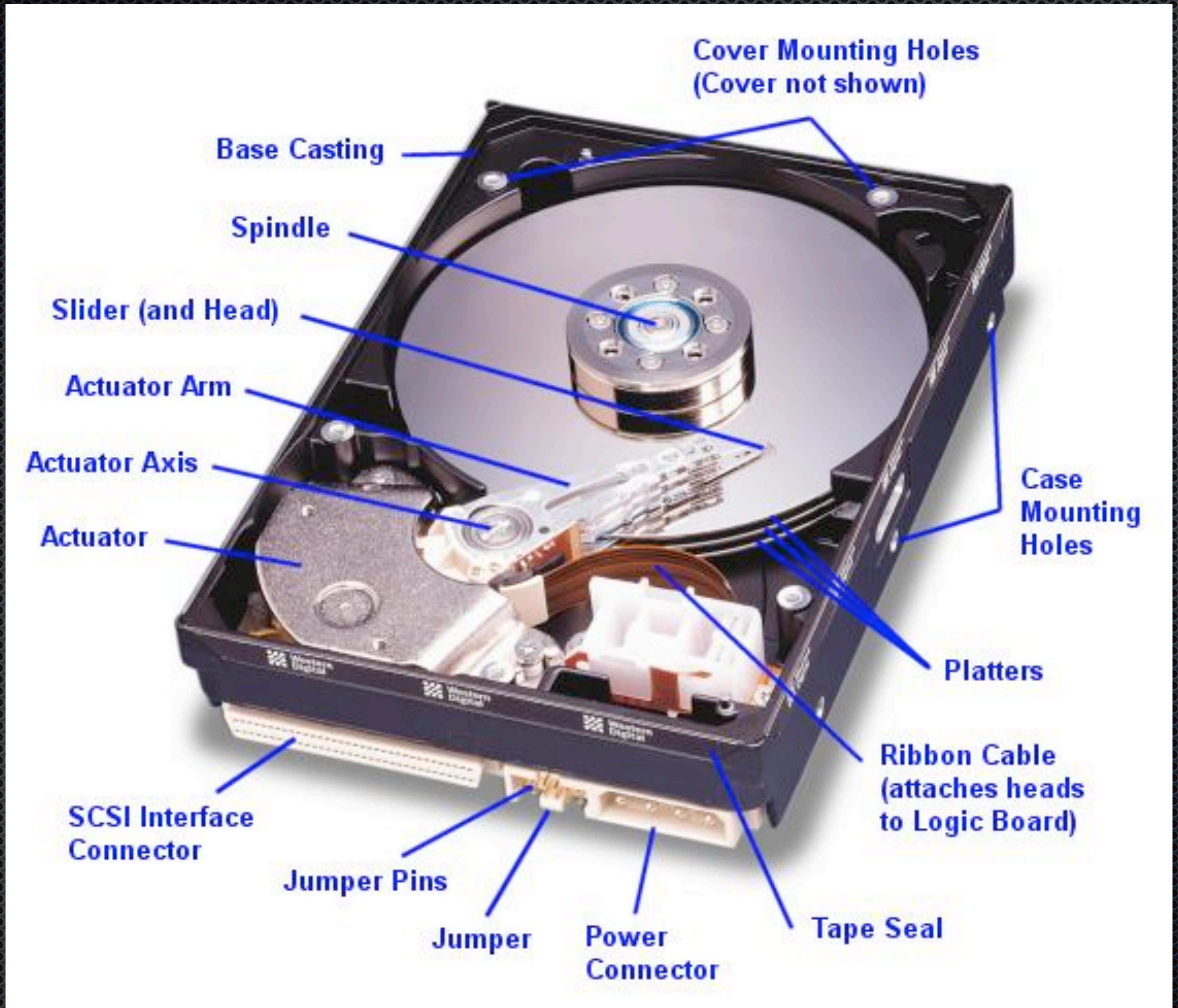
Disk Storage

Nonvolatile bulk memory

Basic Concepts

- Rotating platters
- Moving heads on arms
- Uniform magnetic surface
- Data written as magnetic spots

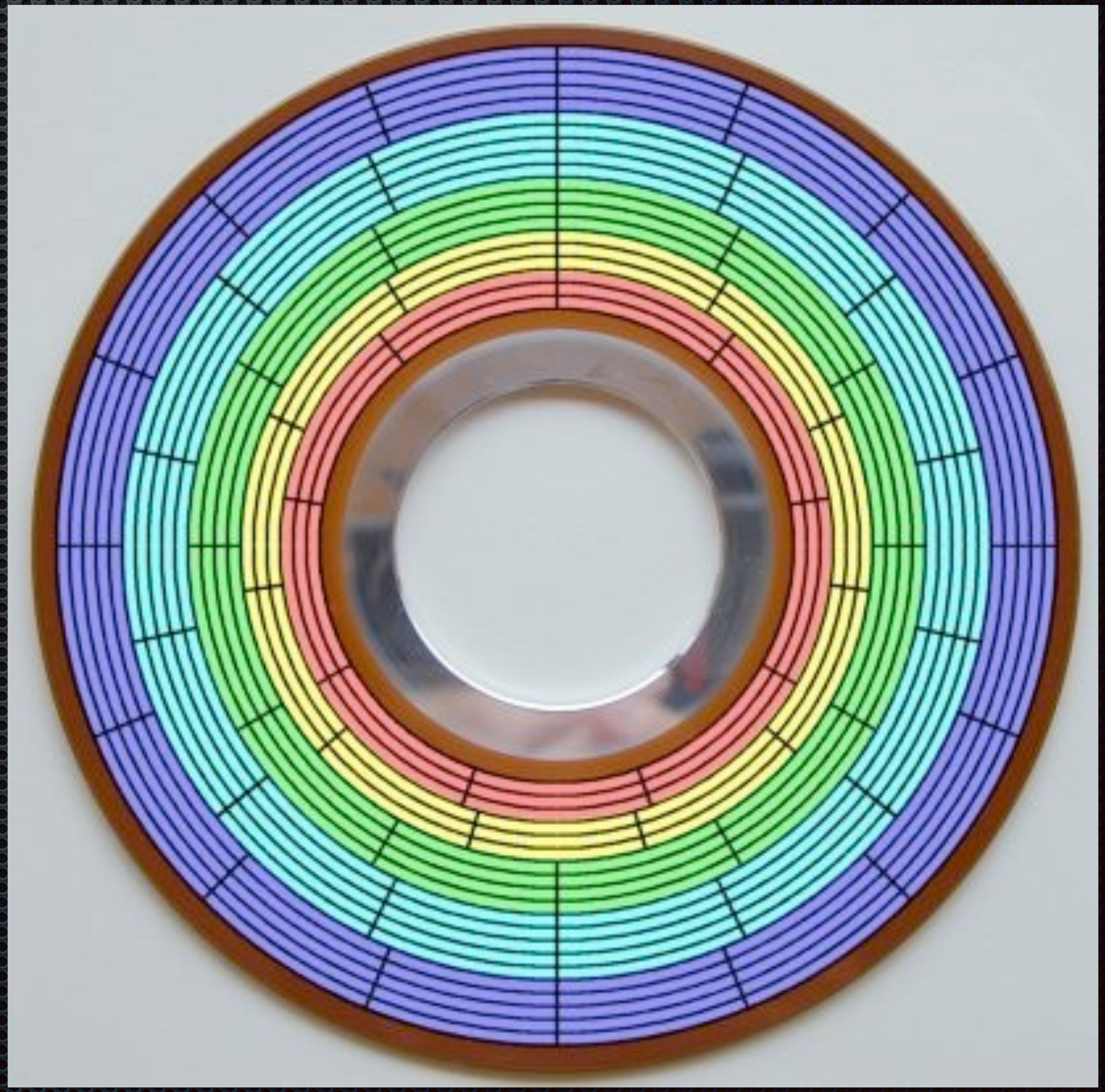
Structure



Data organized in tracks and cylinders

Zoned Bit Recording

- Textbooks refer to tracks with fixed number of sectors
- Modern disks use variable size sectors
- Pack more data on outer, faster-moving tracks
- Disk controller performs logical mapping of fixed sectors to ZBR



Images from storagereview.com

Low-level Formatting

- Done at factory -- not changeable
- Patterns tracks, sectors, servo marks
- Bad sectors identified
- Spare sectors mapped into their place
- Means different disks with identical data, written in the same order, can have different access times

Error Correction

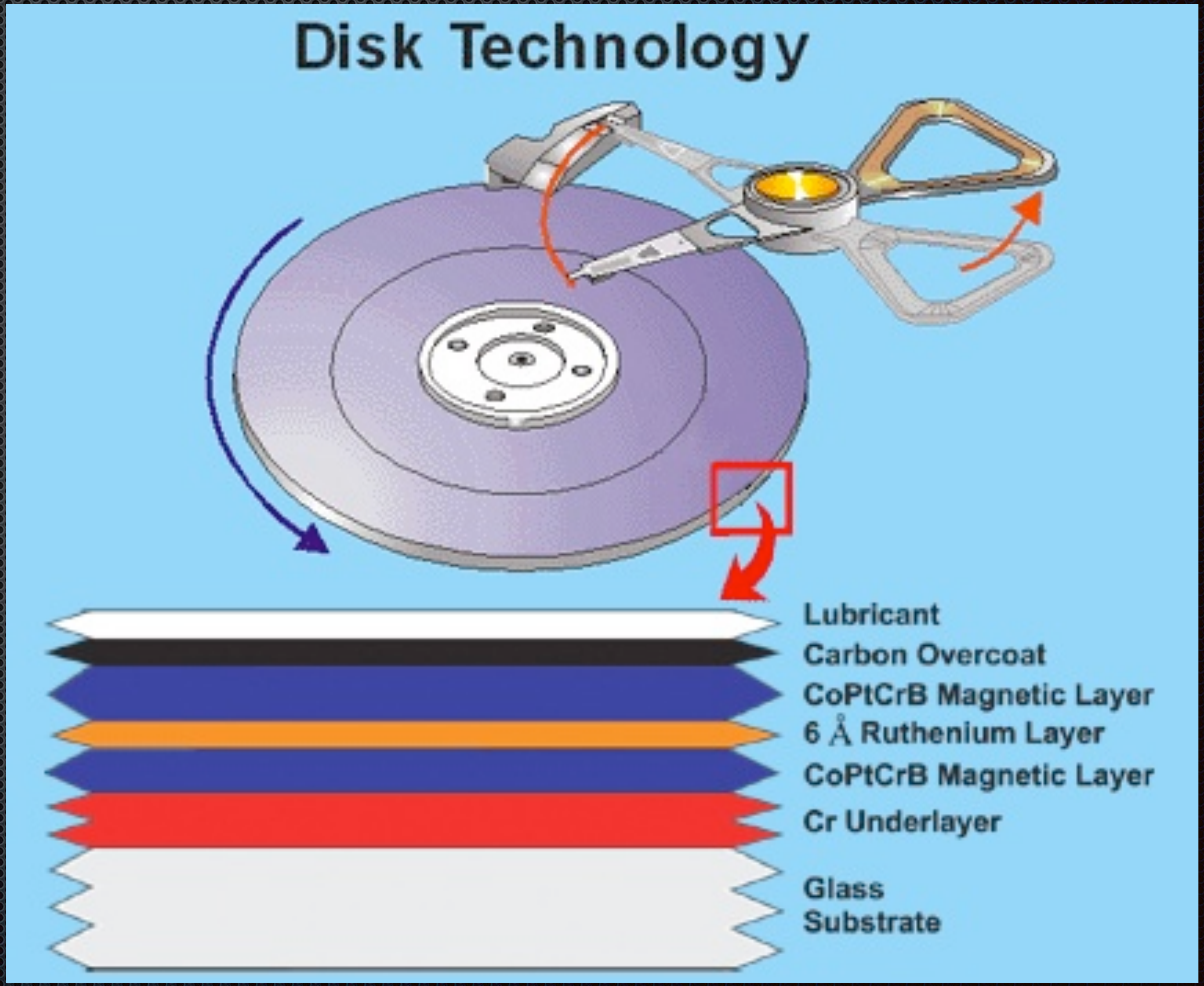
- Read errors are common
- Sectors include error correcting code
- Read and check for error -- if none, good
- If error, apply ECC to fix
- If not fixed, reread, try stronger correction
- If not recoverable, report error

Parameters

- Typically 1 to 10 platters
- 5.25, 3.5, 2.5, 1.8, 1.3, 1.0 inches in diameter
  - Smaller platters: Easier to make, lighter, more rigid, less noise and vibration, faster seek times
- Rotation speed: 5400, 7200 RPM (10K, 15K for older)
- Substrate materials: aluminum or glass

Coating

- Early disks used iron oxide or similar coating
  - Relatively thick, easily damaged, low data density
- Modern disks use a thin film with carbon overcoat and lubricant

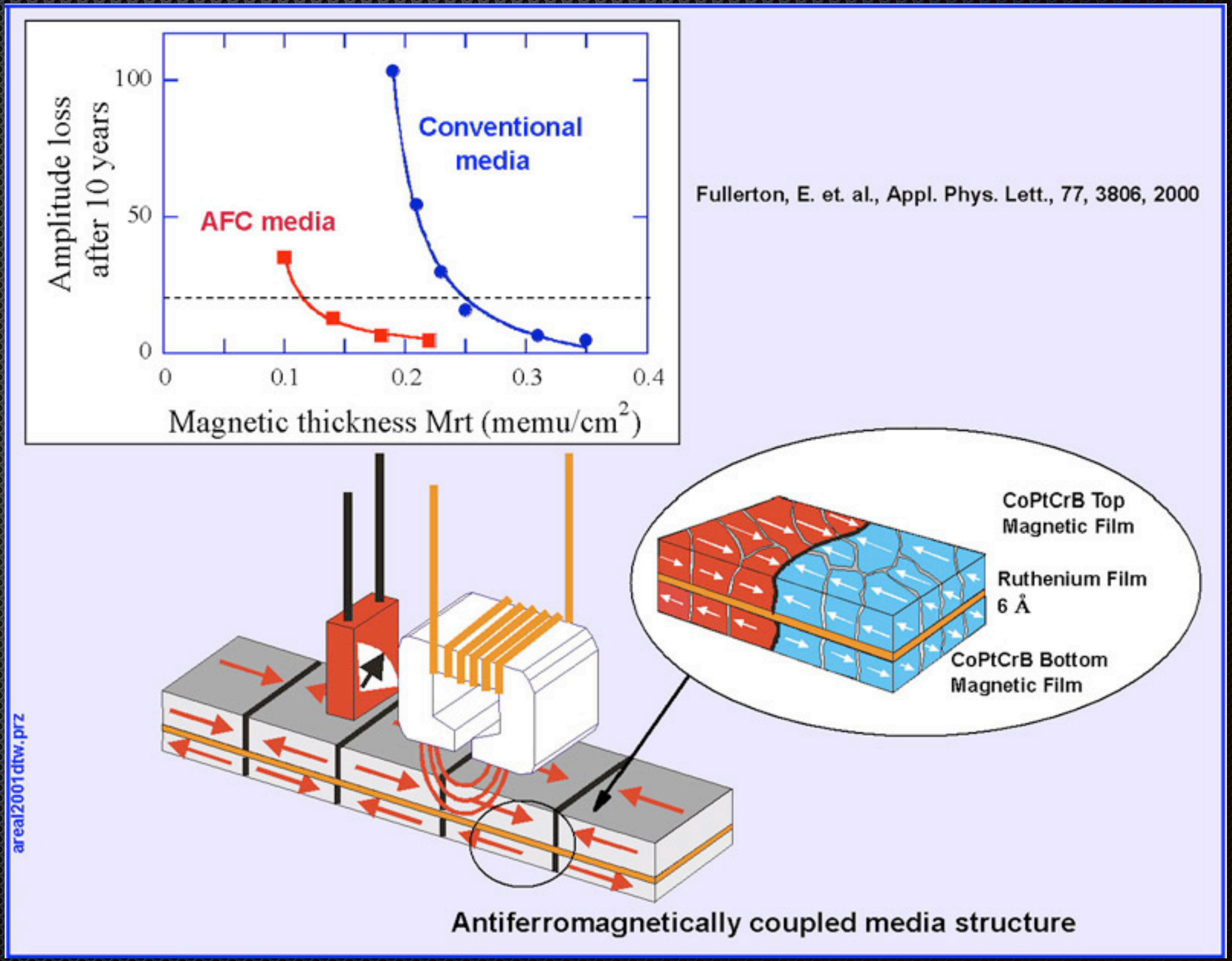


Thin Film

- Thinner enables denser storage -- domains cannot spread out as far
- Grains must be very small
- Must have higher coercivity (resistance to change) and magnetization
- As spot size shrinks, energy to change increases, and approaches thermal limit

Antiferromagnetic Coupling

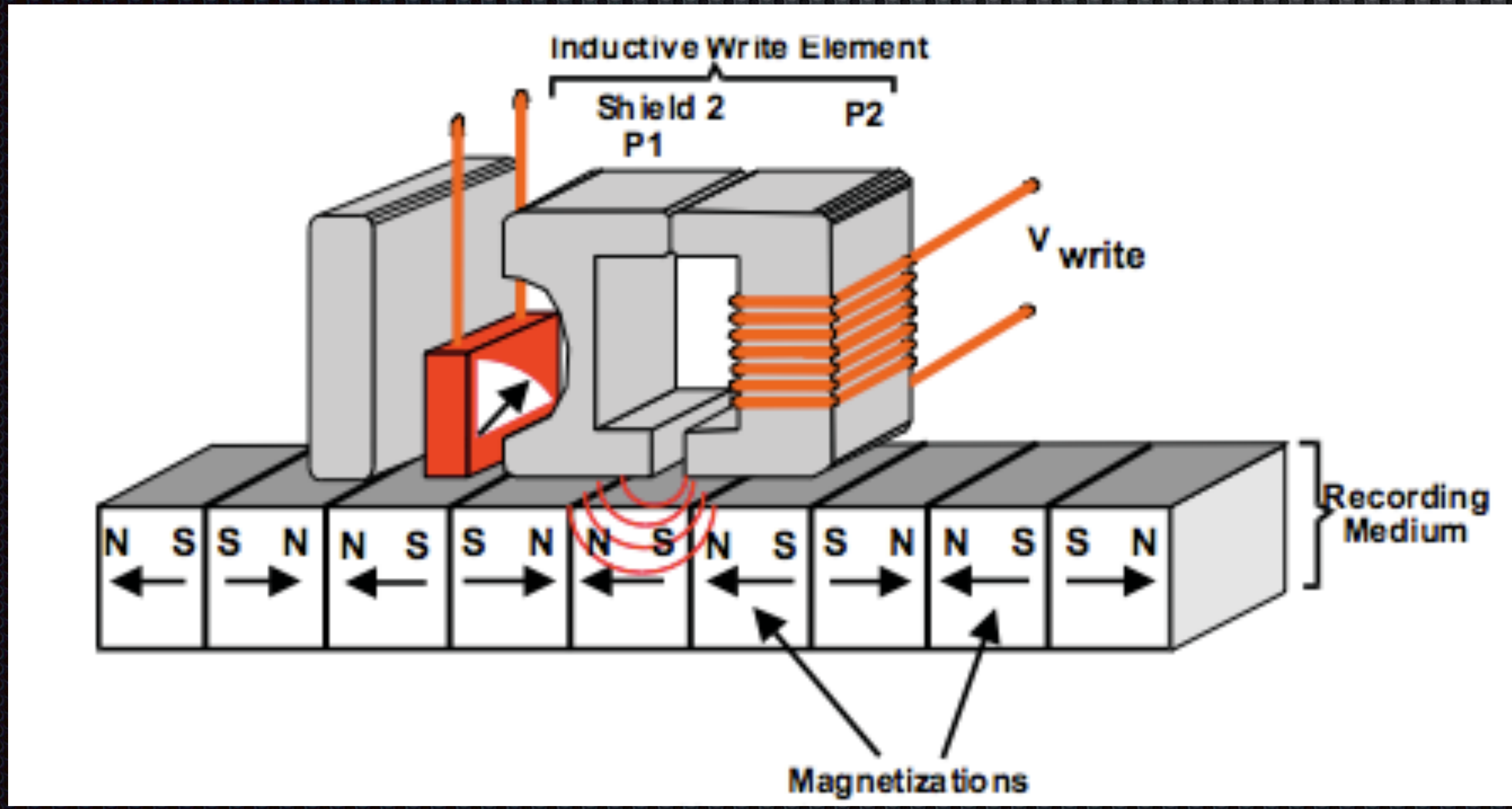
- Coupling layer between magnetic layers
- Effectively makes magnetization layer as thin as coupling layer (a few atoms)
- Allows thicker magnetic layers
- Extends life

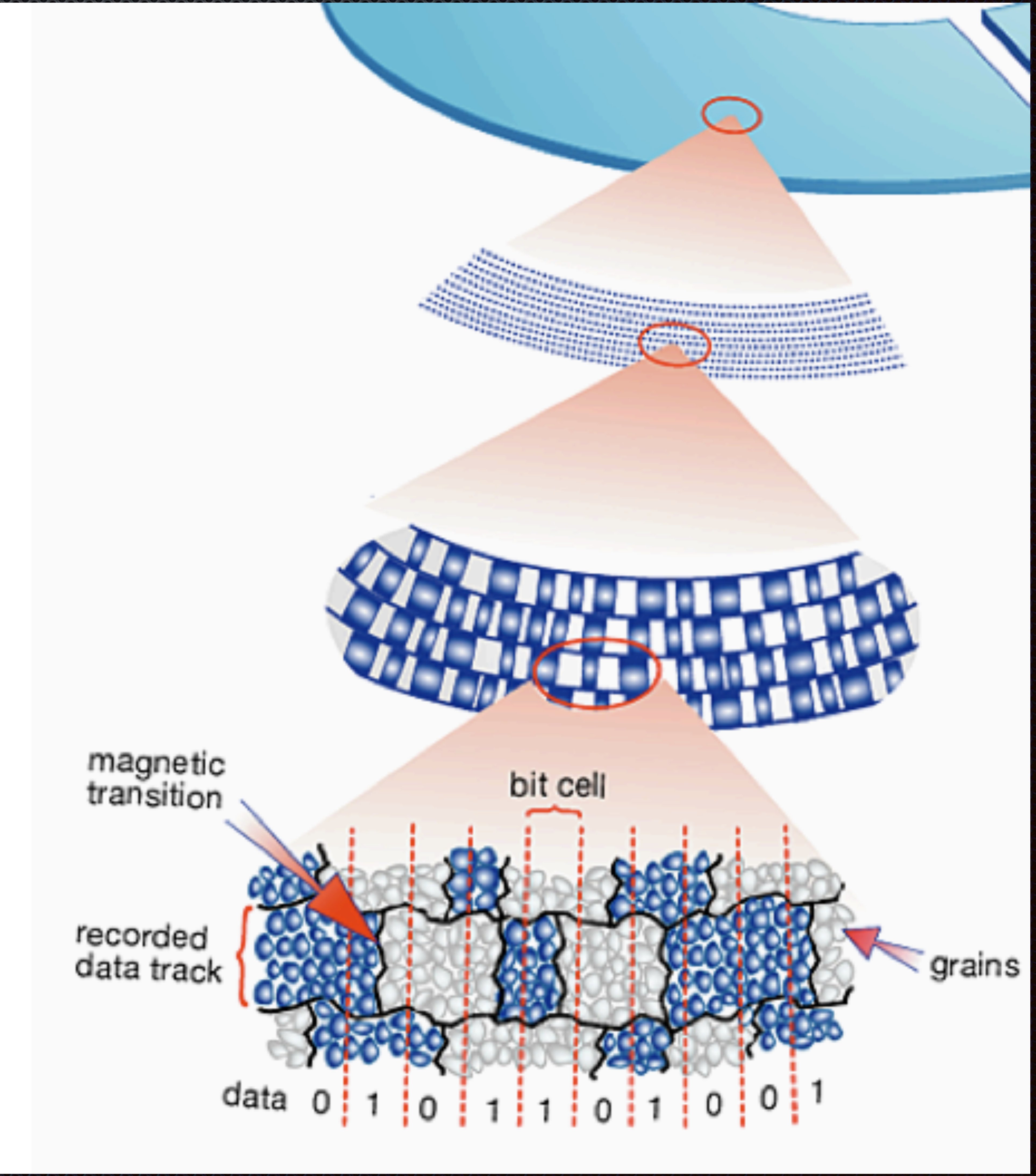


Figures from Hitachi Global Storage Technologies

Longitudinal Recording

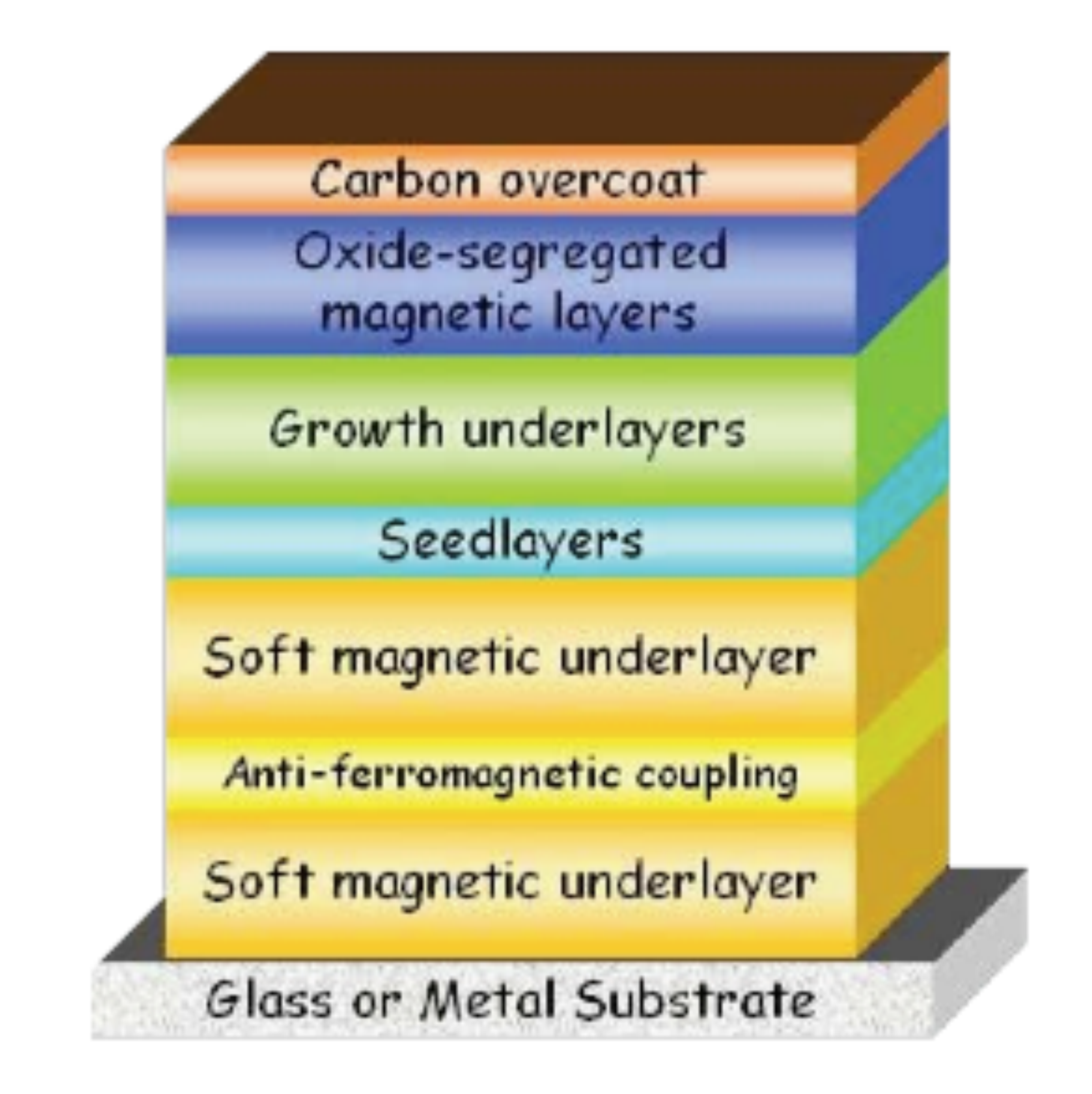
- Spots with same magnetic orientation = 0
- When orientation changes within spot = 1

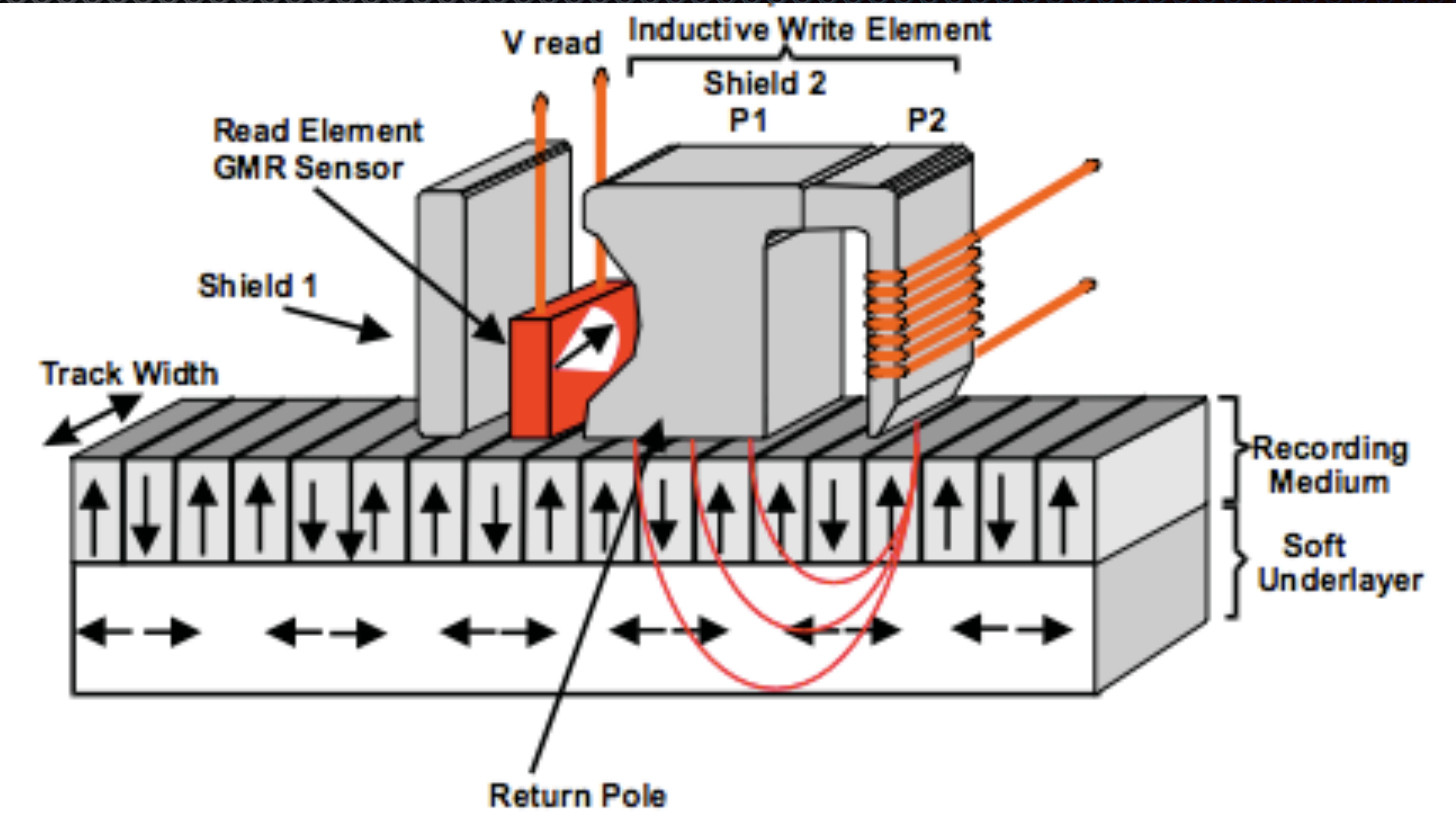




Perpendicular Recording

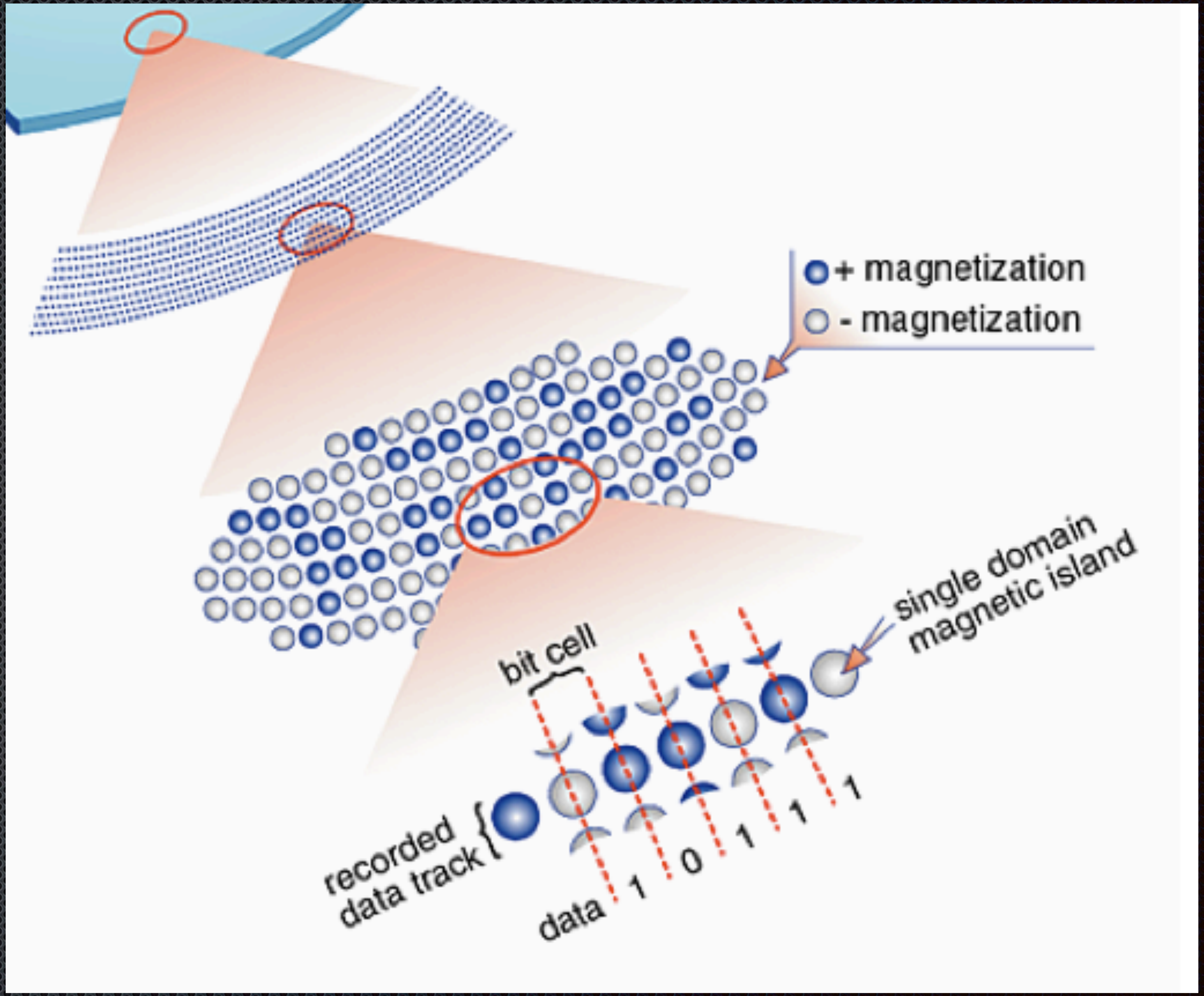
- New film layering with soft underlayer
- New form of write head
- Increases density without reaching thermal limit
- Density will eventually reach point that adjacent domains flip each other

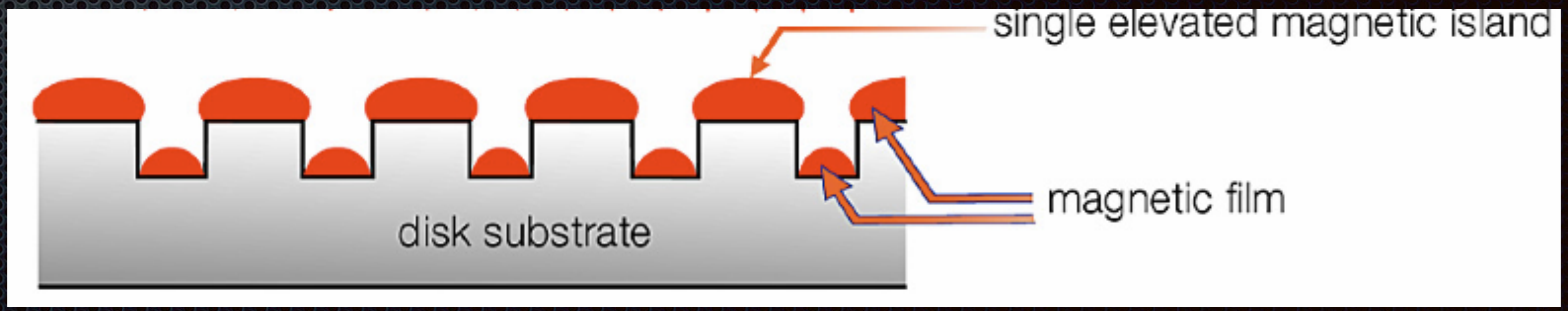




Patterned Recording

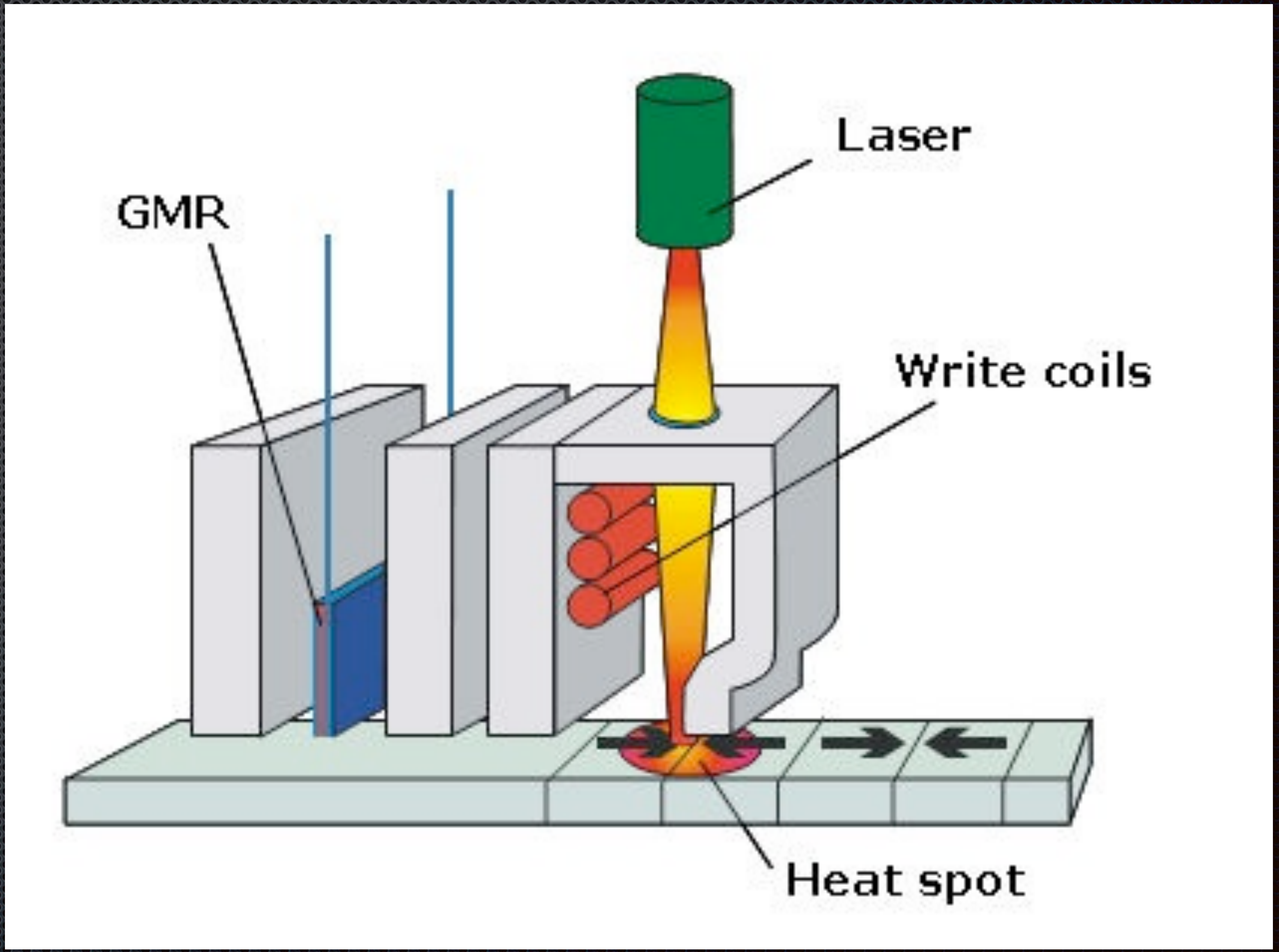
- Use lithography to texture surface for application of film
- Separates domains to avoid interference
- Creates rough surface
- More fabrication steps





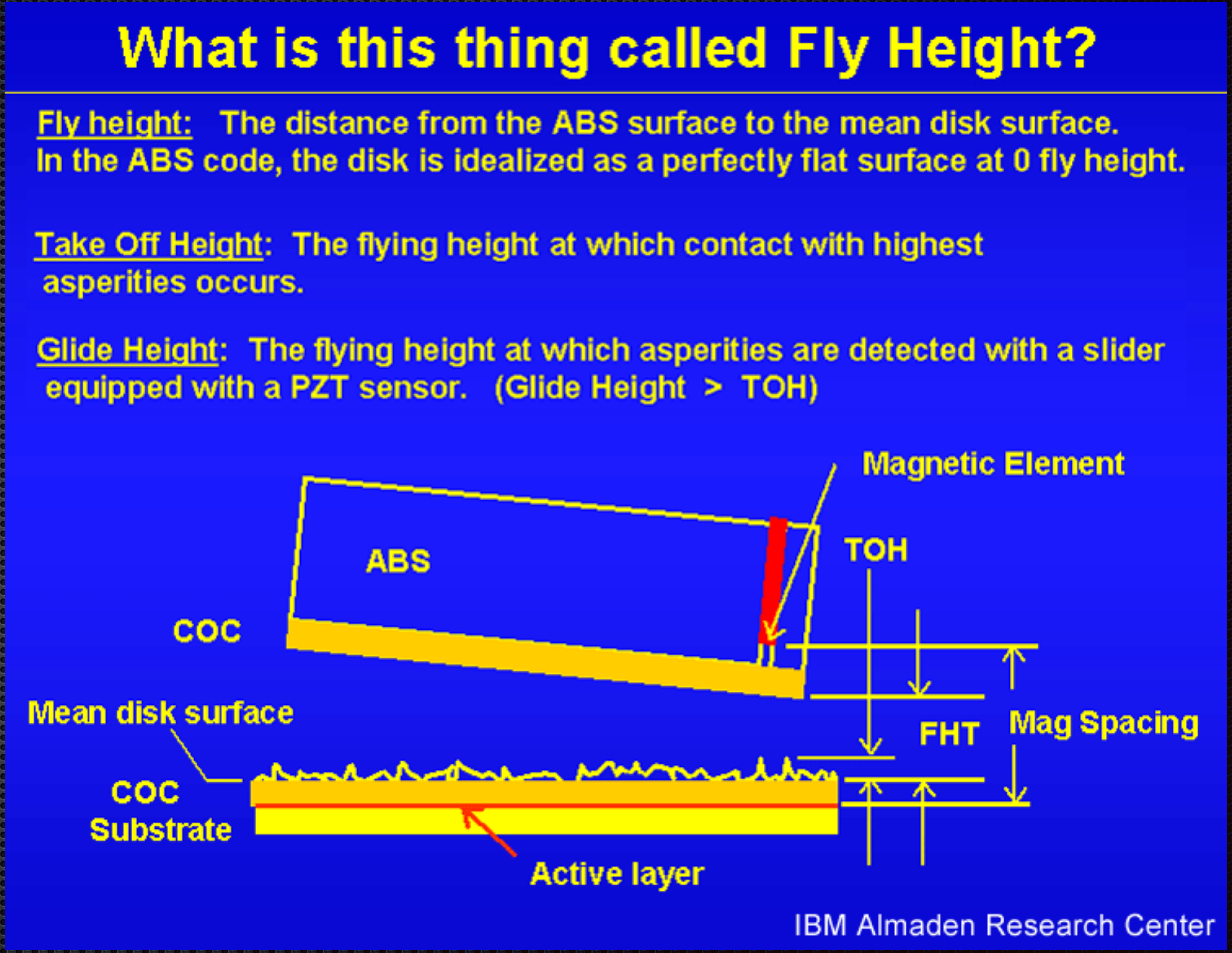
Thermally Assisted Recording

- Use more stable material
- Heat with laser to make temporarily unstable
- Use perpendicular recording to control magnetization before the spot cools



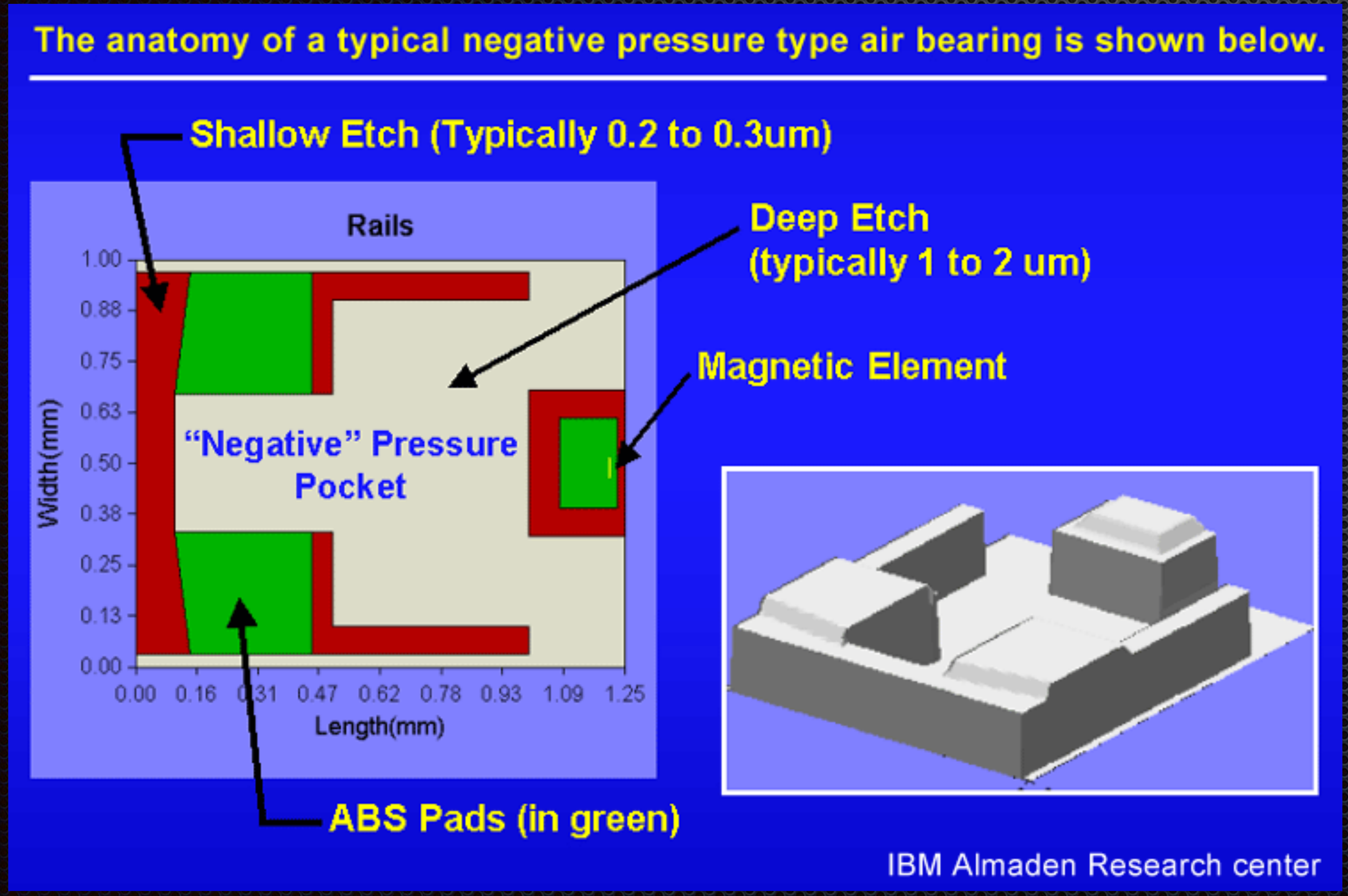
Read Head

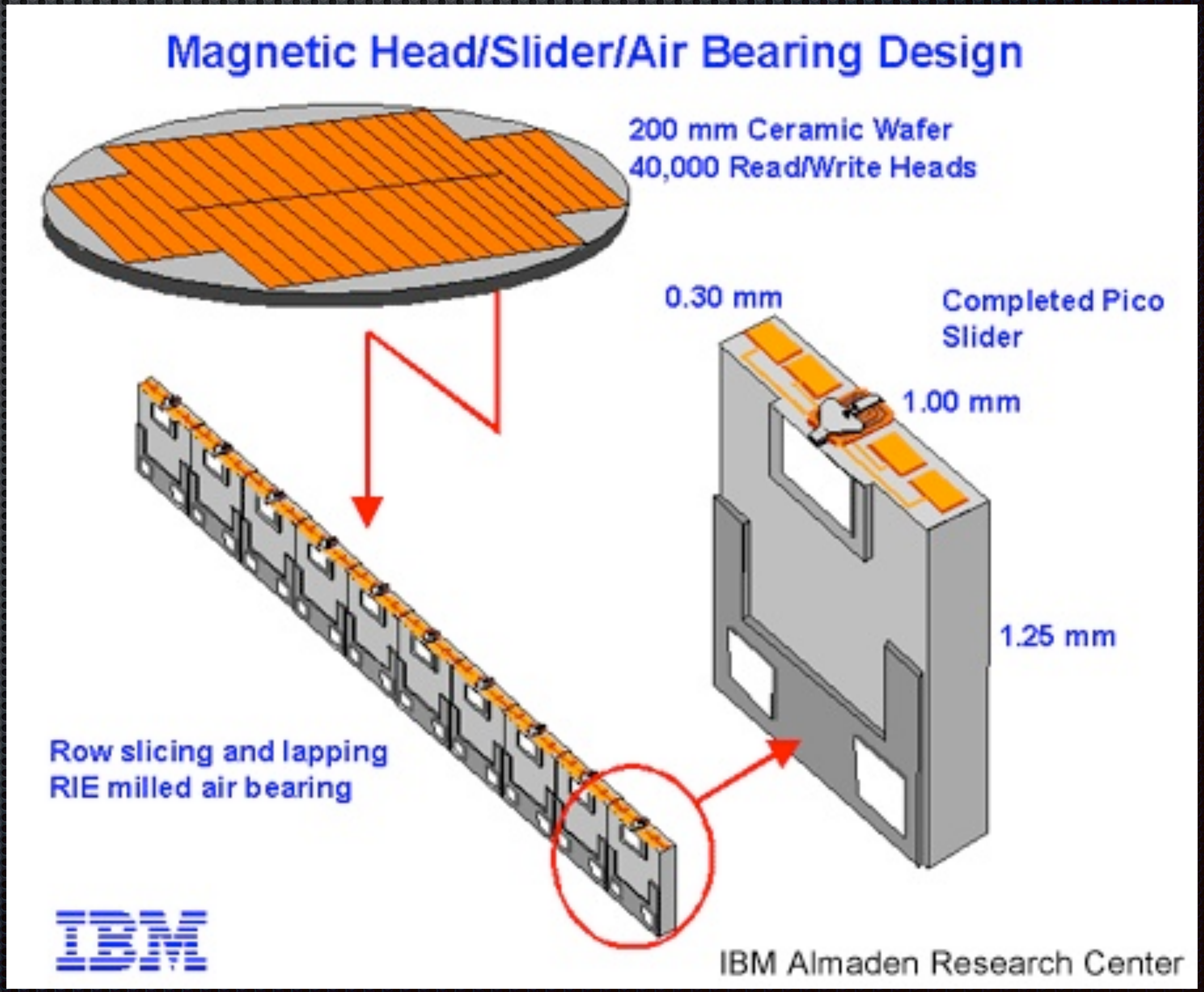
- Flies above spinning surface
- Disk creates airflow
- Lifts head against pressure
- Disk has landing zone for spin-down



Slider

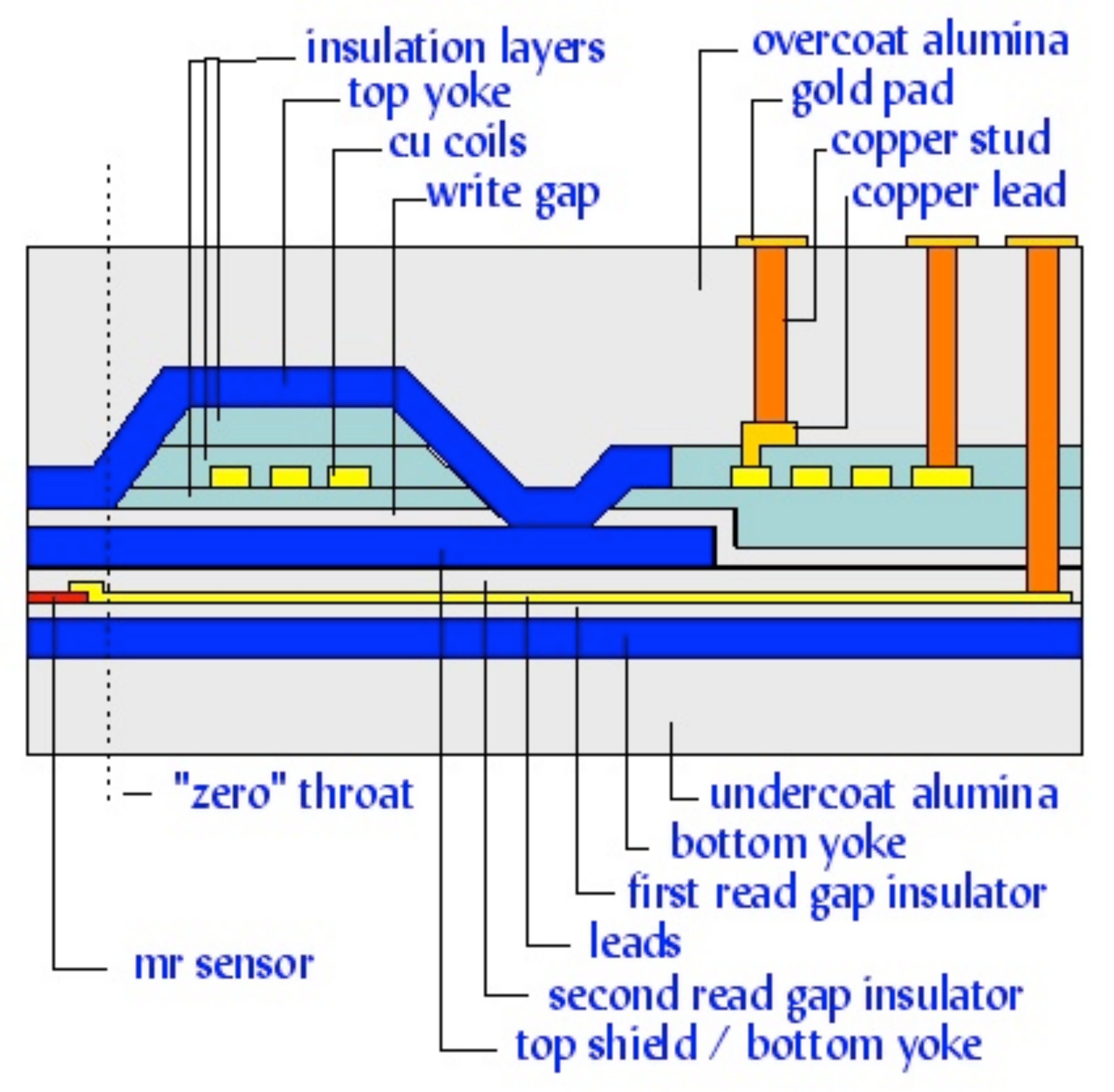
- Aerodynamic shape etched into underside of head to create proper lift and angle
- Electromagnet head attached to edge





Thin Film Head Construction

- Created with lithographic processes
- Copper coils to induce field
- Yoke to concentrate
- Connections to outside

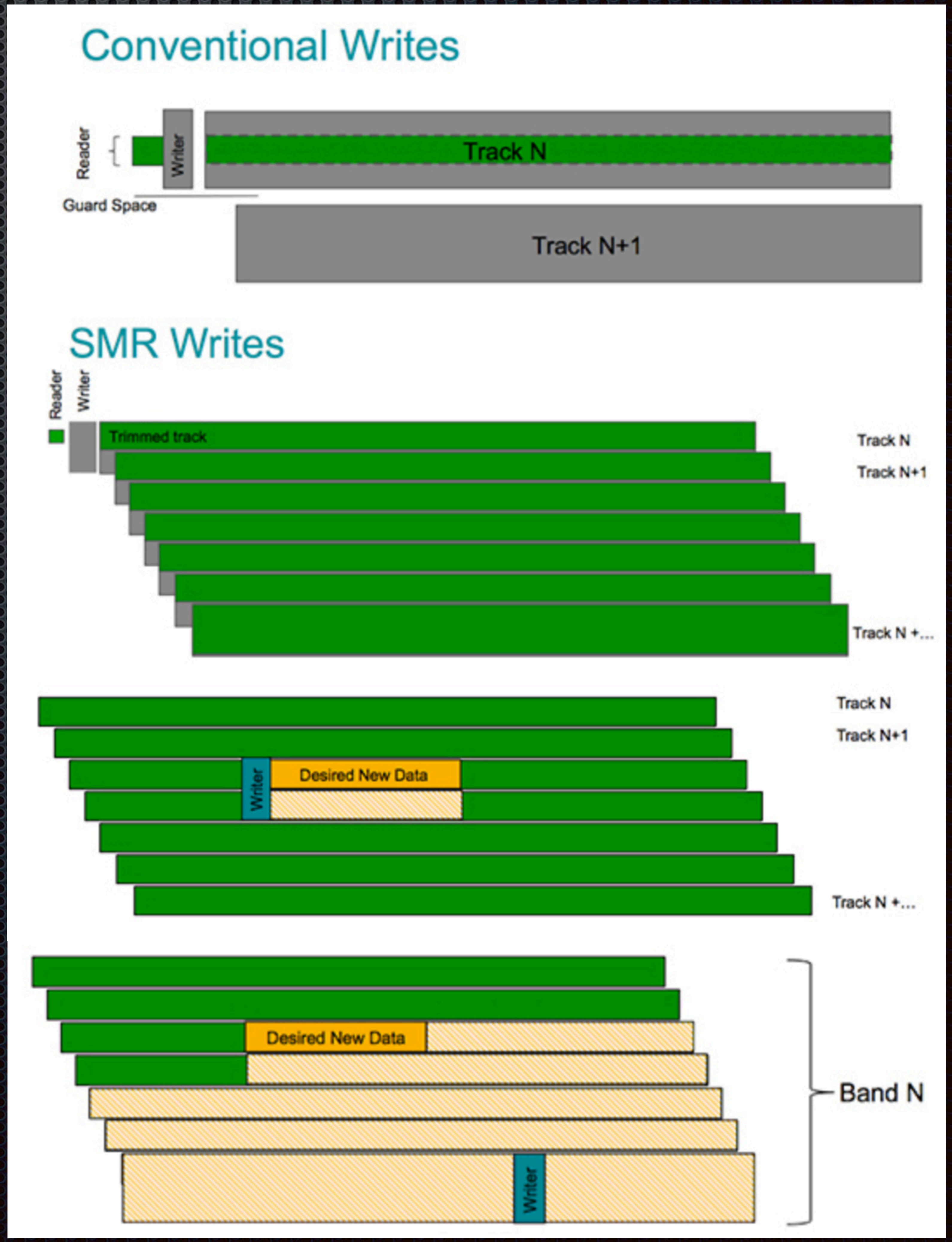


Future

- Projected growth in density of 50% per year (down from 100% per year 10 years ago)
- Superparamagnetic limit probably about 2019
  - Shifting to alternate technology began in 2017
- Current density about 1 Tb/in^2
- Expect growth of 100 before limit is reached
- Will lead to interesting shifts in research focus

Post-SPM Limit Technologies

Source: Seagate



- Two-dimensional recording - enabling closer tracks
  - Offset pair of read heads better distinguishes signal from inter-track cross-talk
- Shingled recording (SMR)
  - Tracks overlap, using TDMR to recover data
  - Writing will overwrite multiple tracks
  - Read out adjacent track and rewrite (recurses)
  - Divide platter into bands where rewriting ends

More Post-SPML Technologies

- Microwave Assisted Magnetic Recording (MAMR) — Western Digital
  - 20 to 40 GHz EM field softens material to reduce coercivity for write head
  - Can be built using conventional head manufacturing
  - 18TB drives in 2020, expected limit of 4Tb/in^2 w/ 40TB planned for 2025
- Heat Assisted Magnetic Recording (HAMR) — Seagate
  - 200 mW laser heats surface to 750F, reducing even higher coercivity
  - Requires new head technology to mount laser
  - 40TB drives planned for 2023, expected limit of 10Tb/in^2
- Heat Dot Magnetic Recording (HDMR) - patterned platter with HAMR, projected to 100Tb/in^2

Disk Power

- Rotational power proportional to $P * R^{2.8} * D^{4.6}$
- P = platter count
- R = rotational speed (RPM)
- D = diameter of platters
- Head movement small in comparison

Seeking

- Time depends on weight of arm, strength of voice coil, distance to seek
- Speedup phase, coasting phase, slowdown phase, settling phase (servo guidance)
- Moving a few tracks is mostly resettling (more common for smaller platters)
- Moving 10s of tracks is speedup/slowdown
- Moving long distance is mainly coasting
- Controller keeps table of seek impulse quantities

Special Cases

- When moving one track (e.g., data continues on next track), essentially same as settle time
- Does not read from cylinder in parallel -- minor track misalignment. Switch to reading same track on another platter requires settling time
- Reading tries to get data before settling, then use ECC
- Write must wait for settling

Reading

- Signal is weak and noisy
- Must be amplified, converted from analog to digital at higher frequency than data bit rate
- Signal processing applied to extract bits from waveform
  - Even more processing for TDMR, combining two signals
- Bits then forwarded to ECC for check/correct

Disk Controller Caching

- RAM or NVRAM buffer for data going to/from disk
- Helps hide latency
- On reading, prefetch extra sectors
- On write, store data until seek/rotation brings sector into place
  - Multiple cached writes enable dynamic scheduling
- Rewrites of high error sectors from cache

Reliability Reducing Factors

- Vibration - greater chance of head crash, stresses on bearings
- Rotation speed, mass of platter assembly - wears spindle bearings
- Temperature (15^{o}C increase = 50% lower life)
- Frequency of access - stresses actuator bearings, more crash opportunities
- Power-down after long run time (bearing lubricant)

Discussion Question

Given a particular recording technology, there are two ways to increase disk capacity: (1) more platters, (2) bigger platters.

What are the tradeoffs for each, and how might they be affected by the change from linear recording to MAMR and HAMR?

Patterson SigMod 1988

A Case for Redundant Arrays of Inexpensive Disks (RAID)

Redundant Arrays Not New

- Mainframe companies had been building redundant disk arrays, with data spread over drives and augmented with ECC (IBM patent, 1978)
- Such arrays were large, expensive, and proprietary
- RAID introduced this technology for easy use with low-cost disks
- Proposed a standard set of configurations
- Showed benefits

Early RAID Array

- PC-style disks
- Mounted on racks
- Custom-built controller

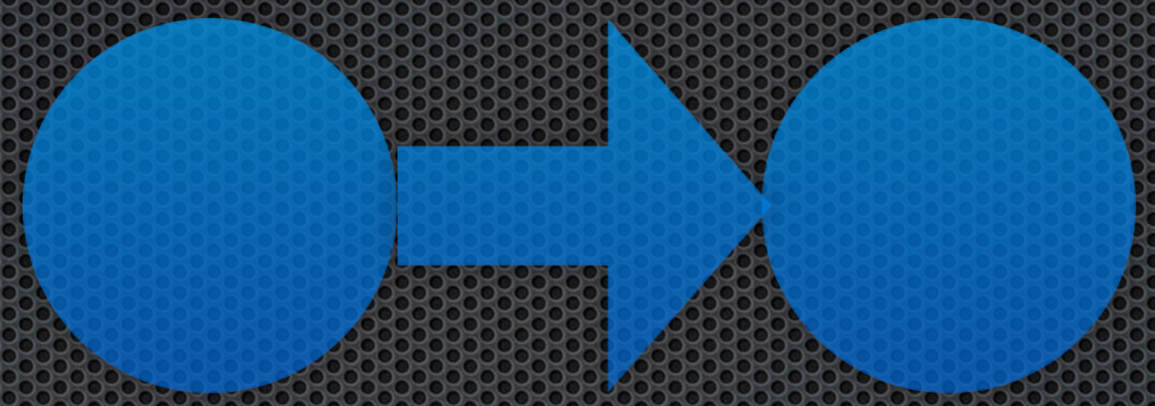
Motivation

- Computers getting faster
- Disks getting denser but not significantly faster
- Disks getting cheap
- Standard interfaces available (SCSI, ATA)
- Disks are unreliable

Arrays of Inexpensive Disks

- Reliability is Mean Time to Failure divided by number of disks in an array
- The more disks, the more failures
- Add redundancy to disk array
  - Extra check disks for ECC

RAID 1



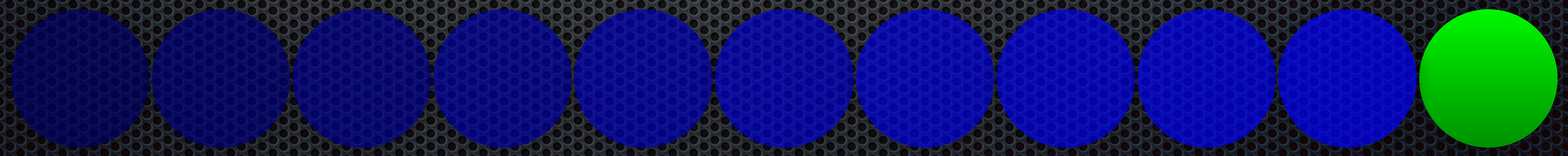
- Mirrored disks
- Data is automatically copied to a second disk
- Good reliability, very inefficient for many faults

RAID 2

- If enough disks, use a Hamming code for error correction
- e.g., 10 data disks need 4 check disks
- Good for larger accesses, slow for many short accesses

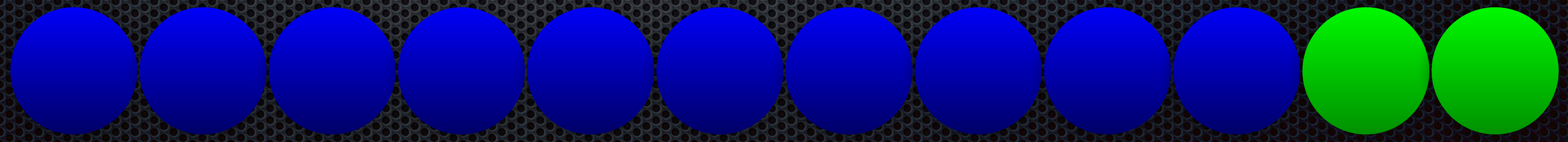
RAID 3

- Single check disk per group
- If error, use ECC to correct, or regenerate ECC if check disk failed
- Similar speed to RAID 2, but less expensive



RAID 4

- Interleave at sector level rather than bit level
- Check sectors locally
- Still have check disks to check/recover sector errors
  - Bottleneck
- Allow parallel accesses
- Better for small transfers



RAID 5

- Distribute check values
- No single check disk, so no bottleneck
- Parallel access
- Good for large and small accesses

Subsequent to Paper

- RAID 0 - no redundancy, just parallel striped disks
- RAID 6 - striped disks with correction for two errors
- RAID 1+0 - Set of mirrors + arrange as striped disks
- RAID 0+1 - Set up 2 stripes with mirrored set
- Hybrid (nested) - Some level + 0 (e.g., 5+0 or 50)
- Software RAID

Evaluation

- RAID drive failures tend to cluster
- If a failure is due to old age, the recovery scan of all disks in the array can push other disks of equal age into failure
- Different RAID levels offer varying benefits
  - Not equivalent to backups

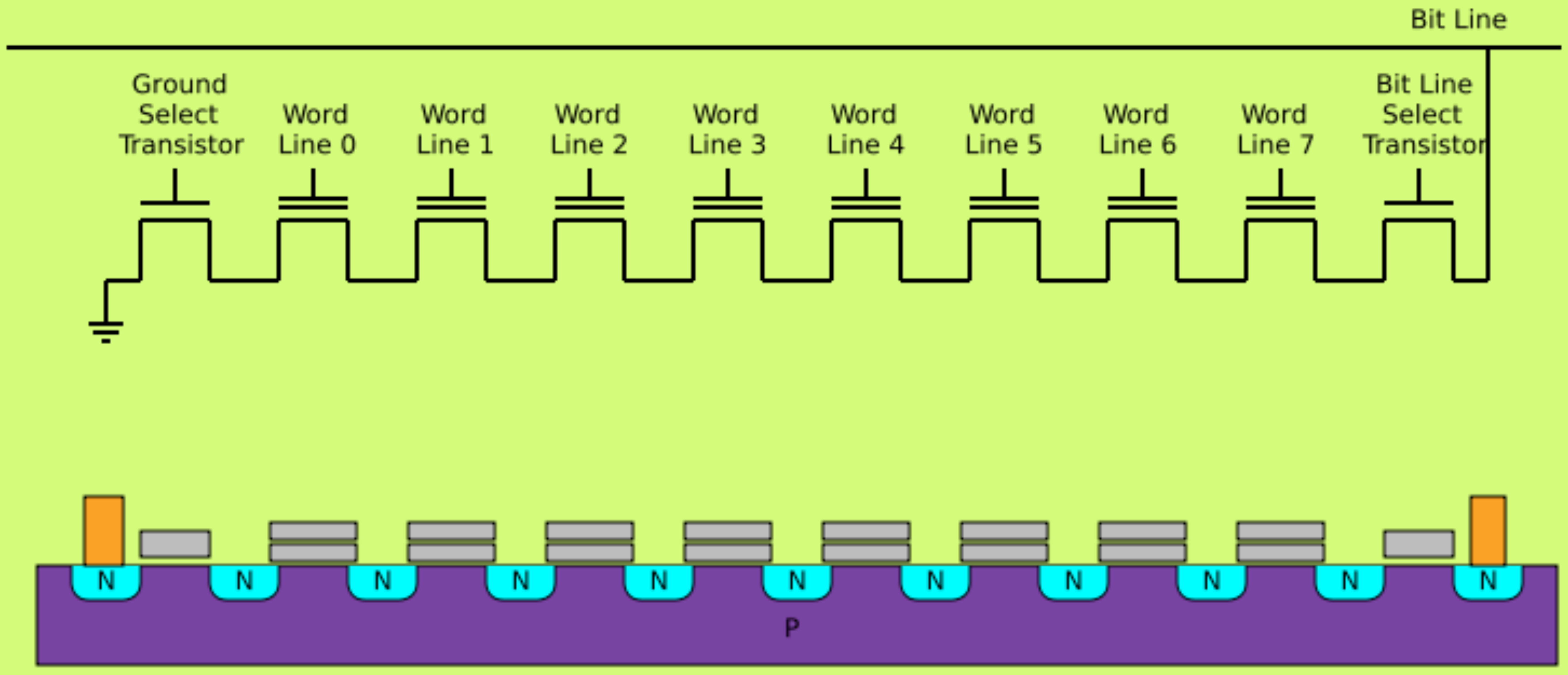
Flash Memory

Sorta like RAM, kinda like disk, but not really

Flash Memory

- Nonvolatile storage (up to a point)
- Traps charge on a floating gate
- Limited endurance (wearout)
- Comes in two forms: NOR (used in some kinds of consumer electronics) and NAND (used as general purpose storage)
- NOR is random access, slower, more expensive
- NAND is cheaper, faster, but not random access

NAND Flash Structure



Source: Wikipedia

NAND Flash Organization

- Arranged in planes, with blocks of pages (typically blocks contain 64 to 128 pages at, 2KB to 8KB per page). Planes can operate in parallel
- Whole pages are written at once by setting 1s to 0s
- Can rewrite pages, so data can effectively be stored in smaller units, though there are limits
- Erasure is by whole blocks only (reset to 1s), slower
- Reads are for whole pages

FTL - Flash Translation Layer

- Indirection table that maps logical to physical addresses
- Hides wear leveling and layout policies
- Also hides buffering, write coalescing, etc.
- Often seen as the point where Flash can be architected

SLC vs. MLC

- Single Level Cell holds a single bit
- Multi Level Cell holds two to four bits
  - MLC stores multiple levels of charge (typically 4, so two bits)
  - Requires incremental writes, levels vary with age
- SLC is faster, more reliable, more expensive
- MLC is slower, less reliable, cheaper

Parameters

	Minimum	Maximum
Endurance	10,000	1,100,000
Rand Read Latency (μs)	12	200
Typ Program Latency (μs)	200	800
Max Program Latency (μs)	500	2,000
Typ Erase Latency (ms)	1.5	2.5
Max Erase Latency (ms)	2	10
Typ Read Power (mW)	30	45
Max Read Power (mW)	60	90
Typ Program Power (mW)	30	45
Max Program Power (mW)	60	90
Typ Erase Power (mW)	30	45
Max Erase Power (mW)	60	90
Typ Idle Power (μW)	30	60
Max Idle Power (μW)	150	300

Where it Fits

- Slower, similar density, more power hungry than RAM
- Faster, physically more compact, lower power, more expensive than hard disk
- Less durable (shorter life) than both, although less sensitive to shock and vibration than hard disk
- Lower shelf life than disk or CD/DVD
- Replacing HDD for mobile, some desktop (or as hybrid HDD w/Flash cache)

Failure Modes

- Wearout due to charge carriers not fully returning on erasure
- SLC wearout in 10,000 to 100,000 erase/write cycles
- MLC wearout in 1,000 to 5,000 cycles
  - Causes permanent failure of bits
- Bit corruption due to nearby reads/writes
  - Causes soft errors that can usually be corrected

MLC

- Useful in high density consumer devices
  - Overwrite a small number of times -- music players, digital camera storage, etc.
- SSD bulk storage for cold files
  - Use RAM and SLC for hot files
  - Need to periodically refresh

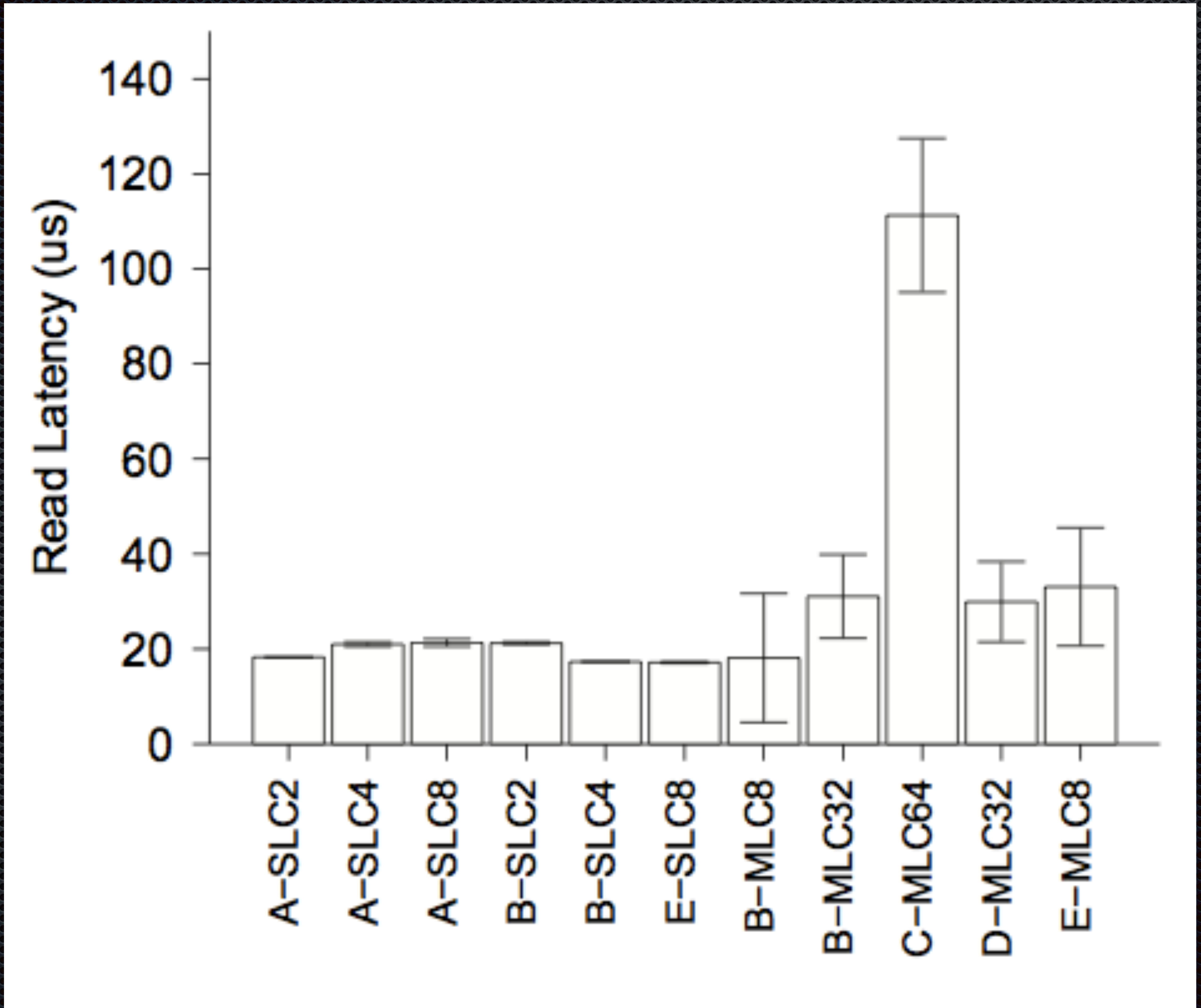
Laura Grupp Micro 2009

Characterizing Flash Memory: Anomalies, Observations, and Applications

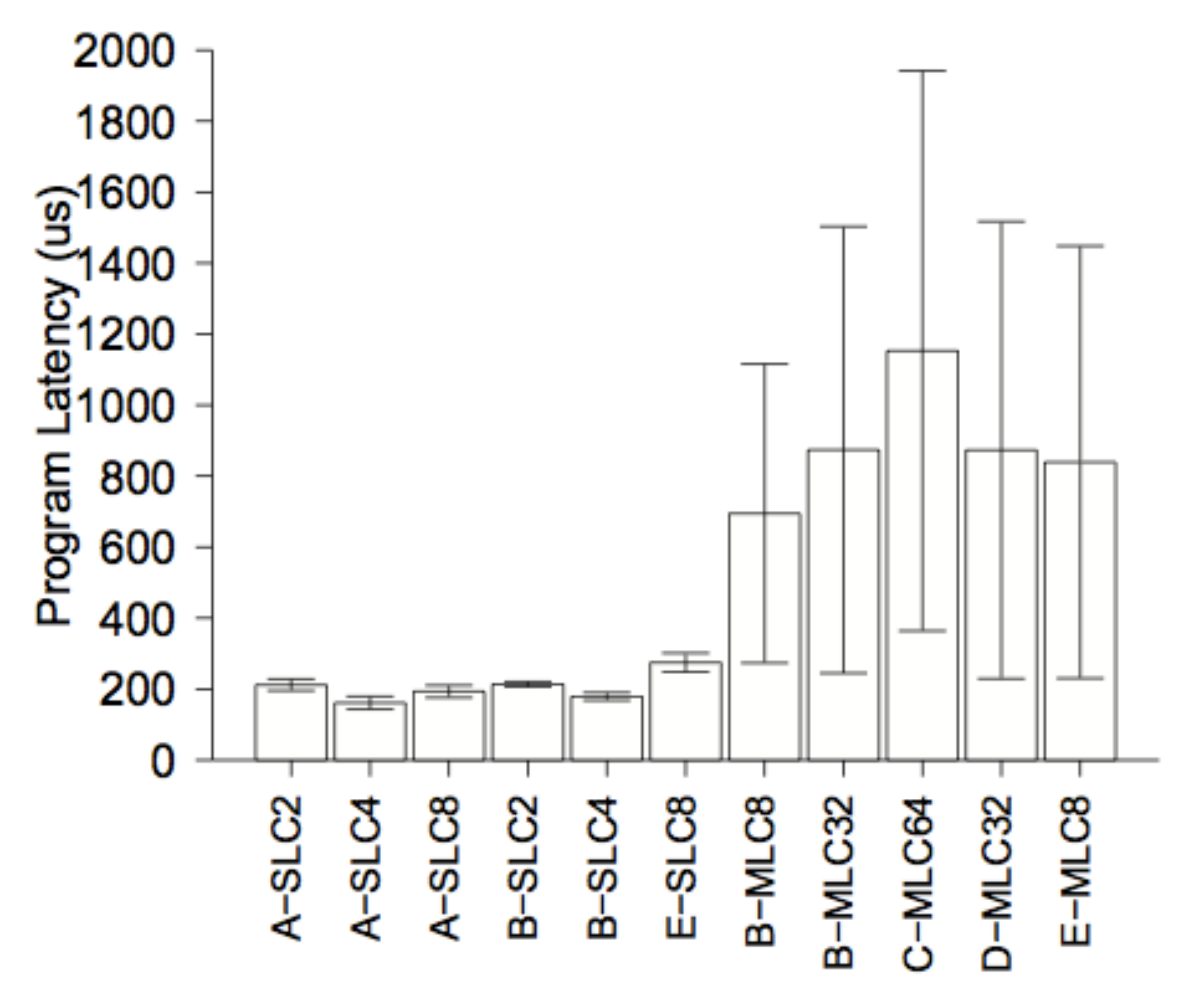
Specifications?

- Manufacturer specifications are purposely vague
- Actual behavior is different
- Behavior across chips varies
- Need to measure actual performance

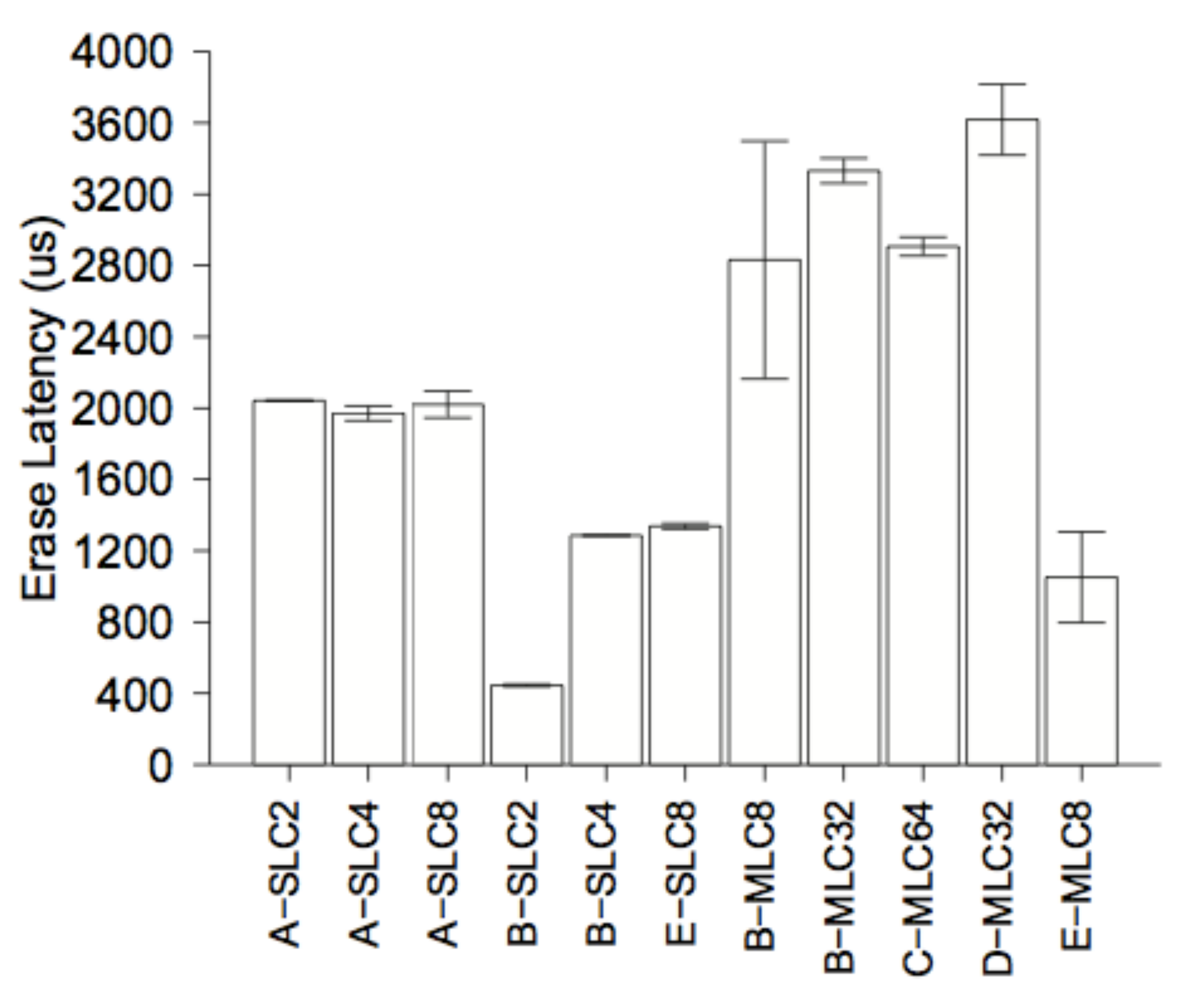
Measured Read Latency



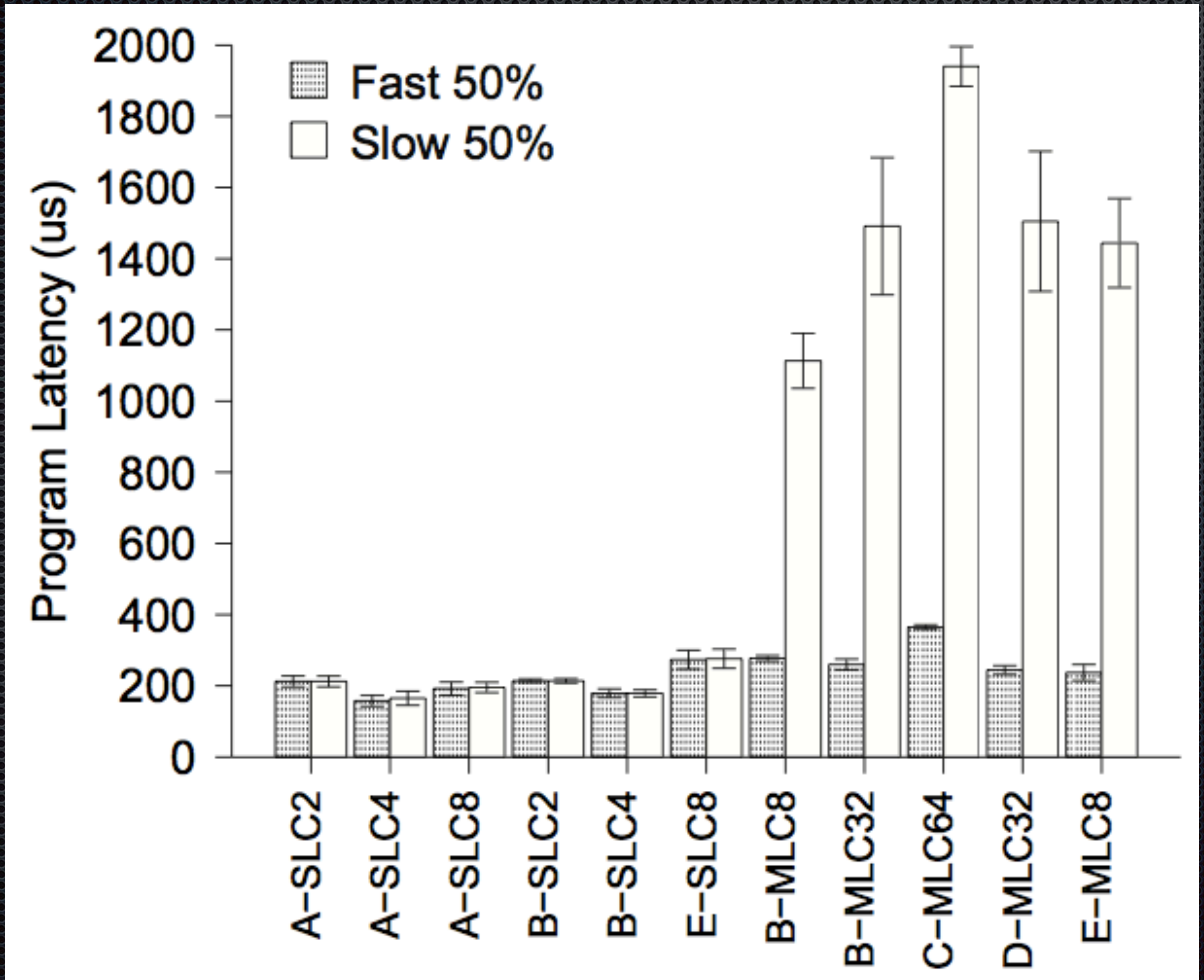
Measured Program Latency



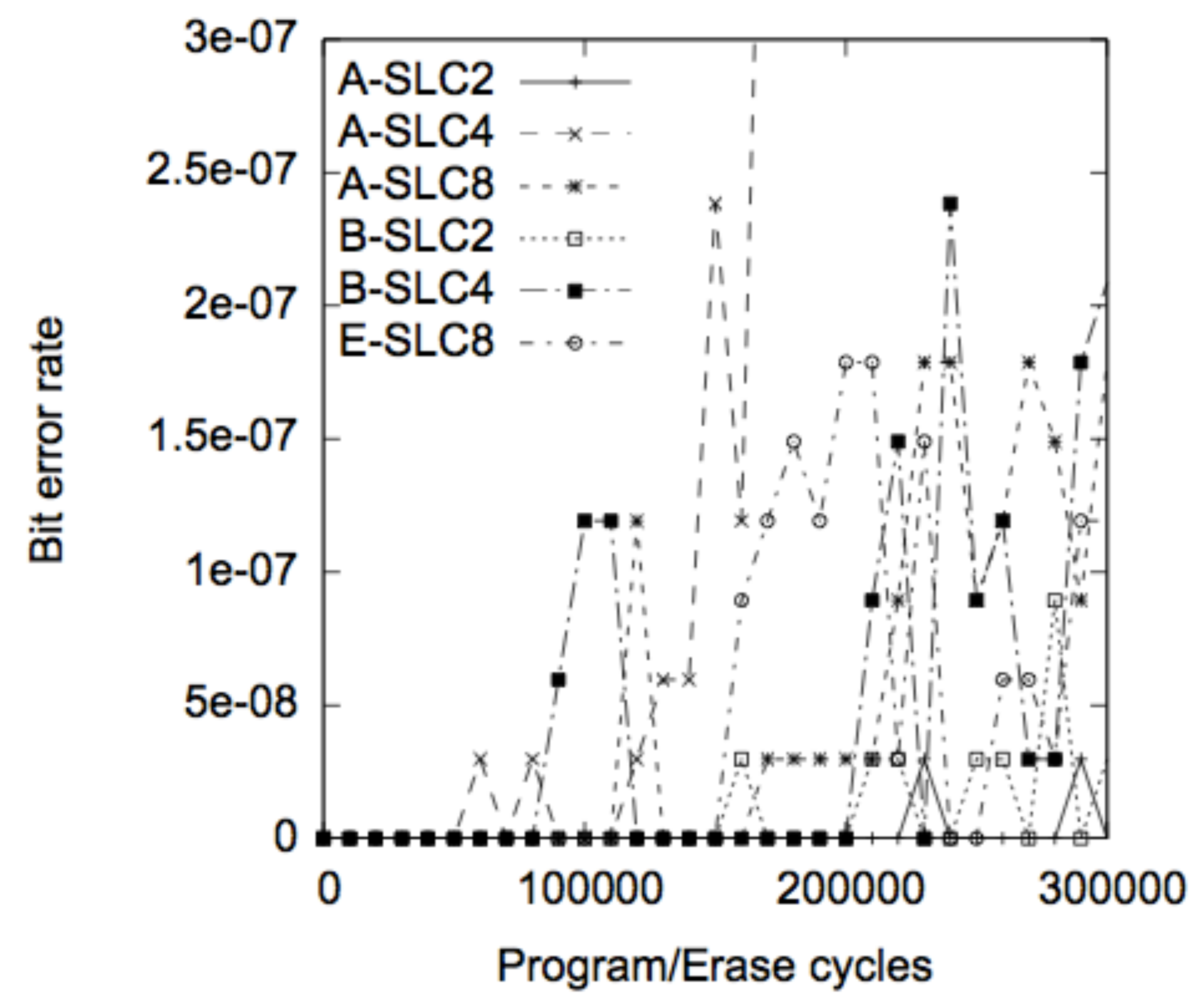
Measured Erase Latency



Program Latency Variance



SLC Measured Endurance



MLC Measured Endurance