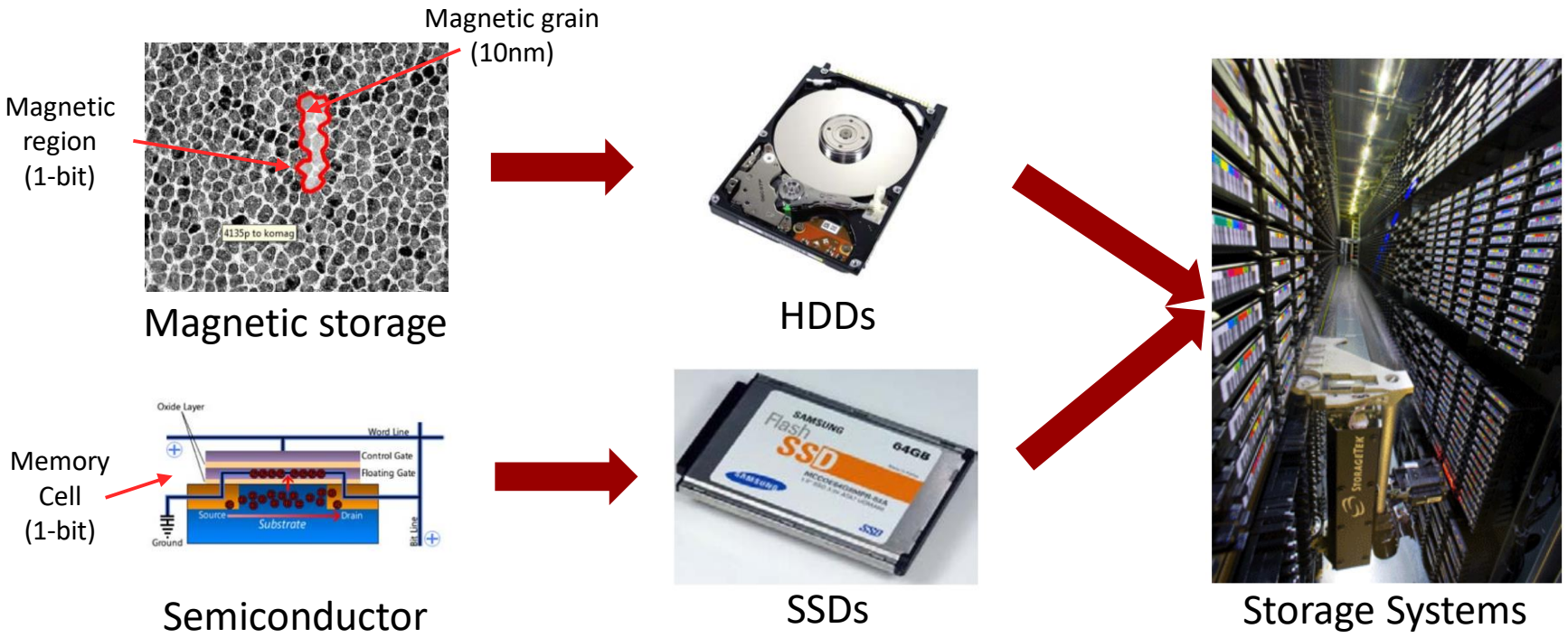
2. Storage Hardware

Special Topics in Computer Systems:
Modern Storage Systems
(IC820-01)

Instructor:

Prof. Sungjin Lee (sungjin.lee@dgist.ac.kr)

Building Storage Systems from Bits



- **Software to aggregate many devices for performance**
- **Software to handle device failures**
  - Erasure codes on blocks, across devices
- **Software to handle software failures**
  - Server failover, write-ahead logging

Type of Storage Devices

High performance

Technology		Latency	Capacity
SRAM/DRAM	L1 CPU Cache	4 cycles (~1 nsec)	32K I, 32K D
	L2 CPU Cache	10 cycles (3 nsec)	256K
	LLC CPU Cache	40 cycles (13 nsec)	30 MB
	DRAM	240 cycles (80 nsec)	32 GB
NVRAM (e.g., PRAM)	Optane DIMM	1200 cycles (400 nsec)	128 GB
	Optane PCIe	30K cycles (10 usec)	1.5 TB
NAND Flash (e.g., SSDs)	SSD Read	150K cycles (50 usec)	32 TB
	SSD Write	1500K cycles (500 usec)	32 TB
Magnetic (e.g., HDDs)	HDD Write min	1500K cycles (500 usec)*	8 TB
	HDD Read min	15000K cycles (5 msec)	8 TB
	HDD Read max	75000K cycles (25 msec)	8 TB

Low cost

Outline

- **Hard Disk Drives**
- Flash and Solid-state Drives
- Storage Class Memory

Hard Disk Drives (HDDs)

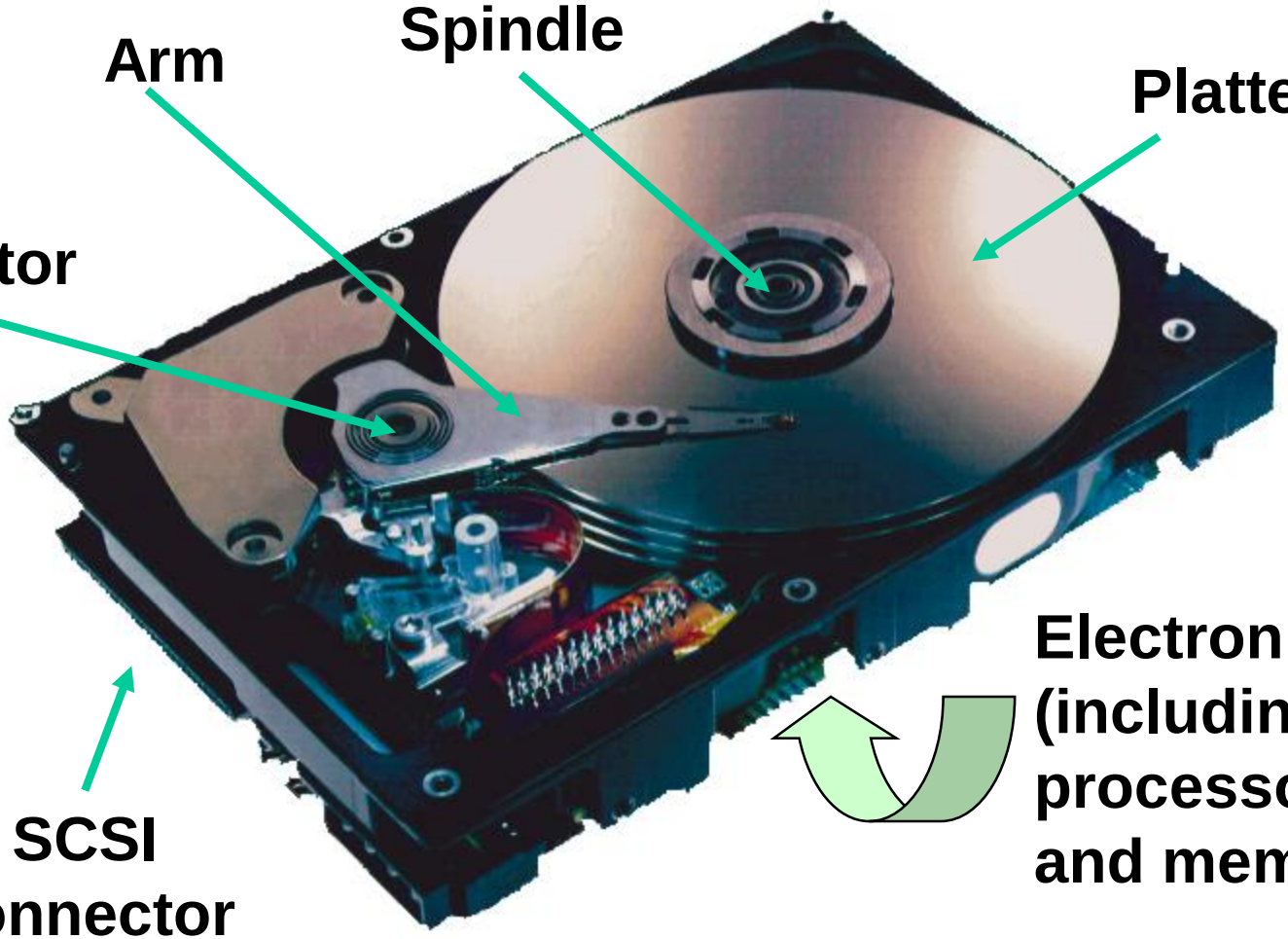
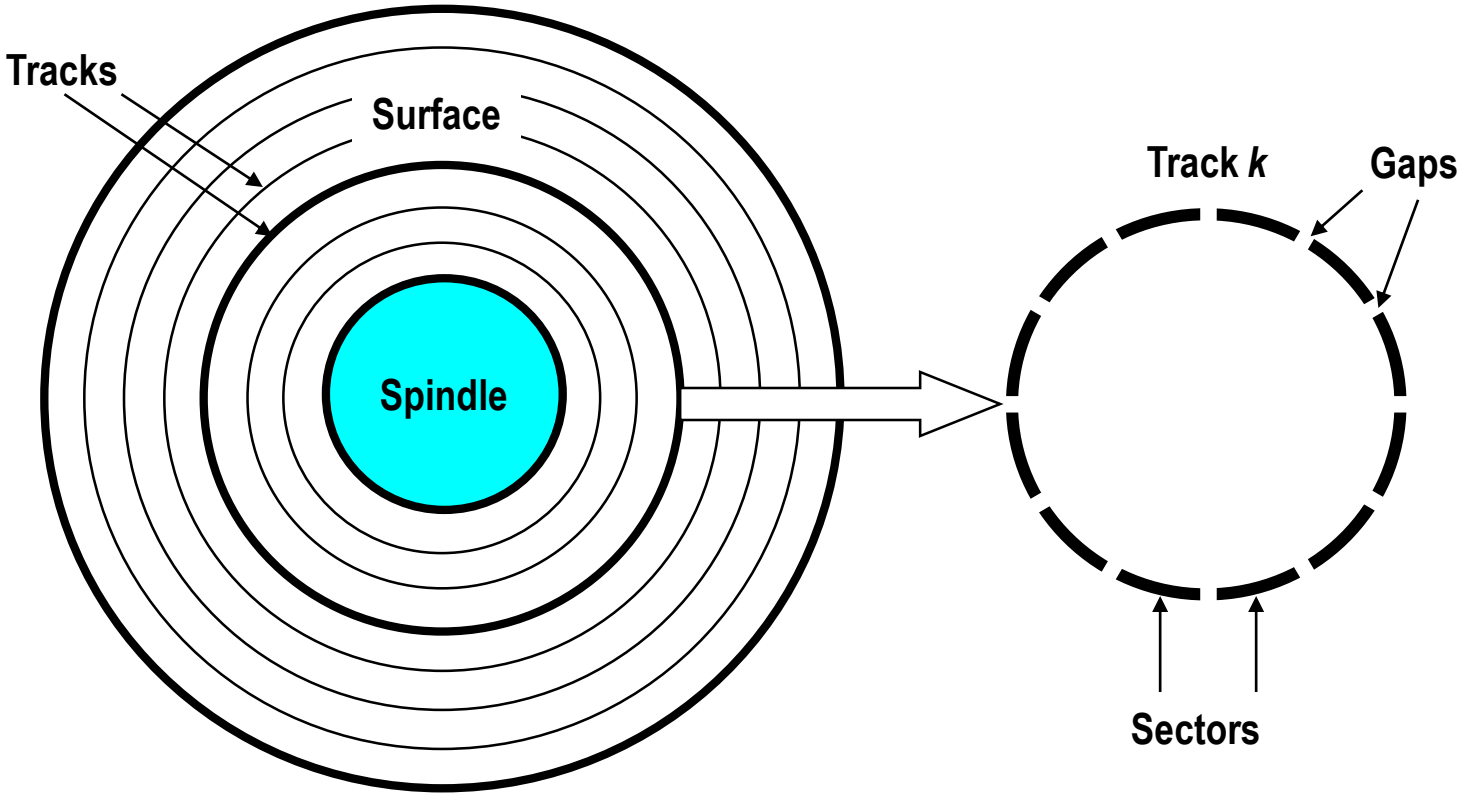


Image courtesy of Seagate Technology

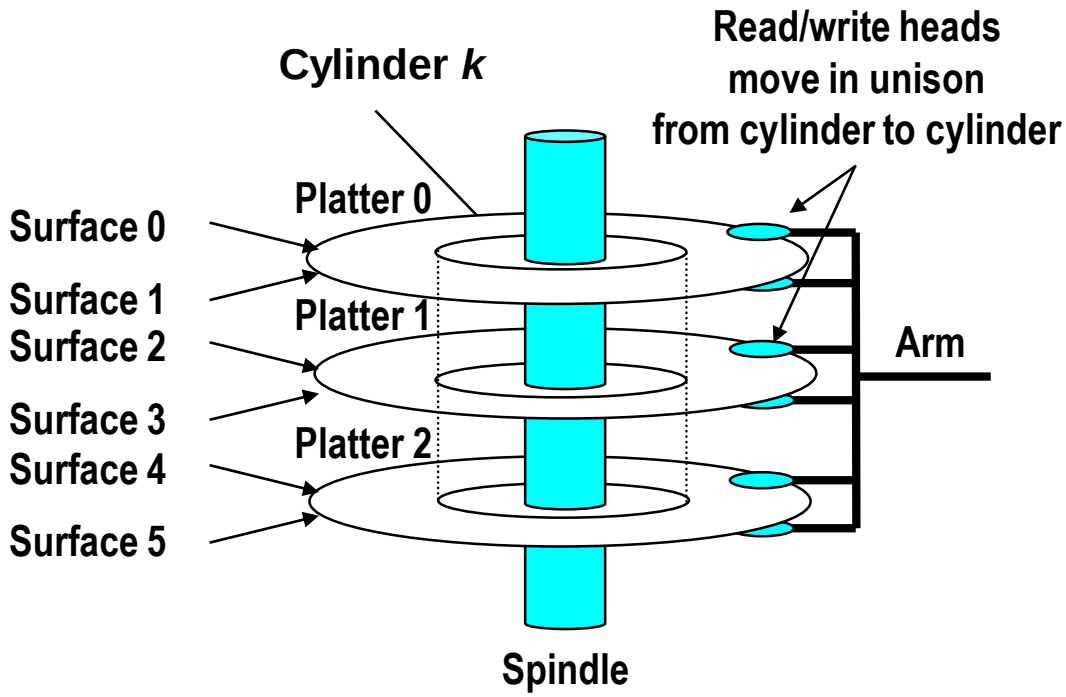
Disk Geometry

- Disks consist of platters, each with two surfaces
- Each surface consists of concentric rings called tracks
- Each track consists of sectors separated by gaps



Disk Geometry (Muliple-Platter View)

- Aligned tracks form a cylinder



Computing Disk Capacity

- **Capacity: maximum number of bits that can be stored**
  - Vendors express capacity in units of gigabytes (GB), where 1 GB = 10^9 Bytes

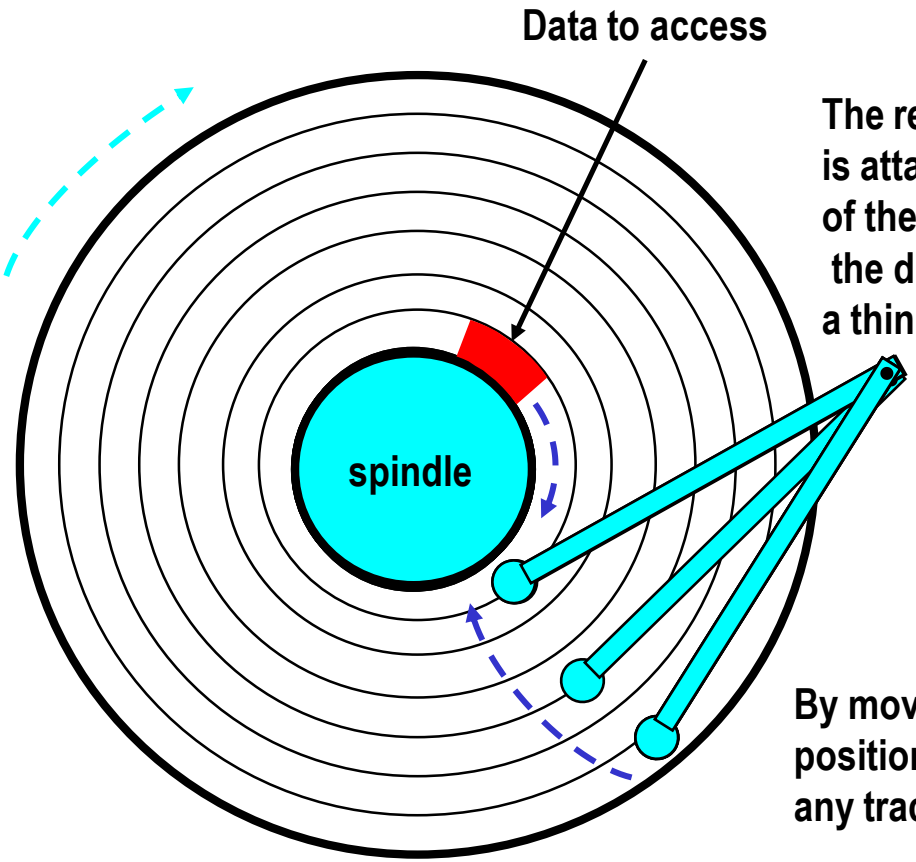
Capacity = (# bytes/sector) x (avg. # sectors/track) x (# tracks/surface) x (# surfaces/platter) x (# platters/disk)

Example:

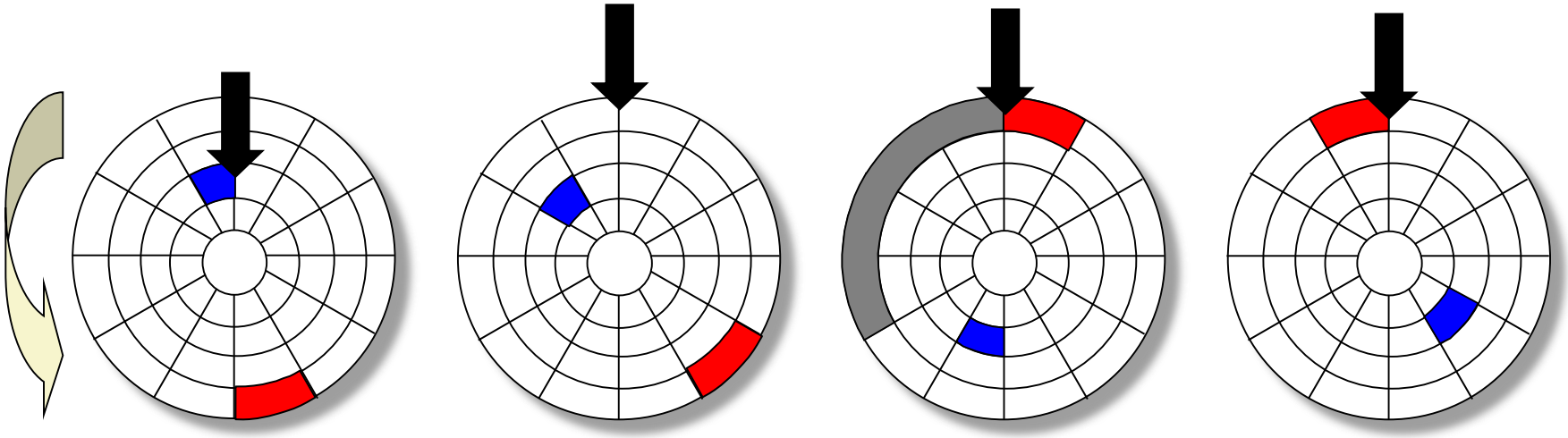
- 512 bytes/sector
- 300 sectors/track (on average)
- 20,000 tracks/surface
- 2 surfaces/platter
- 5 platters/disk

Capacity = 512 x 300 x 20000 x 2 x 5 = 30,720,000,000 = 30.72 GB

Disk Operation (Single-Platter View)



Disk Access – Service Time Components



Disk Access Time

- **Average time to access some target sector approximated by :**
  - Taccess = Tavg seek + Tavg rotation + Tavg transfer
- **Seek time (Tavg seek)**
  - Time to position heads over cylinder containing target sector.
  - Typical Tavg seek is 3—9 ms
- **Rotational latency (Tavg rotation)**
  - Time waiting for first bit of target sector to pass under r/w head.
  - Tavg rotation = 1/2 x 1/RPMs x 60 sec/1 min
  - Typical Tavg rotation = 7200 RPMs
- **Transfer time (Tavg transfer)**
  - Time to read the bits in the target sector.
  - Tavg transfer = 1/RPMs x 1/(avg # sectors/track) x 60 secs/1 min.

Capacity vs Bandwidth

- **Per drive capacity increases by 50% to 100% per year**

2008	2009	2010	2012	2014	2015	2016	2018	2019
500 GB	1 TB	2 TB	4 TB	6 TB	8 TB	10 TB	12 TB	14 TB

8x 3x

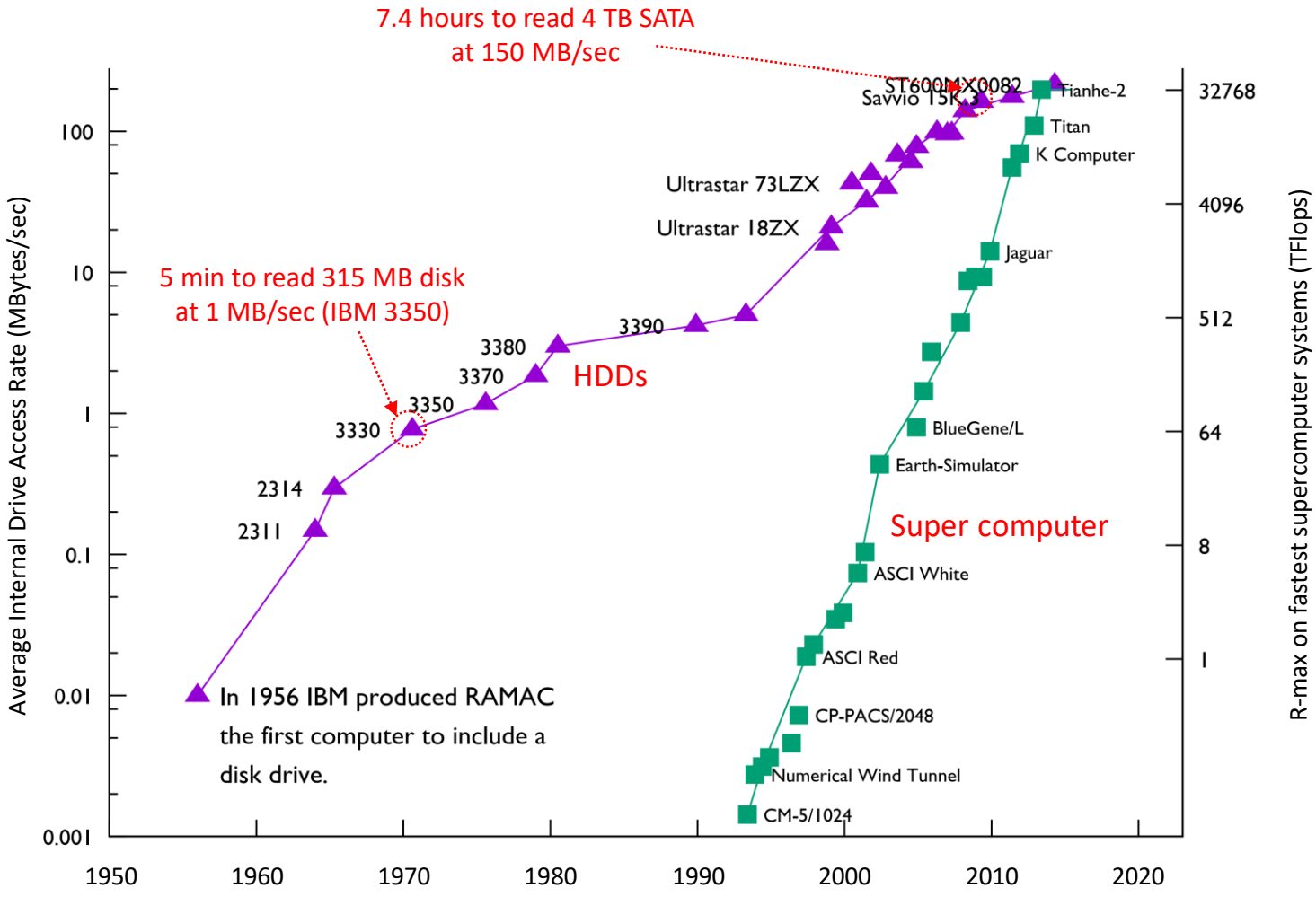
- **Drive interface speed increases by 15-20% per year**

2008	2009	2010	2012
98 MB/s (WD RE2)	113 MB/s (WD RE3)	138 MB/s (WD RE4)	150 MB/s (WD SAS)

1.5x

- **It is getting harder and harder to improve disk performance**

Disk Transfer Rate over Time



- **What does it mean?**

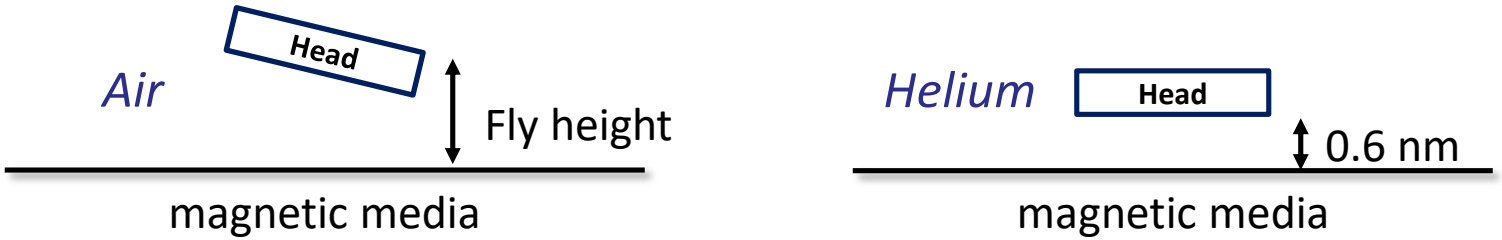
HDD becomes the entires bottleneck!
(사실 지금은 CPU가 bottleneck)

HDD Technologies

- **Hard drive manufactures have roadmaps out to 60TB/disk over the next decade**
  - Approximately. Might take longer. Might be bigger
- **However, advances in drive interface speeds will continue to lag advances in drive capacity**
  - As usual, every new generation of drive capacity will mean it takes longer to read or write the entire device
- **New Technologies**
  - Helium filled drives
  - Heat Assisted Magnetic Recording (HAMR)
  - Bit Patterned Media (BPM)
  - Shingled Magnetic Recording (SMR)

Helium Filled Drives

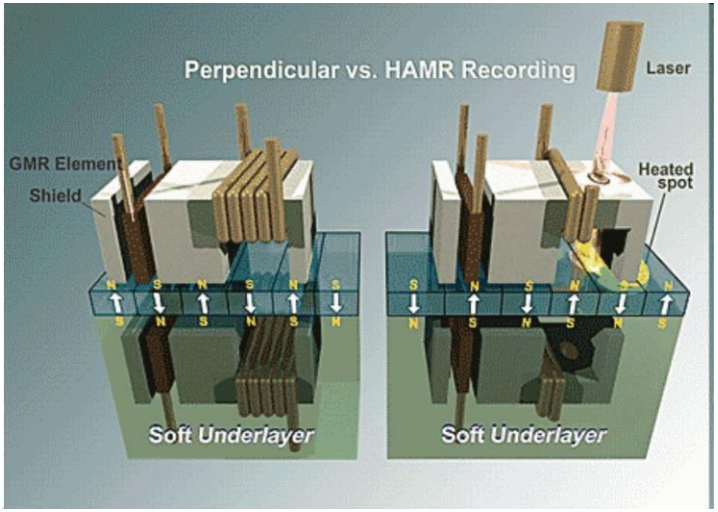
- **The *lightness* of helium gas helps to increase the medium's potential speed and storage density**
- **Helium filled drives have "reduced drag" and "less vibration"**
  - Fly the head closer so the bits can be packed more densely

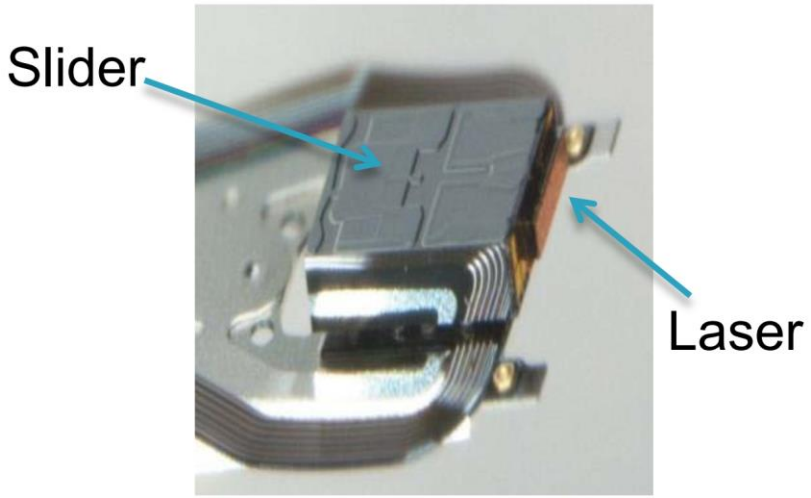


- **Increase number of platters > 5 to increase per-device capacity**
  - 2013: 5 platters, 10 heads, 800 GB/platter yields a 4 TB device
  - 2015: 7 platters, 14 heads, 1200 GB/platter yields an 8 TB device
  - 2019: 9 platters, 18 heads, 1500 GB/platter yields a 14 TB device
- **Keeping the drive sealed is a major challenge**
  - Helium likes to "creep" uphill and through narrow openings

Heat Assisted Magnetic Recording (HAMR)

- **Temporarily heating the disk material during writing**
  - Makes it much more receptive to magnetic effects
  - Allows writing to much smaller regions (i.e., smaller # of magnetic grains)
- **Seagate demonstrated 16 TB drive in December 2018**
  - Western Digital predicts 100 TB in 2032
  - May need patterned media and helium filled drives to reduce drag

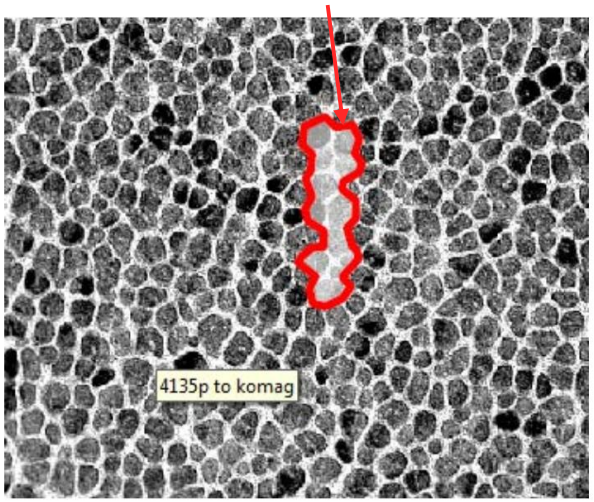




Bit Patterned Media

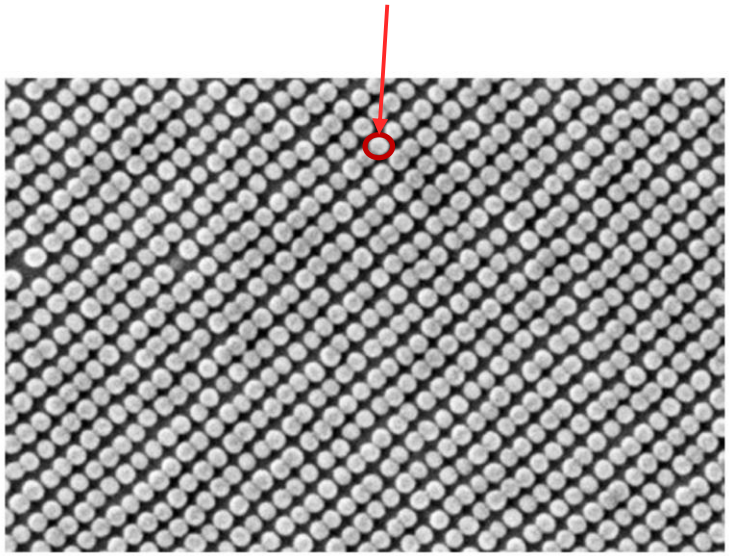
- **Record data in magnetic islands (one bit per island)**
  - It is opposed to current hard disk drive technology, where each bit is stored in 20-30 magnetic grains

Magnetic region (1-bit)
= 20-30 grains



Metal oxide granules

Island

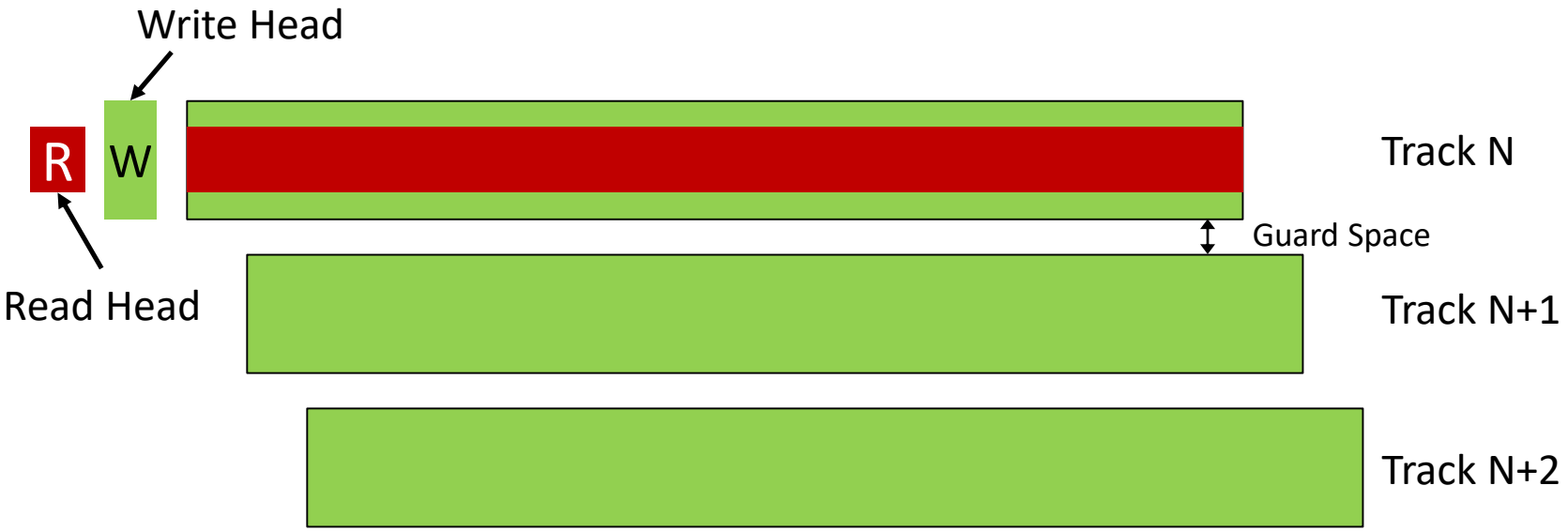


Bit-patterned media

- **HGST (now WD) expects it to be cost effective way to double bit density by the end of this decade**

Shingled Magnetic Recording (SMR)

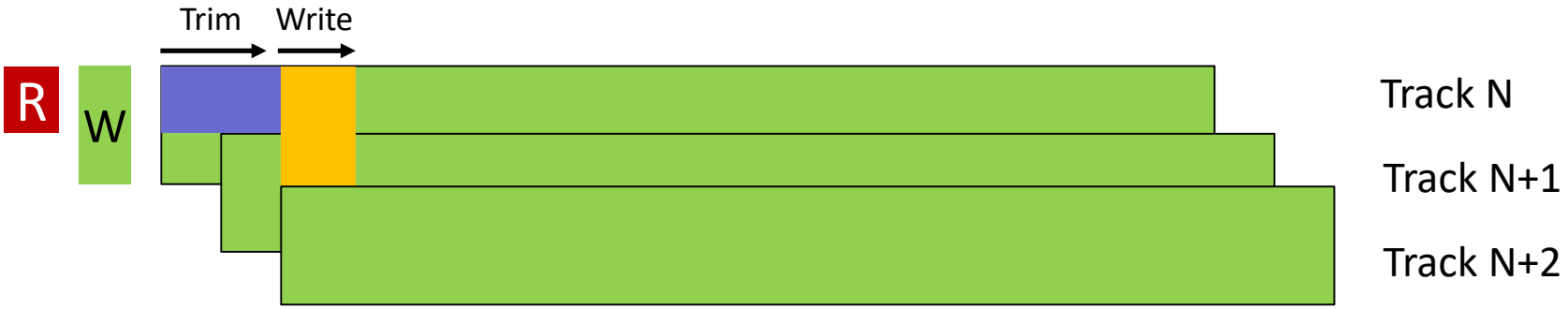
- **Conventional magnetic recording**
  - HDDs have two heads for writing and for readings, respectively
  - The write head is larger than the read head
    - To avoid interference, cross-talk, or misinterpretation when data is read later
  - The track size is thus decided by the width of the write head



Shingled Magnetic Recording (SMR) (Cont.)

- **Overlapping recording tracks are like shingles on your roof**
  - The track width is the same as the read head size

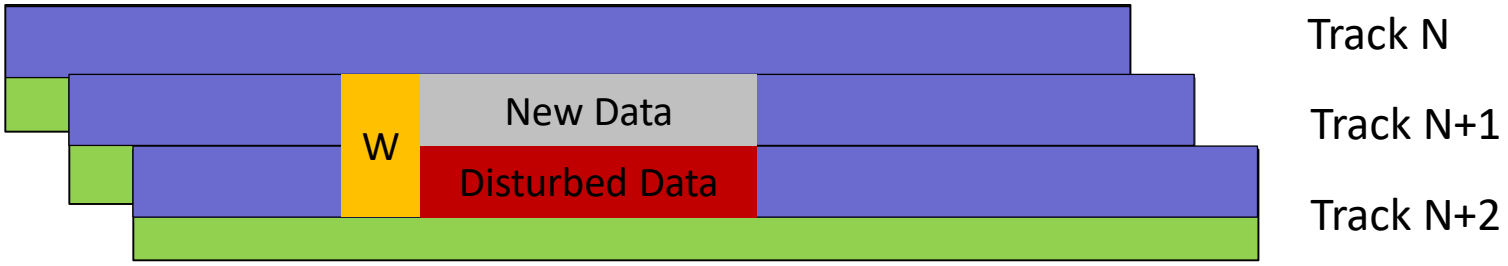
- **How SMR operates?**
  - The write head "writes" the data to the full track
    - That is, both Track N and N+1
  - The drive "trims" the data behind the write head to the width of the read head
  - The read head is able to retrieve the data as usual



- **SMR enables us to put more tracks in the same disk area**

Shingled Magnetic Recording (SMR) (Cont.)

- **Technical challenges in SMR**
  - Advanced signal processing is required (e.g., no guard space)
  - ***Append-only property***
    - Optimized for sequential writes
    - Overwrites writes are prohibited or expensive (redirection)



  - Intelligent software supports are required
    - e.g., Address remapping and LSM-Tree algorithm

Summary

- We have seen how HDDs operate and reviewed emerging techniques, such as SMR, for future HDDs
- Since it is hard to improve disk performance, almost all the techniques attempt to increase capacity
- This means that the gap between CPU and HDD will be even wider!

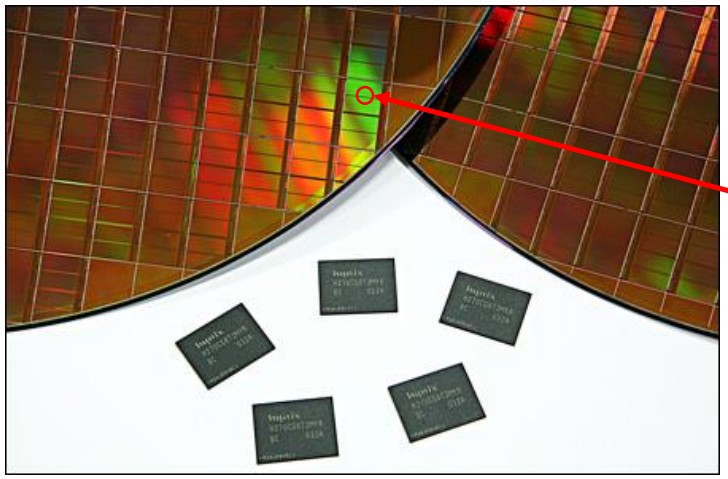
- How can we address such a technical issue?

Outline

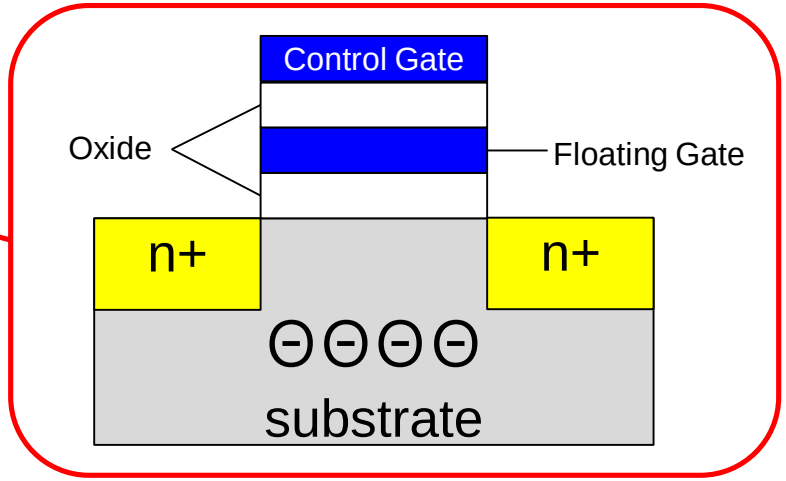
- Hard Disk Drives
- **Flash and Solid-state Drives**
- Storage Class Memory

Flash Memory

- **Fundamentally different from HDDs – It is based on a transistor (a cell) that can be written and read using *electronic circuits***
  - No mechanical parts needed
- **Flash memory is "non-volatile"**
  - One (or more) bits are stored in a ***floating gate transistor*** (a cell) that holds a value *without power supply*



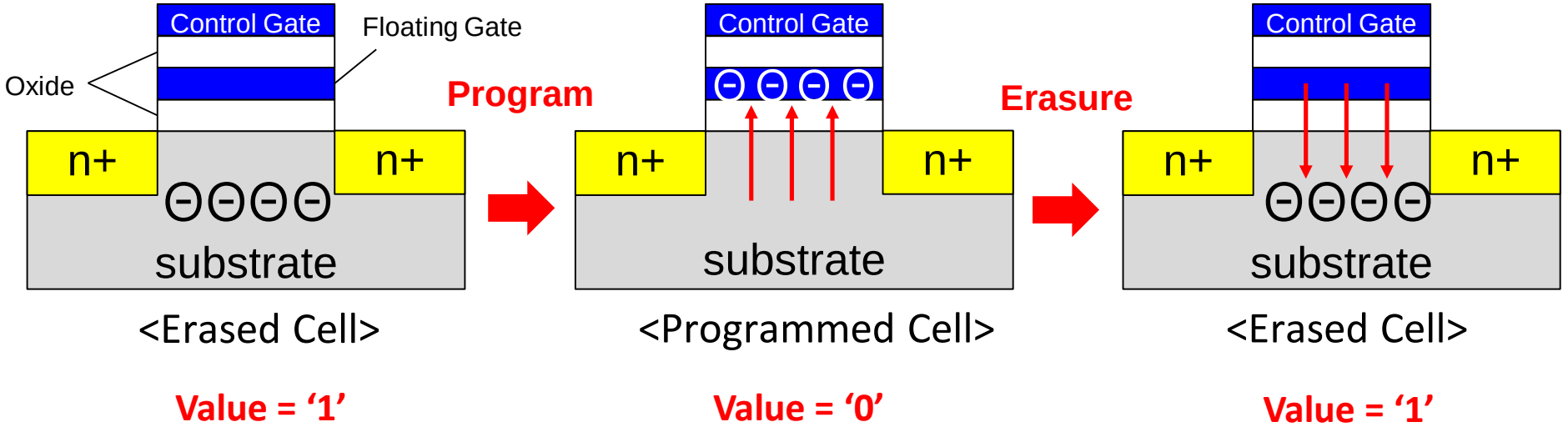
Wafer, Die, and Package



Floating Gate Transistor

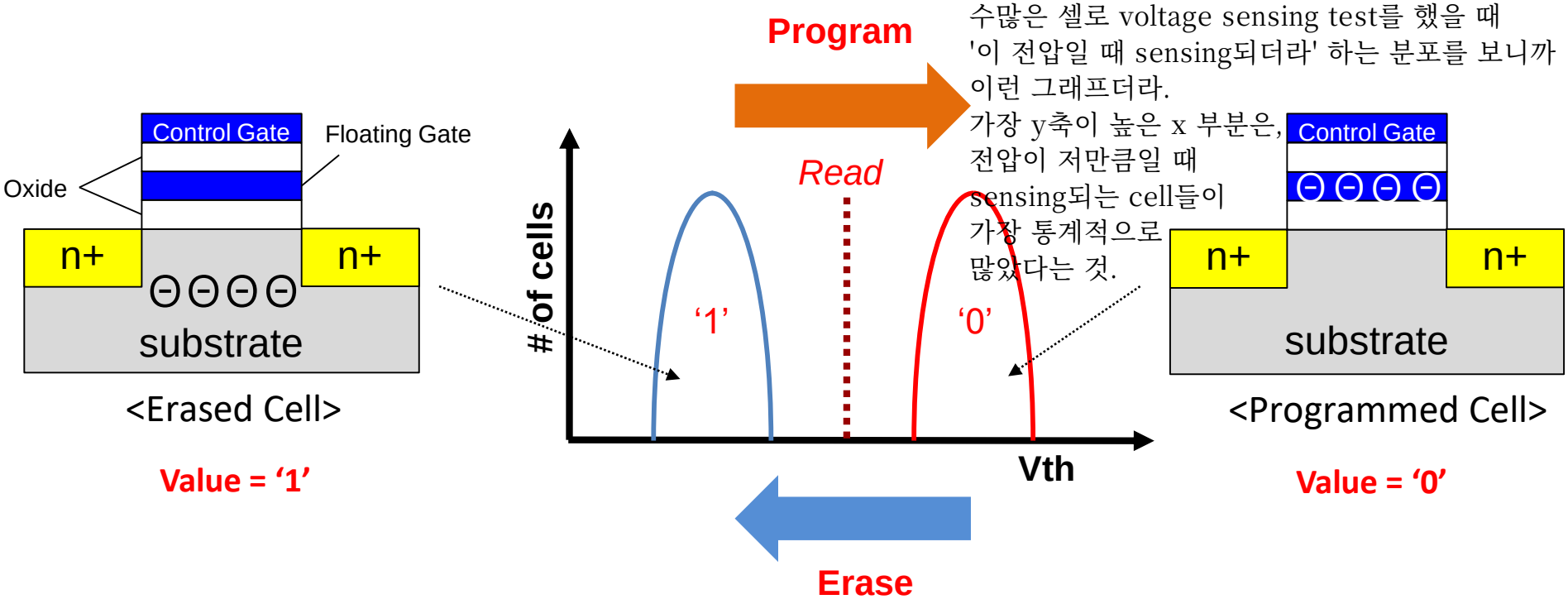
How Value is Stored and Retrieved?

- **A bit value of each cell can be controlled by three operations**
  - Each cell is initially erased, and its value is represented as ‘1’
  - A value is changed to ‘0’ by a ***program*** operation
  - A value is changed to ‘1’ by an ***erasure*** operation
  - A value can be retrieved by a ***read*** operation



How Value is Stored and Retrieved? (Cont.)

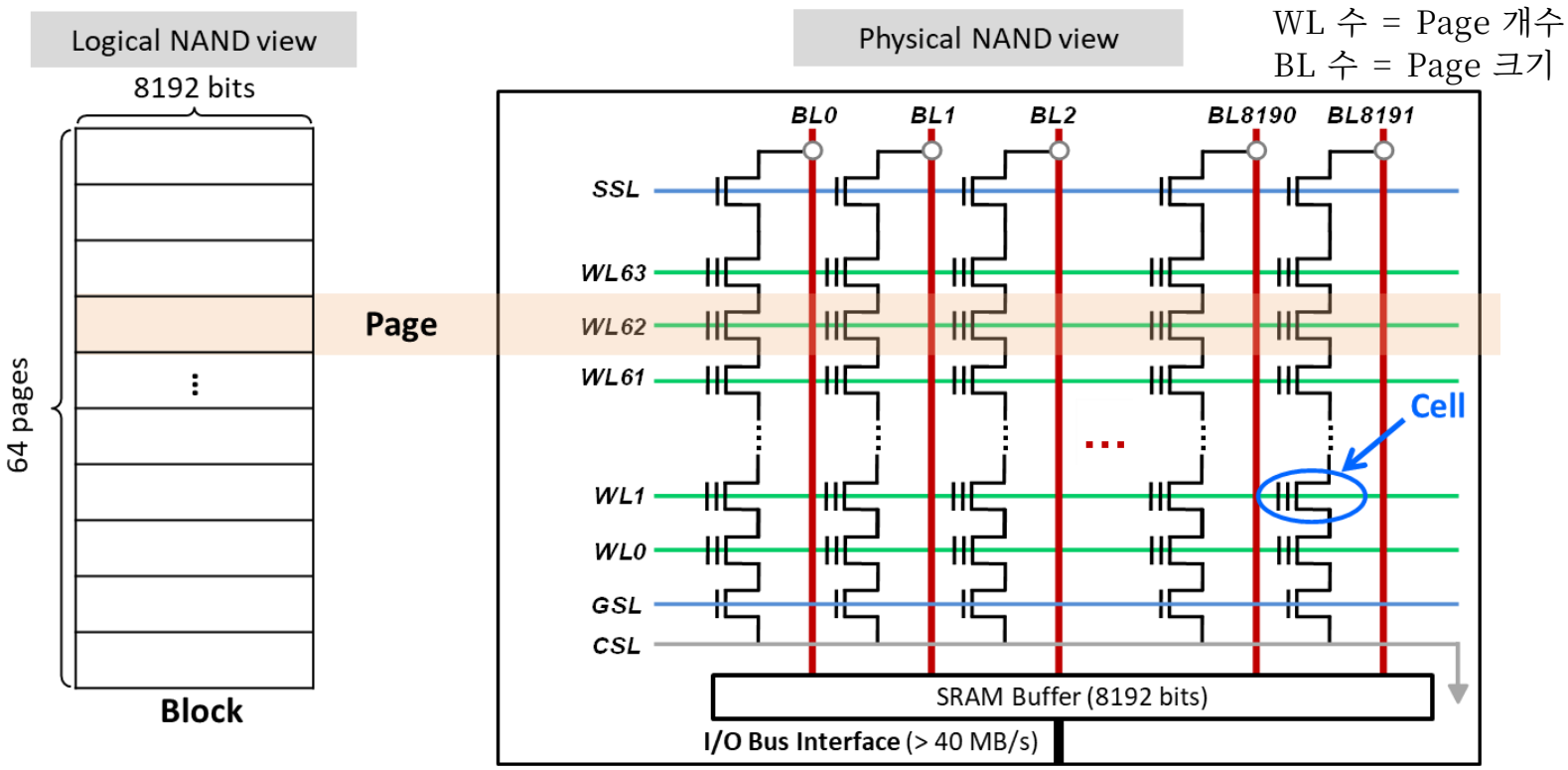
- **Program & read binary data to a flash cell**
  - Data “0” → Program → Shift cell Vth to high → Off state → No current flow
  - Data “1” → Erase → Shift cell Vth to low → On State → Current flow



✓ *Read : Check the current flow*

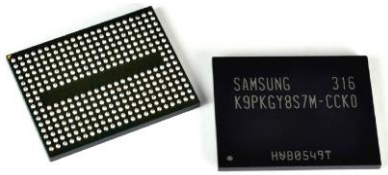
NAND Cell Array

- **A group of NAND cells are written and read simultaneously**
  - This group of cells is called a *page* (4K – 16K cells)
- **A group of pages should be erased together**
  - This group of pages is called a *block* (128 – 256 pages)
- **This asymmetric I/O unit makes *random writes quite expensive***



NAND Flash Chip

- **A group of blocks compose a NAND *flash die***
- **Two or four dies are then packed as a single NAND chips**



<NAND flash chip specification>

Page Read to SRAM	50 μs
Page Program (Write) from SRAM	500 μs
Block Erasure	4 *ms*
Serial Access to SRAM	100 μs (per 4 KB)
Page Size	16 KB
Block Size	4 MB
Die Size	8 Gb
Dies per Package	1, 2, or 4

- **Key properties**
  - Read is 10x faster than write: 50 μs vs 500 μs
  - Erasure is slowest: 4 *ms*
  - Throughput is not so high: 36.4 MB/s for reads (Note: recent SSDs offer 5 GB/s)

즉, NAND chip 자체의 스펙은
disk에 비해 그렇게까지 뛰어난 게 아니다.

How to Improve Performance?

Read throughput (36.4 MB/s) = 16 KB / (50 + 400) *µs*

- **Strategy 1: Reduce the latency of three I/O primitives**
  - The latency of three I/O primitives get longer as NAND cells scale down

	SLC (2D)	MLC (2D)	TLC (2D)	MLC (3D)
Page Program	25 *µs*	50 *µs*	100 *µs*	50 *µs*
Page Read	250 *µs*	500 *µs*	3 *ms*	500 *µs*

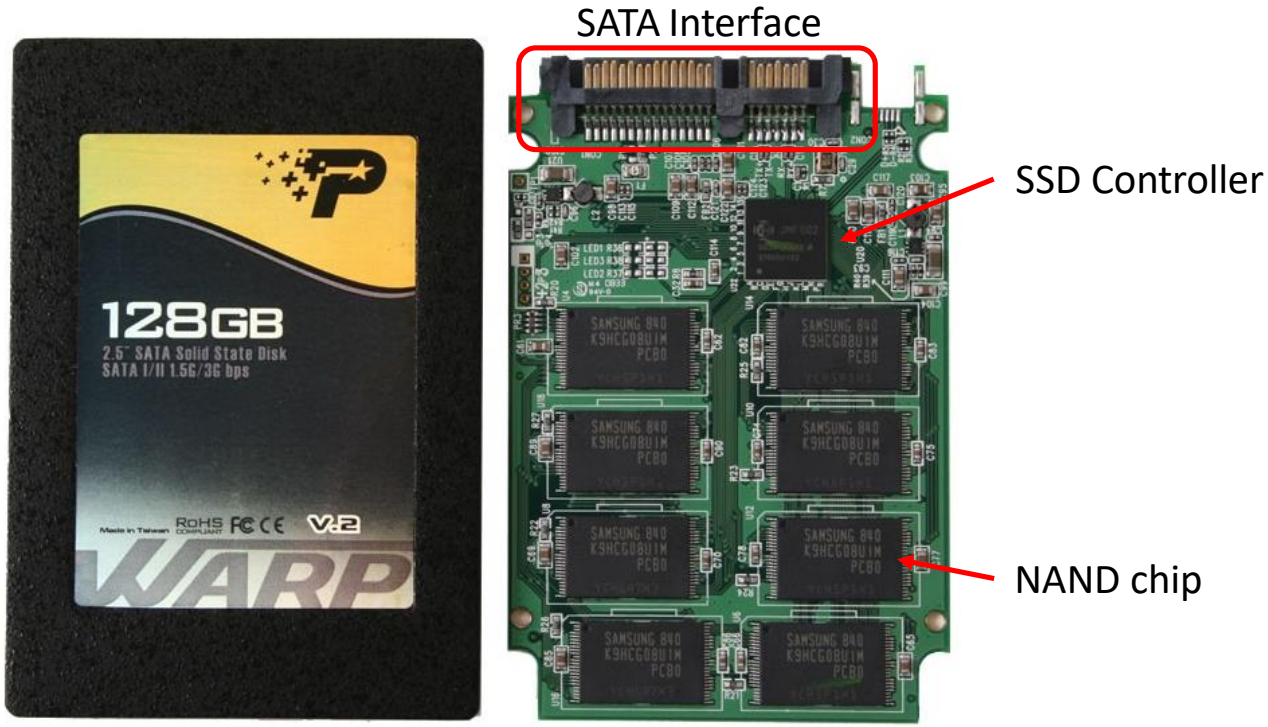
    - 3D NAND improves I/O latency with larger cells

- **Strategy 2: Improve the bandwidth of I/O interface**

	Interface	Throughput (MB/s)
Before 2006	No Standard (SDR)	40 MB/s
2006	NV-SDR	50 MB/s
2008	NV-DDR	200 MB/s
2011	NV-DDR2	533 MB/s
2014	NV-DDR3	800 MB/s

How to Improve Performance? (Cont.)

- **Strategy 3: Aggregate many NAND chips and access them in parallel**
  - Example: 36.4 MB/s x 64-128 NAND dies = **2.3-4.7 GB/s**
  - This is what an SSD controller does!
    - To achieve optimal performance, sophisticated algorithms are needed!



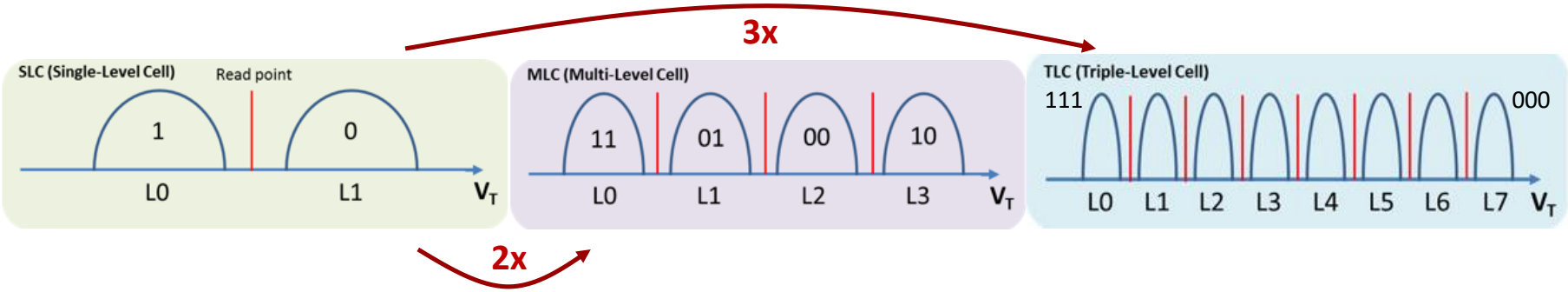
How to Increase Capacity?

- **Strategy 1: Reduce the size of each cell**
  - Put more transistors per same die

No more scaling

Year	2008	2009	2010	2011	2012	2014	2015
Cell Size	42 *n*m	32 *n*m	27 *n*m	21 *n*m	19 *n*m	16 *n*m	14 *n*m

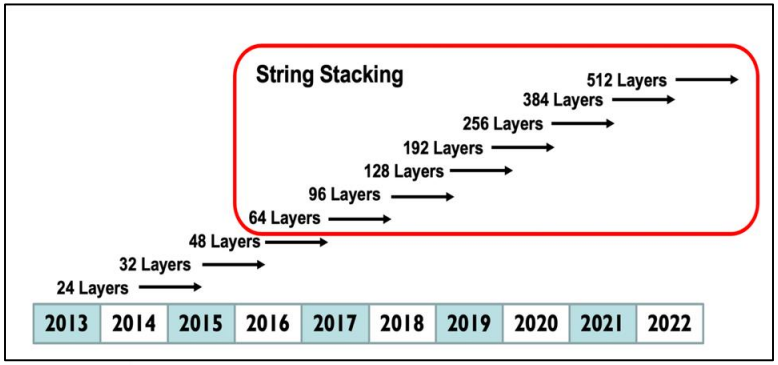
- **Strategy 2: Store more than one bit per cell**
  - Multi-level cell (MLC) technologies enable us to store more than one bit per cell

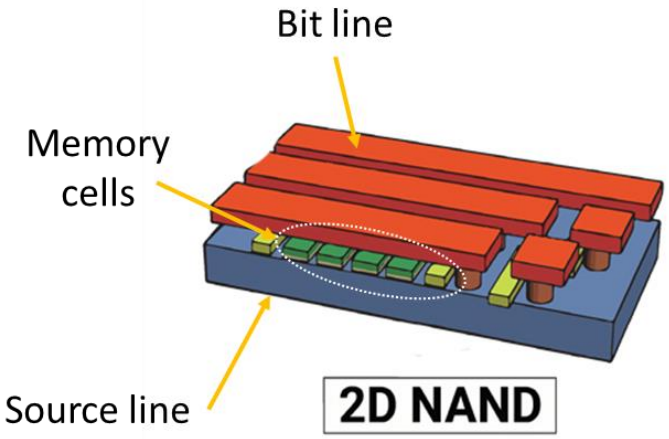


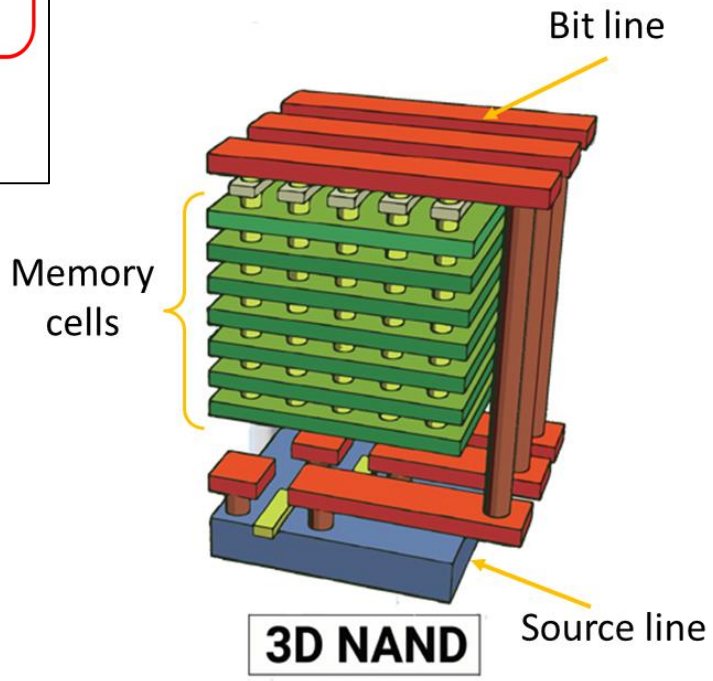
- **Inevitably result in the degradation of speed and reliability!**

How to Increase Capacity? (Cont.)

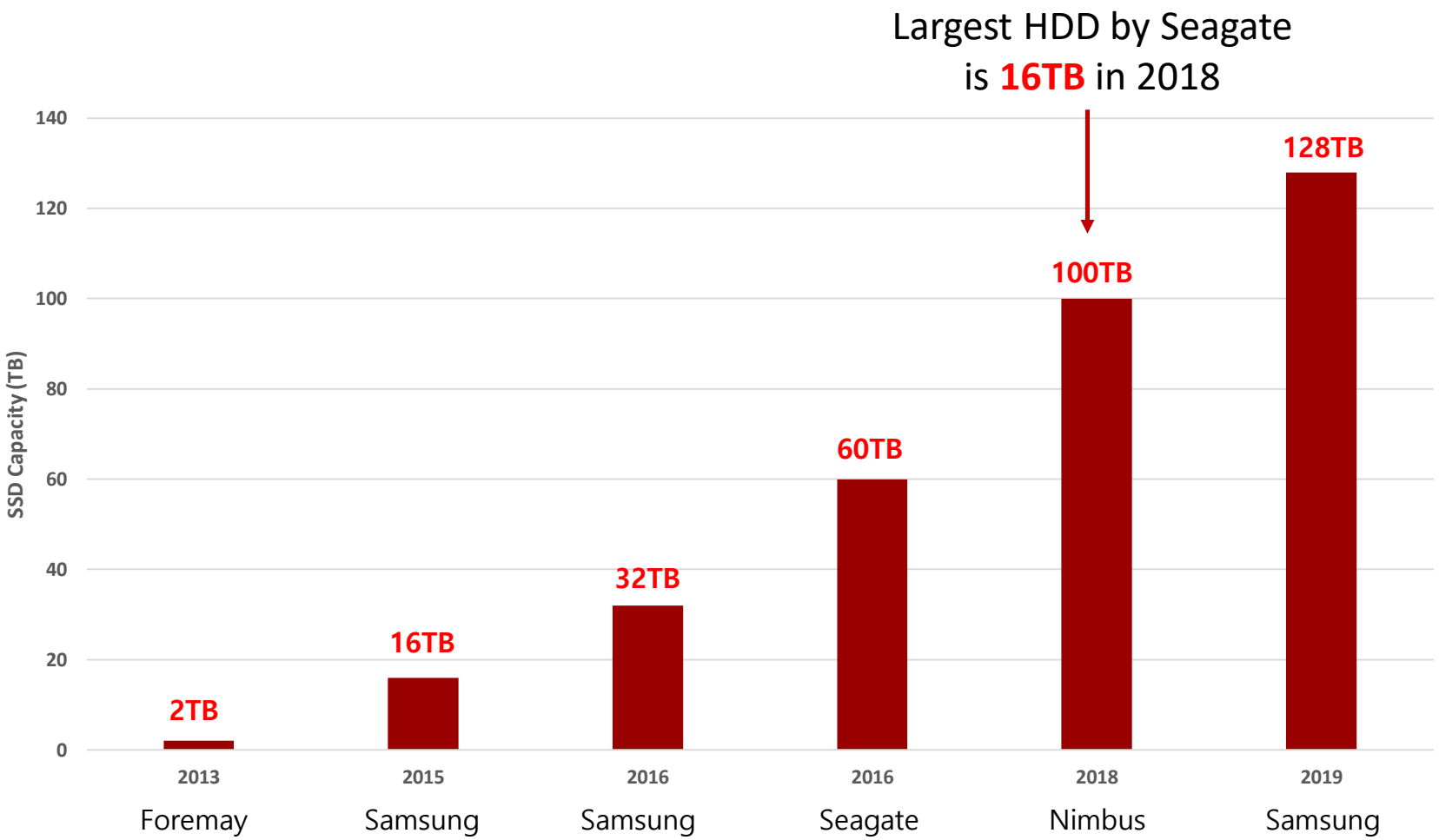
- **Strategy 3: 3D NAND**
  - Stacking cells *vertically* to avoid problems w/ shrinking feature sizes
  - 96 – 128 layered 3D NAND flash is in production







SSD Capacity Trend



SSD Performance Trend

Intel's DC P4618 (6.4 TB)

Sequential Read	6,650 MB/s
Sequential Write	5,350 MB/s
Random Read	1,210,000 IOPS
Random Write	484,000 IOPS
DWPD	3-5 years

Samsung's PM1725b (12.8 TB)

Sequential Read	6,300 MB/s
Sequential Write	3,300 MB/s
Random Read	940,000 IOPS
Random Write	110,000 IOPS
DWPD	3-5 years

Samsung's Z-SSD (800GB)

Sequential Read	3,200 MB/s
Sequential Write	3,200 MB/s
Random Read	750,000 IOPS
Random Write	170,000 IOPS
Read/Write Latency	12 / 16 us
DWPD	30 years

Outline

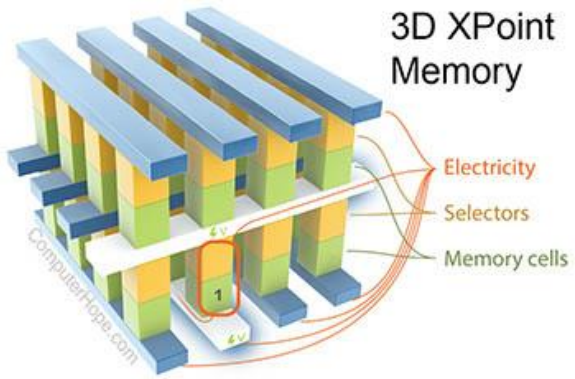
- Hard Disk Drives
- Flash and Solid-state Drives
- **Storage Class Memory**

Storage Class Memory

- **Byte addressable stable memory devices**
  - Phase Change (PCM)
    - Crystalline material heated to two different phases
  - Spin Torque Transfer (STT-RAM, MRAM)
    - Magnetic spin captured in ferrous element
  - ReRAM
    - Resistive RAM
- **SCM will fit a niche between DRAM and NAND**
  - Power advantages over DRAM because there is no “refresh” cycle necessary to preserve logic value
  - Cost disadvantages over NAND because block-addressing makes NAND more compact

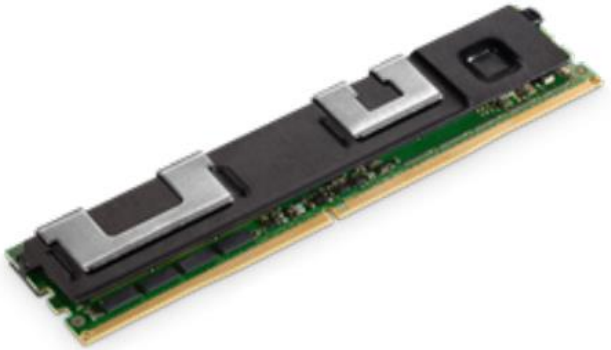
3D X-Point NVRAM

- **New Technology from Intel/Micron**
  - Similar to Phase Change and Resistive Memories
  - Bit addressable
  - No Erase requirement
- **"~1000 faster" than NAND**
  - But still 3x to 5x slower than DRAM
- **Process has a trajectory to catch NAND density**
  - This is a big deal, first parts are 128 Gb
- **Much improved durability**
  - However, even at 100 million write cycles, you could wear it out in seconds w/out wear leveling
- **Products are memory (DIMM) or storage devices (PCIe)**



Intel's Optane DIMM

- Intel's Optane DIMM (512GB in 2019): 400 ns latency and up to 5.3 GB/s throughput



Read Throughput (256B)	5.3 GB/s
Write Throughput (256B)	1.89 GB/s
Read Throughput (64B)	1.4 GB/s
Write Throughput (64B)	0.47 GB/s
Latency	400 ns

Web Pricing, November 3, 2018

Description	$ Price	1yr % chg	$/GB
Intel 3D Xpoint 480 GB SSD PCIe 4x	**602**	-	**$1.310**
SanDisk 32 GB microSD U3/Class 10 SDSQXCG-032G	16	**-45%**	$0.500
Samsung 1TB SDSSDA-1T00-G26	**149**	**-54%**	**$0.160**
Seagate 1TB HDD ST1000DM010	**46**	0%	**$0.045**
Seagate 2TB HDD ST2000DM006	60	-10%	$0.029
Seagate 4TB HDD ST4000DM004	98	0%	$0.024
Seagate 8TB External HDD STGY8000400	150	-33%	$0.018
Seagate 8TB 7200 RPM HDD ST8000DM004	199	-28%	$0.024
Seagate 10TB HDD ST10000VN0004	299	-12%	$0.030
Seagate 12TB HDD ST12000VN0007	399	-	$0.033

End of Chapter 2