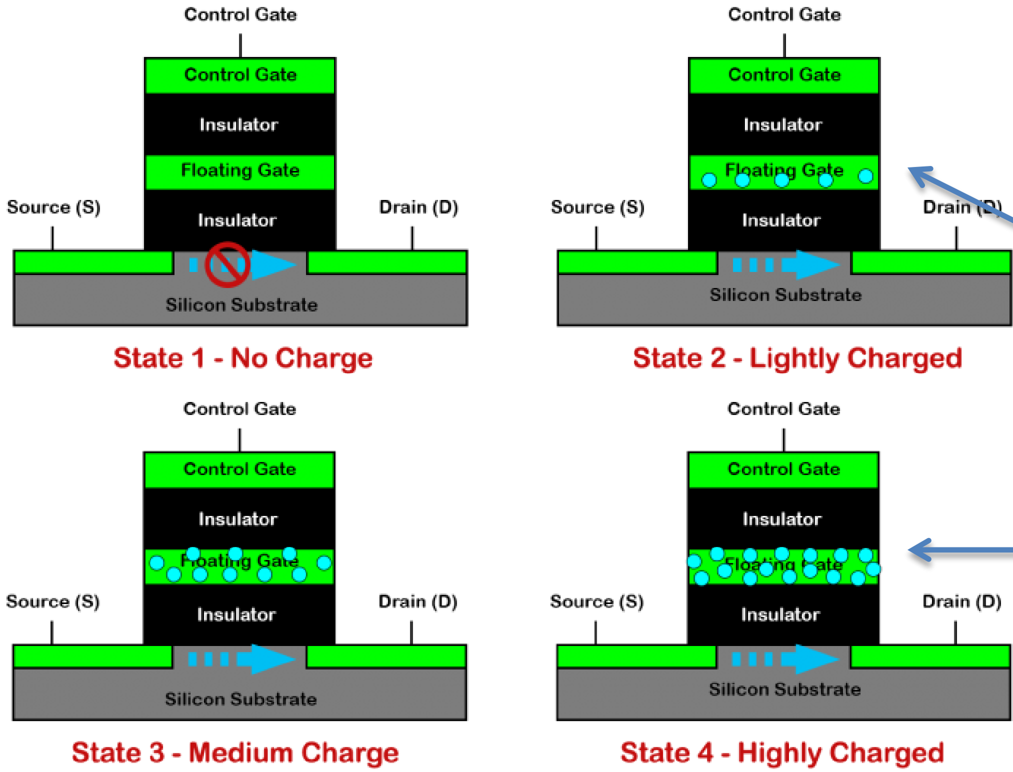
	SLC	MLC	TLC	HDD	RAM
P/E cycles	100k	10k	5k	*	*
Bits per cell	1	2	3	*	*
Seek latency (μs)	*	*	*	9000	*
Read latency (μs)	25	50	100	2000-7000	0.04-0.1
Write latency (μs)	250	900	1500	2000-7000	0.04-0.1
Erase latency (μs)	1500	3000	5000	*	*
Notes	* metric is not applicable for that type of memory				
Sources	P/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/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 -> 0
  - Then erased (E): 0 -> 1

Multi-level cell (MLC)

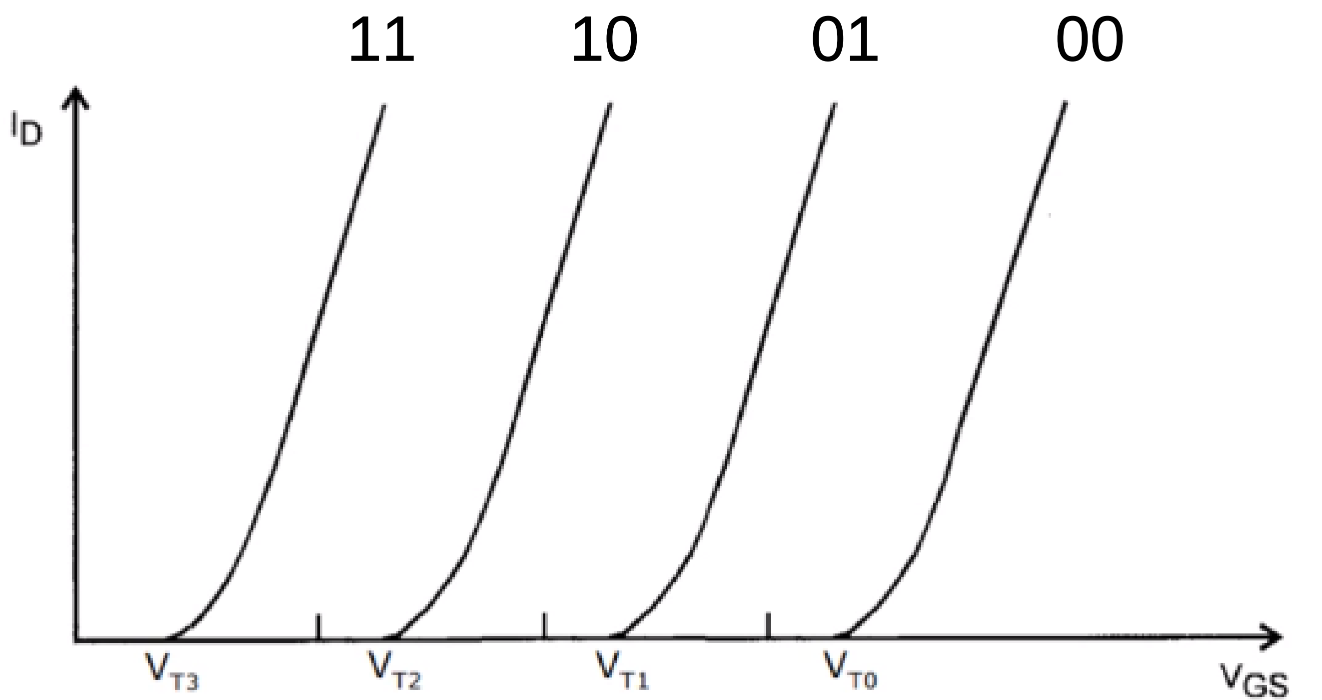


Note different levels of electrons

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

MLC reading

- 2 bits, 3 intermediate voltages



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=msi5GDz9JIw
- 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

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=IcrBqCFLHIY&feature=youtu.be

References

- How Flash Memory Works
  - https://www.youtube.com/watch?v=s7JLXs5es7I