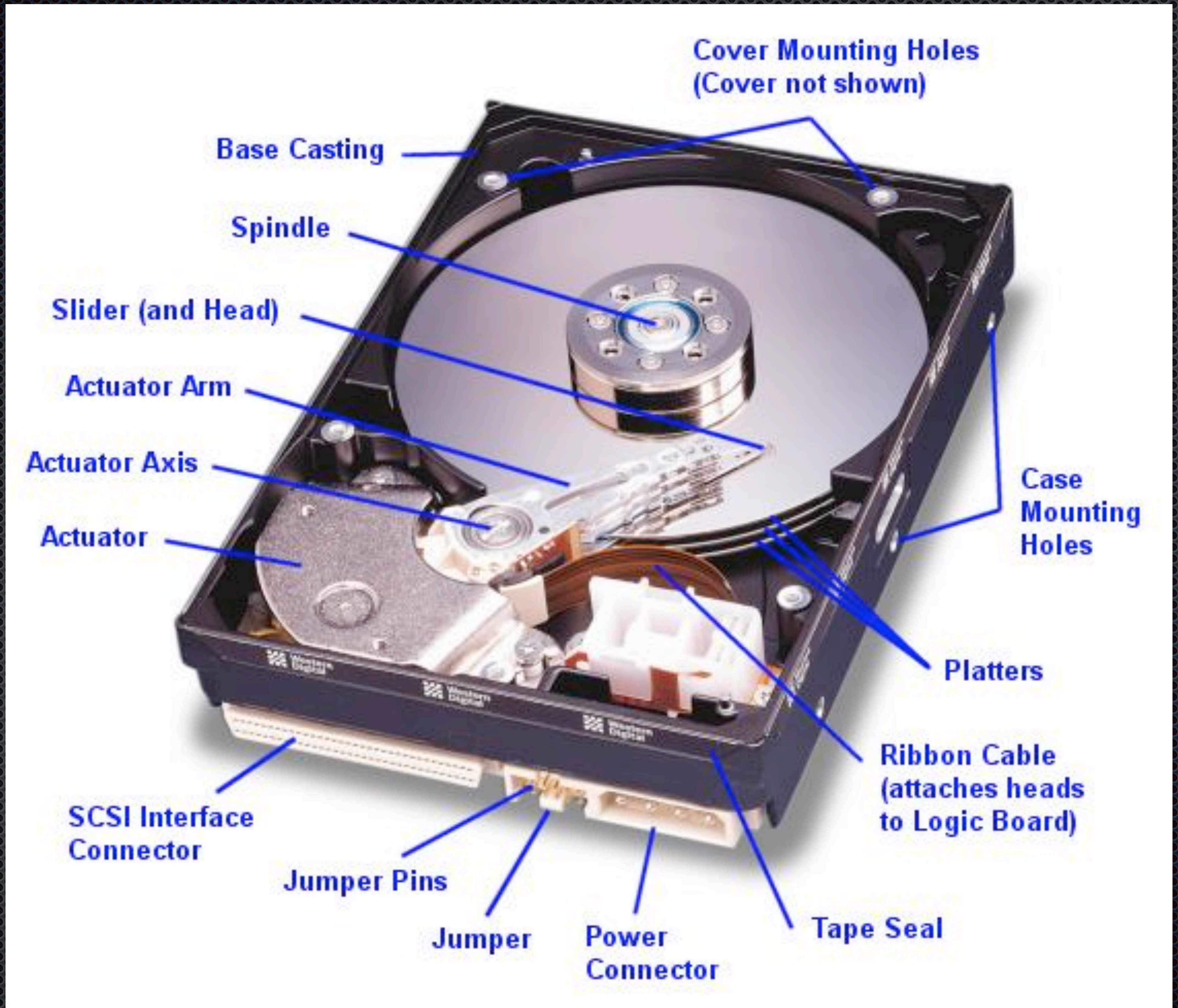
Disk Storage

Nonvolatile bulk memory

Basic Concepts

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

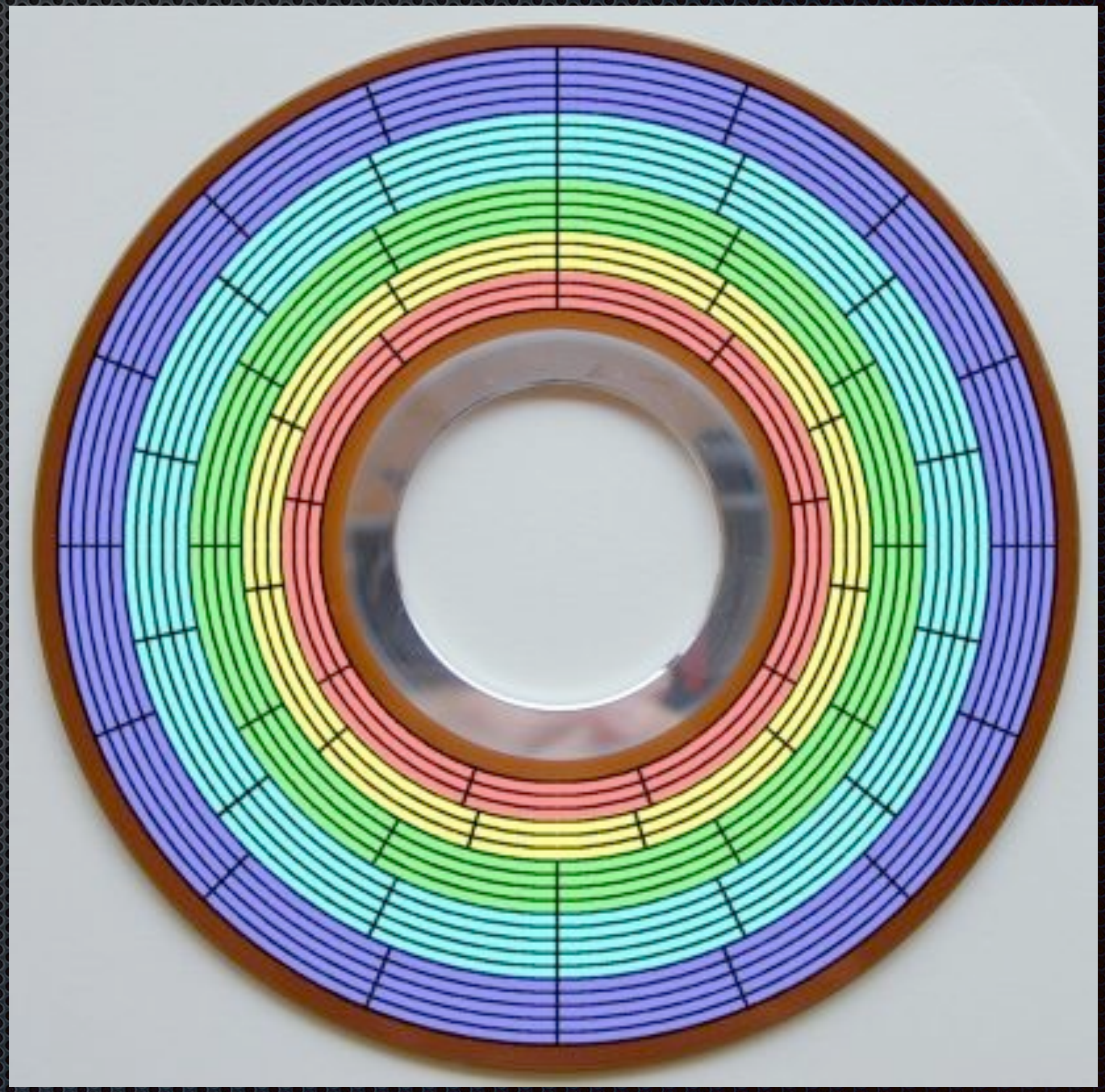
Structure



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

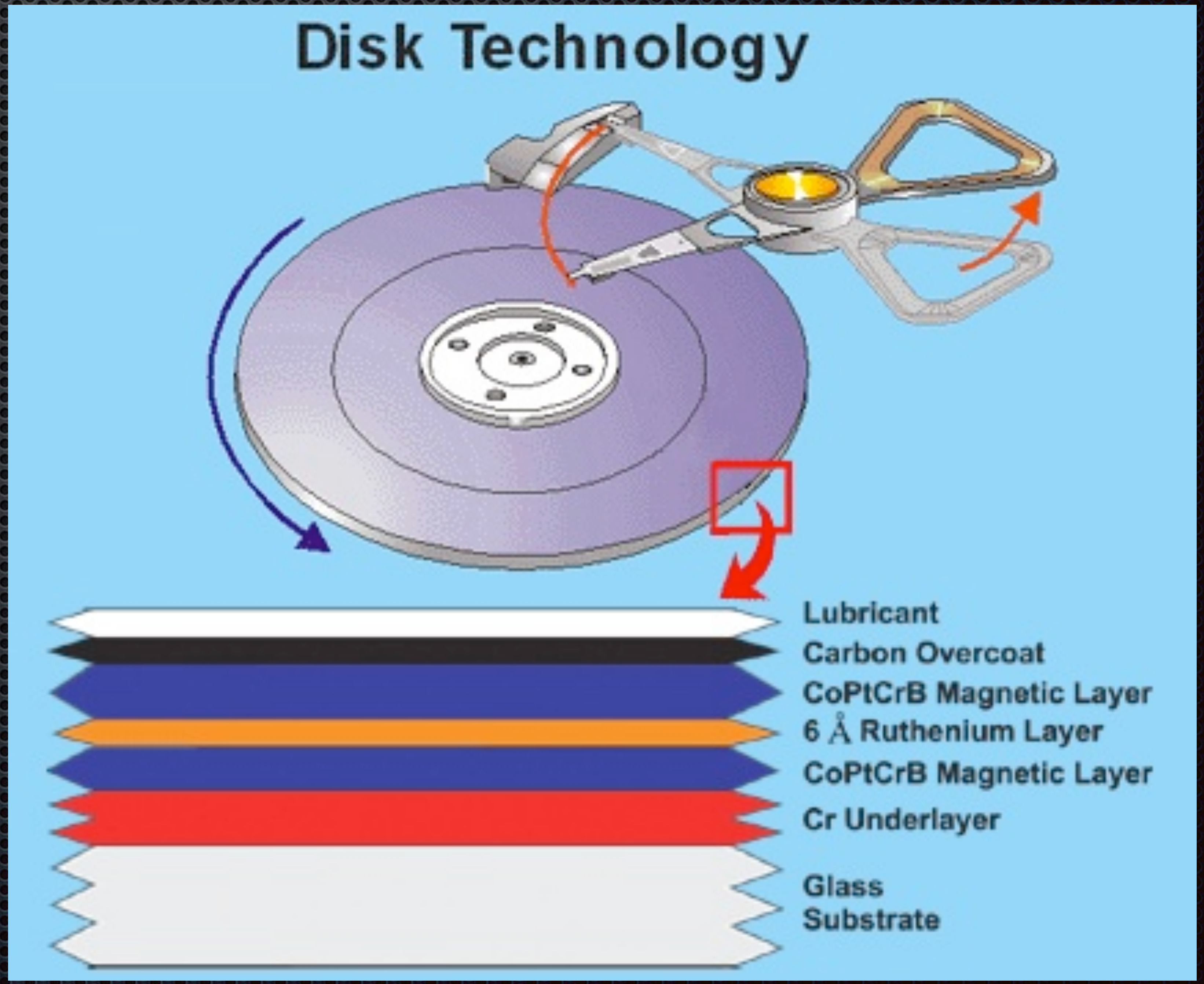
- Read errors are common
- Sectors include error correcting code
- Read and check for error -- if none, good
- If error, apply ECC to fix
- If not fixed, reread, try stronger correction
- If not recoverable, report error

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: 7200, 10,000, 15,000 RPM
- 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

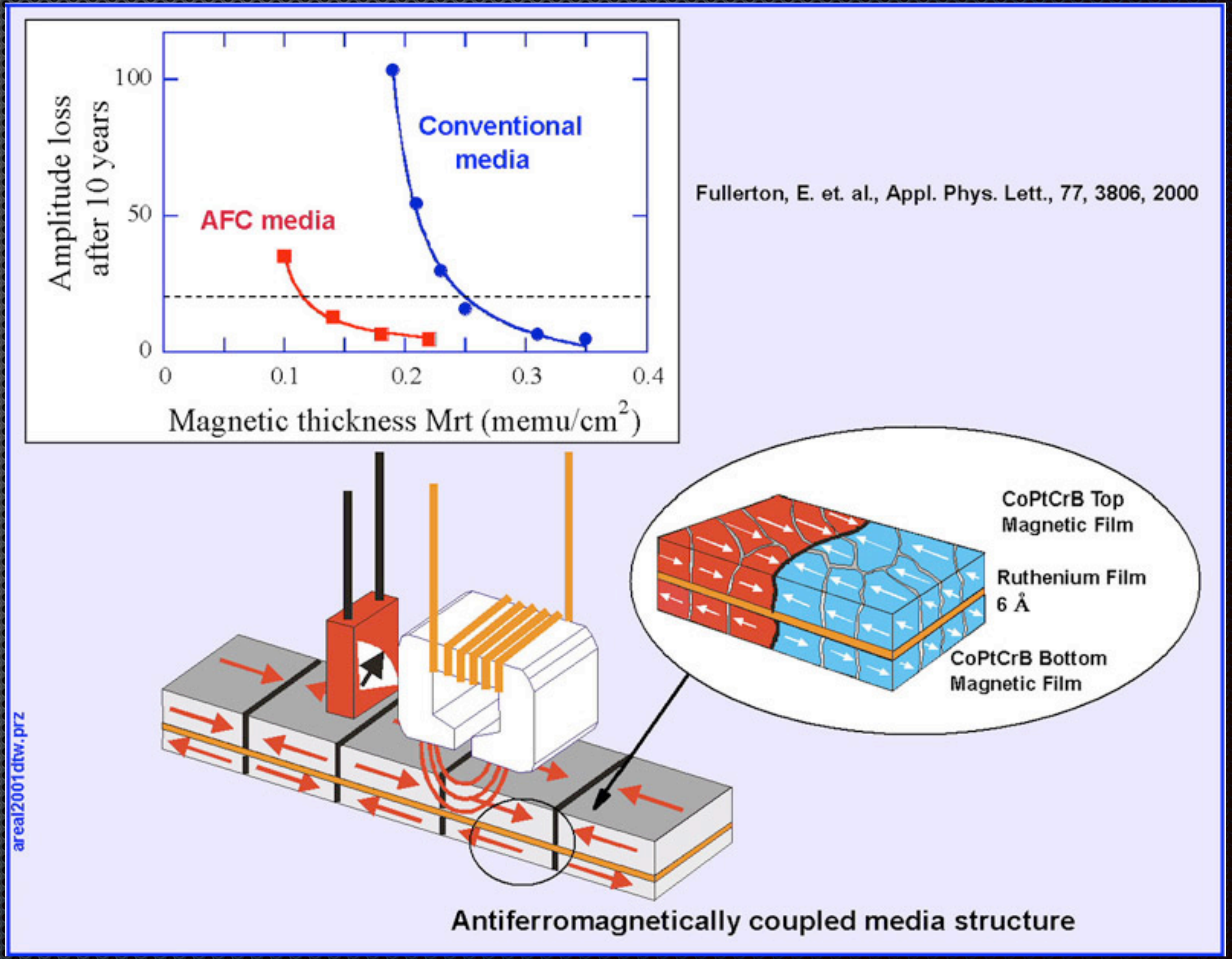


Thin Film

- 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

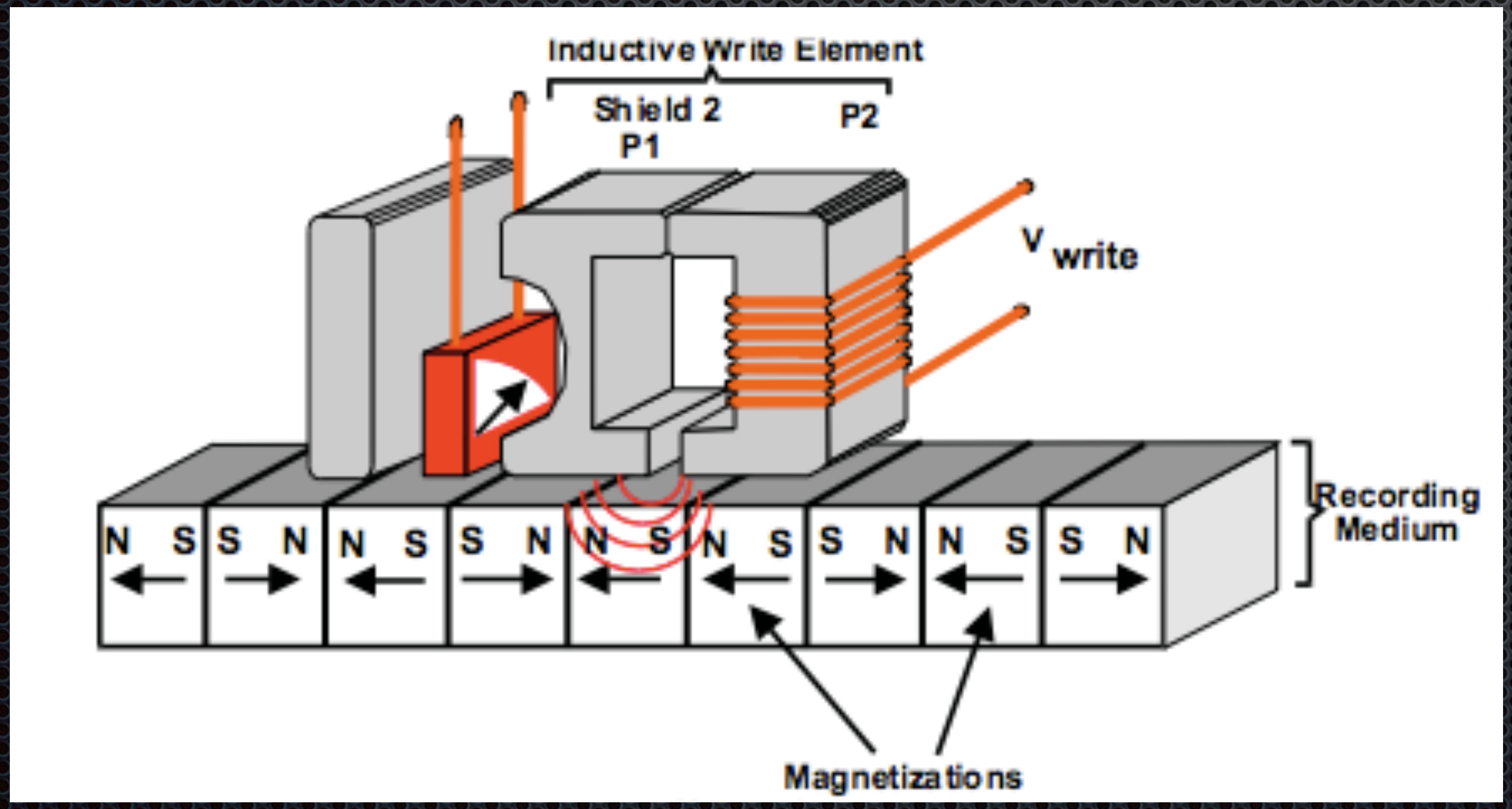
- Coupling layer between magnetic layers
- Effectively makes magnetization layer as thin as coupling layer (a few atoms)
- Allows thicker magnetic layers
- Extends life

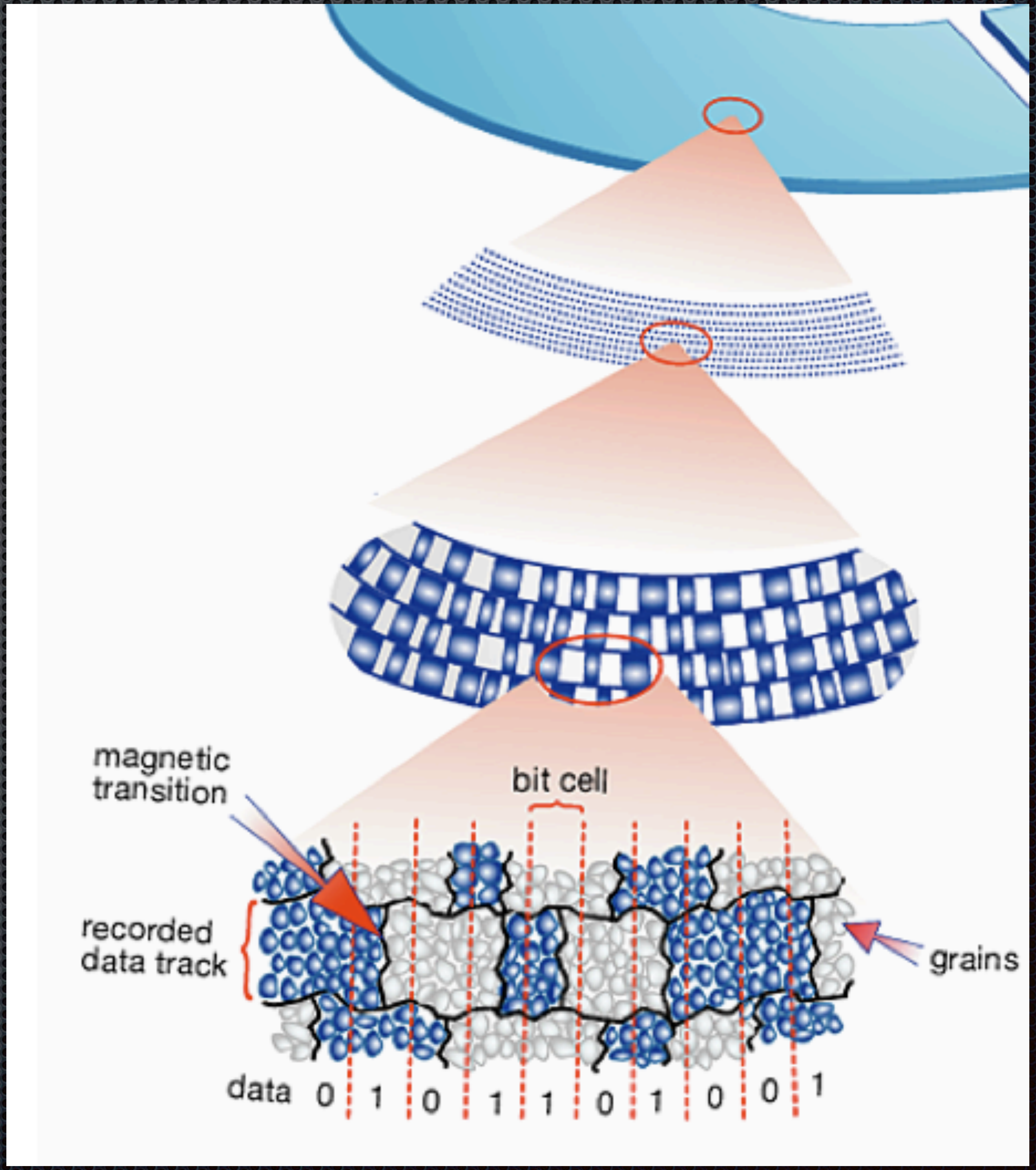


Figures from Hitachi Global Storage Technologies

Longitudinal Recording

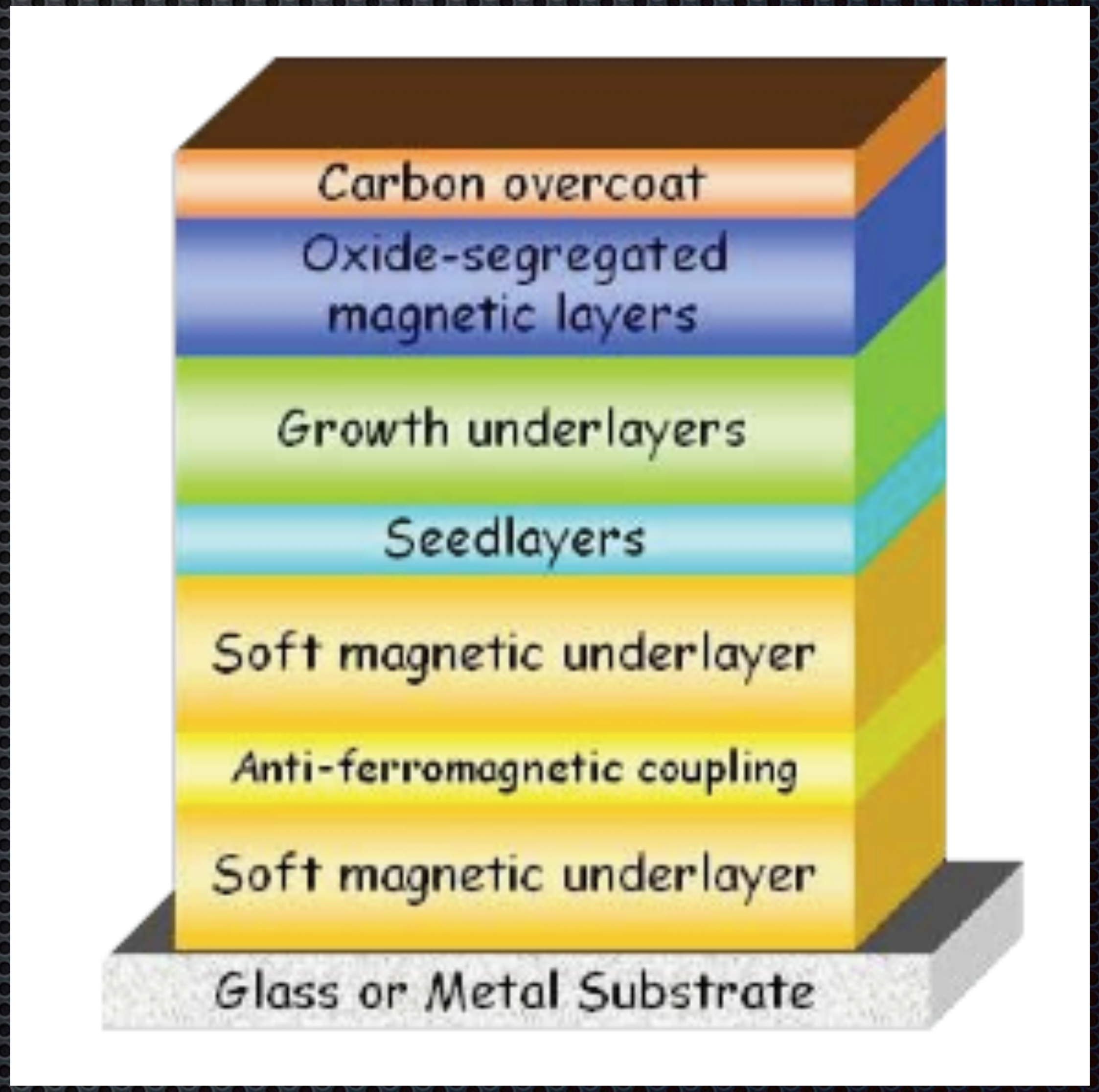
- Spots with same magnetic orientation = 0
- When orientation changes within spot = 1

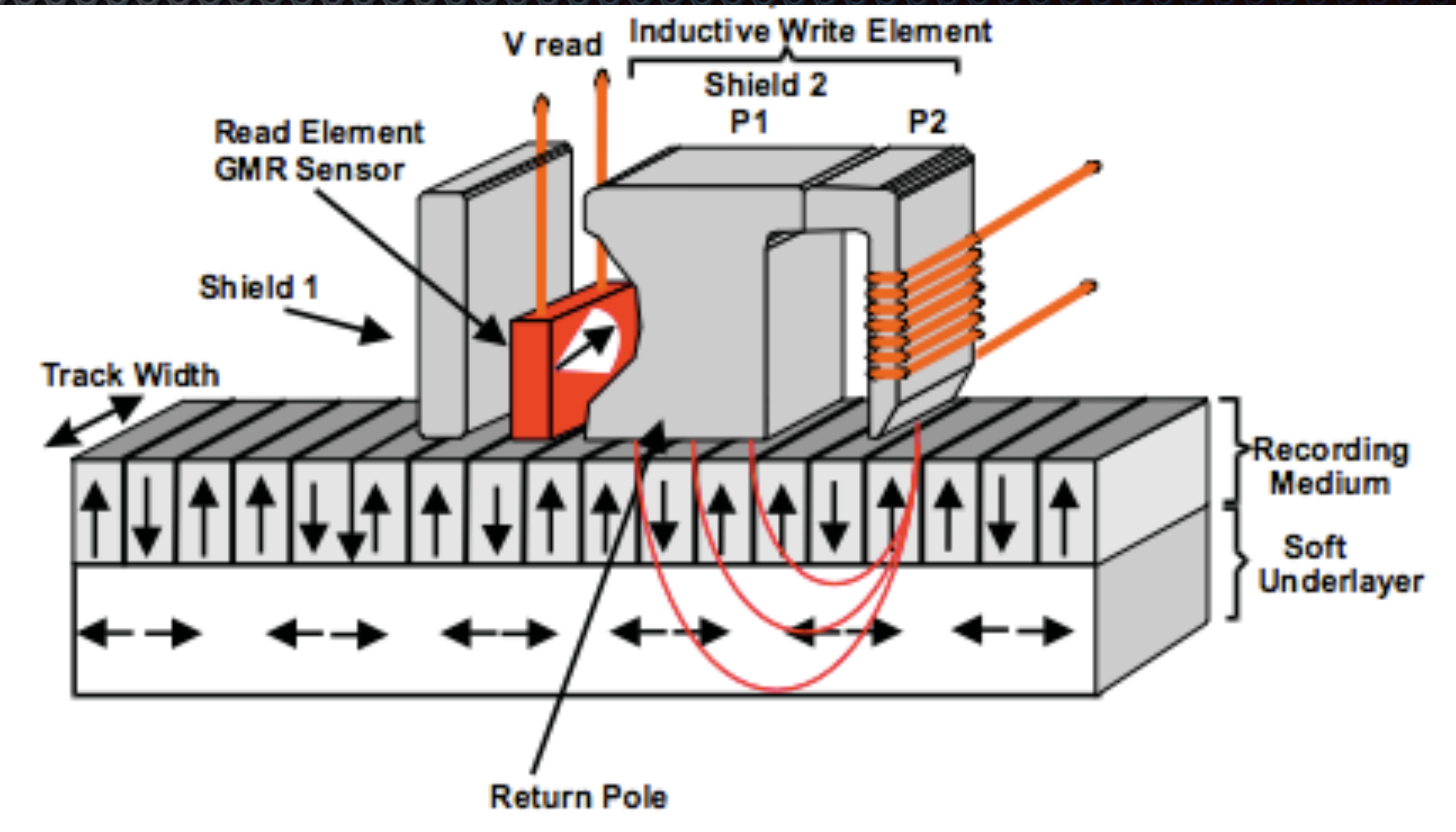




Perpendicular Recording

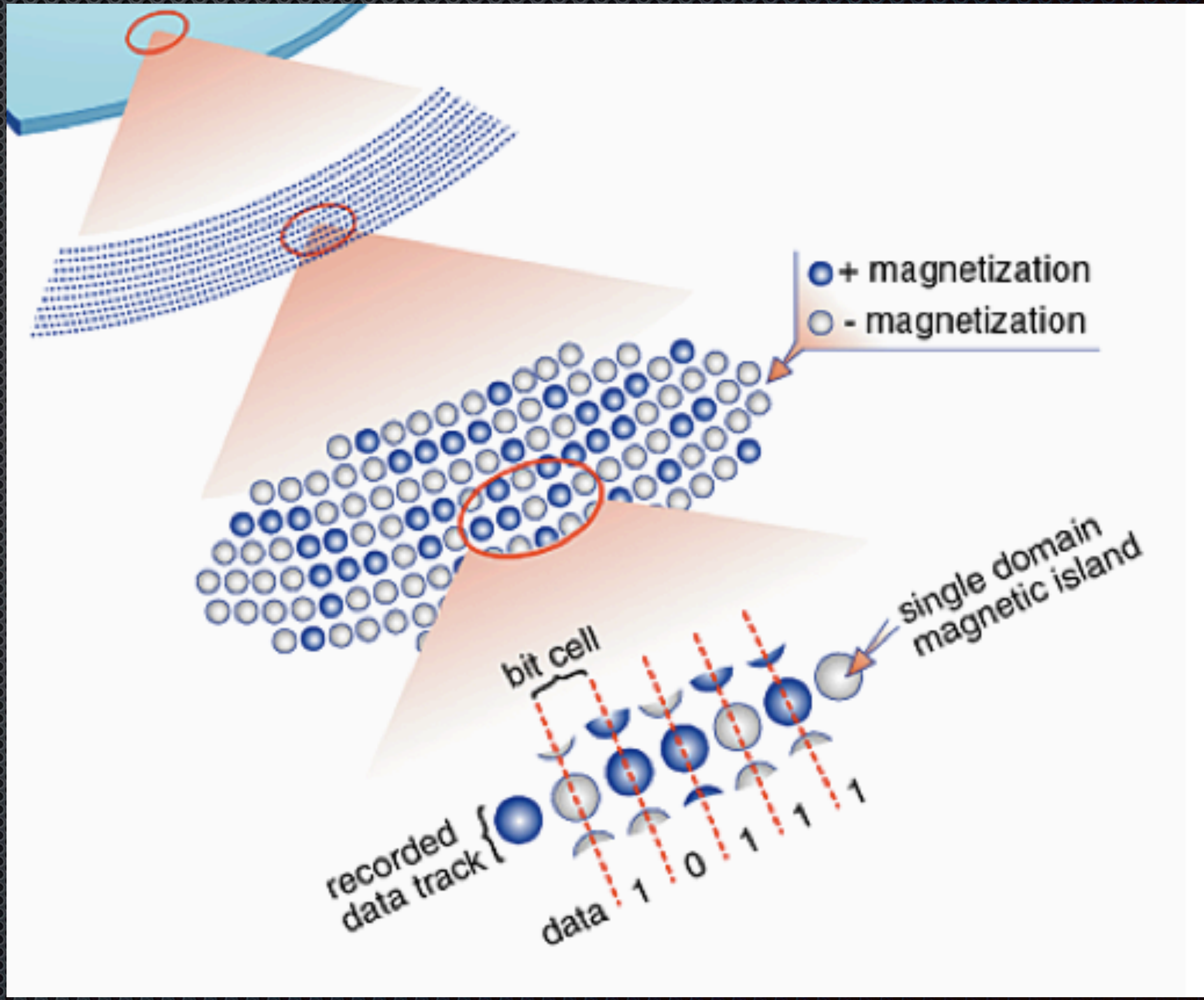
- New film layering with soft underlayer
- New form of write head
- Increases density without reaching thermal limit
- Density will eventually reach point that adjacent domains flip each other

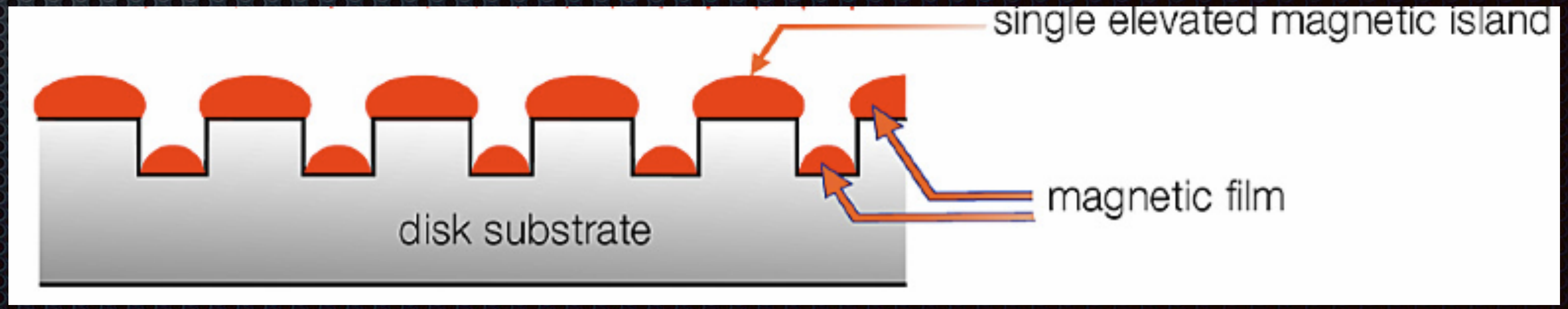




Patterned Recording

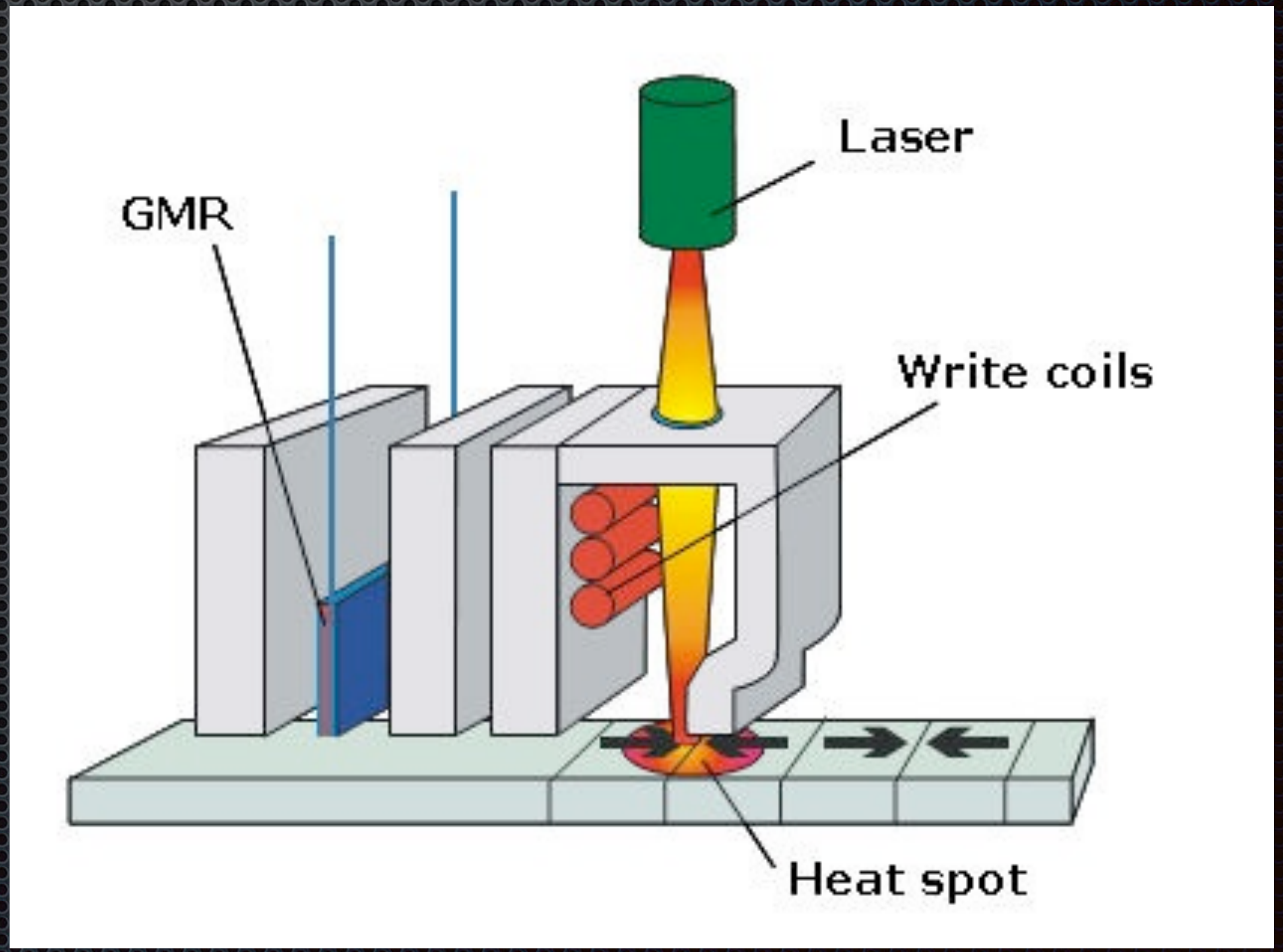
- Use lithography to texture surface for application of film
- Separates domains to avoid interference
- Creates rough surface
- More fabrication steps



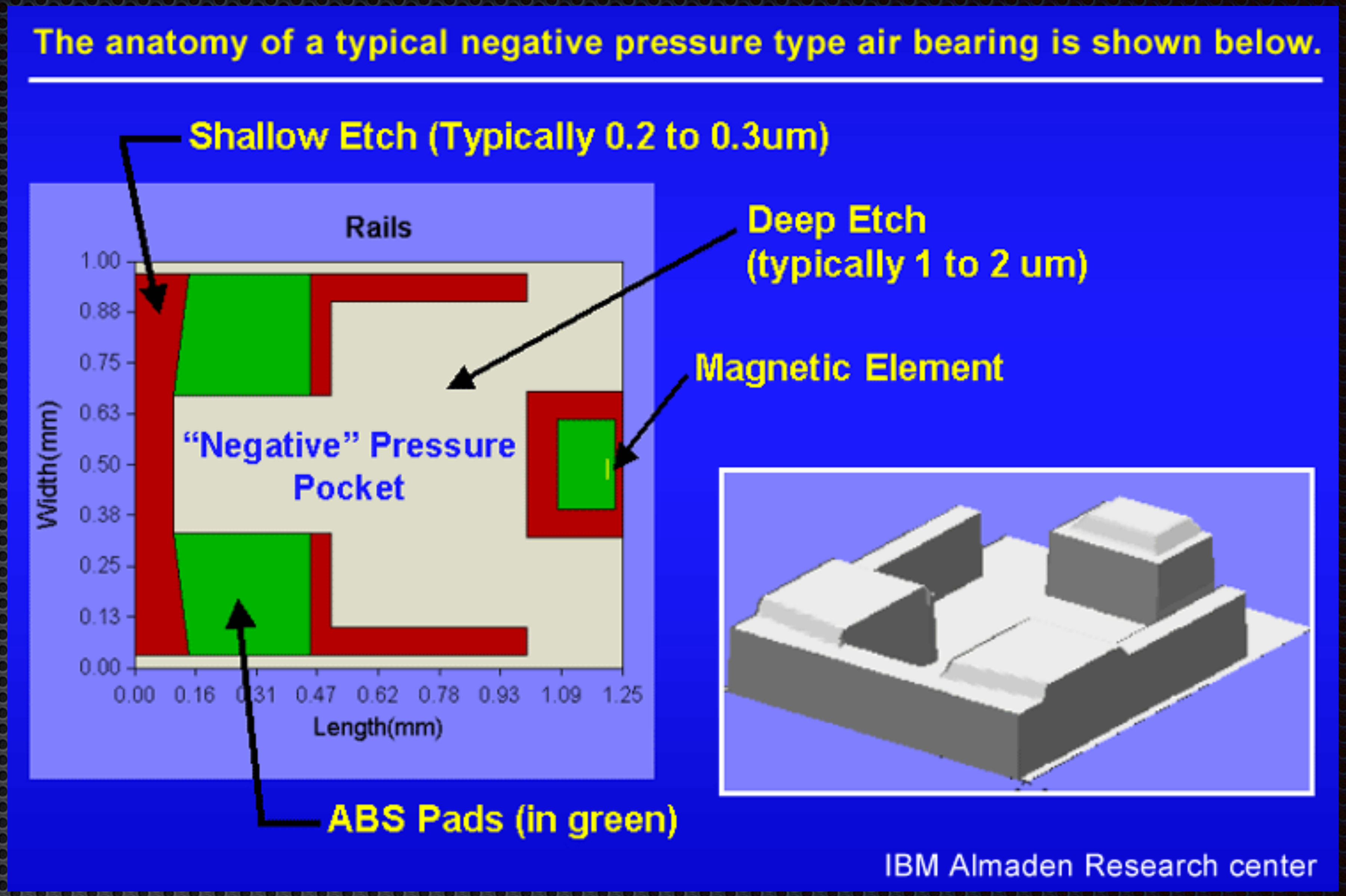


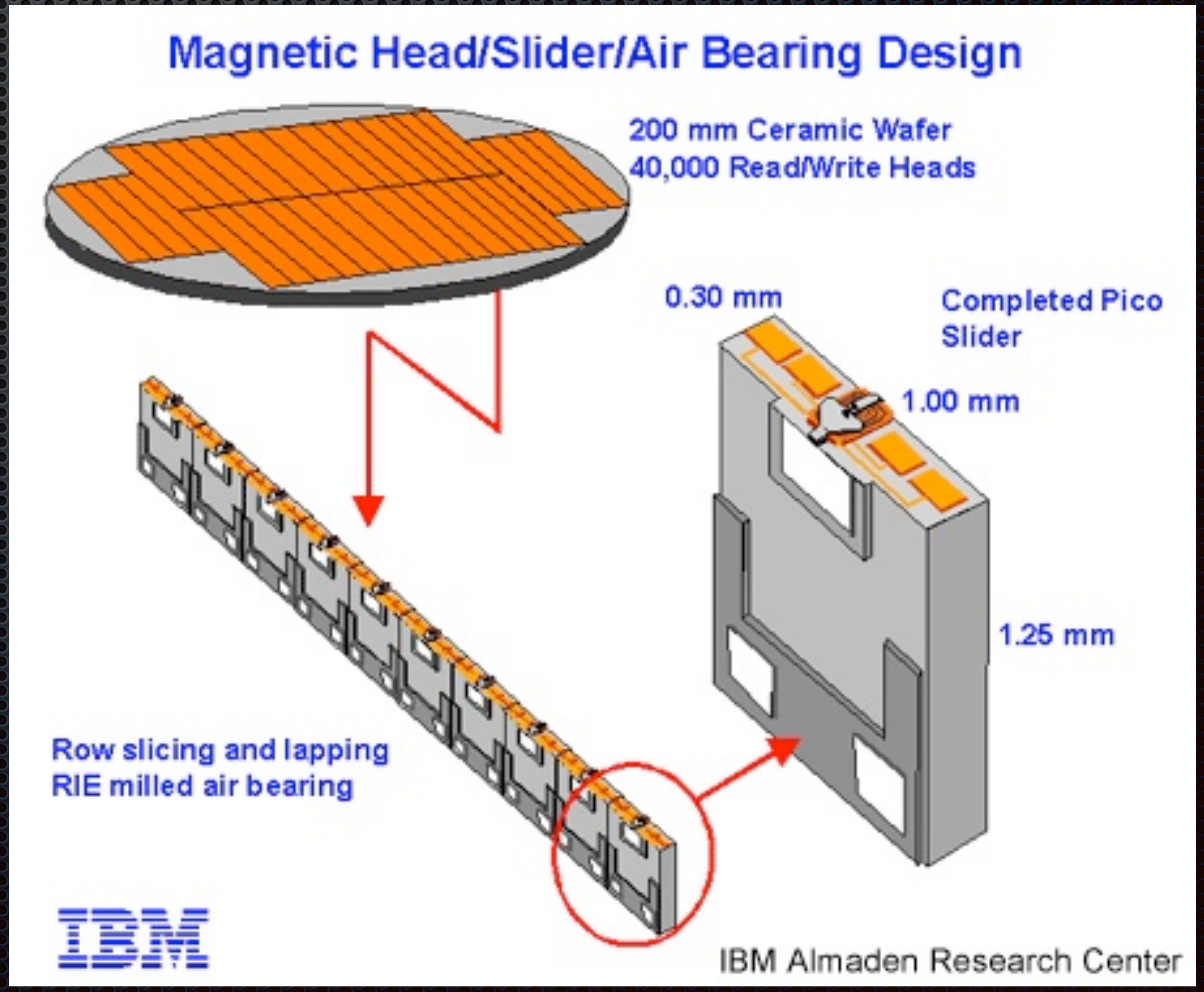
Thermally Assisted Recording

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



Slider

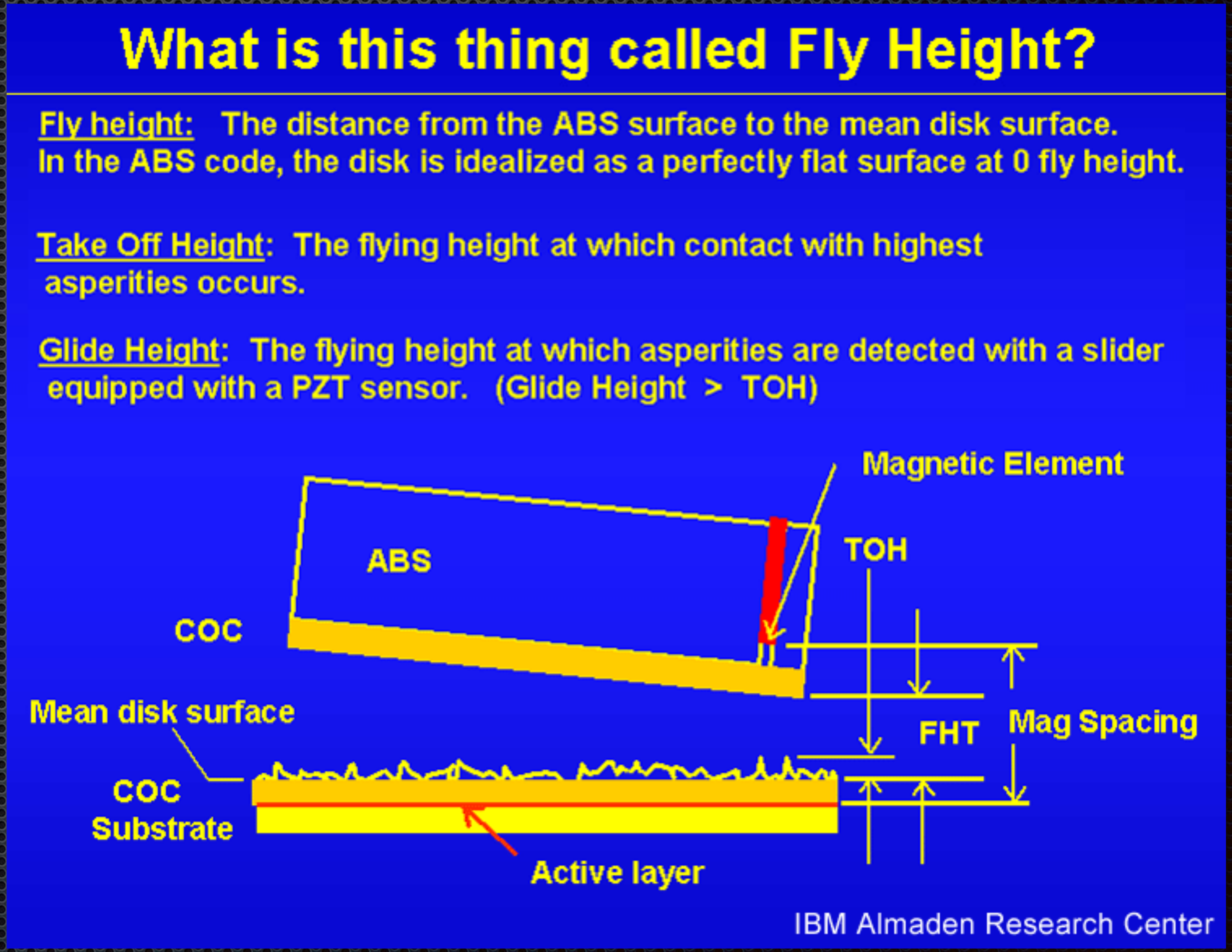




- Aerodynamic shape etched into underside of head to create proper lift and angle
- Electromagnet head attached to edge

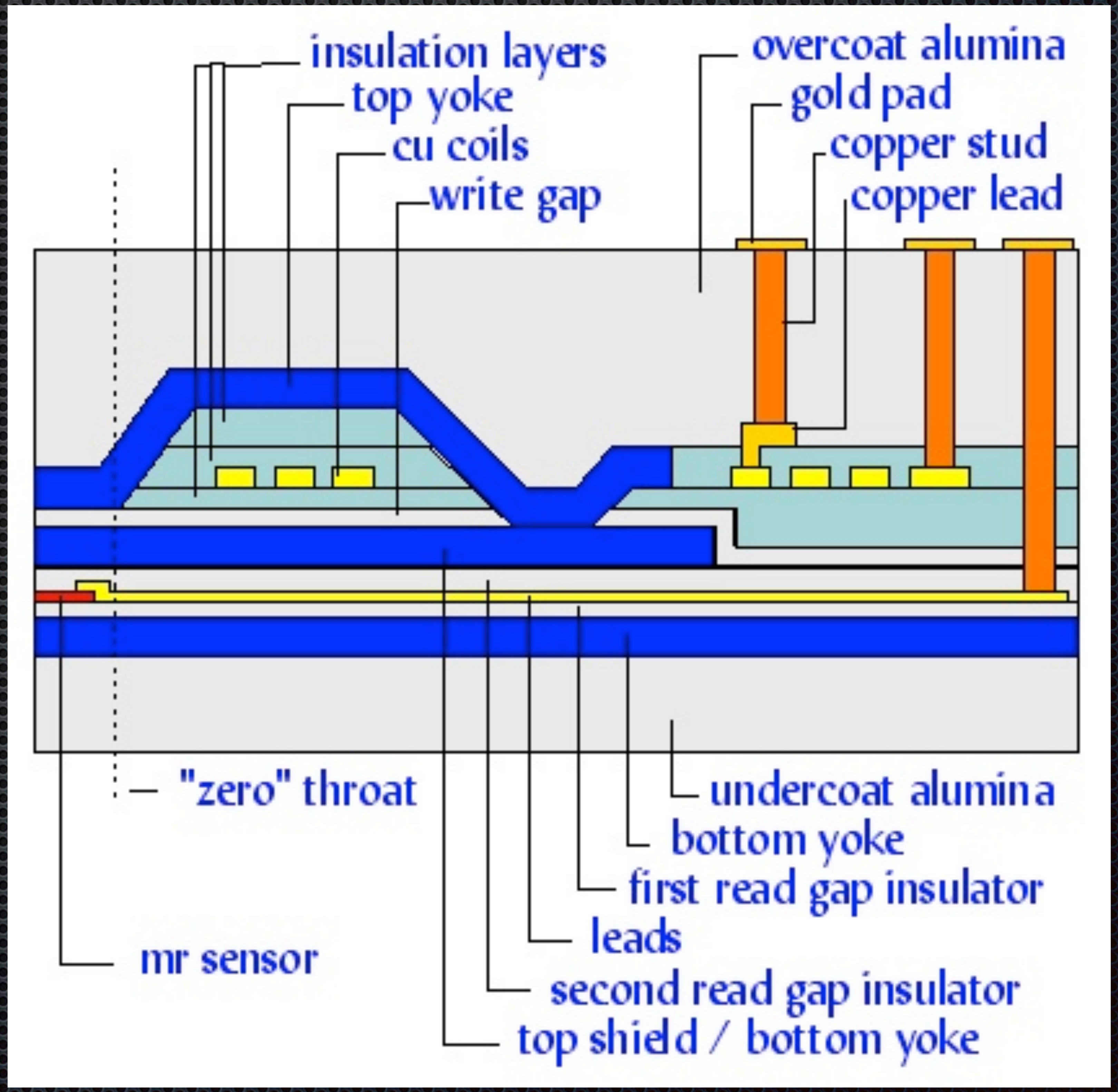
Read Head

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



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
- Current density about 1 Tb/in^2
- Expect growth of 100 before limit is reached
- Will lead to interesting shifts in research focus

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
- Bits then forwarded to ECC for check/correct

Disk Controller Caching

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

Reliability Factors

- Vibration
- Rotation speed, mass of platter assembly
- Temperature (15^{o}C increase = 50% lower life)
- Frequency of access
- Power-down after long run time (bearing lubricant)

Questions? Discussion?

Xue CODES 11

Emerging Non-Volatile Memories: Opportunities and Challenges

PCM

Attributes	DRAM	PCM	NAND
Non-Volatile	No	Yes	Yes
Erase Required	Bit	Bit	Block
Software	Simple	Simple	Complex
Power	~W/GB	100→500mW/die	~100mW/die
Write Bandwidth	~GB/s	1→100 + MB/s/die	10→100MB/s/die
Write Latency	~20-50ns	~1µs	~100µs
Write Energy	~0.1nJ/b	<1nJ/b	0.1-1nJ/b
Read Latency	50ns	50-100ns	10-25µs
Read Energy	~0.1nJ/b	<<1nJ/b	<<1nJ/b
Idle Power	~W/GB	<<0.1W	<<0.1W
Endurance	∞	10^8	$10^5 \rightarrow 10^4$
Data Retention	ms	Not *f* (cycles)	*f* (cycles)

Figure 4: A comparison of PCM with DRAM, NAND[10].

PCM operation

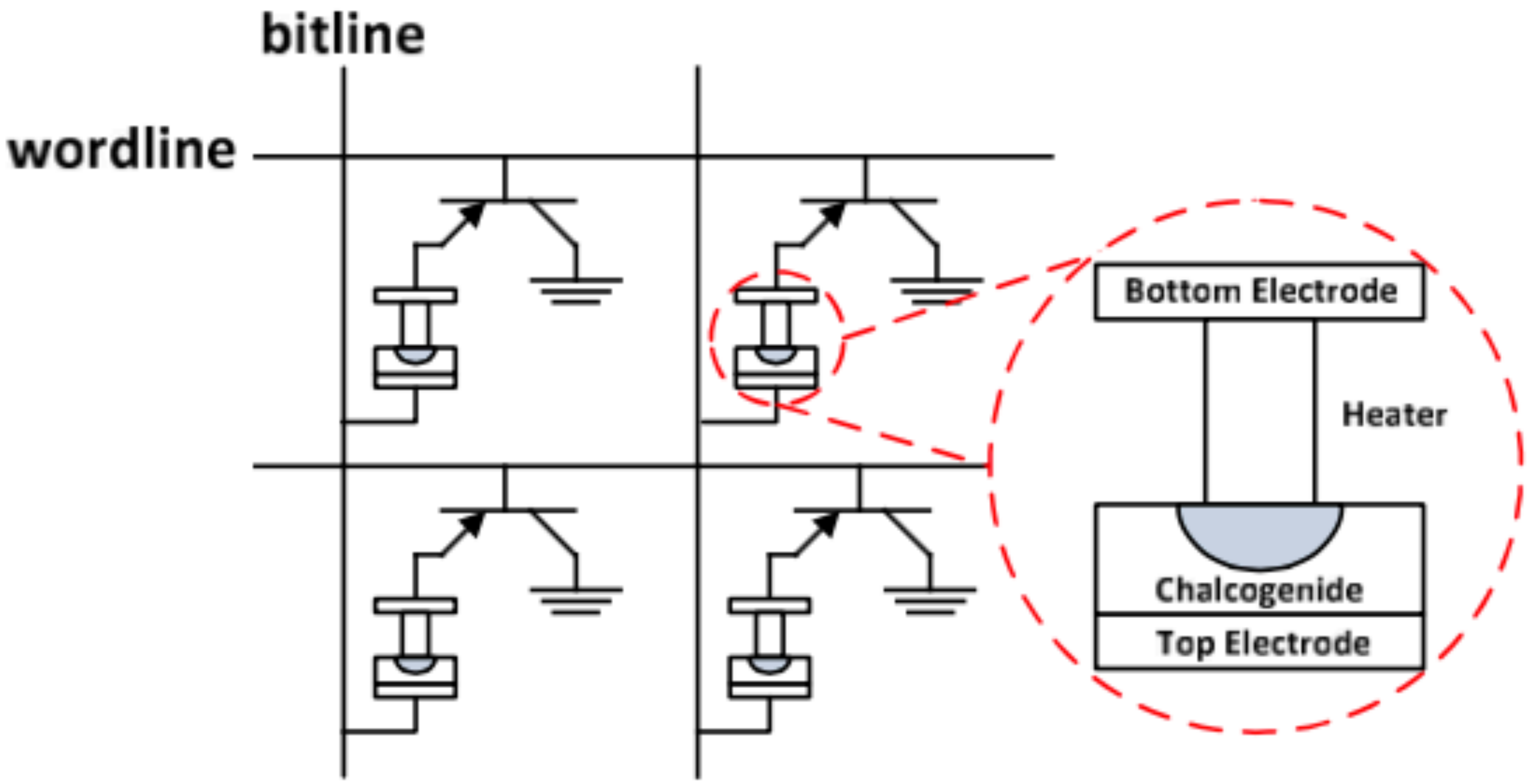


Figure 1: PCM cell array.

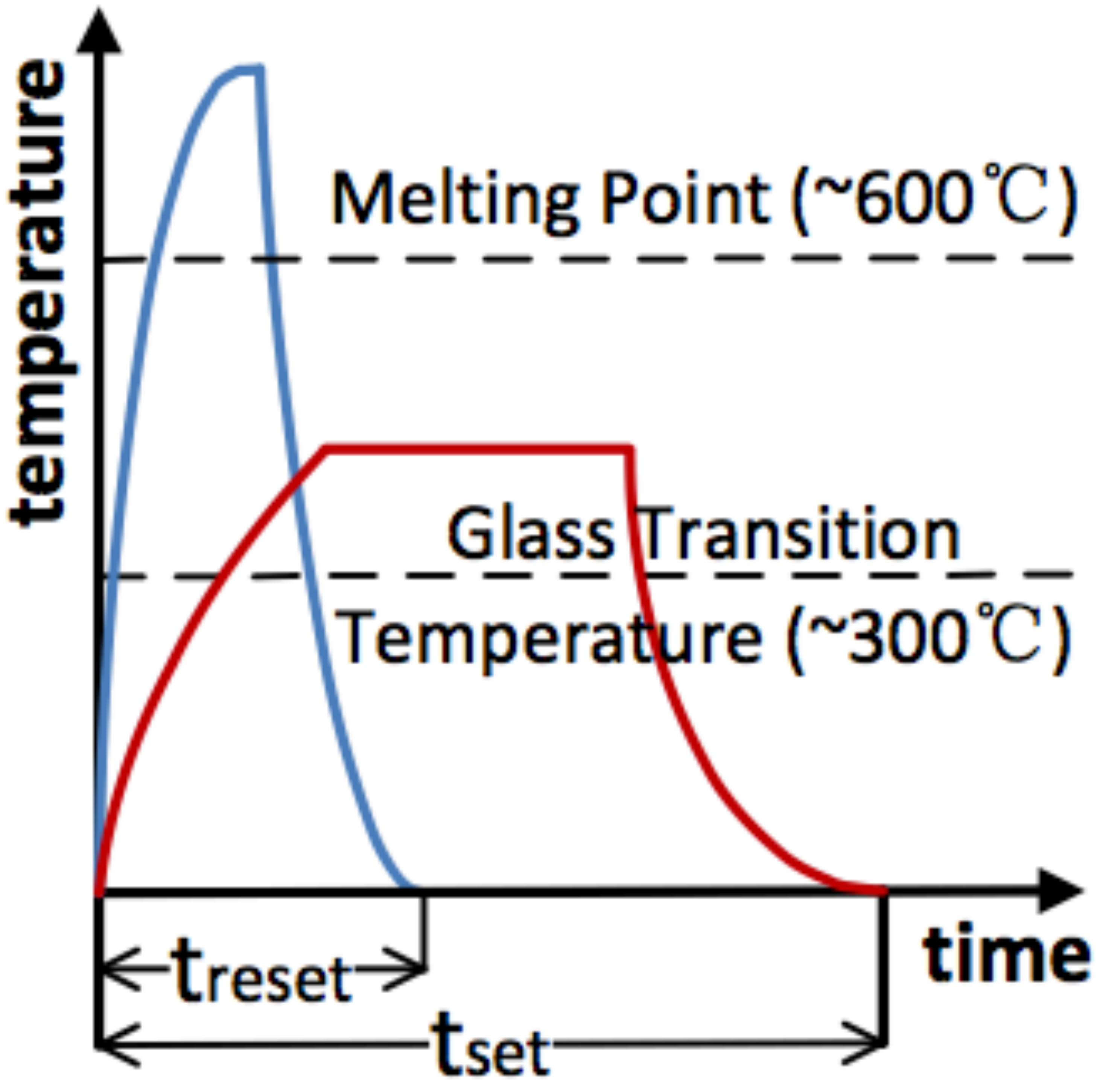


Figure 2: RESET and SET operations..

PCM Summary

- Non-volatile, low power, wearing
- Slow to write (iterative), fast to read
- Potentially higher density than RAM
- Still needs RAM for speed and wear reduction
- Also needs wear-leveling layer, like flash

Spin-Torque-Transfer RAM

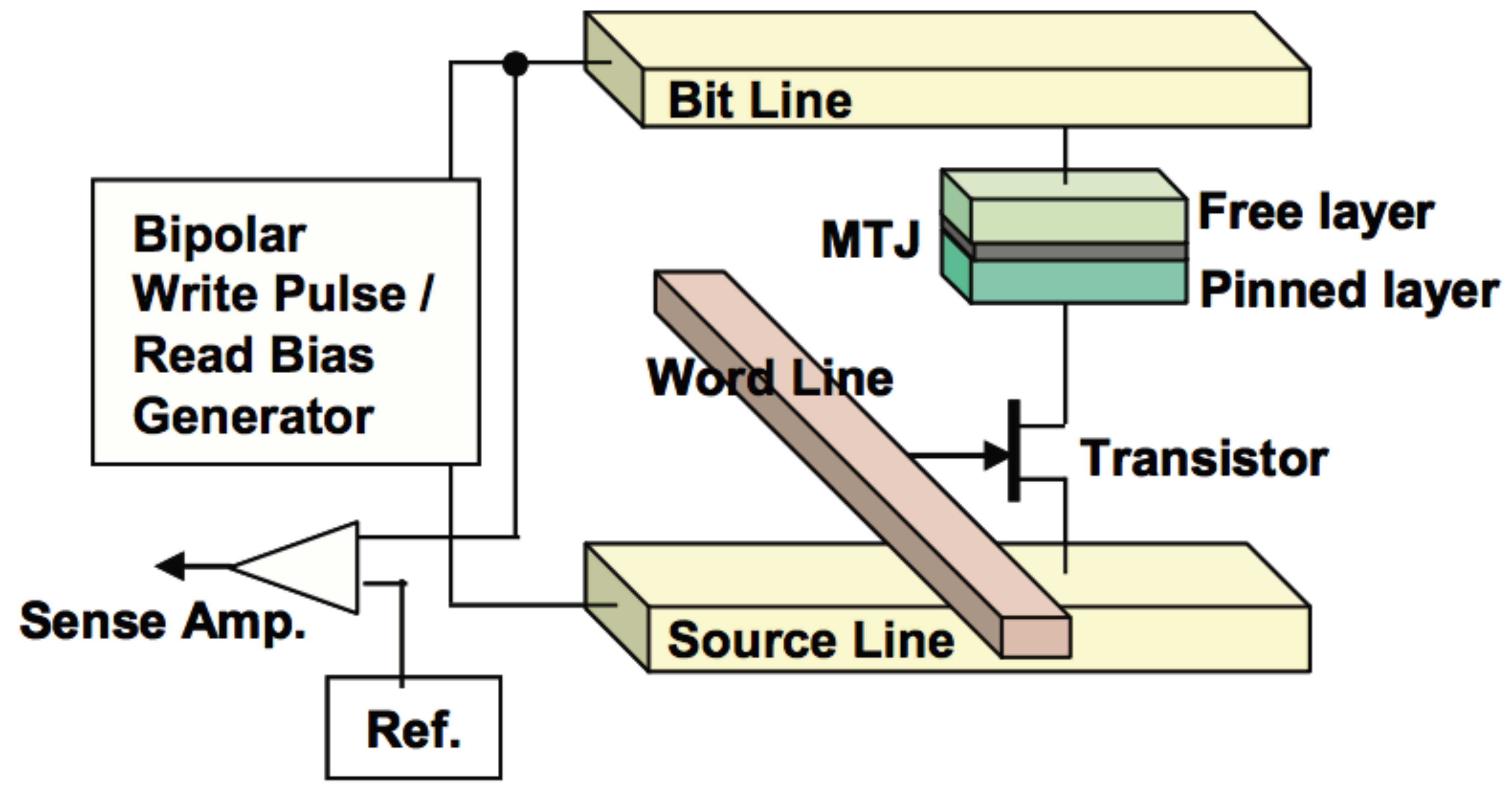


Figure 6: An illustration of a "1T1J' STT-RAM cell.

Cache size	Area	Read Latency	Write Latency	Read Energy	Write Energy	Standby Power
1M SRAM	36.2 mm^2	2.252 ns	2.244 ns	1.074nJ	0.956nJ	1.04W
4M STT-RAM	36.0 mm^2	2.318 ns	6.181 ns	0.858nJ	2.997nJ	0.125W

Magnetic junction changes resistance depending on states of Free and Pinned layers

STT-RAM Summary

- About 1/4 size of SRAM, similar interface
- A few times slower to write, more energy to write
- Reads similar to SRAM, potential for caching
- Lower idle power
- Can be made into multi-level cells, but wear occurs

MemRistors

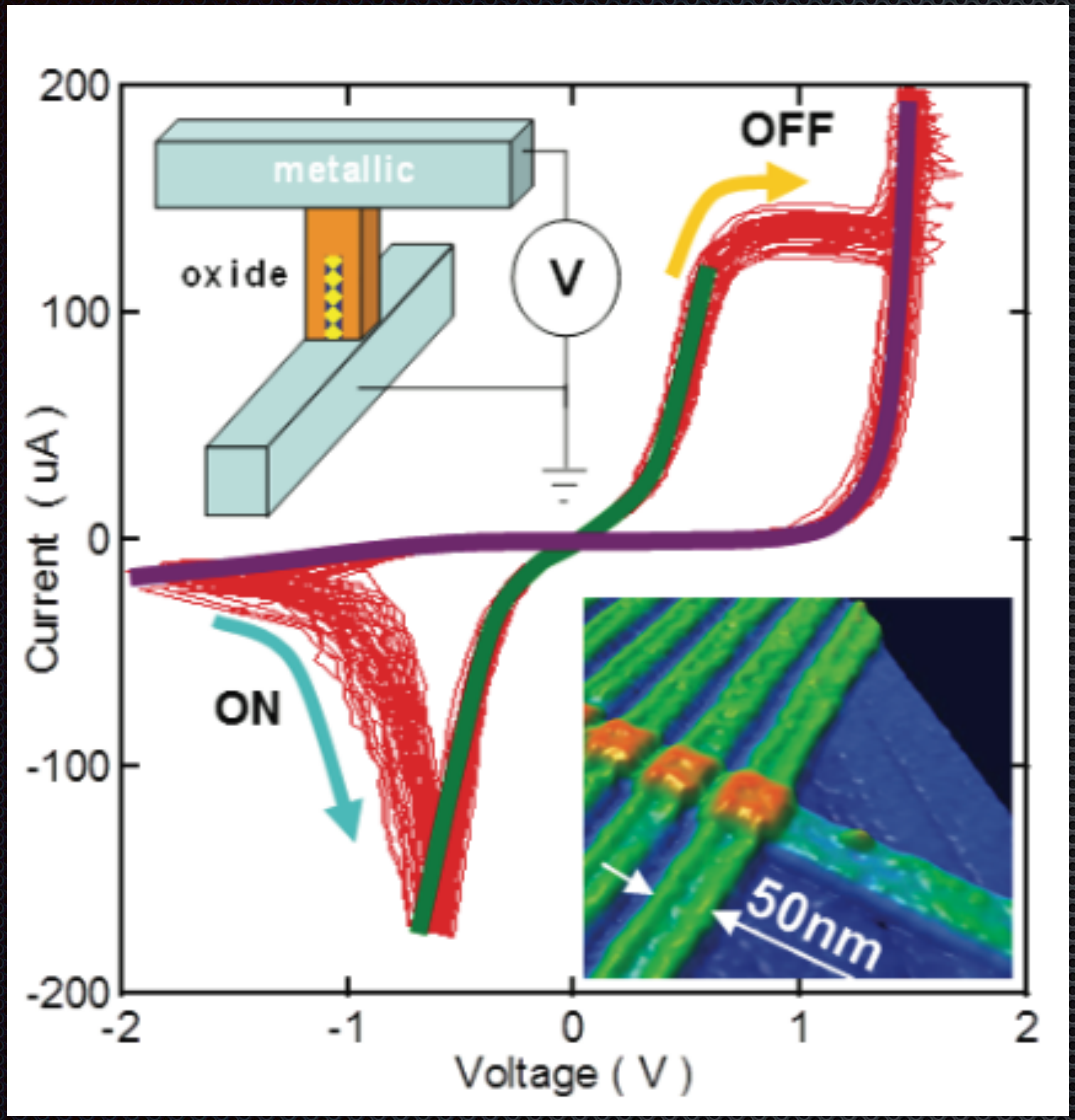


Figure 12: 50 typical current-voltage (I-V) switching loops from a nano device with a $Pt/TiO_{2-x}/TiO_2/Pt$ **stack structure. The atomic force microscopy (AFM) images of the devices are shown as insets. The** ***I-V*** **curves for ON state is conductive and symmetric while those for the OFF state are rectifying, suggesting the role of the metal/oxide interface in the switching. Top inset: schematic of the crosspoint device, showing metallic top and bottom electrodes and the switching oxide. A localized conduction channel made of suboxide with oxygen vacancies is shown in the oxide layer. The growth and retraction of the channel under electric field results in the memristive switching.**

Migrating ionic species result in change in resistance

MemRistors Summary

- Very fast, nonvolatile, low power, low wear
- Challenging to build and operate consistently
- Could be smaller than RAM
- Still in development

Summary

- Emerging memory technologies are nonvolatile
- Idle power is lower
- Writing is often slower than RAM (especially MLC)
- Some technologies have significant wearout

Discussion

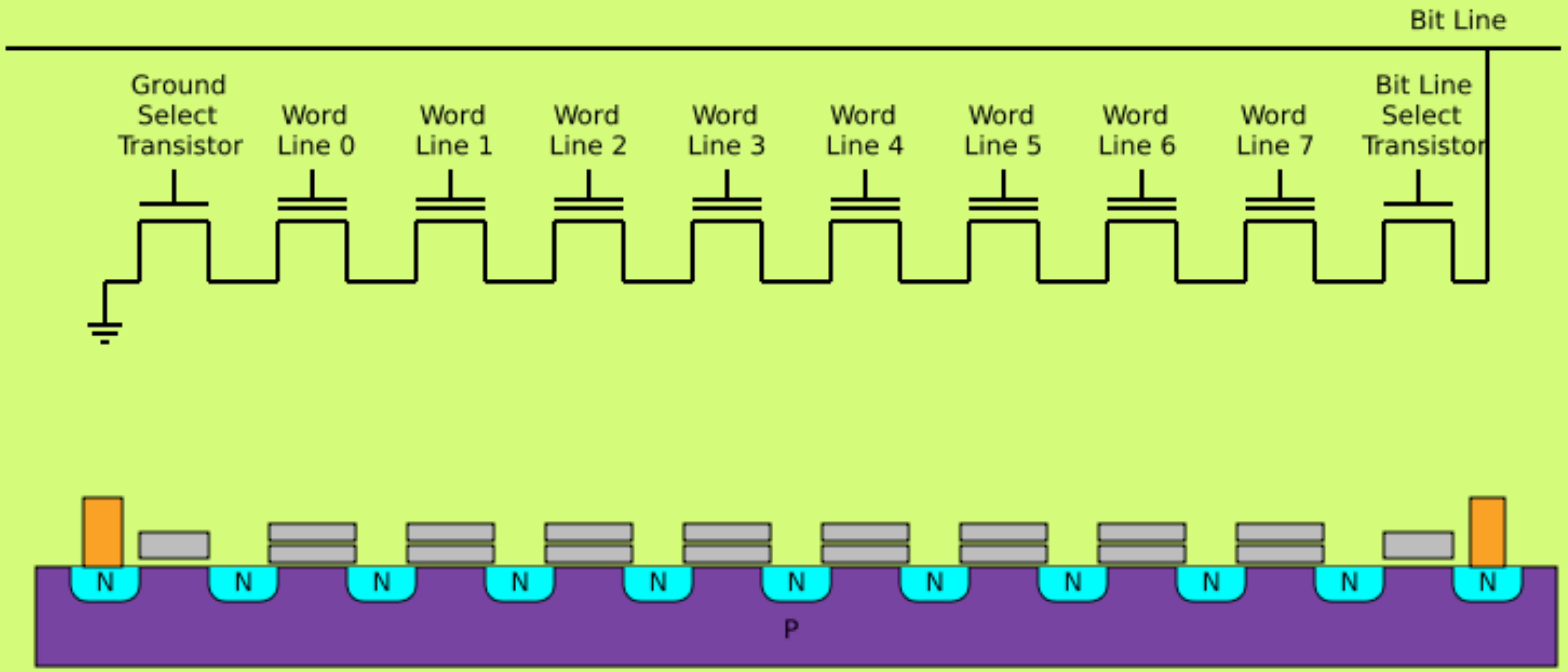
Flash Memory

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

SLC vs. MLC

- Single Level Cell holds a single bit
- Multi Level Cell holds two to four bits
  - MLC stores multiple levels of charge
- 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, more compact, lower power than hard disk
- Less durable than both, although less sensitive to shock and vibration than hard disk
- Lower shelf life than disk or CD/DVD
- Could be new level in memory hierarchy

Failure Modes

- 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

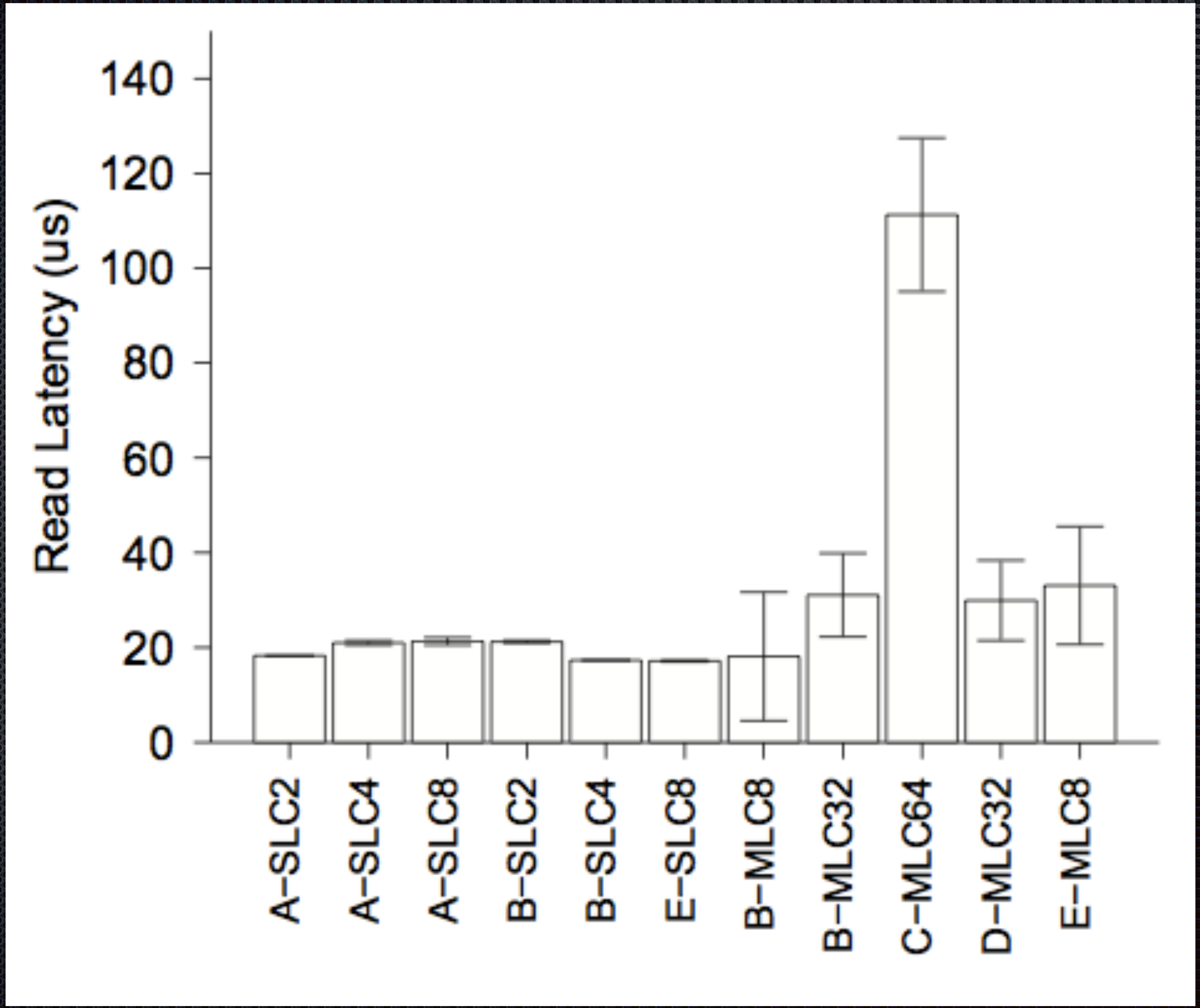
Laura Grupp Micro 2009

Characterizing Flash Memory: Anomalies, Observations, and Applications

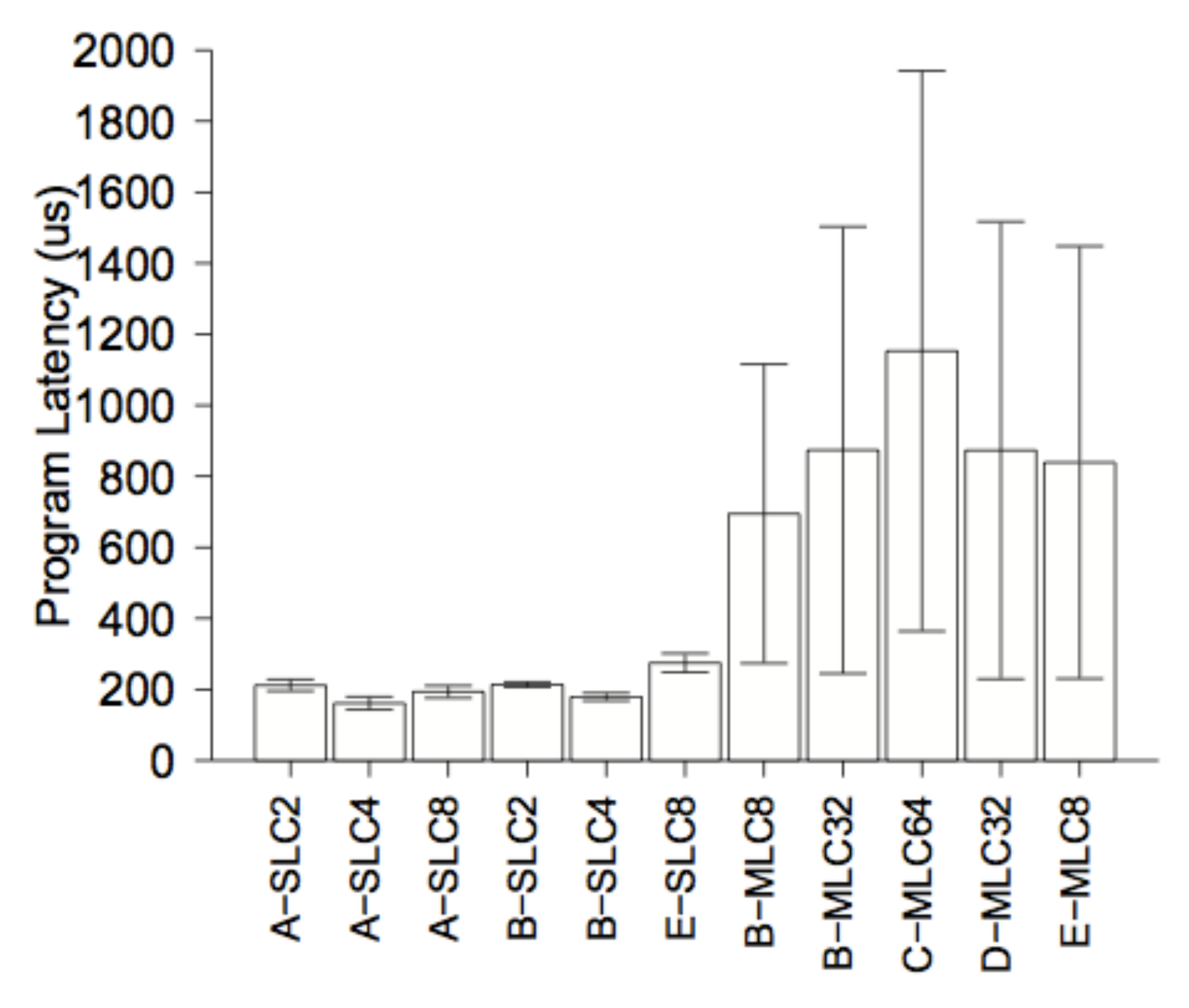
Specifications?

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

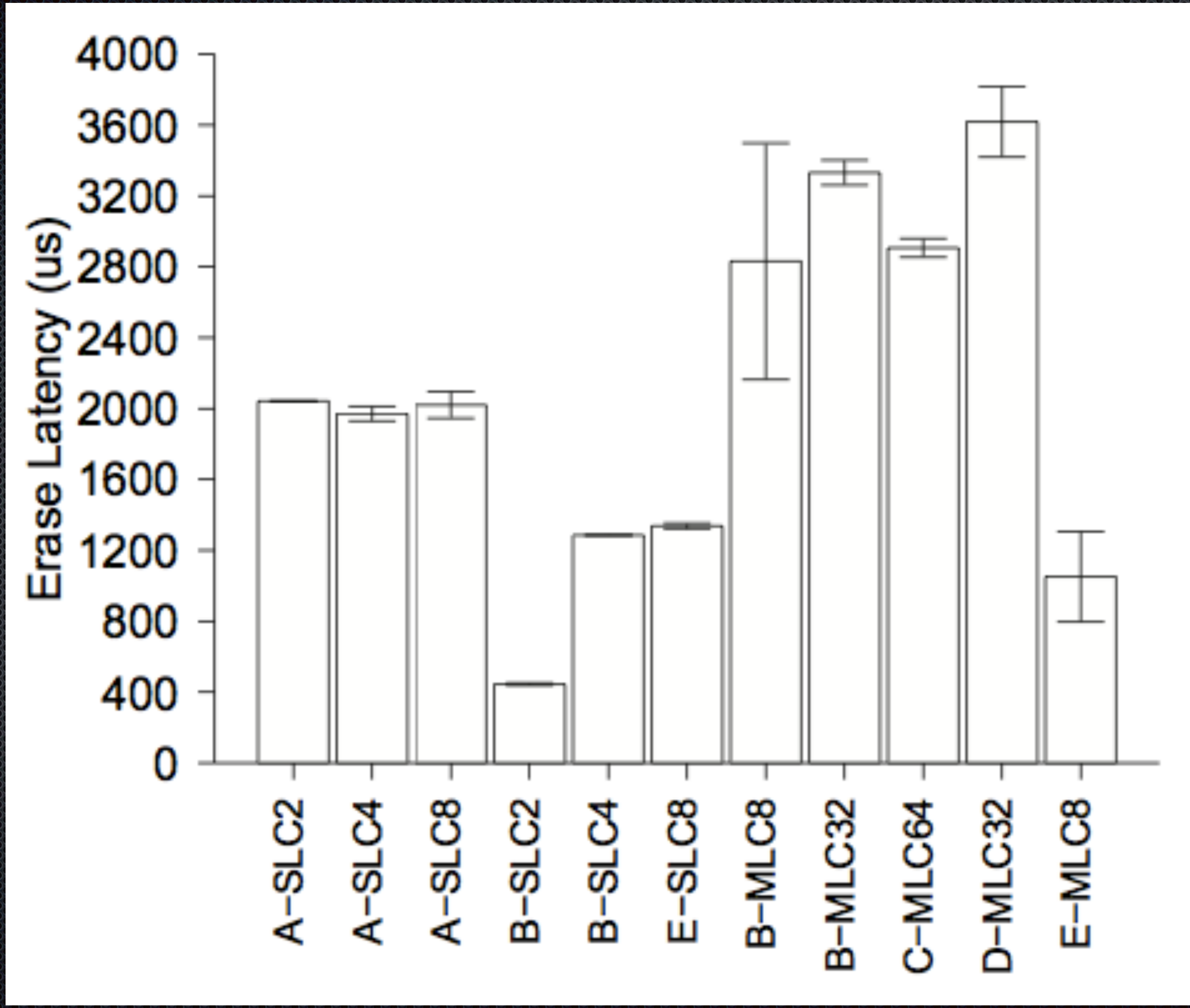
Measured Read Latency



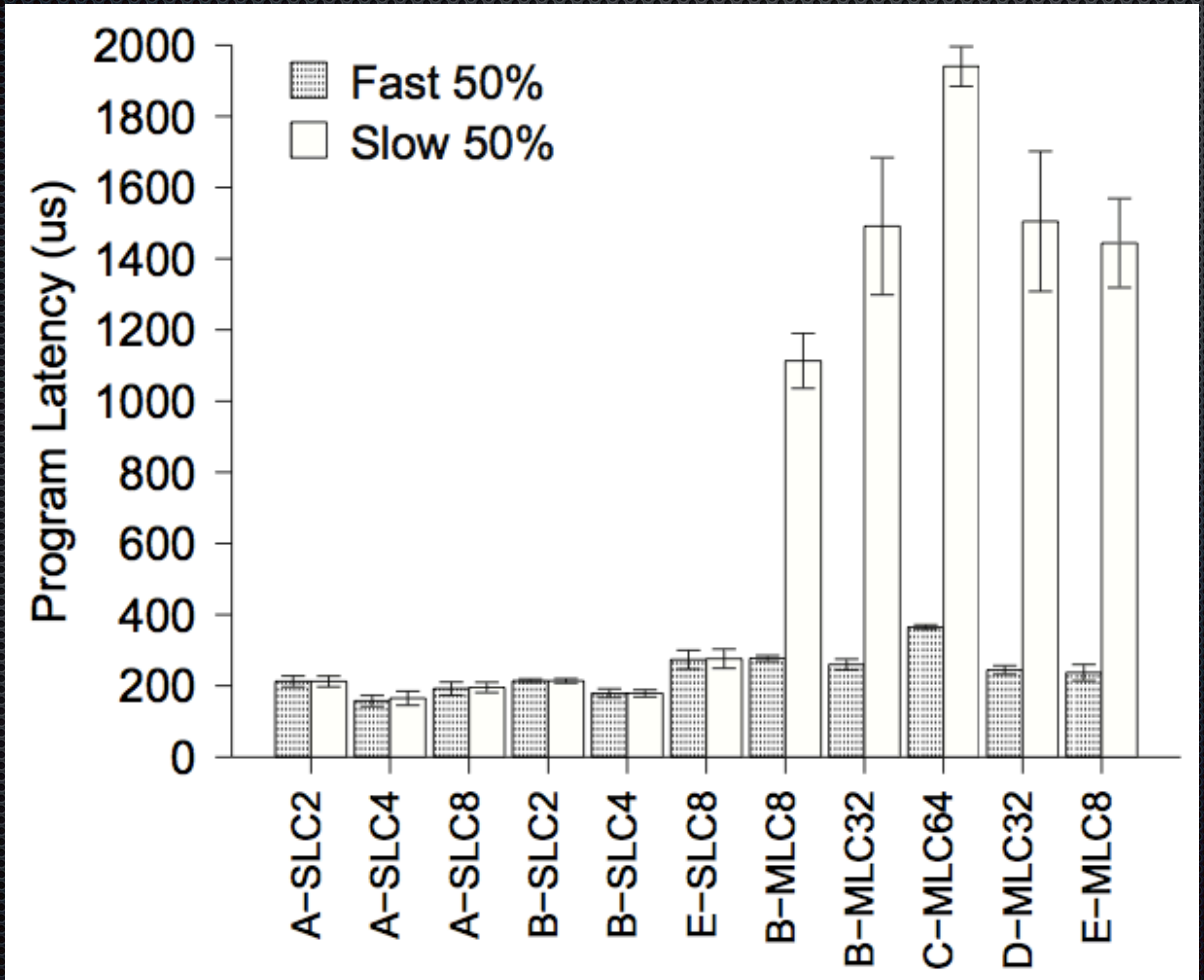
Measured Program Latency



Measured Erase Latency



Program Latency Variance



Program Time with Wear

- Flash becomes easier to program (requires less power) with wear
  - Pre-wearing flash can be used for very low power applications
- SLC is 50% faster, MLC 10-15% faster, near wearout
- Wear also reduces reliability

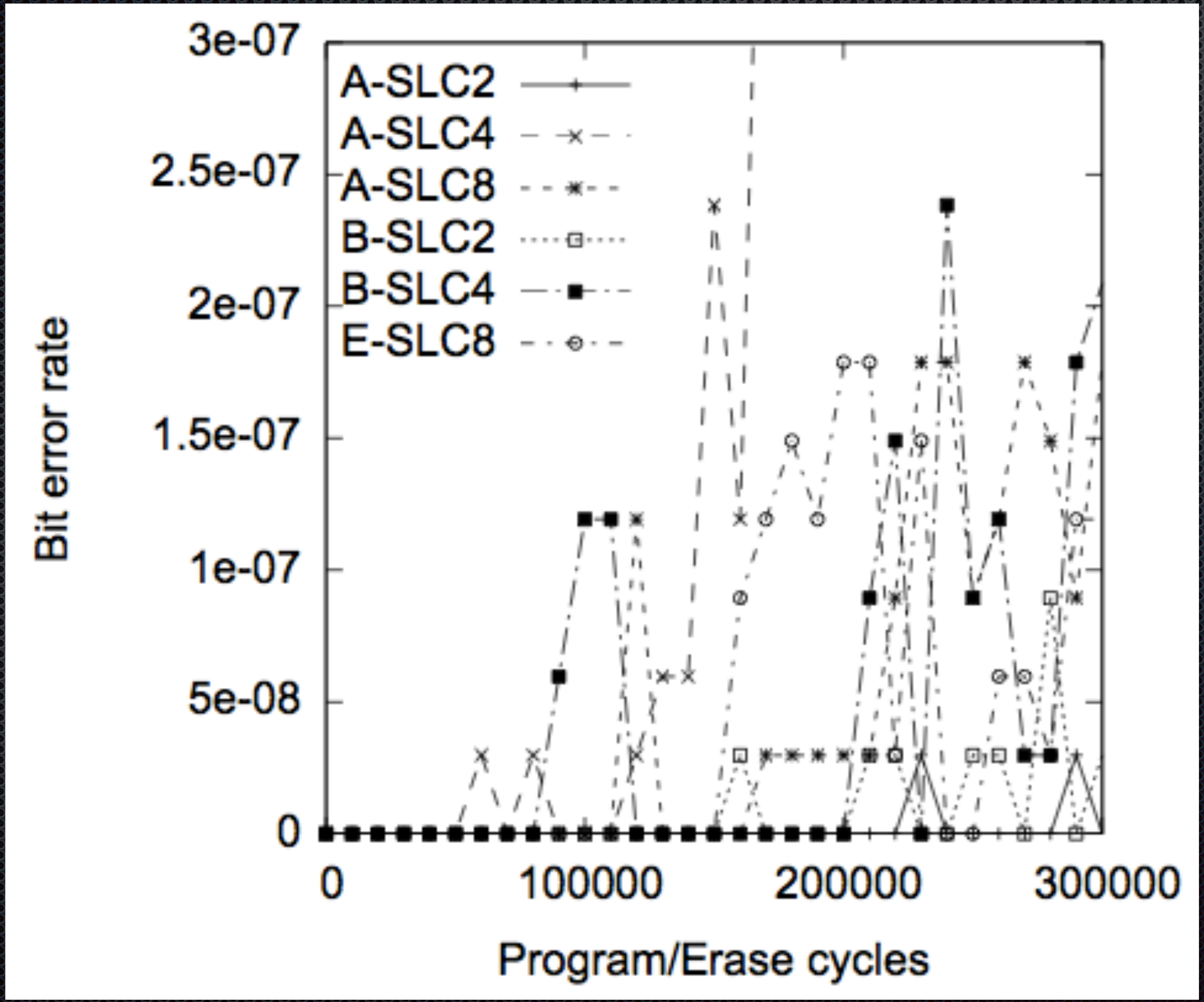
SLC Measured Power

	A-SLC2	A-SLC4	A-SLC8	B-SLC2	B-SLC4	E-SLC8
Peak Read Power in mW (transfer)	35.3 (19.2)	41.1 (18.3)	58.8 (33.1)	27.2 (9.3)	29.9 (8.2)	19.1 (60.8)
Peak Erase Power in mW	30.9	35.5	47.6	25.3	20.0	25.5
Peak Program Power in mW (transfer)	55.2 (43.2)	59.9 (39.2)	78.4 (59.9)	35.0 (13.6)	35.0 (8.4)	56.0 (33.5)
Ave Read Power (mW)	10.3	14.0	21.0	7.4	11.0	18.8
Ave Erase Power (mW)	27.2	38.4	44.4	27.6	22.9	20.8
Ave Program Power (mW)	27.9	32.4	50.1	19.6	20.8	37.5
Idle Power (mW)	2.7	7.1	17.0	2.9	2.9	13.3
Read Energy (nJ/bit)	0.052	0.069	0.088	0.046	0.042	0.0056
Program Energy (nJ/bit)	0.72	0.61	0.97	0.47	0.41	1.01
Erase Energy (nJ/bit)	0.06	0.067	0.093	0.011	0.025	0.031

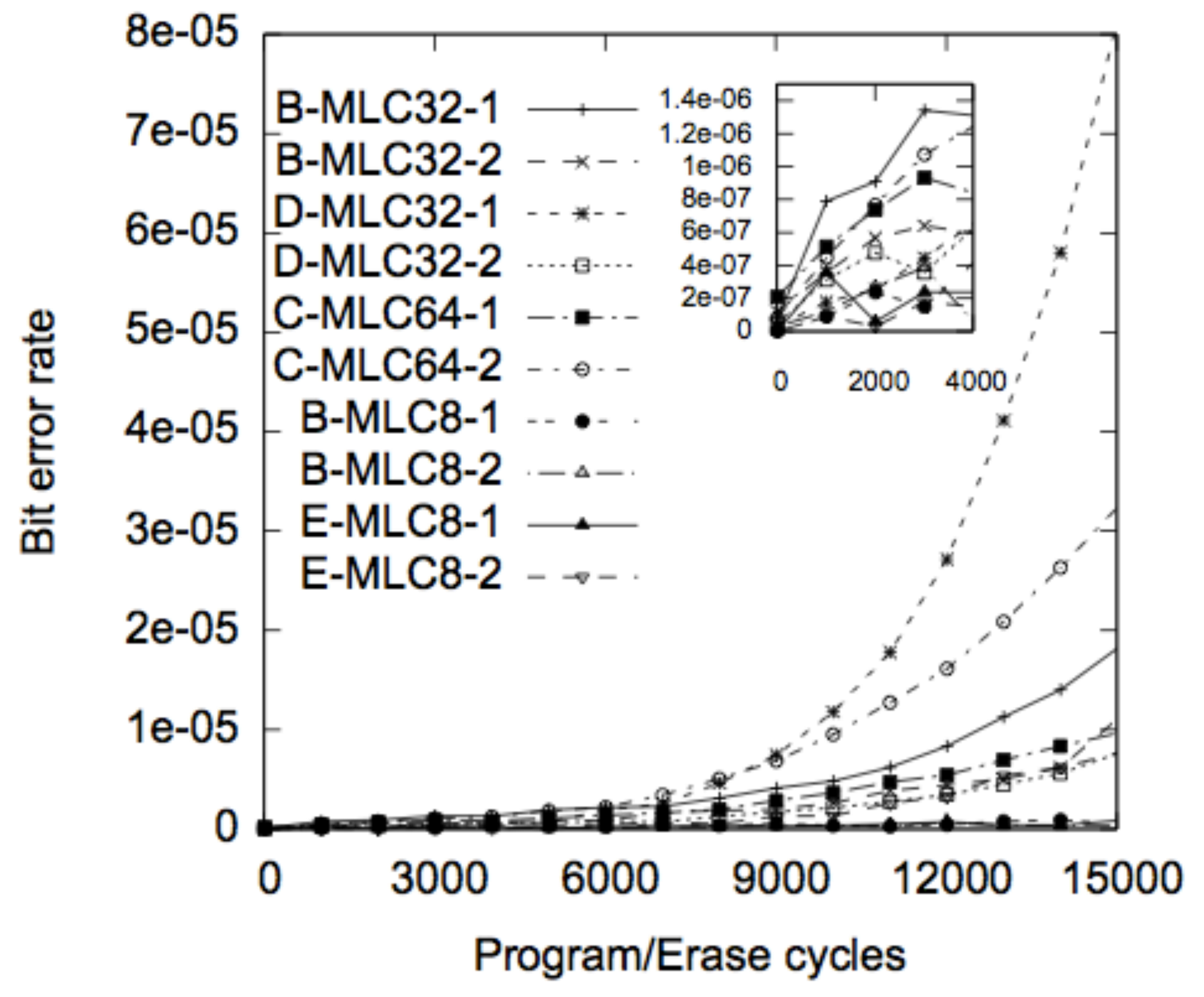
MLC Measured Power

	B-MLC8	B-MLC32	C-MLC64	D-MLC32	E-MLC8	
Peak Read Power in mW (transfer)	54.0 (29.1)	75.9 (41.1)	112.0 (42.8)	66.3 (31.2)	13.4 (39.9)	
Peak Erase Power in mW	42.4	70.6	111.8	57.0	21.3	
Peak Program Power in mW (transfer)	58.9 (22.4)	94.7 (63.1)	132.2 (65.2)	82.3 (31.7)	118.4 (28.5)	
Ave Read Power (mW)	18.1	31.1	41.5	28.3	21.3	
Ave Erase Power (mW)	45.5	53.0	105.0	56.2	23.5	
Ave Program Power (mW)	46.5	52.5	77.0	55.6	40.9	
Idle Power (mW)	12.7	8.5	27.3	11.2	10.2	
Read Energy (nJ/bit)	0.15	0.11	0.19	0.093	0.002	
Fast Program Energy (nJ/bit)	1.09	0.96	0.66	0.79	0.46	
Slow Program Energy (nJ/bit)	3.31	3.30	2.86	2.84	2.07	
Erase Energy (nJ/bit)	0.070	0.056	0.038	0.051	0.0057	

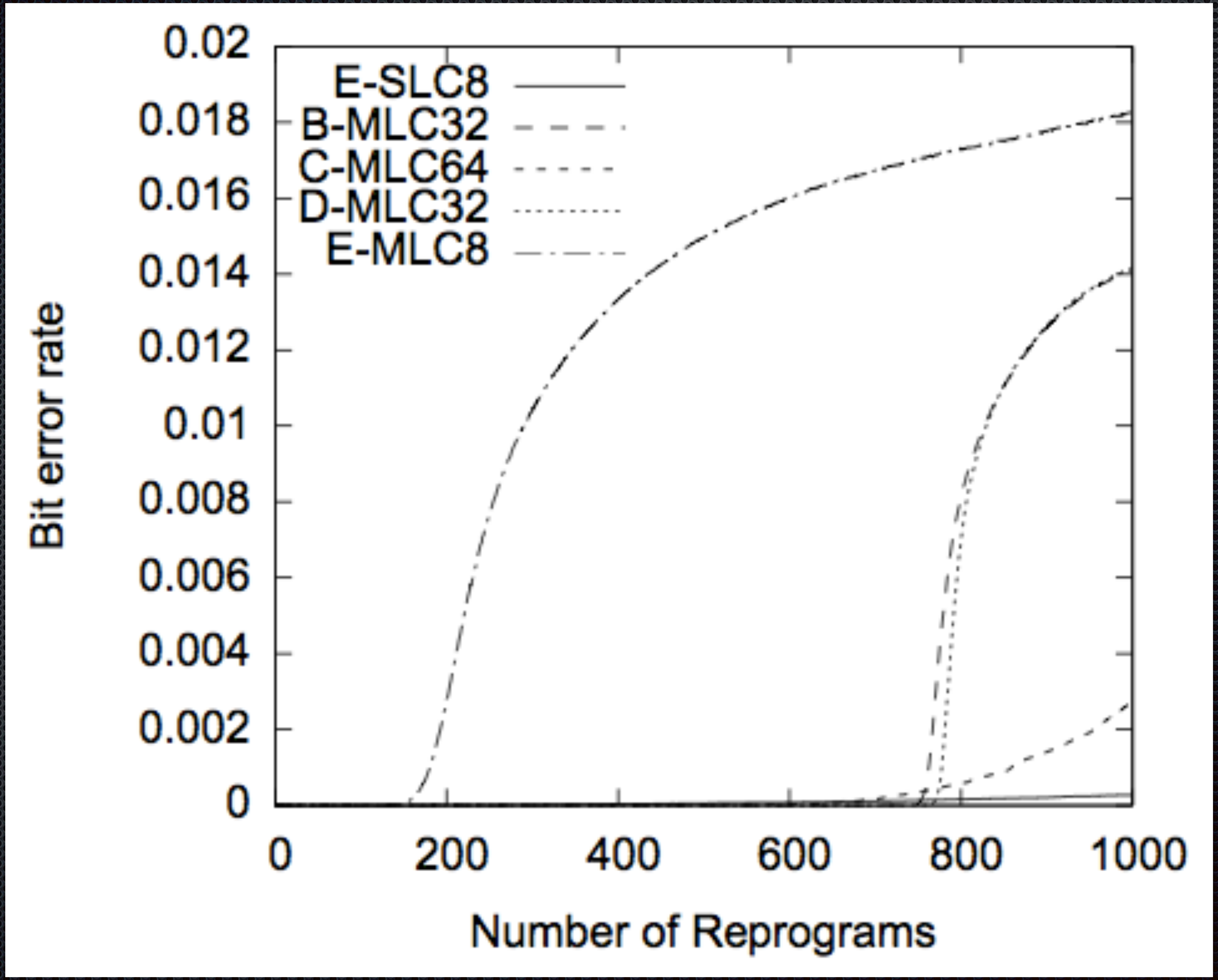
SLC Measured Endurance



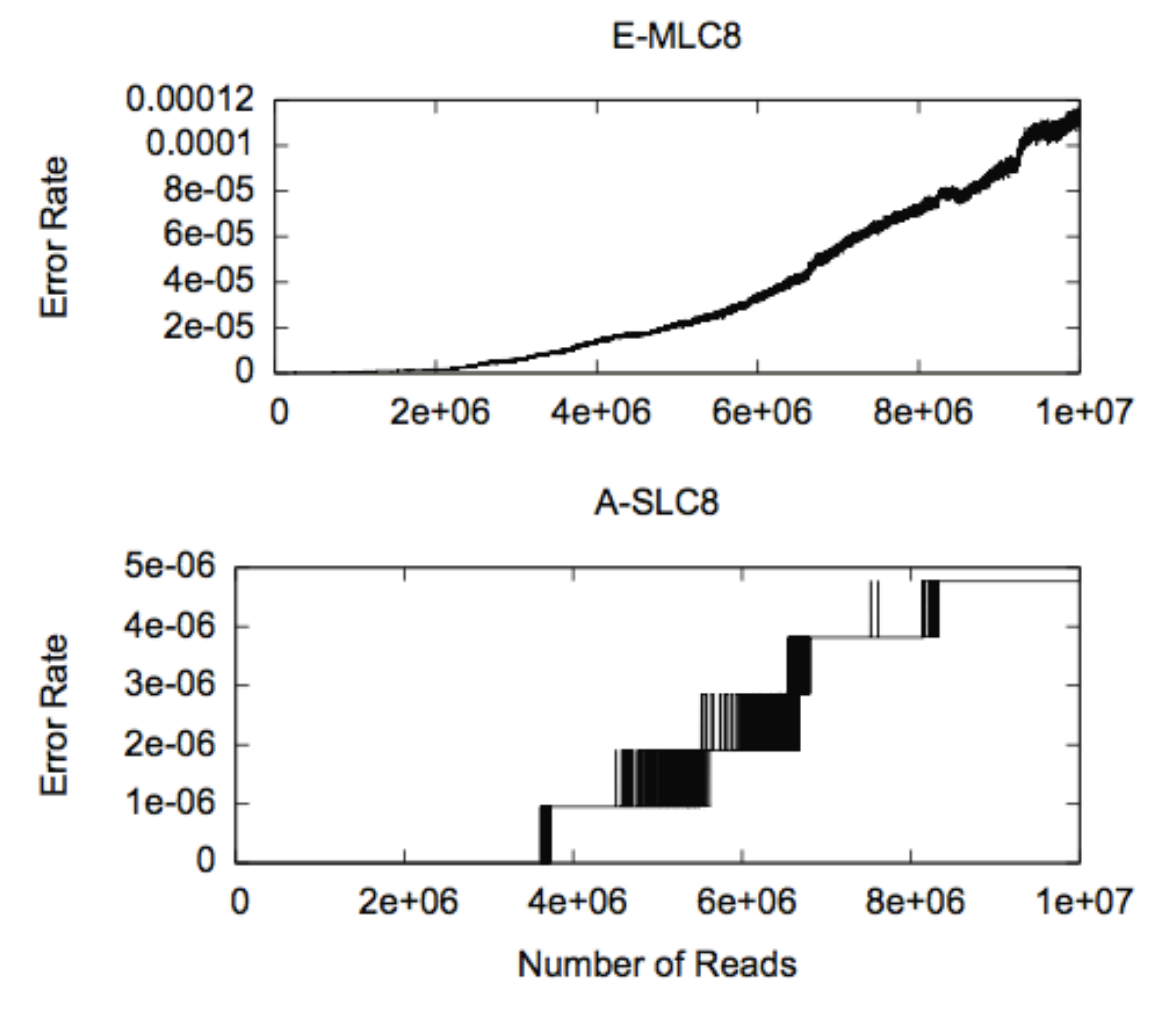
MLC Measured Endurance



Errors Due to Reprograms



Read Disturb Errors



FTL - Flash Translation Layer

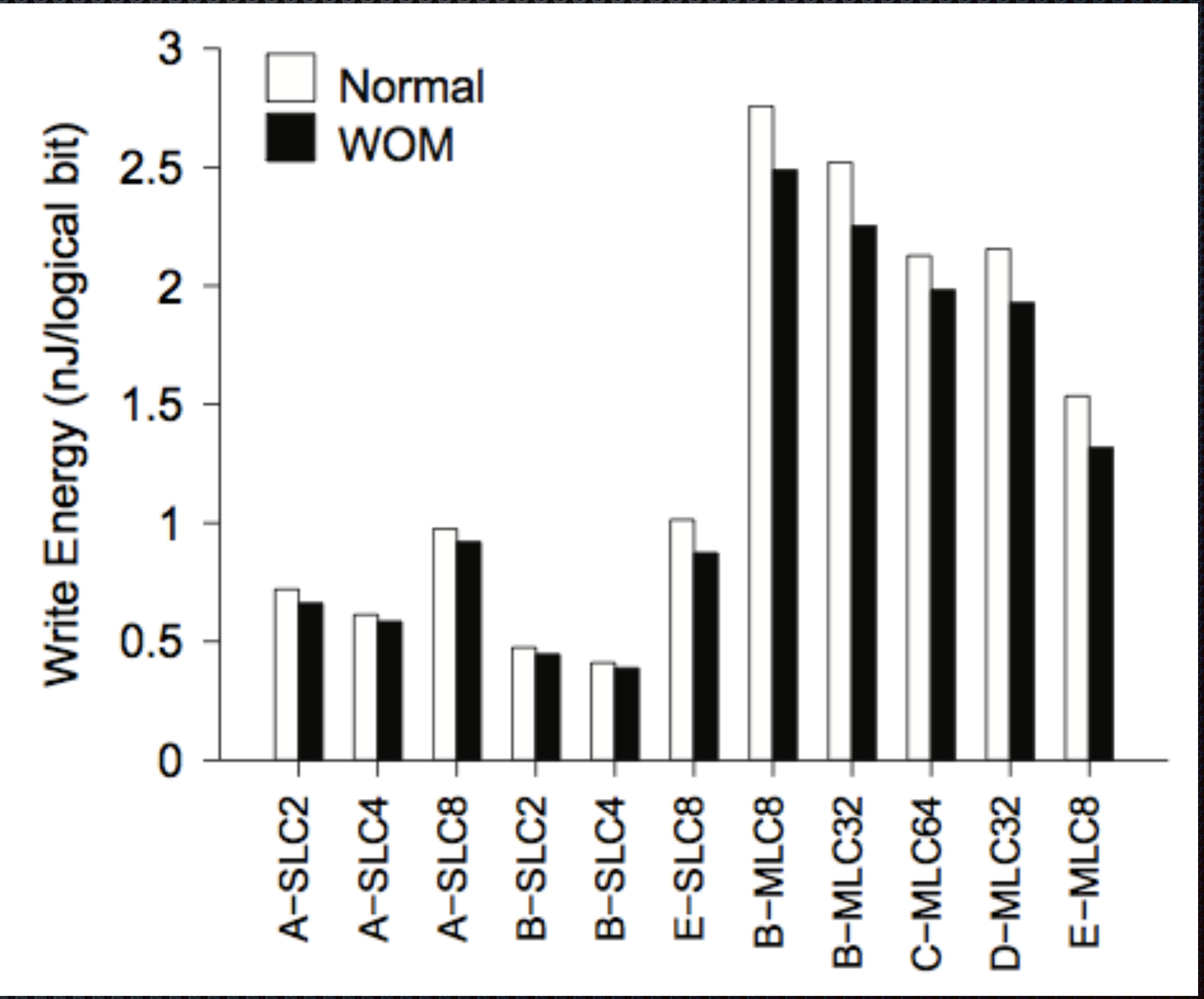
- Indirection table that maps logical to physical addresses
- Hides wear leveling and layout policies
- Also hides buffering, write coalescing, etc.
- Often seen as the point where Flash can be architected

Mango FTL Layer

- Tries to use faster MLC pages preferentially to reduce power and increase performance
- Slight improvement, but 55% reduction in endurance

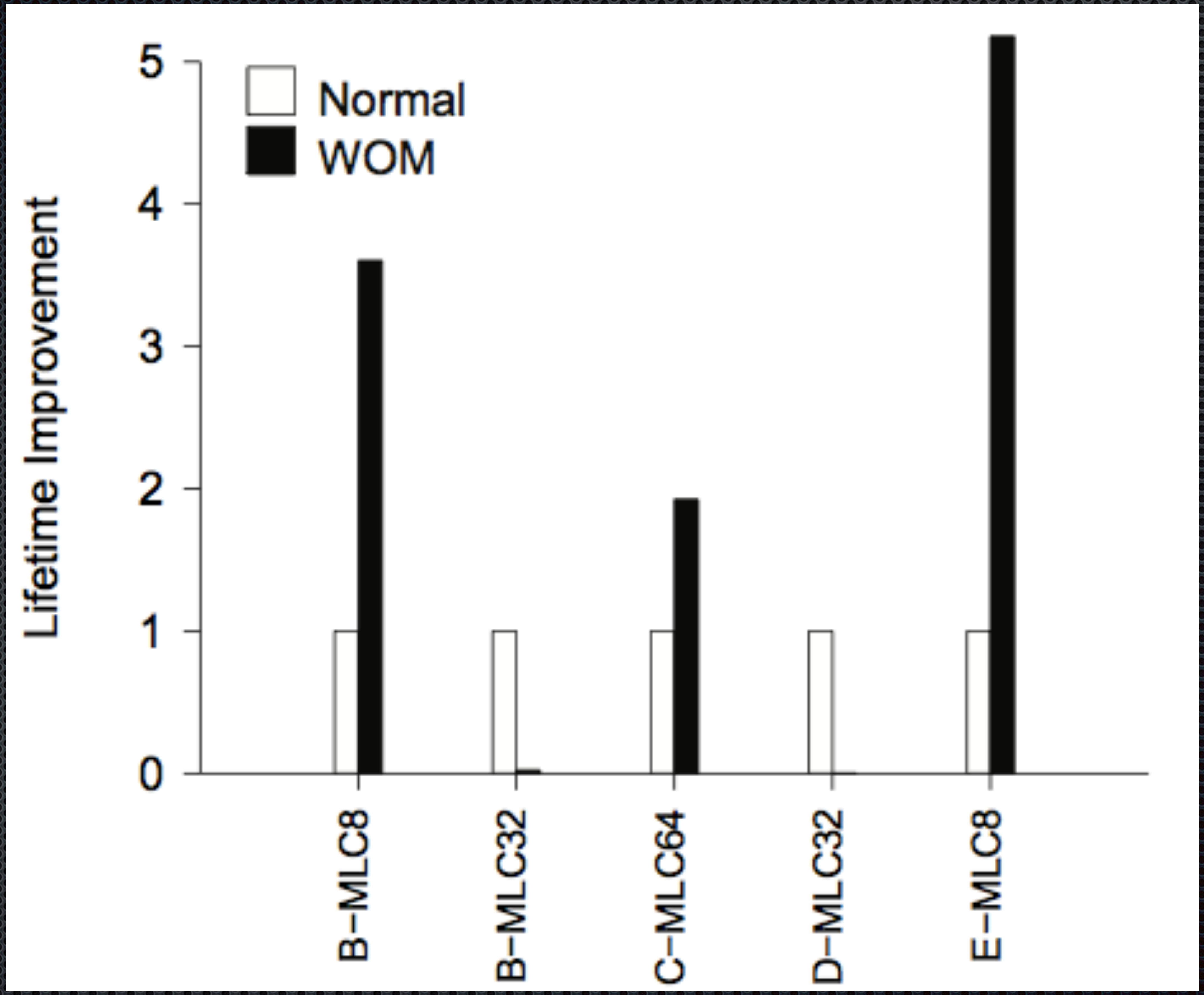
Write-Only Memory Code

Logical bits	First generation	Second generation
00	111	000
01	110	001
10	101	010
11	011	100



Reduce power by avoiding half of erasures

WOM Codes



Increase endurance

Discussion