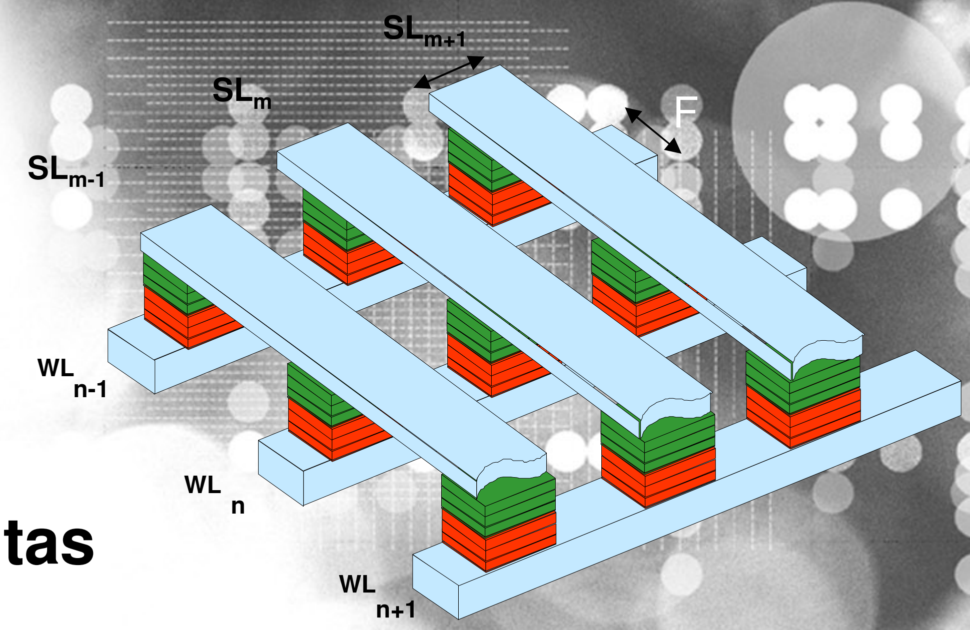
IBM Almaden Research Center

Storage Class Memory, Technology and Use



Rich Freitas

© 2008 IBM Corporation

Agenda

- **Introduction**
- **Storage Class Memory Technologies**
- **Using Storage Class Memories in Systems**
- **Impact on Systems**

© 2008 IBM Corporation

Definition of **S**torage **C**lass **M**emory **SCM**

- **A new class of data storage/memory devices**
  - many technologies compete to be the 'best' SCM
- **SCM features:**
  - Non-volatile (~ 10 years)
  - Fast Access times (~ DRAM like)
  - Low cost per bit more (DISK like – by 2015)
  - Solid state, no moving parts
- **SCM *blurs the distinction* between**
  - MEMORY (*fast, expensive, volatile*) and
  - STORAGE (*slow, cheap, non-volatile*)

© 2008 IBM Corporation

Some Terminology Clarification

- **SCM = Storage Class Memory**
  - –SCM describes a *technology*, not a *use*
  - –FLASH is an early example of SCM
- **NVRAM = Non Volatile RAM**
  - –SCM is one example of NVRAM
  - –Other NVRAM types: DRAM+battery or DRAM+disk combos
- **SSD = Solid State Disk**
  - –Use of NVRAM for *block oriented* storage applications

© 2008 IBM Corporation

Criteria to judge a SCM technology

- **Device Capacity** [GigaBytes]
  - Closely related to cost/bit [$/GB]
- **Speed**
  - Latency (= access time) Read & Write [nanoseconds]
  - Bandwidth Read & Write [GB/sec]
- **Random Access or Block Access** -
- **Write Endurance= #Writes before death** -
- **Read Endurance= #Reads "** -
- **Data Retention Time** [Years]
- **Power Consumption** [Watt]

© 2008 IBM Corporation

Even more Criteria

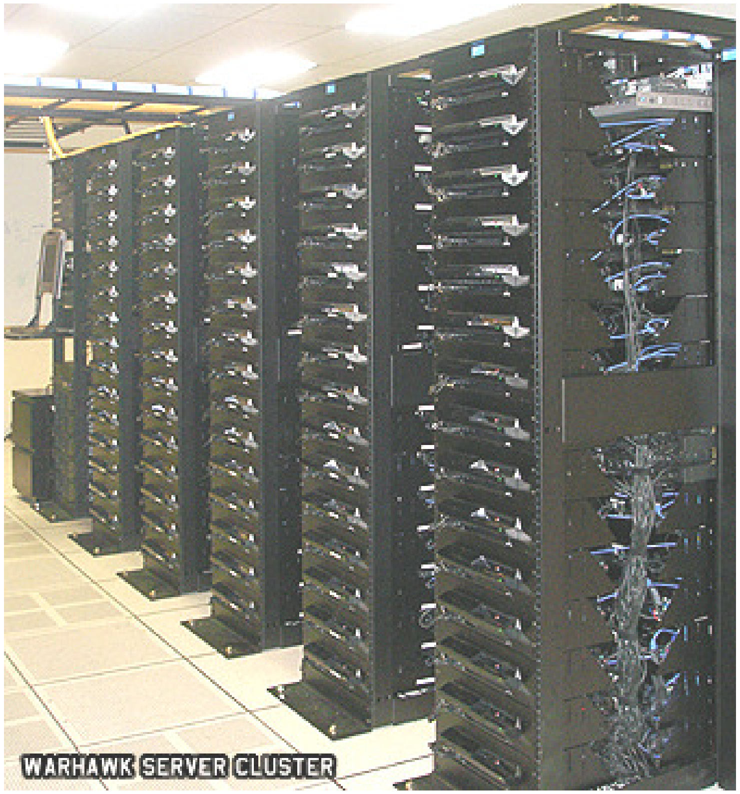
- **Reliability (MTBF)** [Million hours]
- **Volumetric density** [TeraBytes/liter]
- **Power On/Off transit time** [sec]
- **Shock & Vibration** [g-force]
- **Temperature resistance** [^oC]
- **Radiation resistance** [Rad]

~ 16 criteria! This makes the SCM problem so hard

© 2008 IBM Corporation

System Targets for SCM

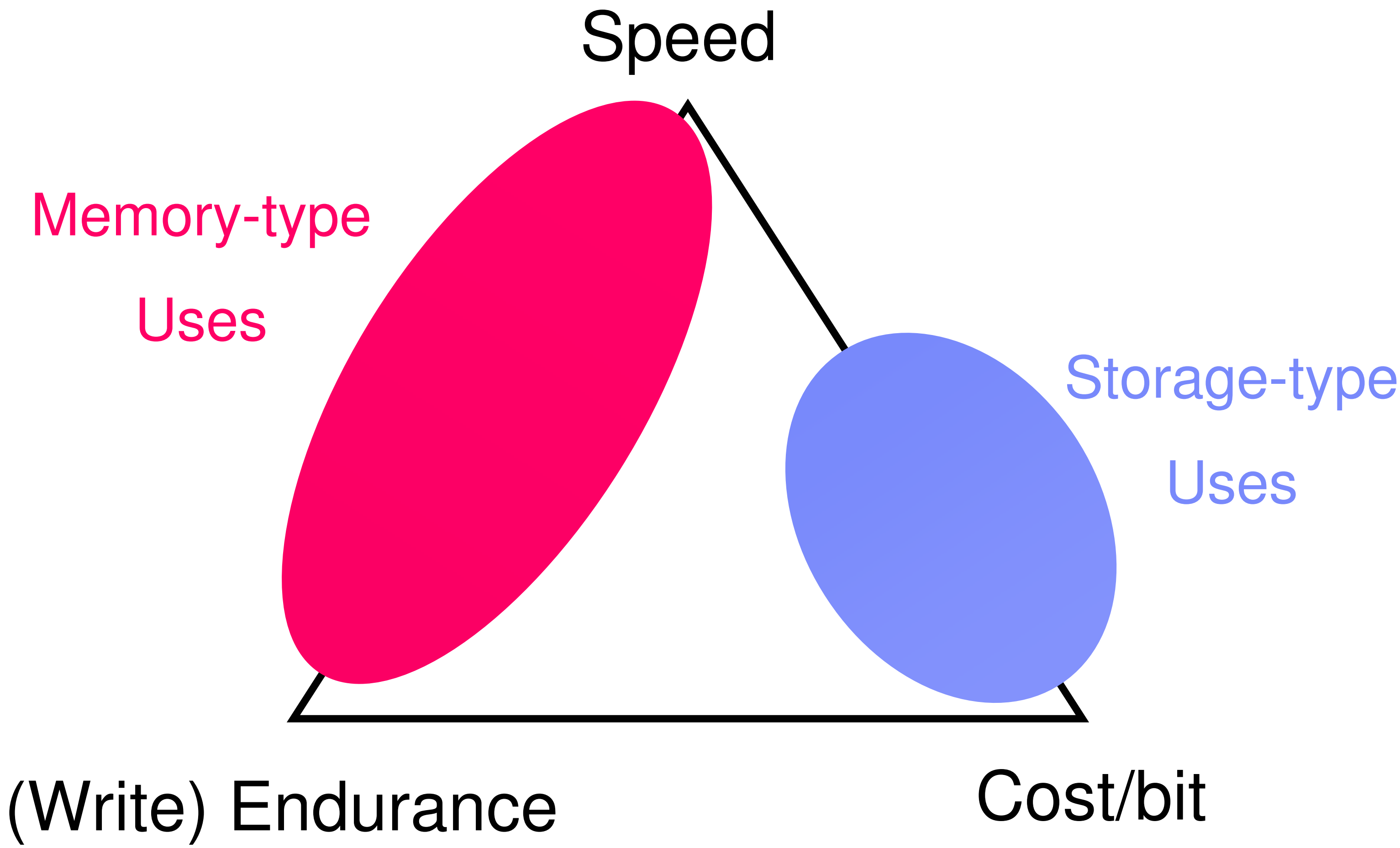
Billions!



Mobile ✓ Desktop X Datacenter ✓

© 2008 IBM Corporation

SCM Design Triangle

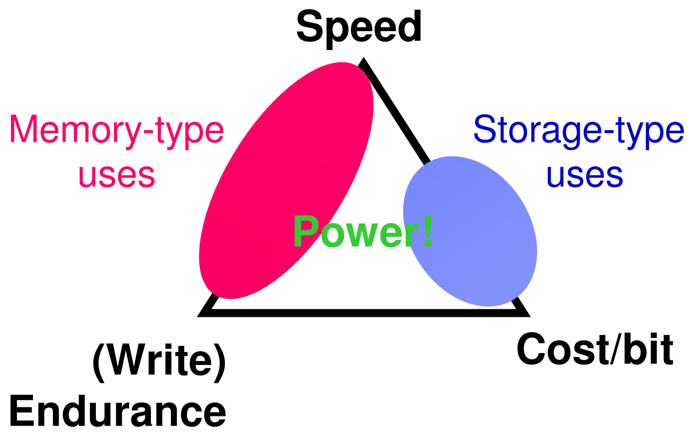


Power!

© 2008 IBM Corporation

Storage Class Memory

A solid-state memory that **blurs the boundaries** between storage and memory by being **low-cost**, **fast**, and **non-volatile**.



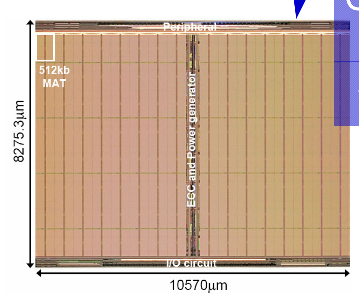
- **SCM system requirements for Memory (Storage) apps**
  - No more than 3-5x the **Cost** of enterprise HDD (< $1 per GB in 2012)
  - **<200nsec (<1 µsec) Read/Write/Erase time**
  - >100,000 **Read I/O operations** per second
  - **>1GB/sec (>100MB/sec)**
  - **Lifetime** of **10^8 – 10^{12}** write/erase cycles
  - 10x lower **power** than enterprise HDD

© 2008 IBM Corporation

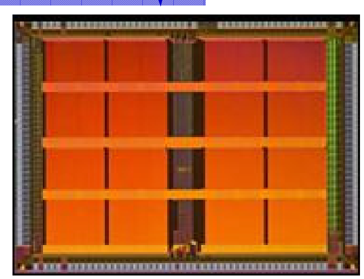
Emerging Memory Technologies

Memory technology remains an active focus area for the industry

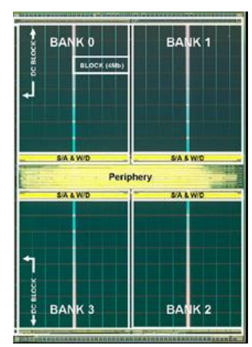
FLASH Extension	FRAM	MRAM	PCRAM	RRAM	Solid Electrolyte	Polymer/ Organic
Trap Storage	Ramtron	IBM	Ovonyx	IBM	Axon	Spansion
Saifun *NROM*	Fujitsu	Infineon	BAE	Sharp	Infineon	Samsung
Tower	STMicro	Freescale	Intel	Unity		TFE
Spansion	TI	Philips	STMicro	Spansion		MEC
Infineon	Toshiba	STMicro	Samsung	Samsung		Zettacore
Macronix	Infineon	HP	Elpida			Roltronics
Samsung	Samsung	NVE	IBM			Nanolayer
Toshiba	NEC	Honeywell	Macronix			
Spansion	Hitachi	Toshiba	Infineon			
Macronix	Rohm	NEC	Hitachi			
NEC	HP	Sony	Philips			
Nano-x'tal	Cypress	Fujitsu				
Freescale	Matsushita	Renesas				
Matsushita	Oki	Samsung				
	Hynix	Hynix				
	Celis	TSMC				
	Fujitsu					
	Seiko Epson					



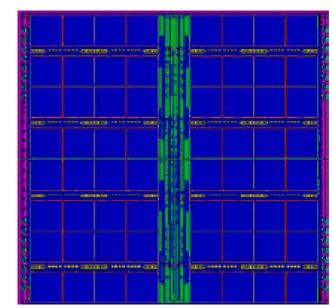
64Mb FRAM (Prototype)
0.13um 3.3V



4Mb MRAM (Product)
0.18um 3.3V



512Mb PRAM (Prototype)
0.1um 1.8V

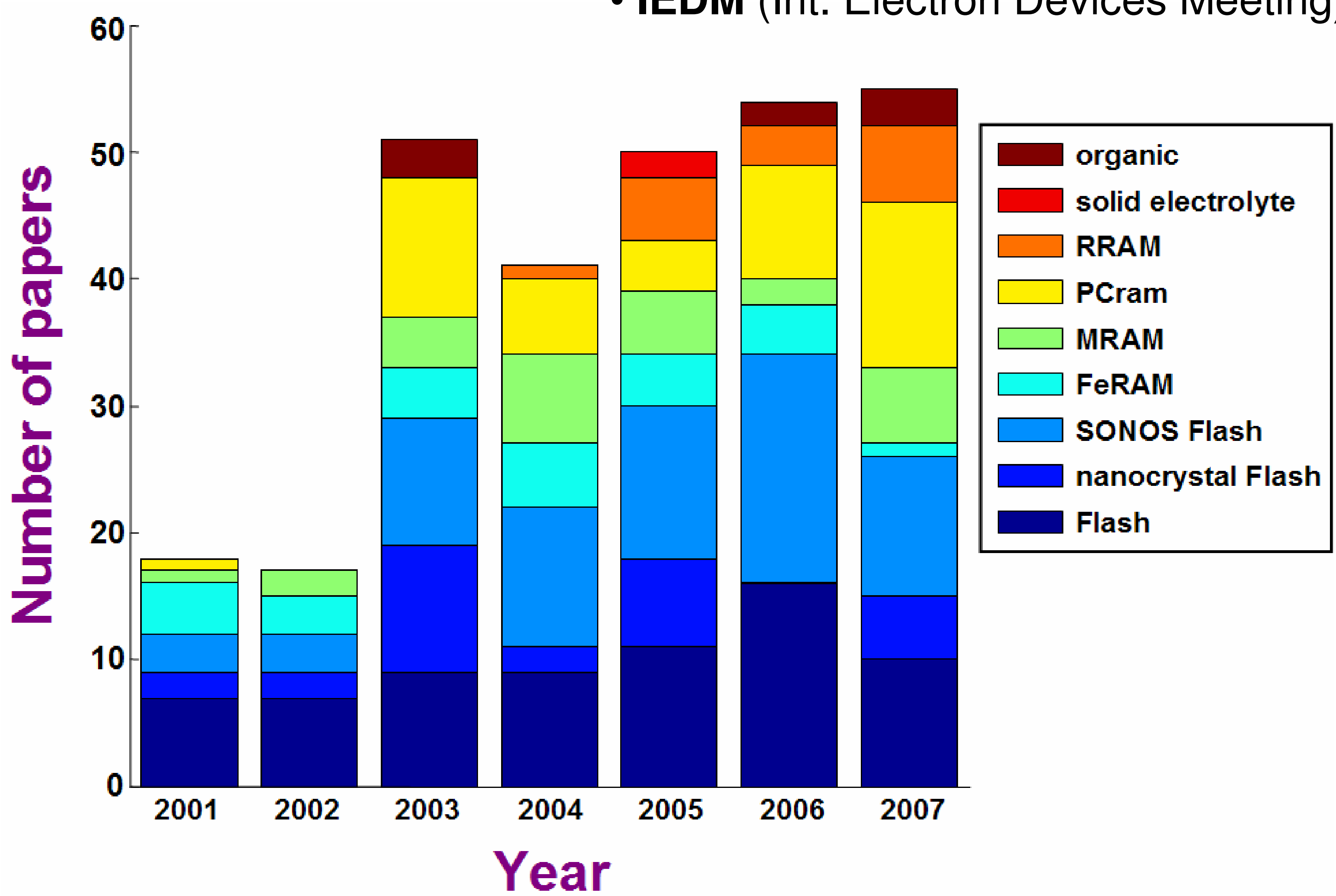


4Mb C-RAM (Product)
0.25um 3.3V

© 2008 IBM Corporation

Research interest

Papers presented at
- Symposium on **VLSI Technology**
- **IEDM** (Int. Electron Devices Meeting)

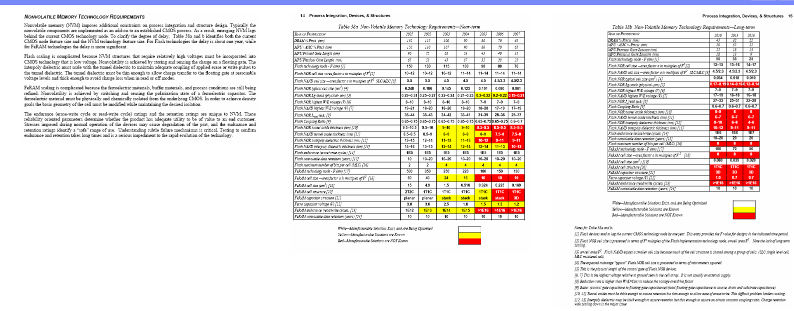


 © 2008 IBM Corporation

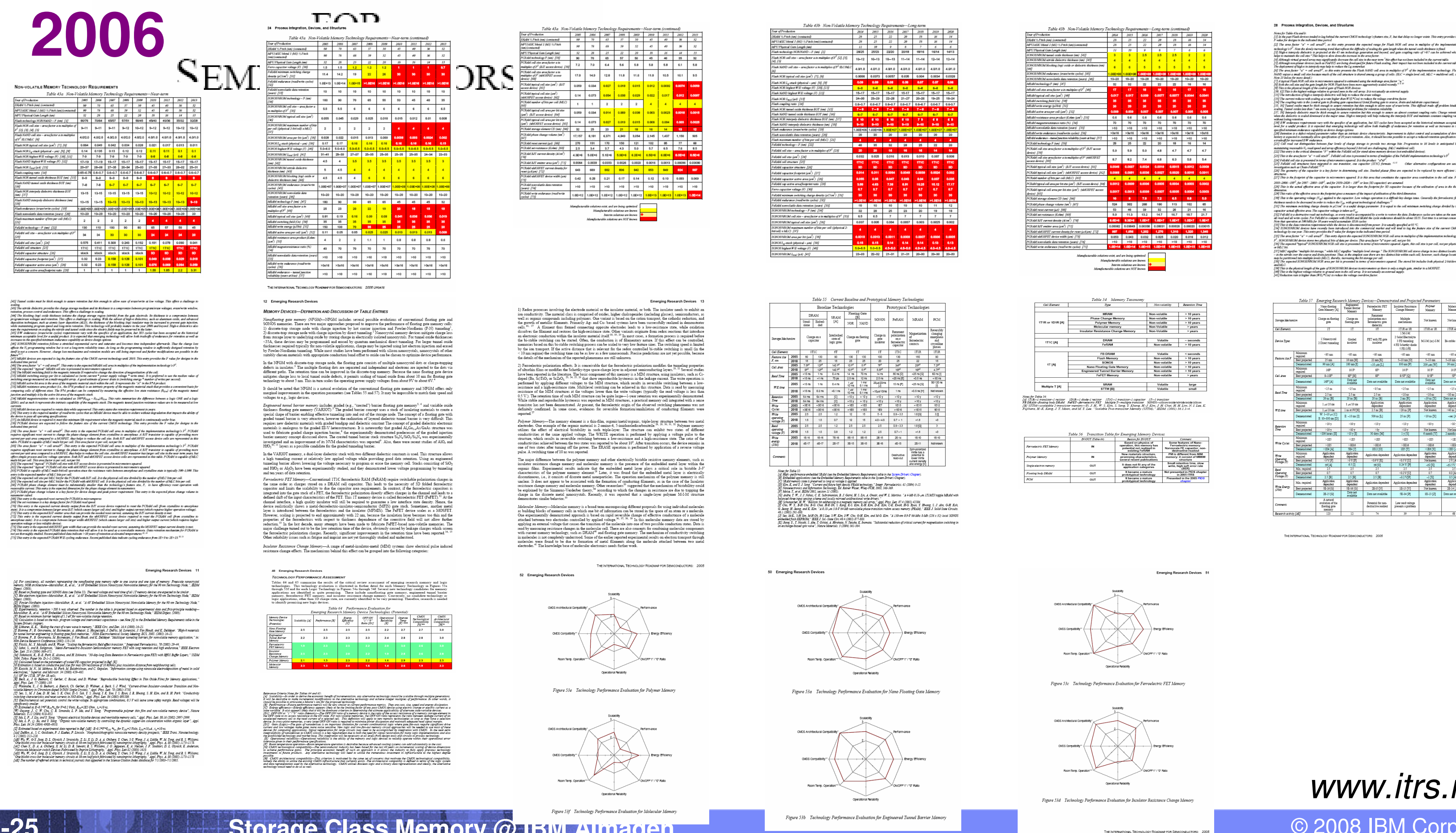
Industry interest in non-volatile memory

2001

INTERNATIONAL TECHNOLOGY ROADMAP FOR SEMI…ORS



2006



www.itrs.net

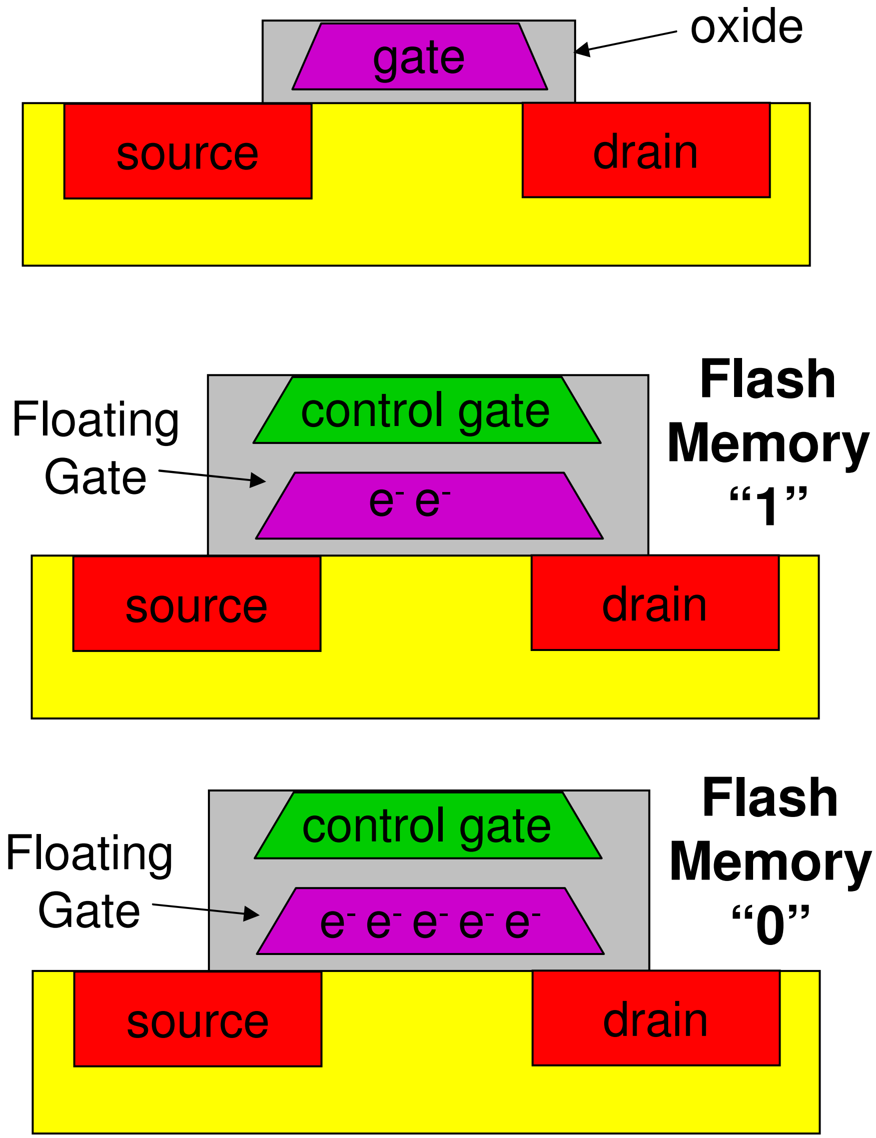
© 2008 IBM Corporation

Candidate device technologies

- **Improved Flash**
- **FeRAM (Ferroelectric RAM)**
  - **FeFET**
- **MRAM (Magnetic RAM)**
  - **Racetrack memory**
- **RRAM (Resistive RAM)**
  - **Organic & polymer memory**
- **Solid Electrolyte**
- **PC-RAM (Phase-change RAM)**

 © 2008 IBM Corporation

What is Flash?



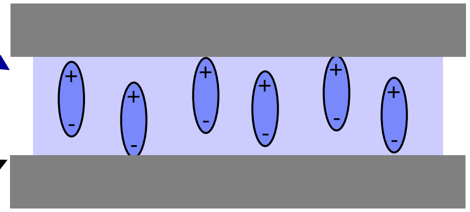
- Based on MOS transistor
- Transistor gate is redesigned
  - Charge is placed or removed near the "gate"
  - The threshold voltage V_{th} of the transistor is shifted by the presence of this charge
  - The threshold Voltage shift detection enables non-volatile memory function.

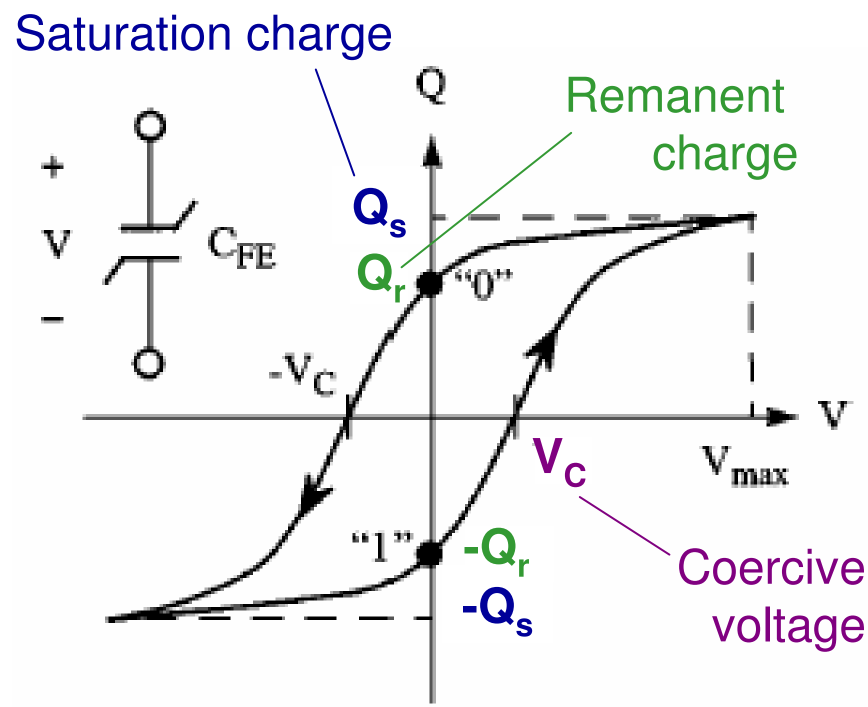
© 2008 IBM Corporation

FeRAM (Ferroelectric RAM)

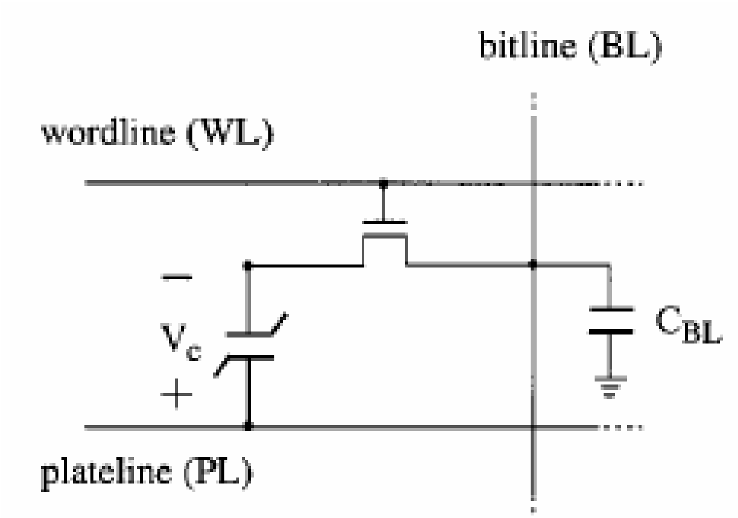
ferroelectric material such as lead zirconate titanate (Pb(ZrxTi1-x)O) or PZT

metallic electrodes





need select transistor – "half-select" perturbs

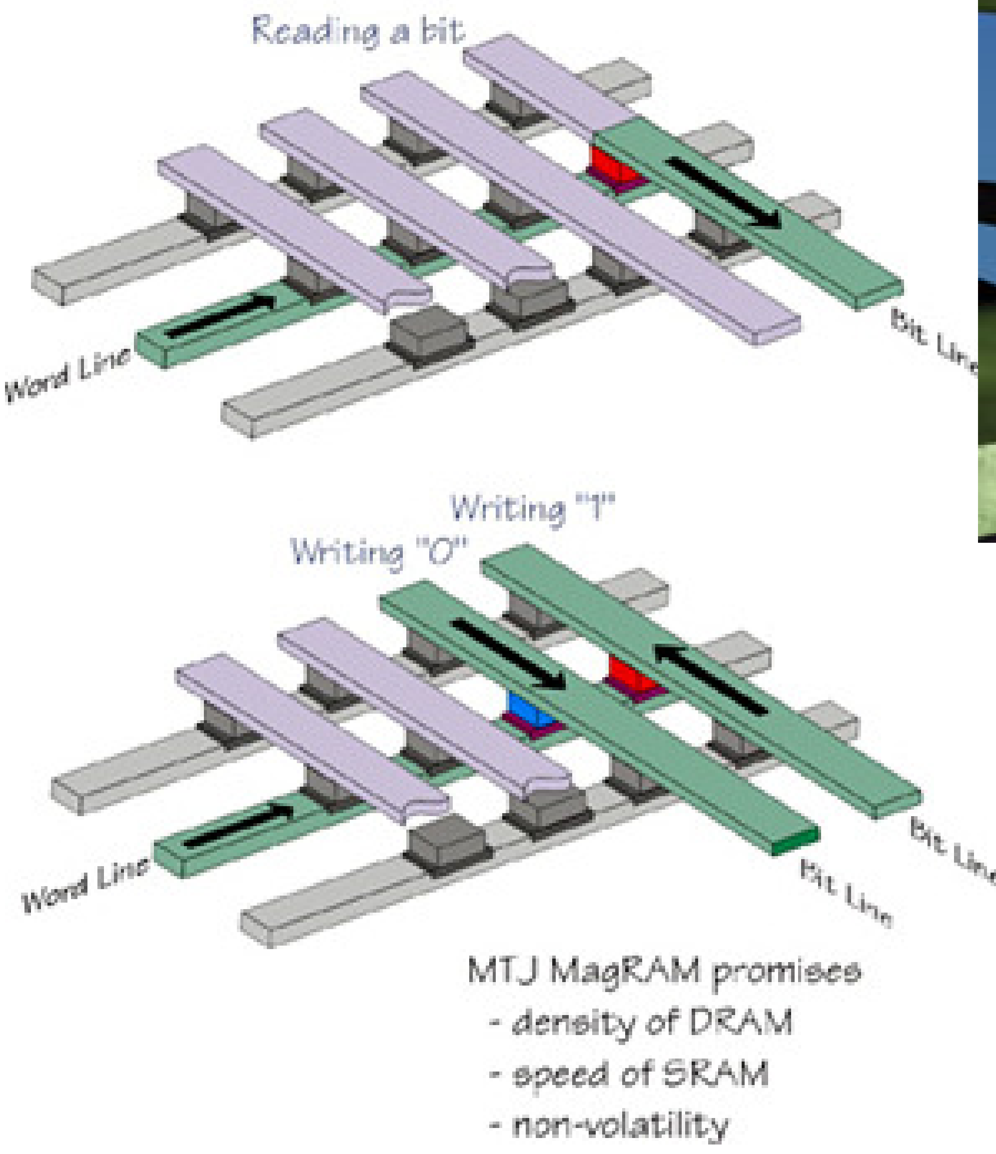


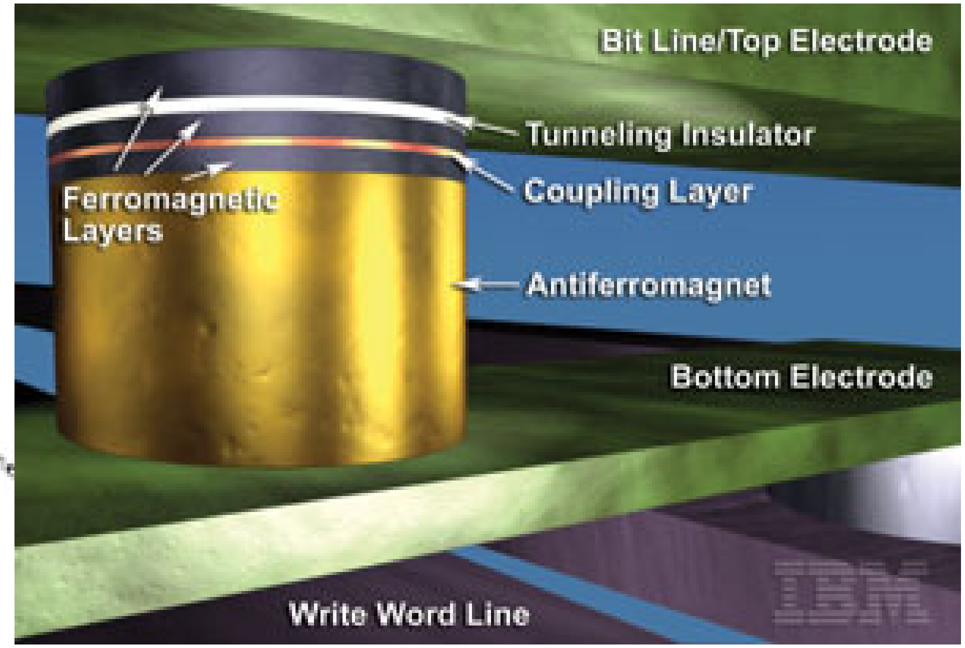
[Sheikholeslami:2000]

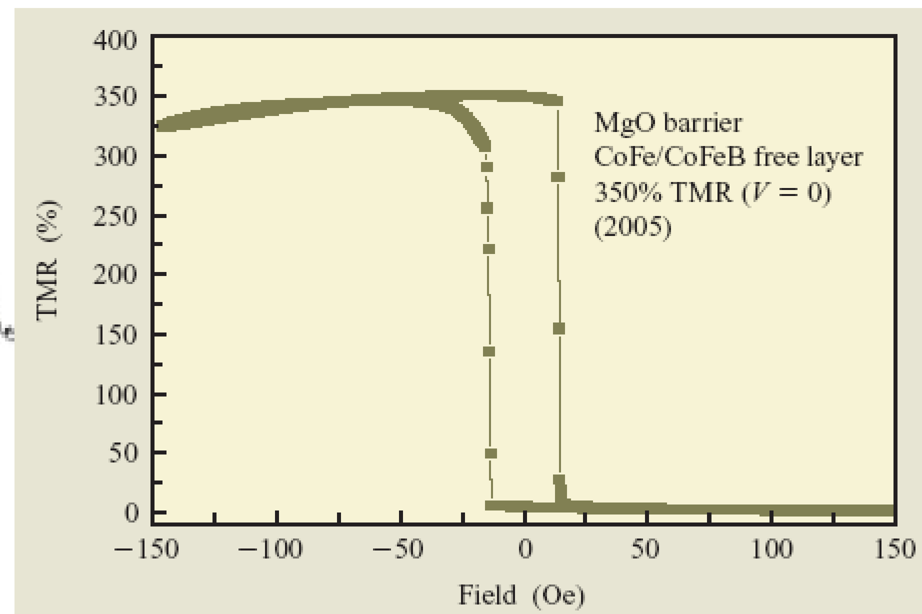
- perovskites (ABO_3) = 1 family of FE materials
- destructive read → forces need for high write endurance
- inherently fast, low-power, low-voltage
- first demonstrations ~1988

© 2008 IBM Corporation

MRAM (Magnetic RAM)





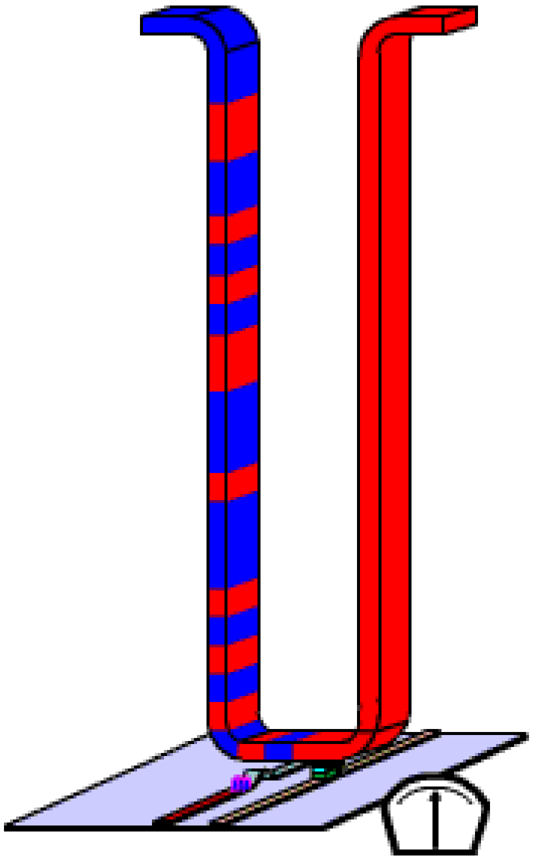


[Gallagher:2006]

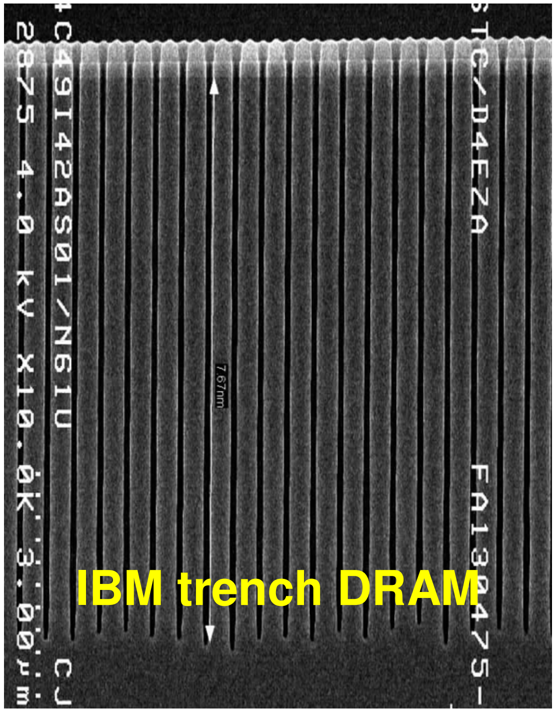
© 2008 IBM Corporation

Magnetic Racetrack Memory

a 3-D shift register



- Data stored as pattern of magnetic domains in long nanowire or “racetrack” of magnetic material.
- Current pulses move domains along racetrack
- Use deep trench to get many (**10-100**) bits per $4F^2$

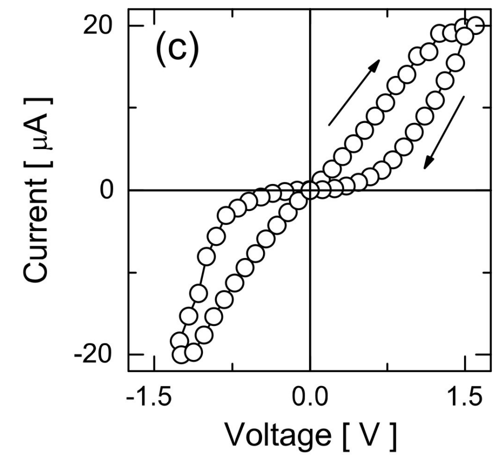


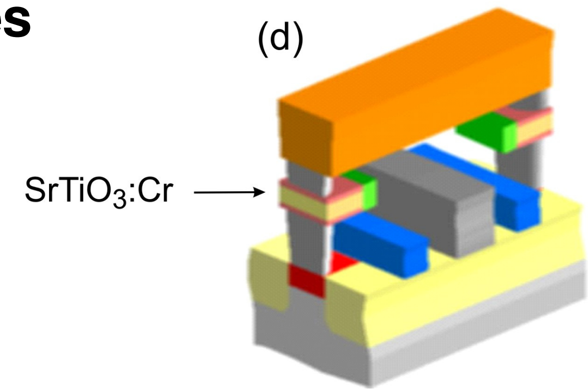
Magnetic Race Track Memory
S. Parkin (IBM), *US patents*
6,834,005 (2004) & 6,898,132 (2005)

© 2008 IBM Corporation

RRAM (Resistive RAM)

- Numerous examples of materials showing hysteretic behavior in their I-V curves
- Mechanisms not completely understood, but major materials classes include
  - metal nanoparticles(?) in **organics**
    - could they survive high processing temperatures?
  - oxygen vacancies(?) in **transition-metal oxides**
    - forming step sometimes required
    - scalability unknown
    - no ideal combination yet found of
      - low switching current
      - high reliability & endurance
      - high ON/OFF resistance ratio
  - metallic filaments in **solid electrolytes**





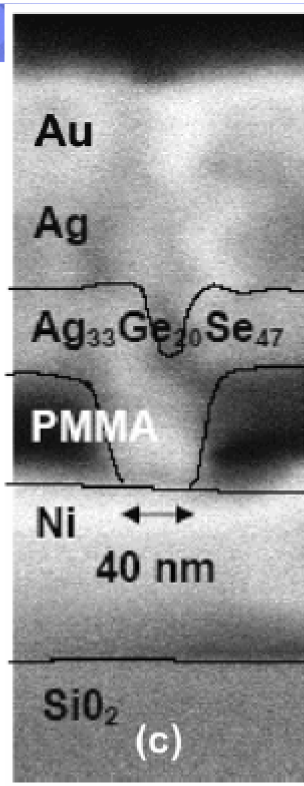
[Karg:2008]

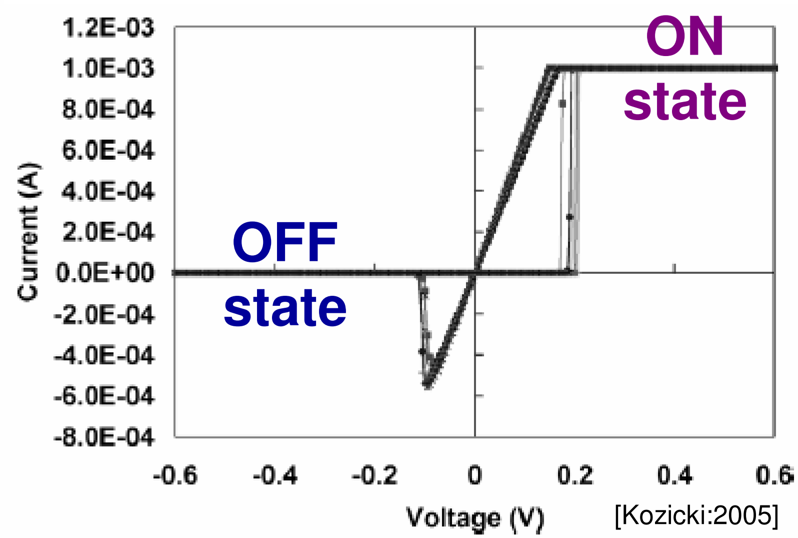
© 2008 IBM Corporation

Solid Electrolyte

Resistance contrast by forming a metallic filament through insulator sandwiched between an inert cathode & an oxidizable anode.

- Ag and/or Cu-doped Ge_xSe_{1-x}, Ge_xS_{1-x} or Ge_xTe_{1-x}
- Cu-doped MoO_x
- Cu-doped WO_x
- $RbAg_4I_5$ system





[Kozicki:2005]

Advantages

- Program and erase at very low voltages & currents
- High speed
- Large ON/OFF contrast
- Good endurance demonstrated
- Integrated cells demonstrated

Issues

- Retention
- Over-writing of the filament
- Sensitivity to processing temperatures (for GeSe, < 200^{o}C)
- Fab-unfriendly materials (Ag)

© 2008 IBM Corporation

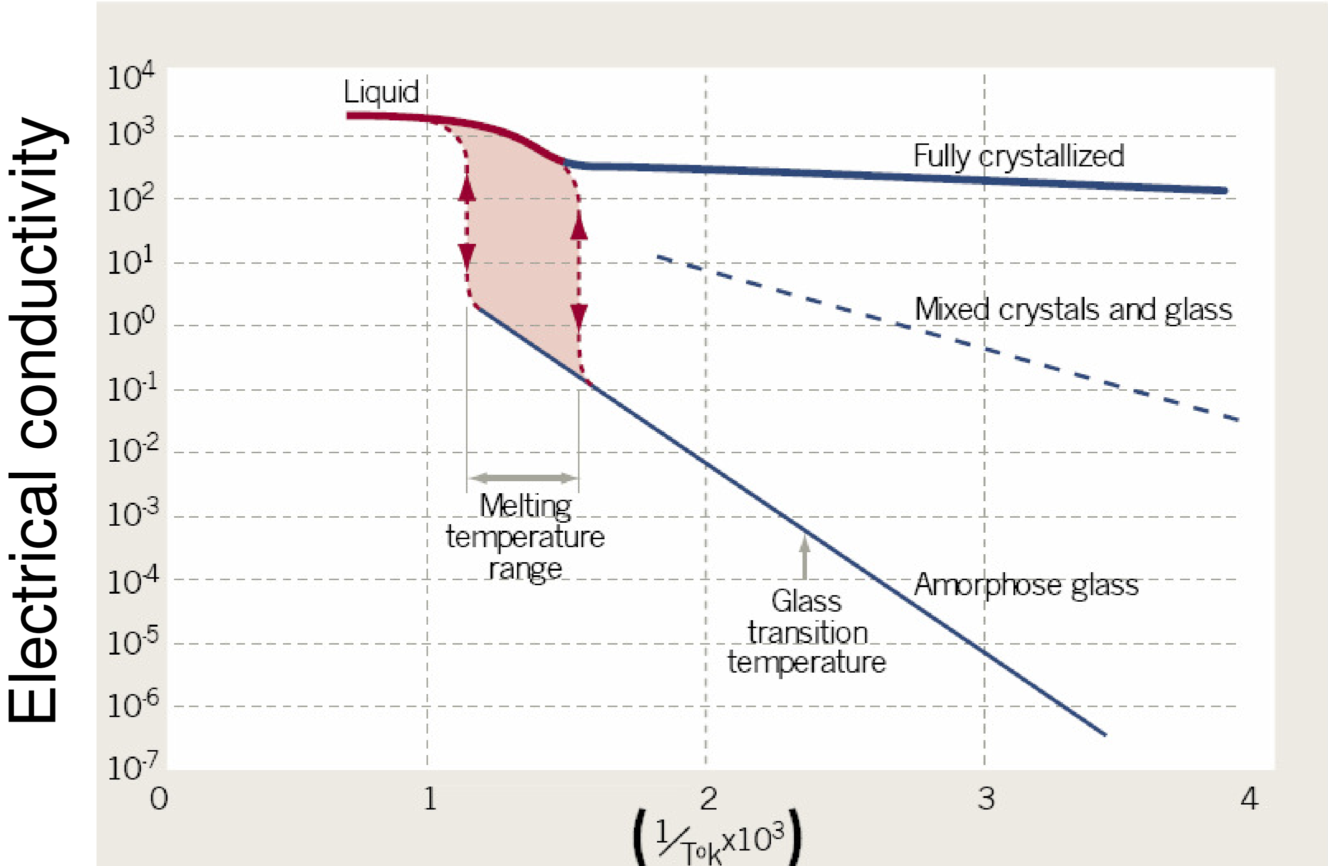
Candidate device technologies

- **Improved Flash**
  - **little improvement expected in write endurance or speed**
- **FeRAM – commercial product but difficult to scale!**
  - **FeFET – old concept, with many roadblocks**
- **MRAM – commercial product, also difficult to scale!**
  - **Racetrack memory – new concept w/ promise, still at point of early basic physics research**
- **RRAM – few demos showing real CMOS integration**
  - **Organic & polymer memory – temperature compatibility?**
- **Solid Electrolyte – shows real promise if tradeoff between retention & overprogramming can be solved…**
- **PC-RAM (Phase-change RAM)**

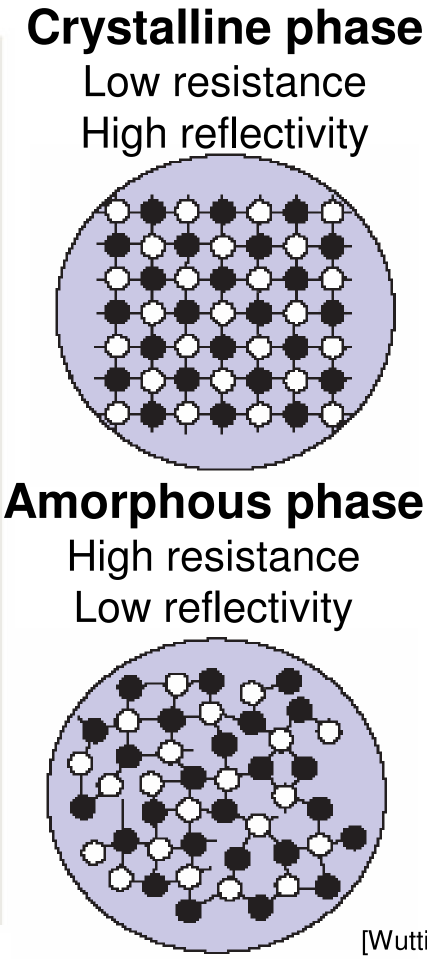
© 2008 IBM Corporation

History of Phase-change memory

- late 1960's – Ovshinsky shows reversible electrical switching in disordered semiconductors
- early 1970's – much research on mechanisms, but everything was too slow!



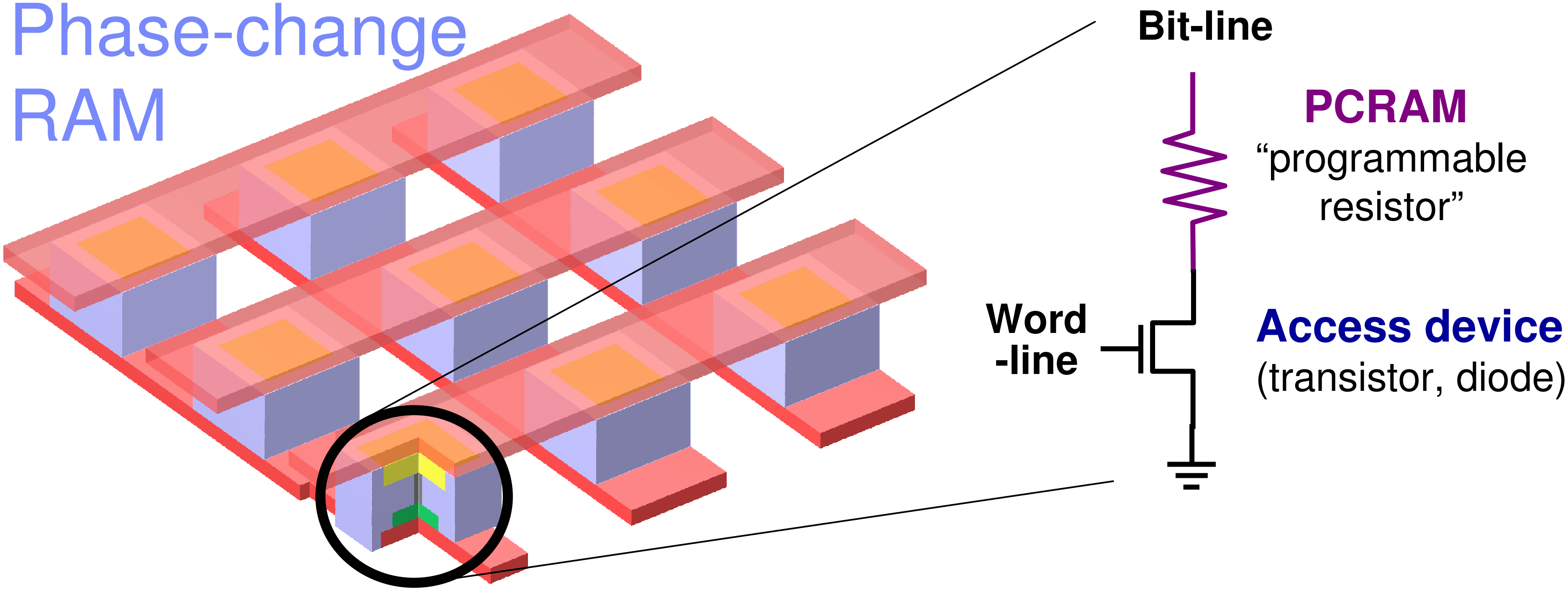
[Neale:2001]

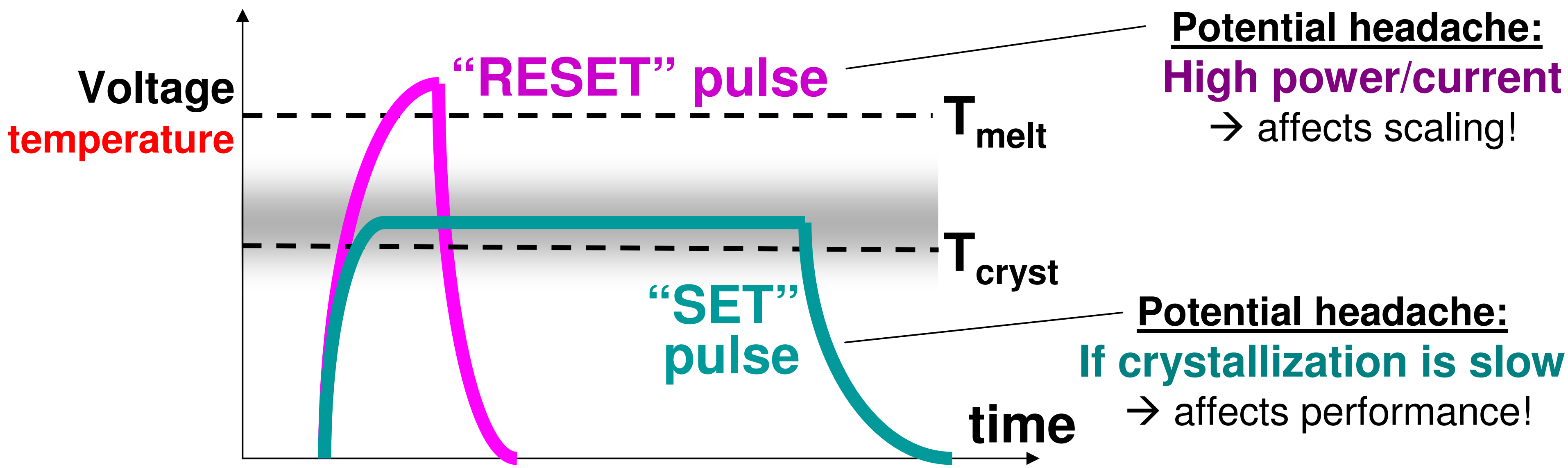


[Wuttig:2007]

© 2008 IBM Corporation

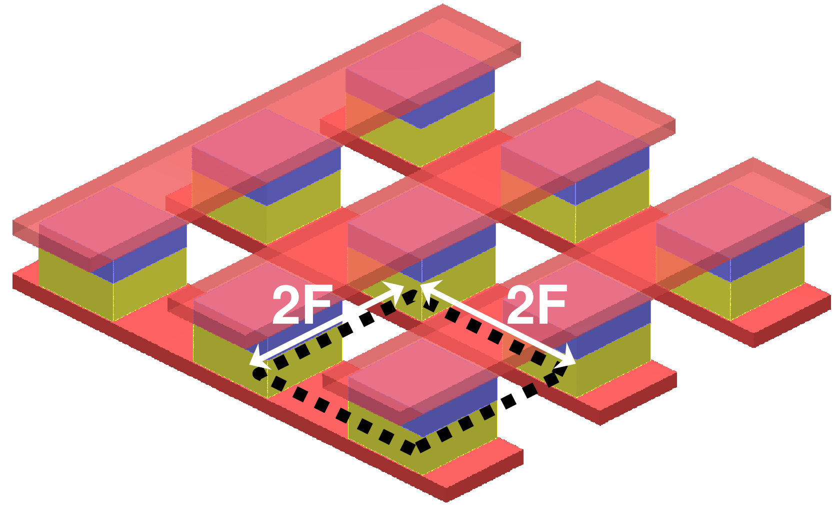
Phase-change RAM





 © 2008 IBM Corporation

Density is key



effective areal density.

Device	Critical feature-size **F**	Area (F^2)	**Density** (Gbit /sq. in)
Hard Disk	**100 nm** (MR width)	**0.5**	**125**
DRAM	**90 nm** (half pitch)	**8.0**	**10**
NAND (2 bit)	**90 nm** (half pitch)	**3.0**	**26**
NAND (1 bit)	**73 nm** (half pitch)	**4.7**	**26**
Blue Ray	**210 nm** (λ /2)	**1.5**	**12**

[Fontana:2004]

© 2008 IBM Corporation

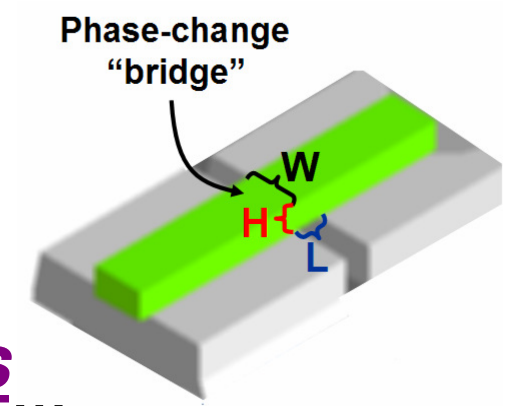
Phase-Change Nano-Bridge

- Prototype memory device with ultra-thin (**3nm**) films – Dec 2006

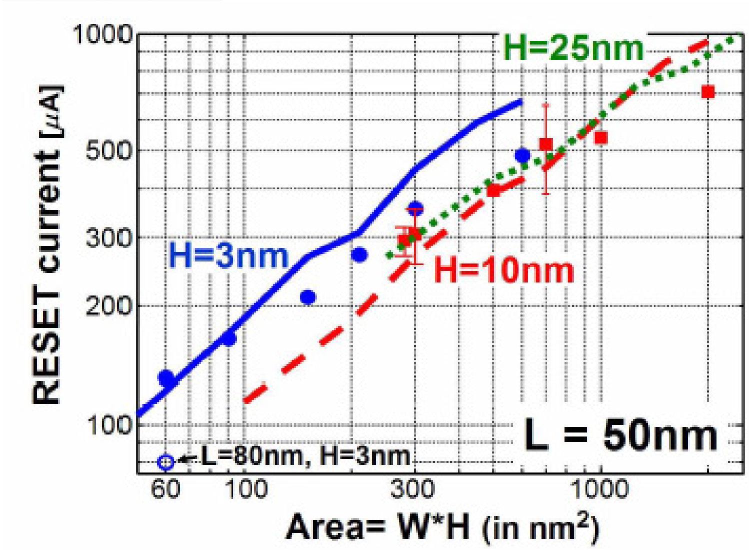
- 3nm * 20nm → $60nm^2$
  ≈ Flash roadmap for **2013**

→ phase-change *scales*

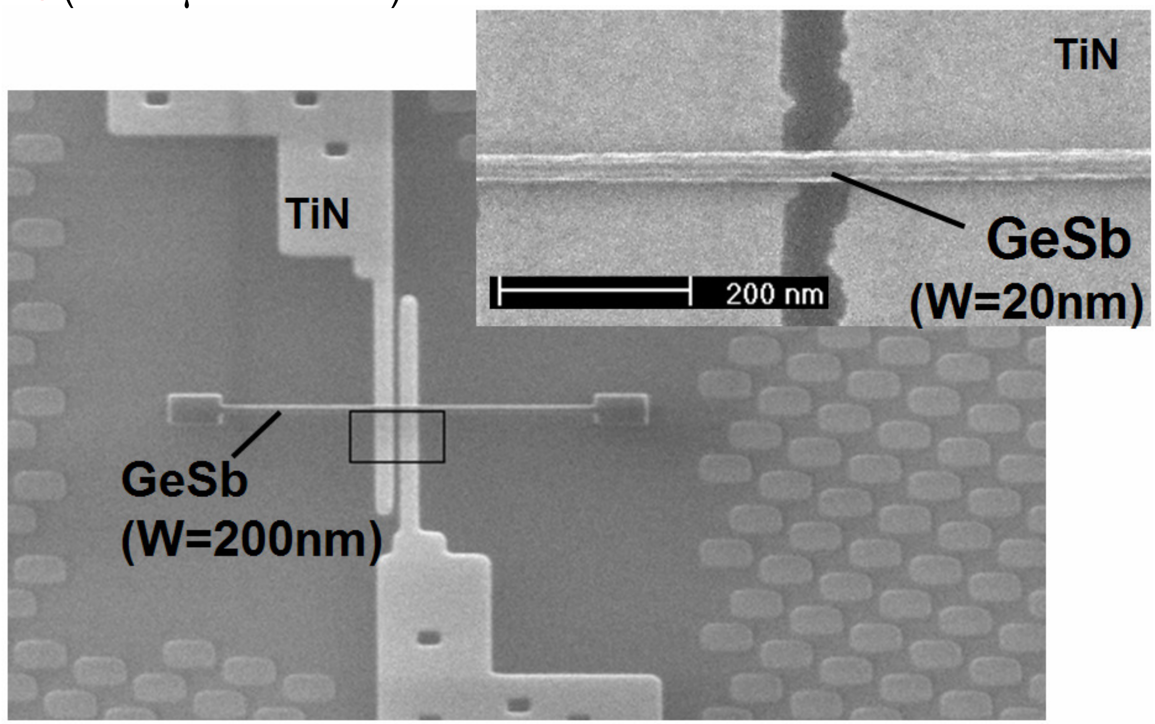
- **Fast** (<100ns SET)
- **Low current** (< 100µA RESET)



W defined by lithography
H by thin-film deposition



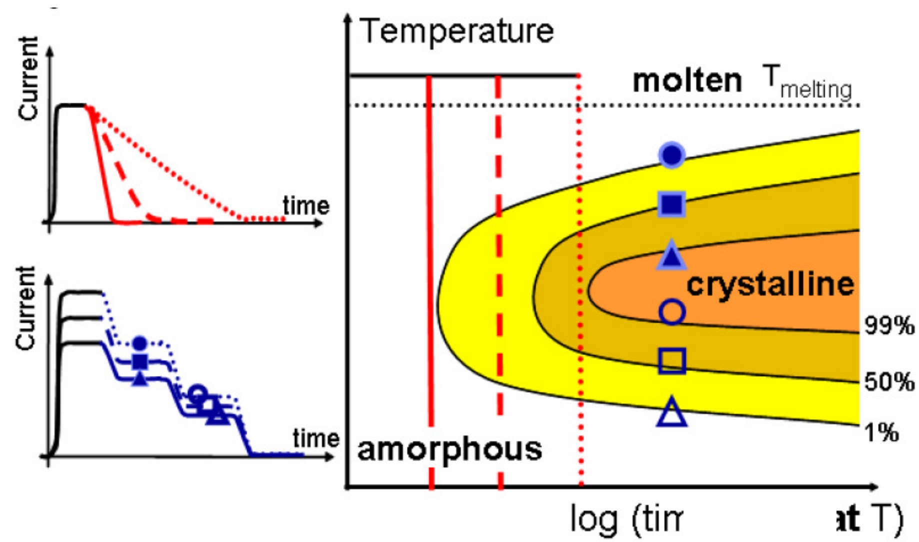
Current scales with area

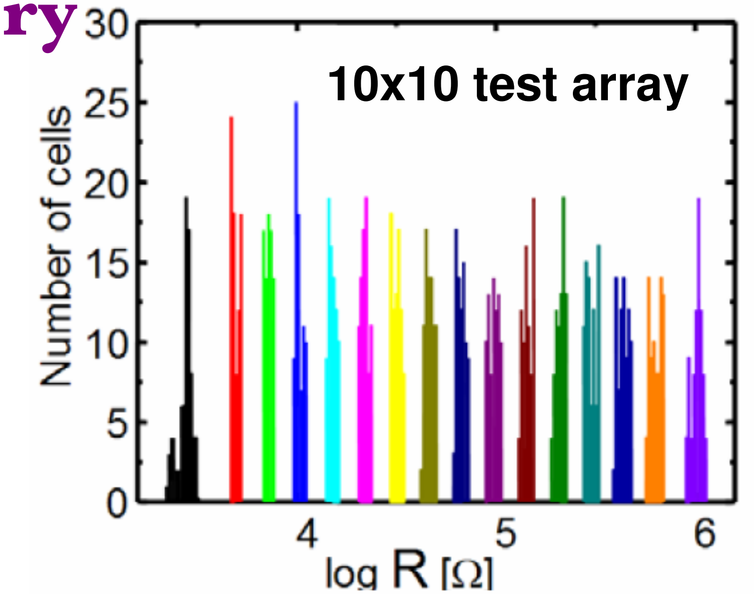


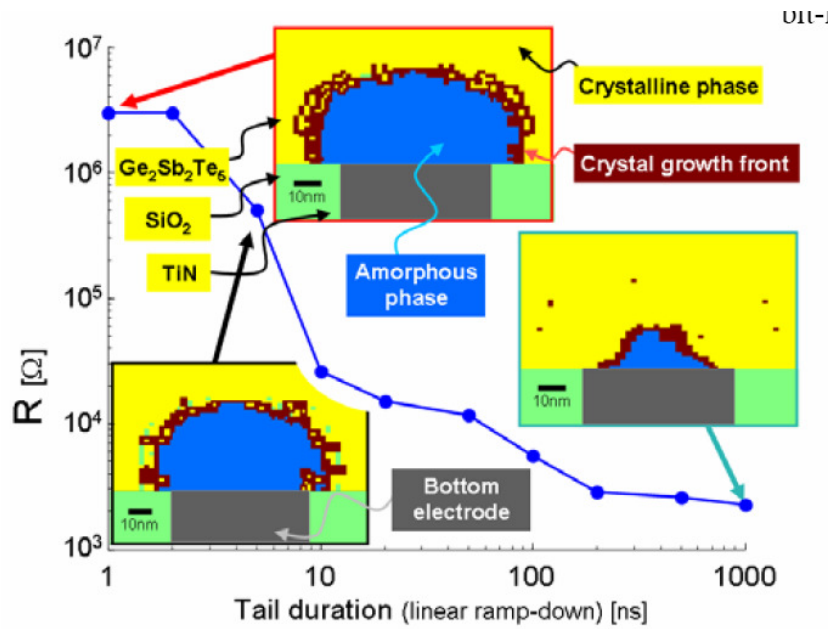
[Chen:2006]

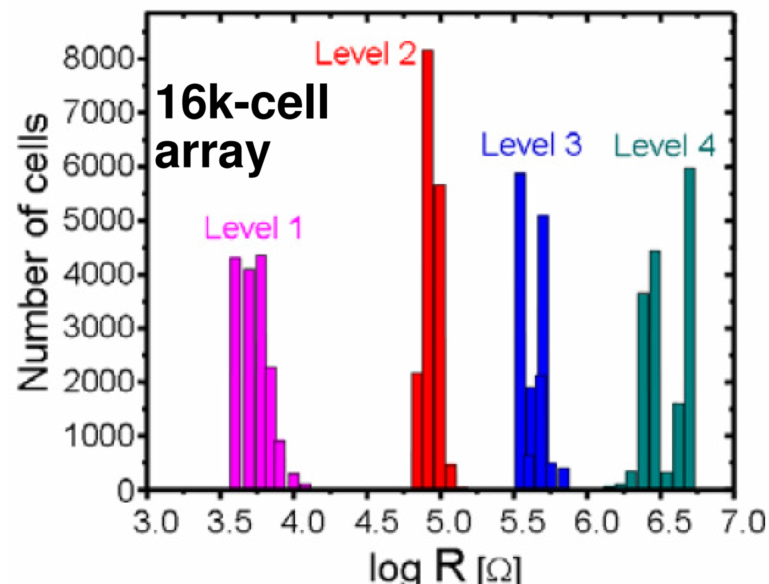
© 2008 IBM Corporation

Multi-level phase-change memory





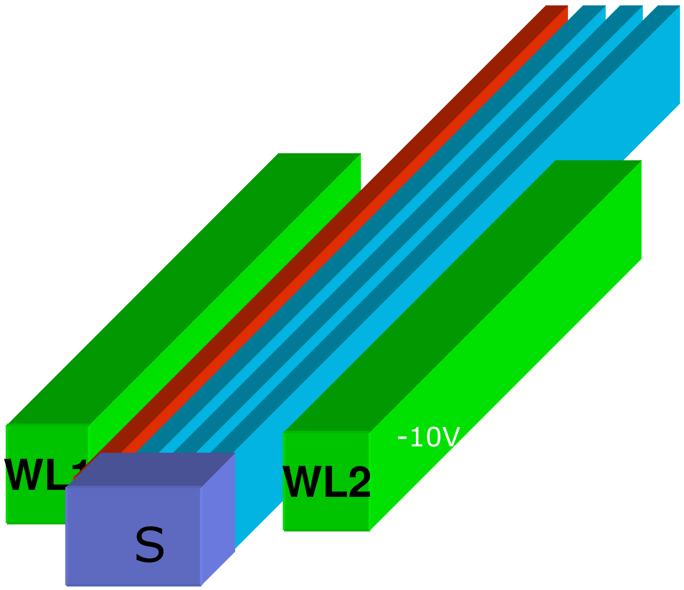


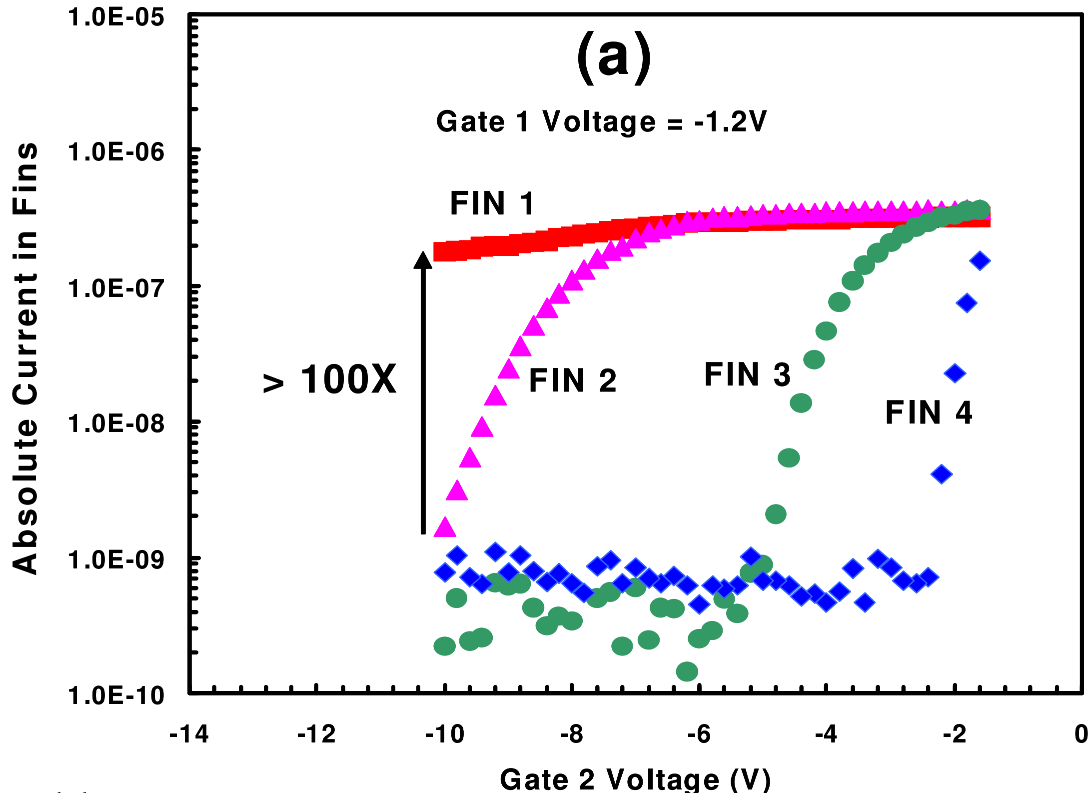


[Nirschl:2007]

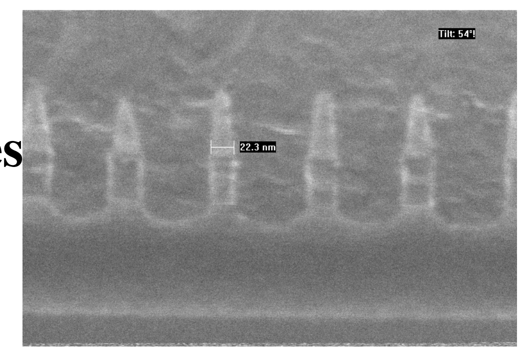
© 2008 IBM Corporation

Micro-Nanoscale Decoder





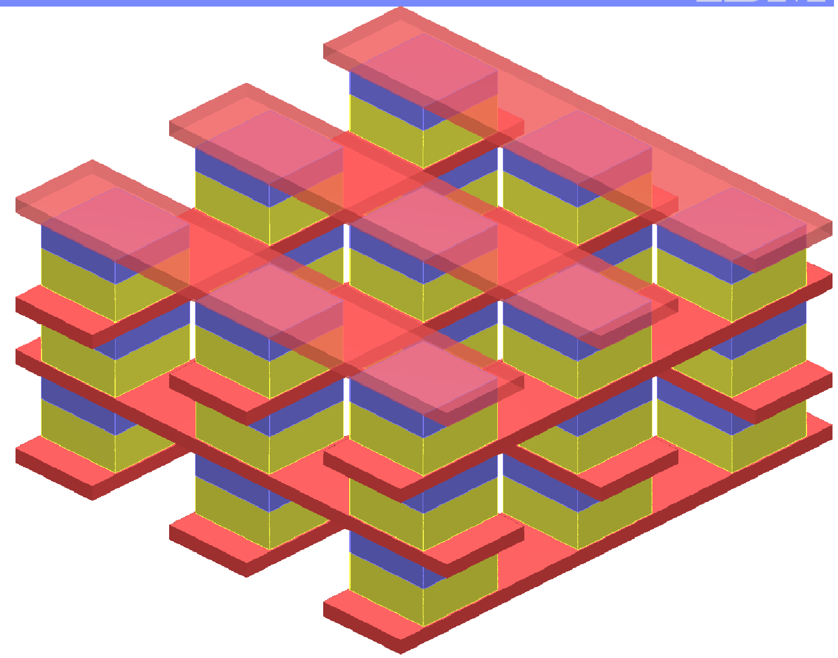
- **Sub-lithographic feature** is selected by moving depletion across the fine structure
- Modulating signal brought in by **lithographically-defined lines**
- Fins down to **sub-20 nm** have been addressed



[Gopalakrishnan:2005]

© 2008 IBM Corporation

3-D stacking

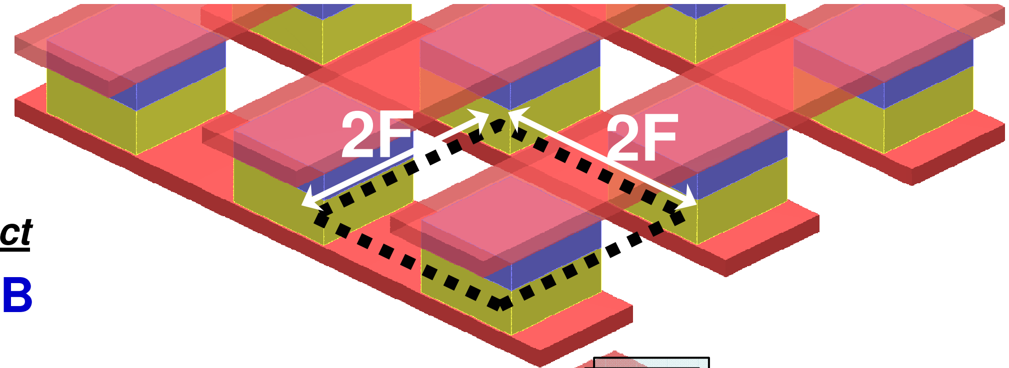


- Stack multiple layers of memory above the silicon in the CMOS back-end
- NOT the same as 3-D packaging of multiple wafers requiring electrical vias through-silicon
- Issues with temperature budgets, yield, and fab-cycle-time
- Still need access device within the back-end
  - re-grow single-crystal silicon (hard!)
  - use a polysilicon diode (but need good isolation & high current densities)
  - get diode functionality somehow else (nanowires?)

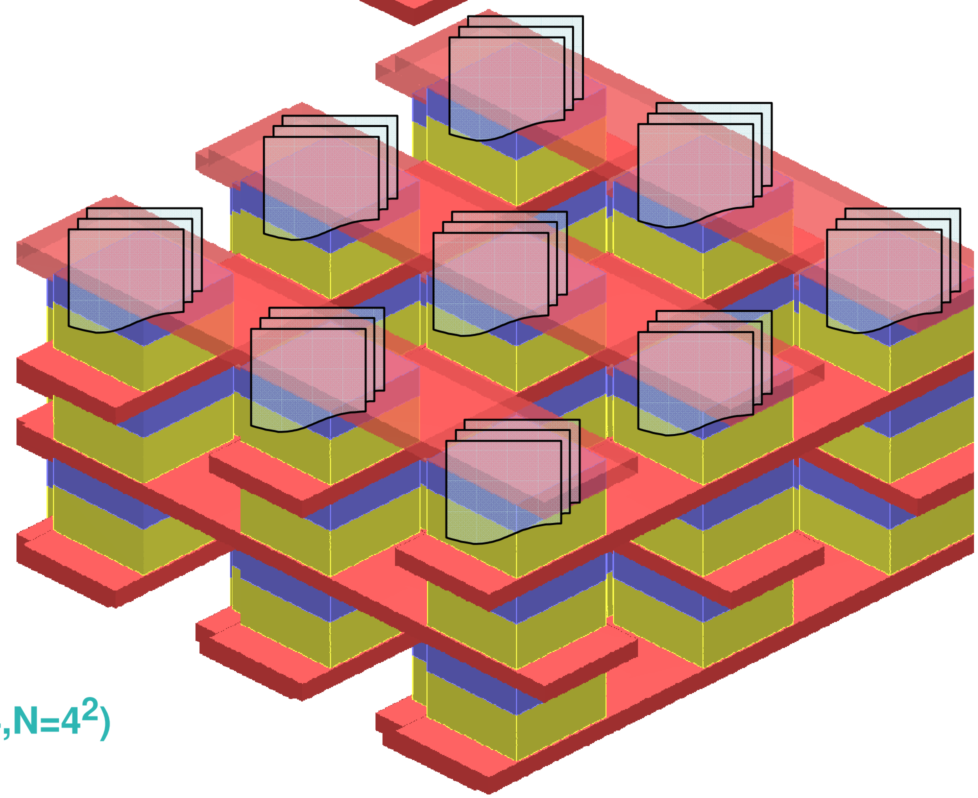
© 2008 IBM Corporation

Paths to ultra-high density memory

At the 32nm node in 2013,
MLC NAND Flash
(already M=2 → $2F^2$!)
is projected* to be at…



		density		product
	2x	**43** Gb/cm²	→	**32GB**
if we could shrink $4F^2$ by…	**4x**	**86** Gb/cm²	→	**64GB**
		e.g., **4 layers of 3-D** (L=4)		
	16x	**344** Gb/cm²	→	**256GB**
		e.g., **8 layers of 3-D, 2 bits/cell** (L=8,M=2)		
	64x	**1376** Gb/cm²	→	**~1 TB**
		e.g., **4 layers of 3-D, 4x4 sublithographic** ($L=4,N=4^2$)		



* 2006 ITRS Roadmap

© 2008 IBM Corporation

If you could have SCM, why would you need anything else?

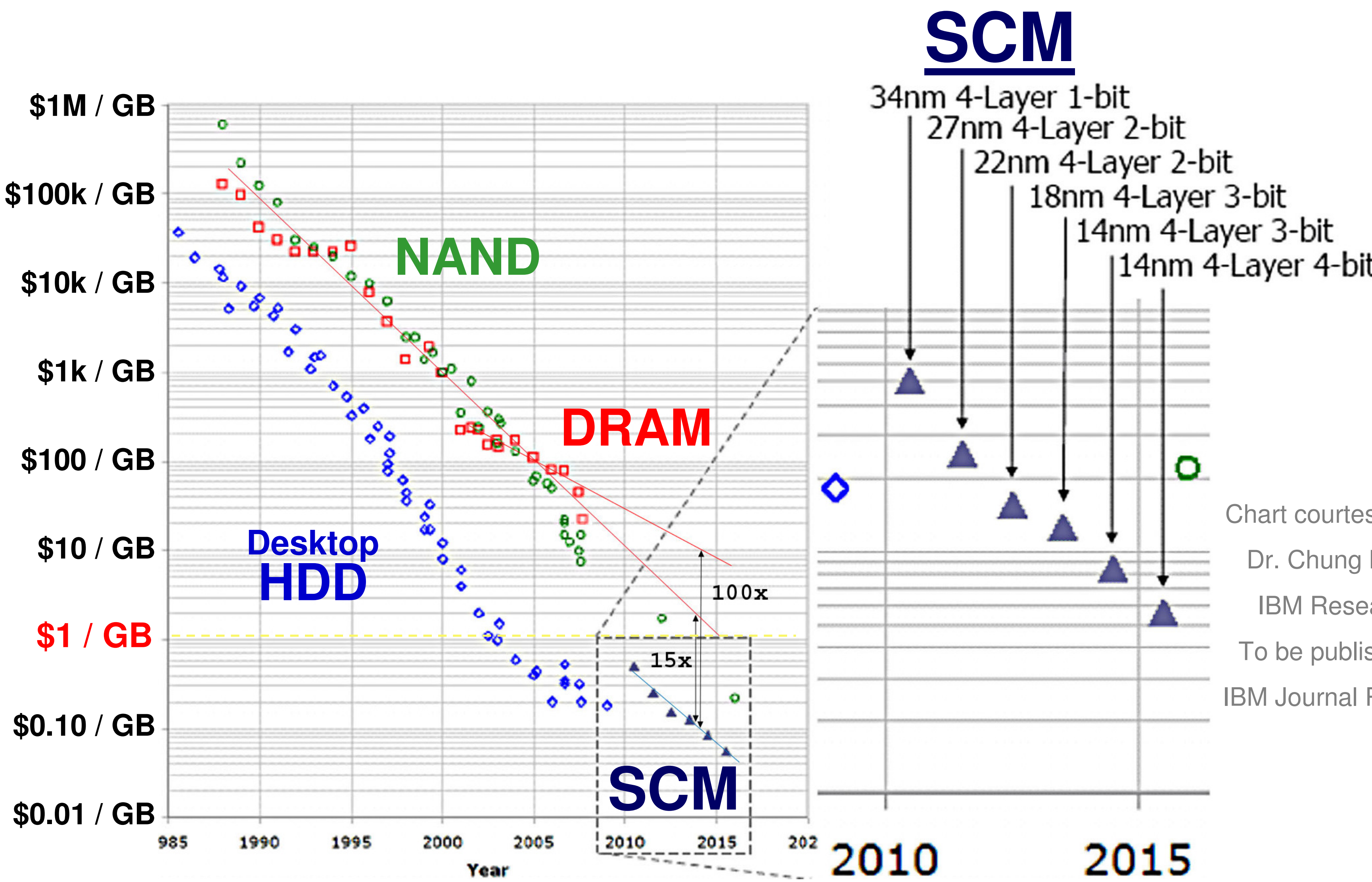


Chart courtesy of Dr. Chung Lam IBM Research To be published IBM Journal R&D

 © 2008 IBM Corporation

In comparison…

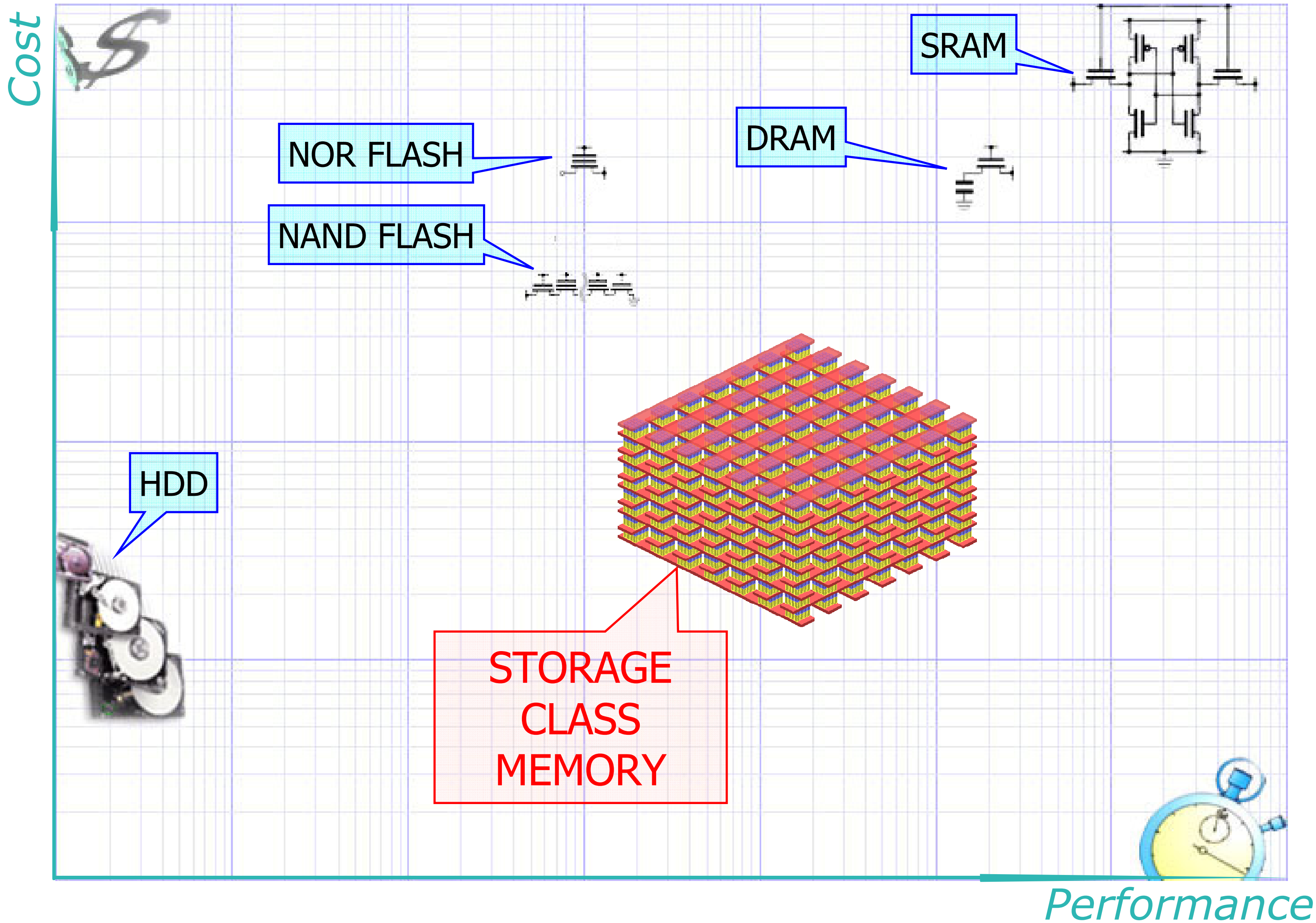
	Flash	SONOS Flash	Nanocrystal Flash	FeRAM	FeFET
Knowledge level	product	advanced development	development	product	basic research
Smallest demonstrated cell	**$4F^2$** ($2F^2$ per bit)	**$4F^2$** ($1F^2$ per bit)	$16F^2$ (@90nm)	$15F^2$ (@130nm)	—
Prospects for… …scalability	**poor**	**maybe** (enough stored charge?)	**unclear** (enough stored charge?)	**poor** (integration, signal loss)	**unclear** (difficult integration)
…fast readout	yes	yes	yes	yes	yes
…fast writing	**NO**	**NO**	**NO**	yes	yes
…low switching Power	yes	yes	yes	yes	yes
…high endurance	**NO**	**poor** (1e7 cycles)	**NO**	yes	yes
…non-volatility	yes	yes	yes	yes	**poor** (30 days)
…MLC operation	yes	yes	yes	**difficult**	**difficult**

© 2008 IBM Corporation

	MRAM	Racetrack	PCRAM	RRAM	solid electrolyte	organic memory
Knowledge level	product	basic research	advanced development	Early development	development	basic research
Smallest demonstrated cell	**$25F^2$** @180nm	—	**$5.8F^2$** (diode) **$12F^2$** (BJT) @90nm	—	**$8F^2$** @90nm ($4F^2$ per bit)	—
Prospects for… …scalability	**poor** (high currents)	**unknown** (too early to know, good potential)	**promising** (rapid progress to date)	**unknown**	**promising** (filament-based, but new materials)	**unknown** (high temp-eratures?)
…fast readout	yes	yes	yes	yes	yes	**sometimes**
…fast writing	yes	yes	yes	**sometimes**	yes	**sometimes**
…low switching Power	**NO**	**uncertain**	**poor**	**sometimes**	yes	**sometimes**
…high endurance	yes	should	yes	**poor**	**unknown**	**poor**
…non-volatility	yes	**unknown**	yes	**sometimes**	**sometimes**	**poor**
…MLC operation	**NO**	yes **(3-D)**	yes	yes	yes	**unknown**

© 2008 IBM Corporation

How does SCM compare to existing technologies?



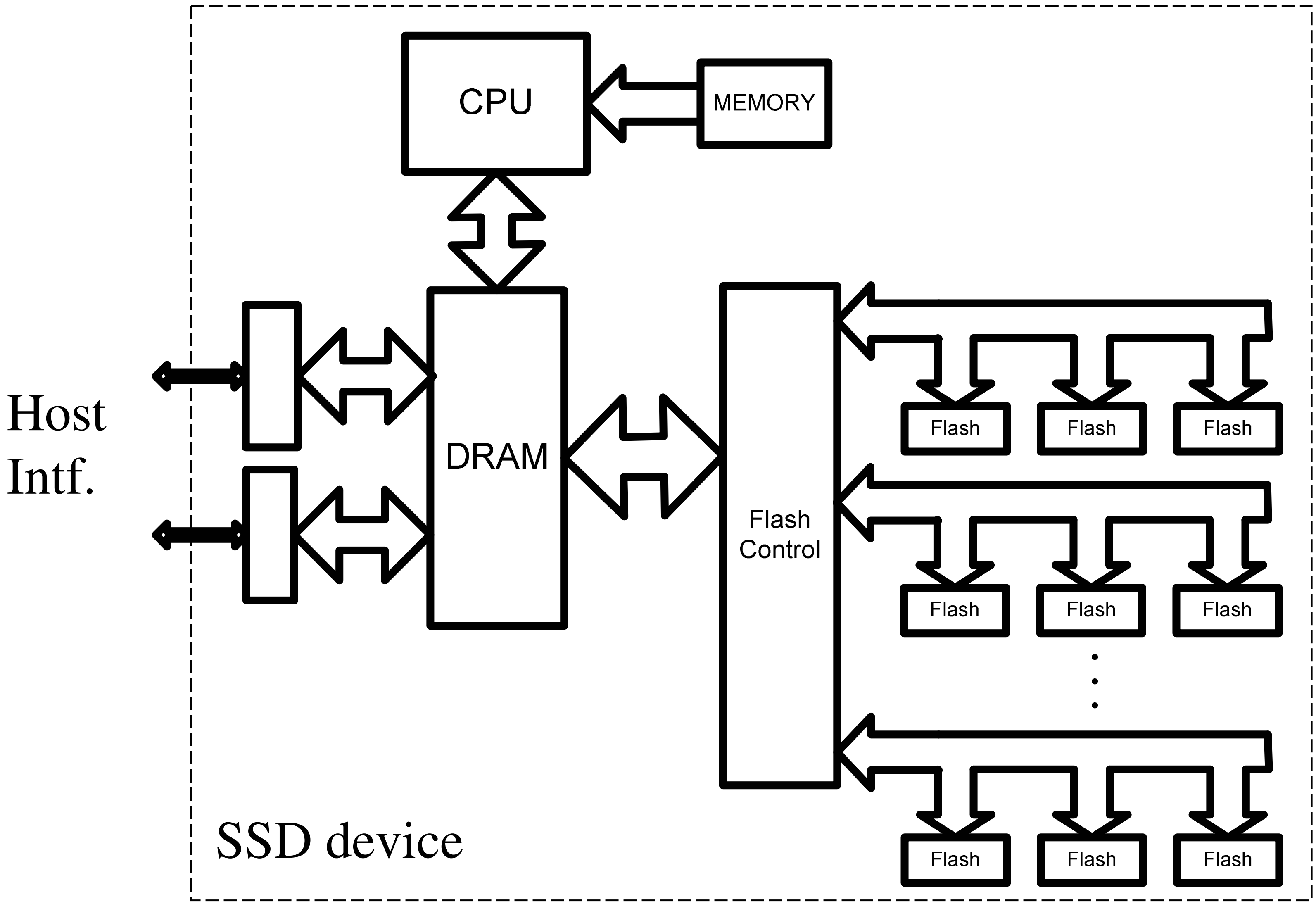
 © 2008 IBM Corporation

Challenges with SCM

- **Asymmetric performance**
  - –Flash: writes much slower than reads
  - –Not as pronounced in other technologies
- **Write endurance**
  - –Many SCM technologies “wear out” on writes
  - –Flash is an example
- **Bad blocks**
  - –Devices are shipped with bad blocks
  - –Blocks wear out, etc.

© 2008 IBM Corporation

SCM: Flash Storage Design



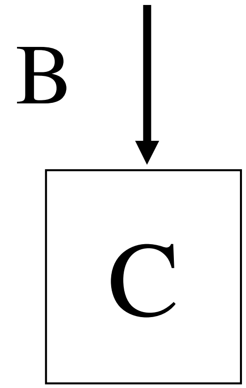
© 2008 IBM Corporation

Write endurance

- **In many SCM technologies writes are cumulatively destructive**
- **For Flash it is the program/erase cycle**
- **Current commercial flash and SCM varieties**
  - Single level cell (SLC) → 10^5 writes/cell
  - Multi level cell (MLC) → 10^4 writes/cell
  - PCM → ~10^8 writes/cell
- **Wear leveling**

© 2008 IBM Corporation

Fill-times and Life-times of SCM devices



$\mathbf{T_{fill} = C/B}$ **(Fill Time)**

= time to write all C Bytes, given bandwidth B

T_{fill} ~ 1 sec for DRAM , ~ 10,000 seconds for disks

Without any wear-leveling, $\mathbf{T_{life} = T_{fill}}$ = very bad

---

(Perfect) Wear-leveling improves $\mathbf{T_{life}}$ by Write Endurance Number **E**

$$\mathbf{T_{life} = E \cdot T_{fill} = E \cdot C/B}$$

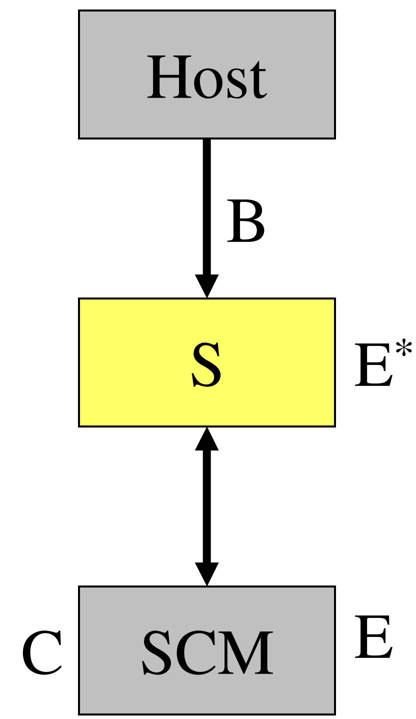
From seconds to years!

Write Endurance — Capacity — Bandwidth

© 2008 IBM Corporation

Lifetime model (more details)

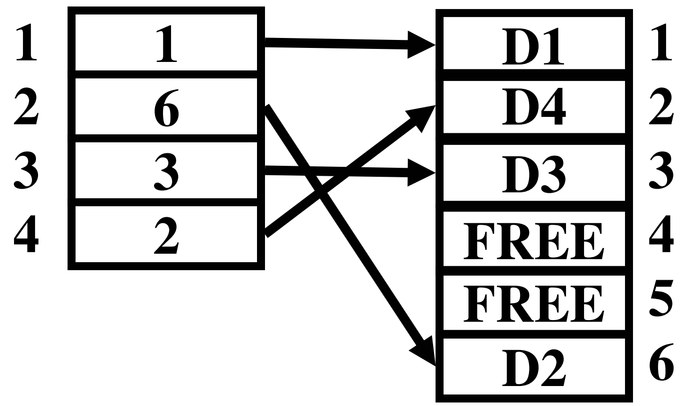
- **S are system level management 'tools' providing an effective endurance of $E^* = S(E) = E$**
  - E is the Raw Device endurance and
  - E^* is the *effective Write Endurance*
- **S includes**
  - Static and dynamic wear leveling of efficiency $q < 1$
  - Error Correction and bad block management
  - Overprovisioning
  - Compress, de-duplicate & write elimination…
  - $E^* = E * q * f(\text{error correction}) * g(\text{overprovisioning}) * h(\text{compress})\ldots$
  - With S included, $T_{life}(\text{System}) = T_{fill} * E^*$



© 2008 IBM Corporation

Dynamic wear leveling

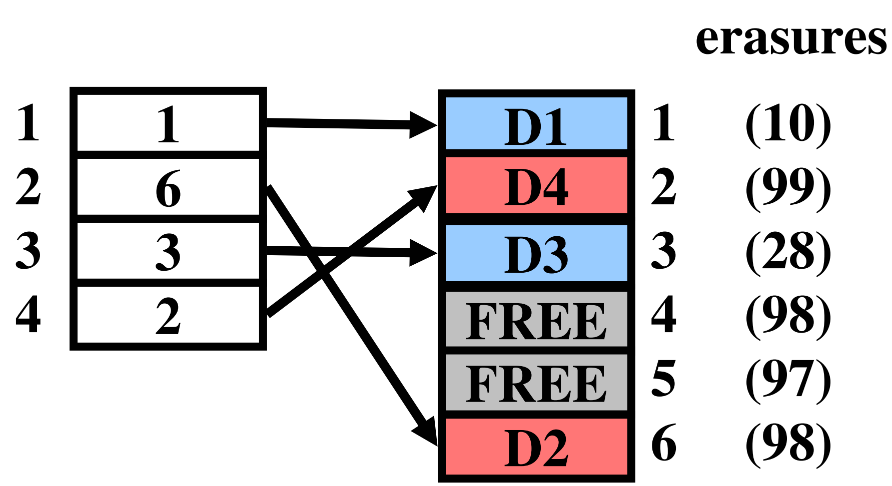
- **Frequently written data – logs, updates, etc.**
- **Maintain a set of free, erased blocks**
- **Logical to physical block address mapping**
- **Write new data of free block**
- **Erase old location and add to free list.**



© 2008 IBM Corporation

Static wear leveling

- Infrequently written data – OS data, etc
- Maintain count of erasures per block
- Goal is to keep counts "near" each other
- Simple example: move data from hot block to cold block
  - Write LBA 4
  - D1 → 4
  - 1 now FREE
  - D4 → 1



© 2008 IBM Corporation

Paths Forward for SCM

- **direct disk replacement with SCM packaged as a SSD**
- **PCIe card that supports a high bandwidth local or direct attachment to a processor.**
- **PCIe connected drawer that provides a large scale sharable storage system**
- **design the storage system or the computer system around SCM from the start**

© 2008 IBM Corporation

SCM module ‘Specs’ in 2020

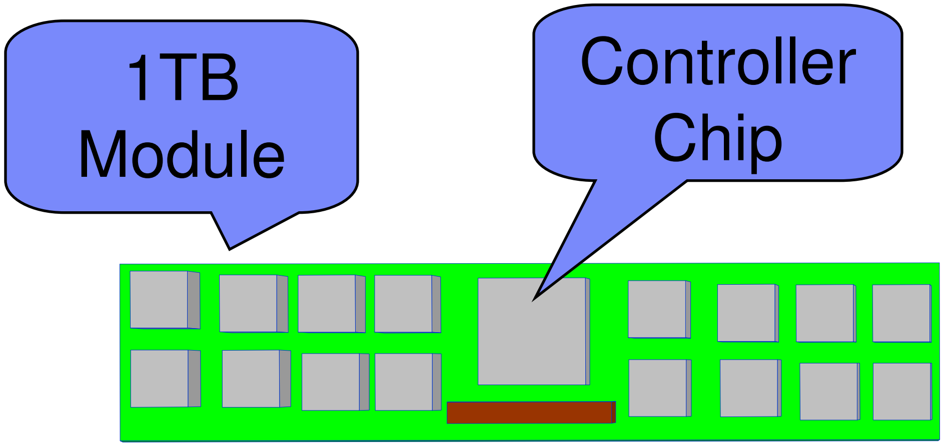
- **SCM modules <u>may be</u> block oriented storage devices**

Capacity	**1 TB**
Read or Write Access Time	**<1 us**
Data Rate	**>1GB/s**
Sustained transaction rate –1us + 4K / 1GB/s = 5us	**200,000 IOPS**
Sustained bandwidth –4KB/5us = >800MB/s	**800MB/s**

© 2008 IBM Corporation

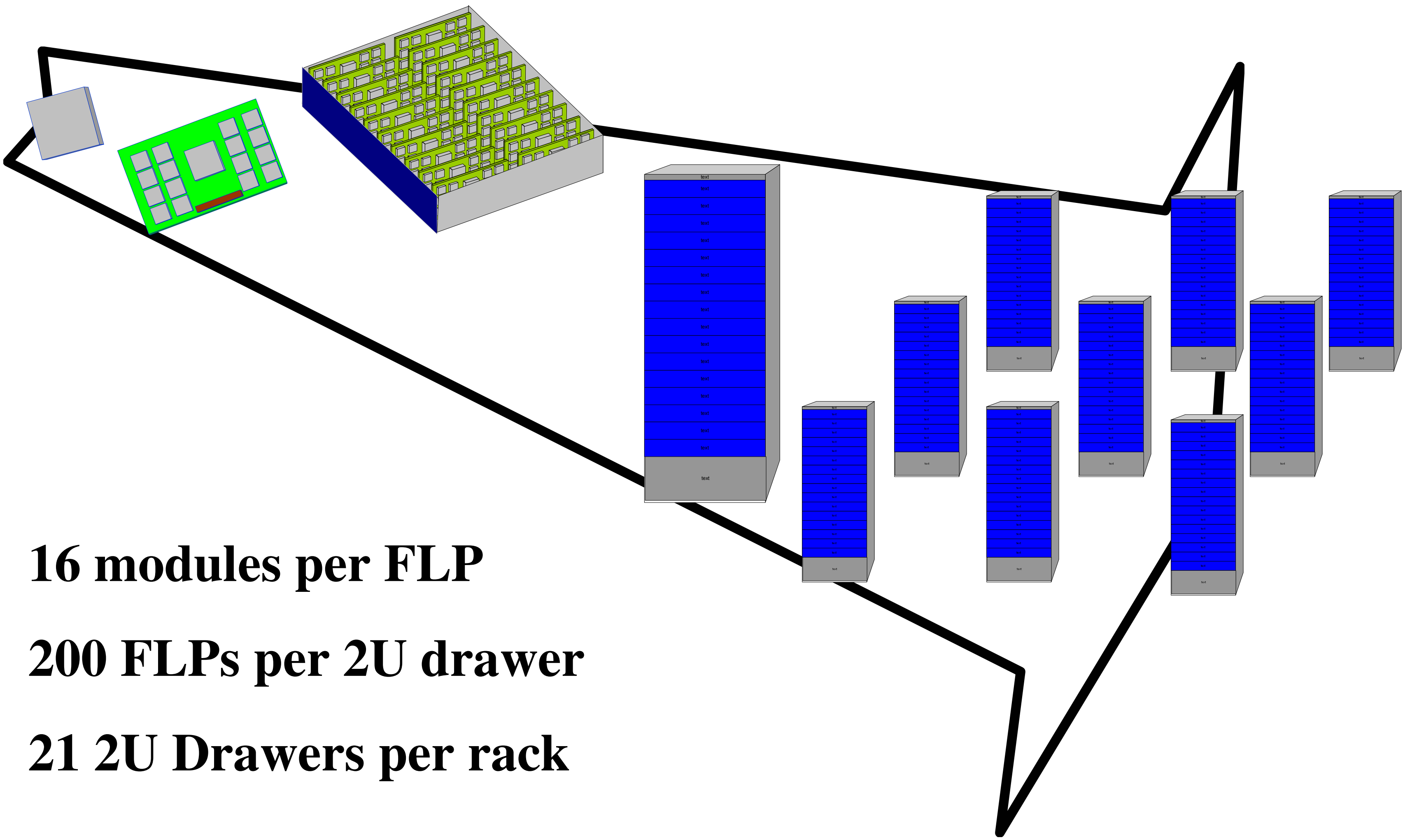
Basic 2020 Storage Package

- **Nonvolatile memory first level package (FLP) (think DIMM)**
- **FLP controller works in concert with other FLP controllers to manage performance, reliability and power**
  - modules checked by controller
  - Redundancy across first level package
  - Detects and attempts to resolve failures
  - Wear leveling
- **16 modules**
  - 1 TB → 16 TB
  - 800 MB/s → 12.8 GB/s
  - 200 kIOPS → 8 MIOPS



© 2008 IBM Corporation

2020 SCM Storage System Package



16 modules per FLP

200 FLPs per 2U drawer

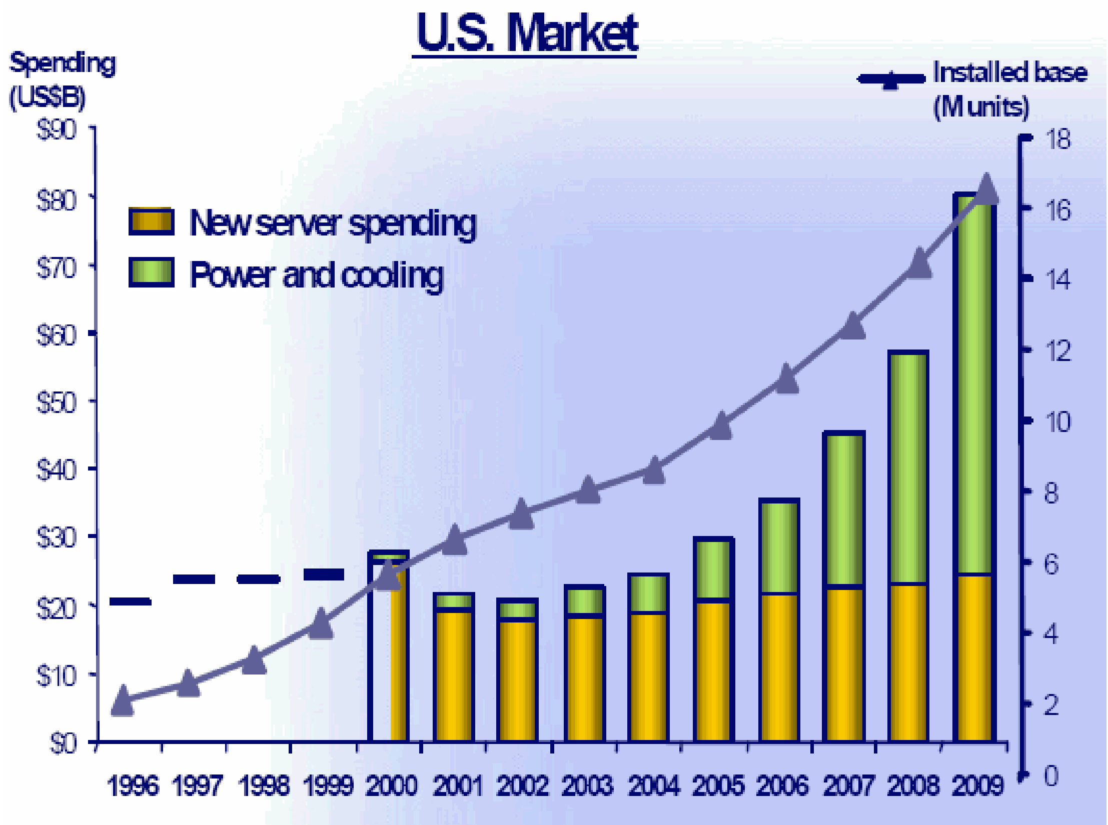
21 2U Drawers per rack

© 2008 IBM Corporation

Power & space in the server room

The cache/memory/storage hierarchy is rapidly becoming the **bottleneck for large systems**.

We know how to create MIPS & MFLOPS cheaply and in abundance,
but **feeding them with data** has become
the performance-limiting *and* most-expensive part of a system (in **both $ and Watts**).



Source IDC: 2006, Document # 201722, "The Impact Of Power and Cooling On Data Center Infrastructure", John Humphreys, Jed Scaramella

Extrapolation to 2020

(at 70% CGR → need **2 GIOP/sec**)

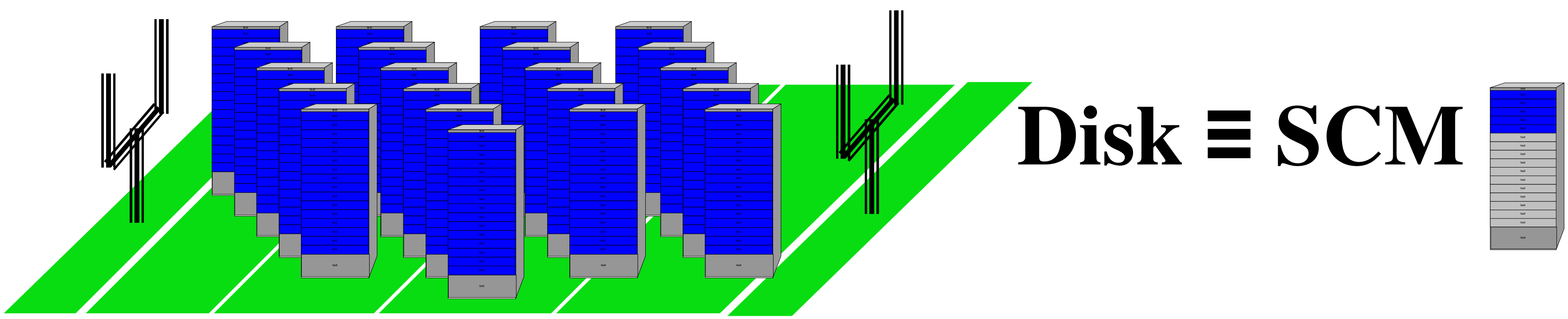
- **5 million HDD**
  - **16,500** sq. ft. !!
  - **22 Megawatts**

R. Freitas and W. Wilcke, *Storage Class Memory: the next storage system technology* –to appear in "Storage Technologies & Systems" special issue of the IBM Journal of R&D.

© 2008 IBM Corporation

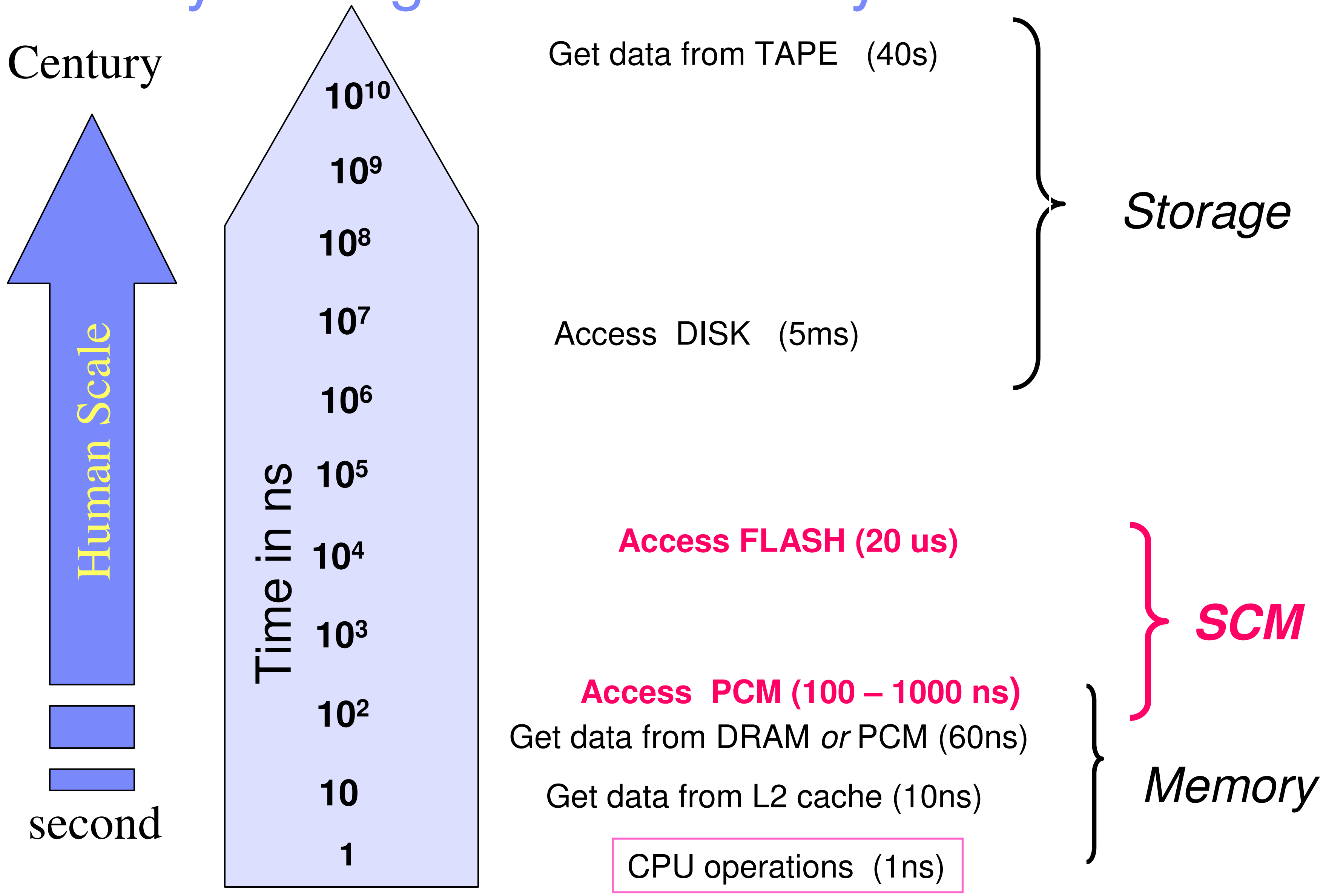
Results of Extrapolation

	Compute centric		Data centric	
	disk	SCM	disk	SCM
Devices	1.3 M Disks	406 K modules	5 M Disks	8 K modules
space	4500 sq.ft.	85 sq. ft.	16,500 sq.ft.	12 sq. ft.
power	6,000 kW	41 kW	22,000 kW	1 kW



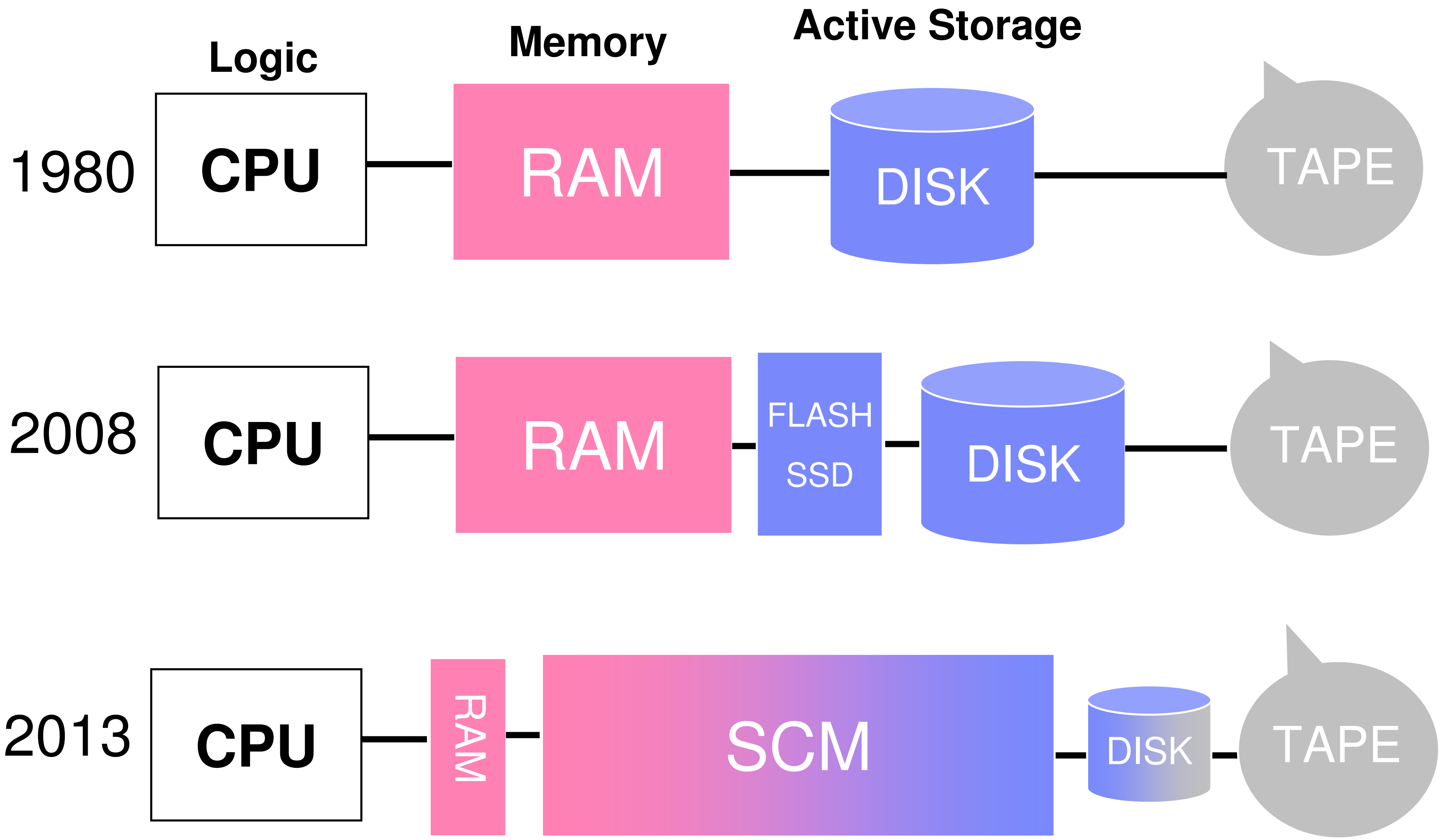
© 2008 IBM Corporation

Memory/Storage Stack Latency Problem



© 2008 IBM Corporation

SCM in a large System



© 2008 IBM Corporation

Shift in Systems and Applications

	▪ ***DRAM – Disk – Tape***	▪ ***DRAM – SCM – Disk – Tape***
Main Memory:	– Cost & power constrained – Paging not used – Only one type of memory: volatile	– Much larger memory space for same power and cost – Paging viable – Memory pools: different speeds, some persistent
Storage:	– Active data on disk – Inactive data on tape – SANs in heavy use	– Active data on SCM – Inactive data on disk/tape – DAS ??
Applications:	– Compute centric – Focus on hiding disk latency	– Data centric comes to fore – Focus on efficient memory use and exploiting persistence

© 2007 IBM Corporation

Summary

- **SCM in the form of Flash and PCM are here today and real. Others will follow.**
- **SCM will have a significant impact on the design of current and future systems and applications**

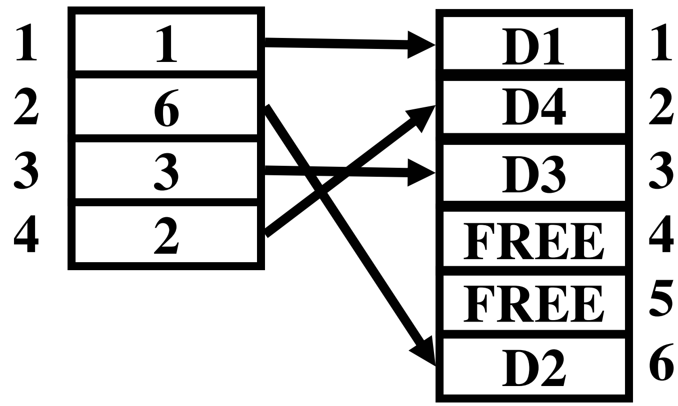
© 2007 IBM Corporation

Questions

© 2007 IBM Corporation

Dynamic wear leveling

- **Frequently written data – logs, updates, etc.**
- **Maintain a set of free, erased blocks**
- **Logical to physical block address mapping**
- **Write new data of free block**
- **Erase old location and add to free list.**

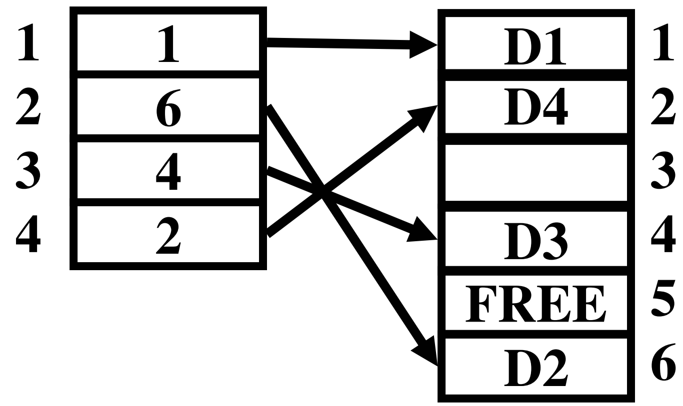


(animated)

© 2008 IBM Corporation

Dynamic wear leveling

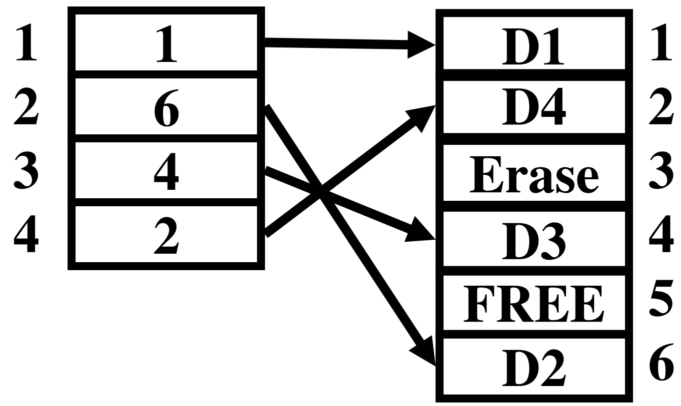
- **Frequently written data – logs, updates, etc.**
- **Maintain a set of free, erased blocks**
- **Logical to physical block address mapping**
- **Write new data of free block**
- **Erase old location and add to free list.**



© 2008 IBM Corporation

Dynamic wear leveling

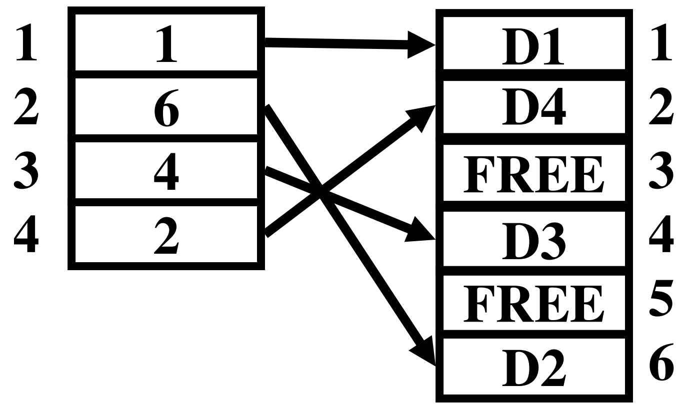
- **Frequently written data – logs, updates, etc.**
- **Maintain a set of free, erased blocks**
- **Logical to physical block address mapping**
- **Write new data of free block**
- **Erase old location and add to free list.**



© 2008 IBM Corporation

Dynamic wear leveling

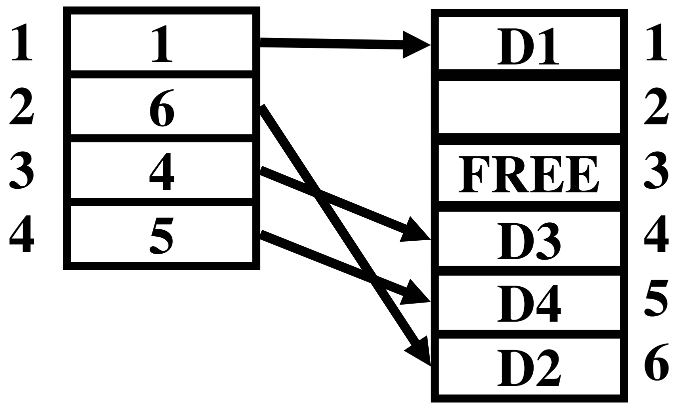
- **Frequently written data – logs, updates, etc.**
- **Maintain a set of free, erased blocks**
- **Logical to physical block address mapping**
- **Write new data of free block**
- **Erase old location and add to free list.**



© 2008 IBM Corporation

Dynamic wear leveling

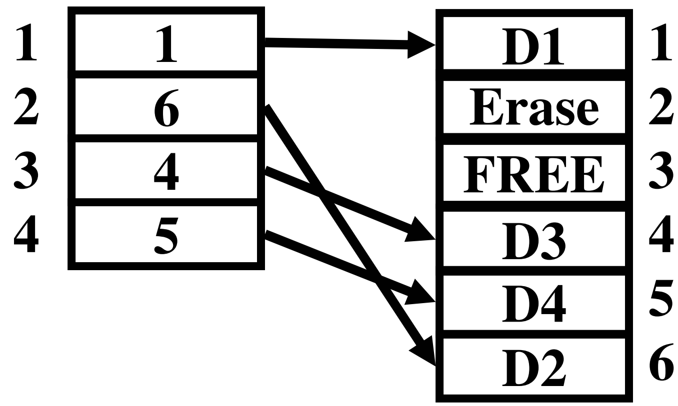
- **Frequently written data – logs, updates, etc.**
- **Maintain a set of free, erased blocks**
- **Logical to physical block address mapping**
- **Write new data of free block**
- **Erase old location and add to free list.**



© 2008 IBM Corporation

Dynamic wear leveling

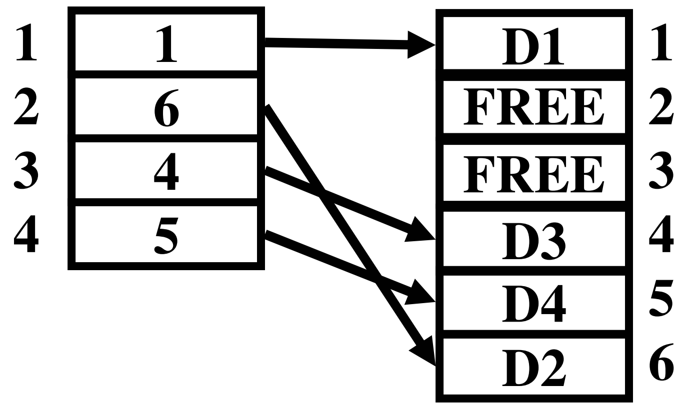
- **Frequently written data – logs, updates, etc.**
- **Maintain a set of free, erased blocks**
- **Logical to physical block address mapping**
- **Write new data of free block**
- **Erase old location and add to free list.**



© 2008 IBM Corporation

Dynamic wear leveling

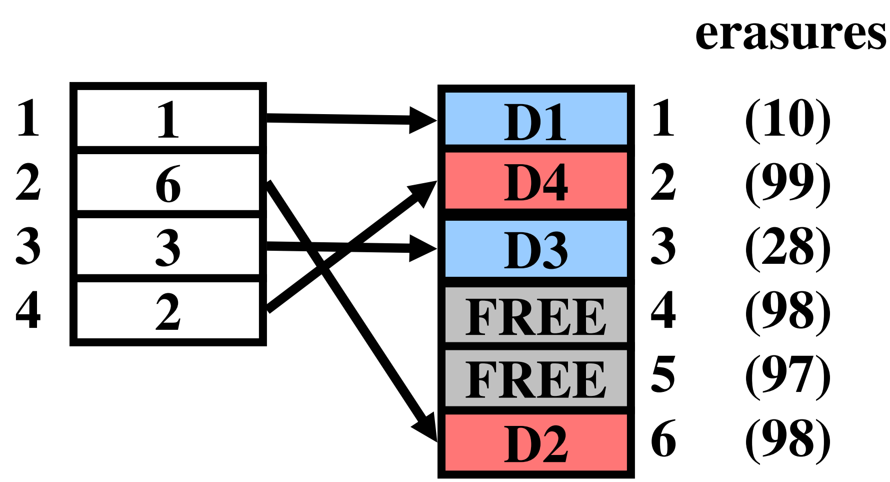
- **Frequently written data – logs, updates, etc.**
- **Maintain a set of free, erased blocks**
- **Logical to physical block address mapping**
- **Write new data of free block**
- **Erase old location and add to free list.**



© 2008 IBM Corporation

Static wear leveling

- Infrequently written data – OS data, etc
- Maintain count of erasures per block
- Goal is to keep counts "near" each other
- Simple example: move data from hot block to cold block
  - Write LBA 4
  - D1 → 4
  - 1 now FREE
  - D4 → 1

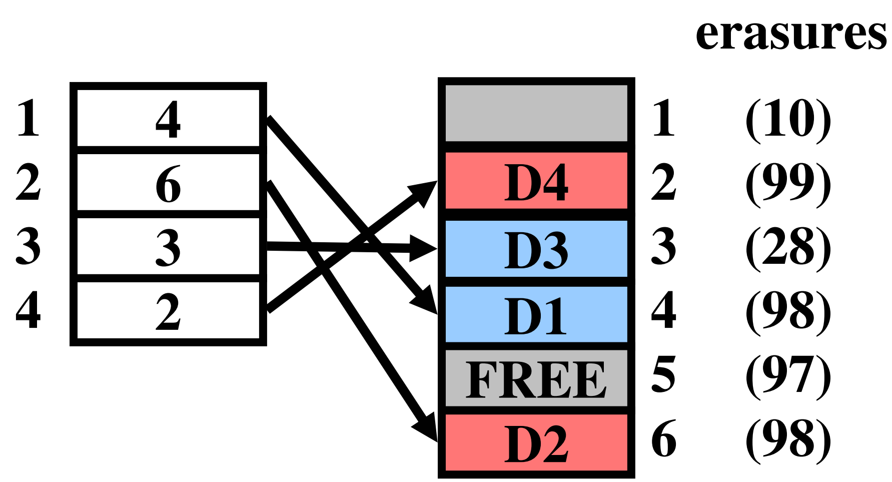


(animated)

© 2008 IBM Corporation

Static wear leveling

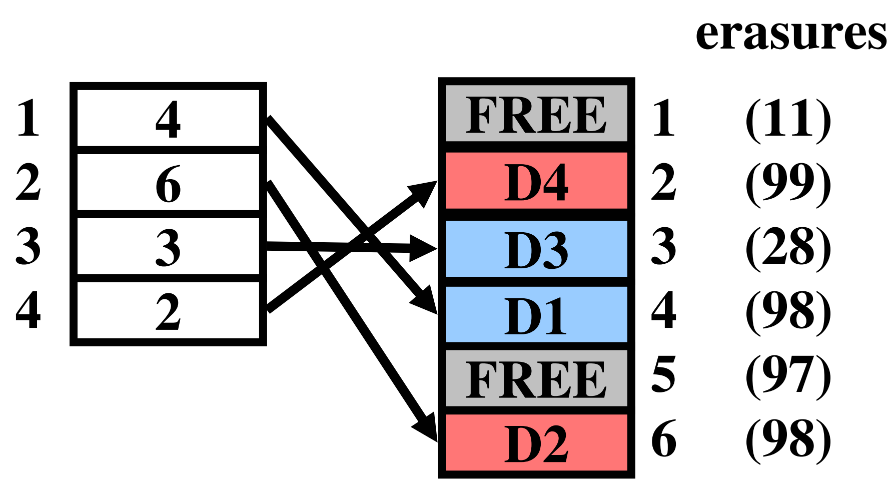
- Infrequently written data – OS data, etc
- Maintain count of erasures per block
- Goal is to keep counts "near" each other
- Simple example: move data from hot block to cold block
  - Write LBA 4
  - D1 → 4
  - 1 now FREE
  - D4 → 1



© 2008 IBM Corporation

Static wear leveling

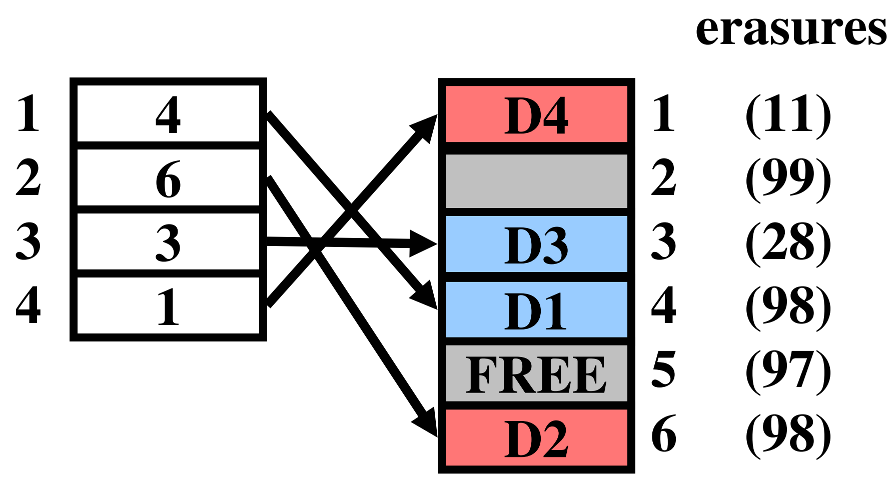
- Infrequently written data – OS data, etc
- Maintain count of erasures per block
- Goal is to keep counts "near" each other
- Simple example: move data from hot block to cold block
  - Write LBA 4
  - D1 → 4
  - 1 now FREE
  - D4 → 1



© 2008 IBM Corporation

Static wear leveling

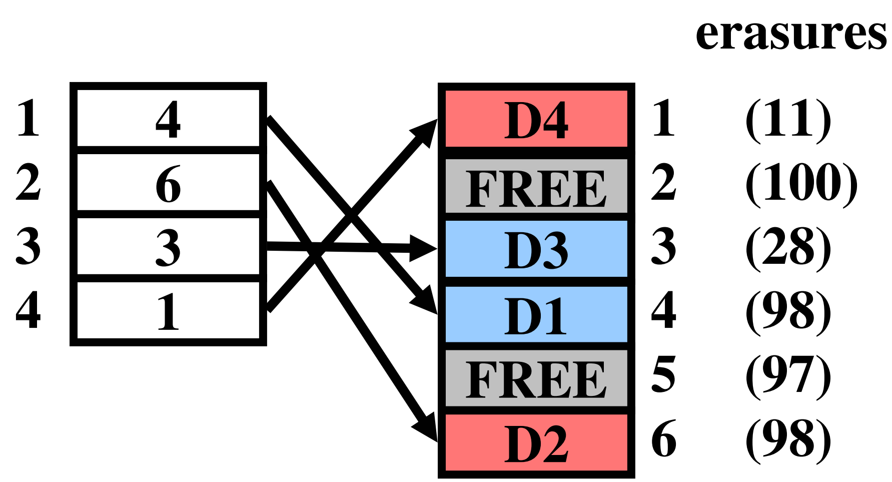
- Infrequently written data – OS data, etc
- Maintain count of erasures per block
- Goal is to keep counts "near" each other
- Simple example: move data from hot block to cold block
  - Write LBA 4
  - D1 → 4
  - 1 now FREE
  - D4 → 1



© 2008 IBM Corporation

Static wear leveling

- **Infrequently written data – OS data, etc**
- **Maintain count of erasures per block**
- **Goal is to keep counts "near" each other**
- **Simple example: move data from hot block to cold block**
  - **Write LBA 4**
  - **D1 → 4**
  - **1 now FREE**
  - **D4 → 1**



© 2008 IBM Corporation

SCM device requirements

▪ Desired attributes

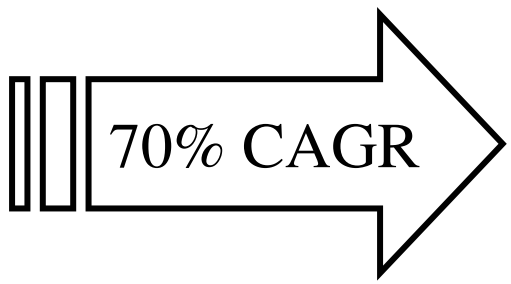
- high **performance** (>1 GB/sec data rate, < 200nsec access time)
- low active & standby **power** (100mW ON power, 1mW standby)
- high read/write **endurance** (10^8 – 10^{12} cycles)
- **non-volatility**
- **compatible** with existing technologies
- continuously **scalable**
- lowest **cost per bit** (target: cost of Enterprise HDD)

© 2008 IBM Corporation

2020 Comparison

- **Extrapolate Disk and SCM solutions to 2020**
- **HPC compute centric and data centric applications**

0.4 TB/s
2 MIOP/s



0.4 PB/s
2 GIOP/s

(10,000 disks)

TODAY

2020

© 2008 IBM Corporation

Disk Assumptions for 2020

- **Enterprise disk: 1.8" diameter**
- **Sustained bandwidth of 300MB/s**
- **400 IOP/s**
- **4 Watts**
- **256 drives packaged in a standard 4U (7 inch high) rack drawer.**
- **Ten such 4U drawers will be packaged in a standard 19-inch rack.**

© 2008 IBM Corporation

Managing bad blocks

- **Flash chips have up to 2% bad blocks when shipped from factory**
- **Bad blocks are indicated within the chip**
- **System must maintain list**
- **Block failures detected on writes**
- **Add new bad blocks to list**

© 2008 IBM Corporation

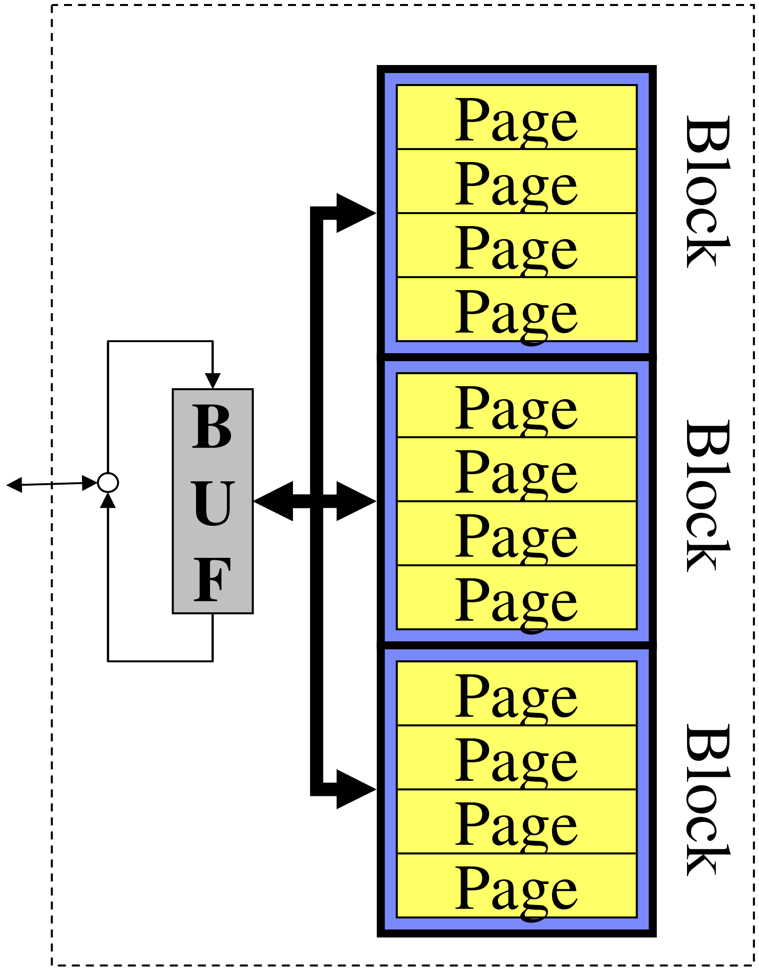
Storage Technology Summary

- **Disk drives are the current technology**
  - –Areal density growth has flatten off to ~40% CAGR
  - –Bandwidth improvement is ~10% CAGR
  - –Access time improvement is ~5% CAGR
- **NVRAMs appearing as contenders**
  - –Flash making its move now
  - –Other SCM technologies in the wings

© 2008 IBM Corporation

NAND Flash Device

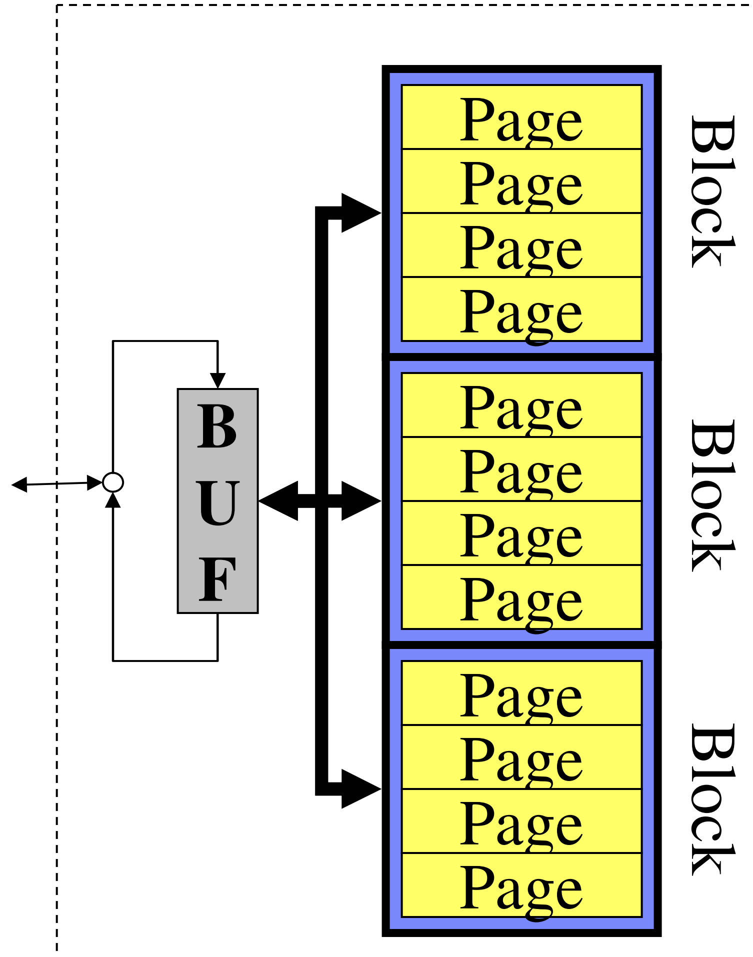
- Chip size: 12mmx20mm
- Power ≈ 100mW
- Interface: byte wide
- Page
  - **2112 Bytes**
  - **Moving to 4224 Bytes**
- Block = 128 Pages



© 2008 IBM Corporation

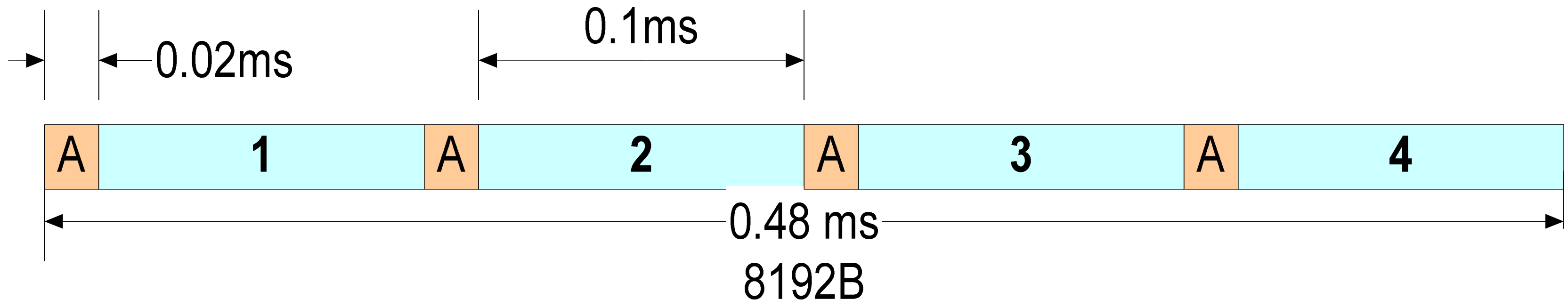
NAND Flash Operations

- Read copies Page into BUF and streams data to host
  - **Read 20us access,**
  - **20 MB/s transfer rate – sustained**
  - **Moving to 40 MB/s**
- Write streams data from host into BUF
  - **6 MB/s transfer rate sustained**
  - **20 MB/s burst → 40 MB/s**
- Program copies BUF into Page
  - **Program 2 KB / 4 KB page: 0.2 ms**
- Erase clears all Pages in a Block to "1"s
  - **Erase 128 KB block: 1.5 ms**

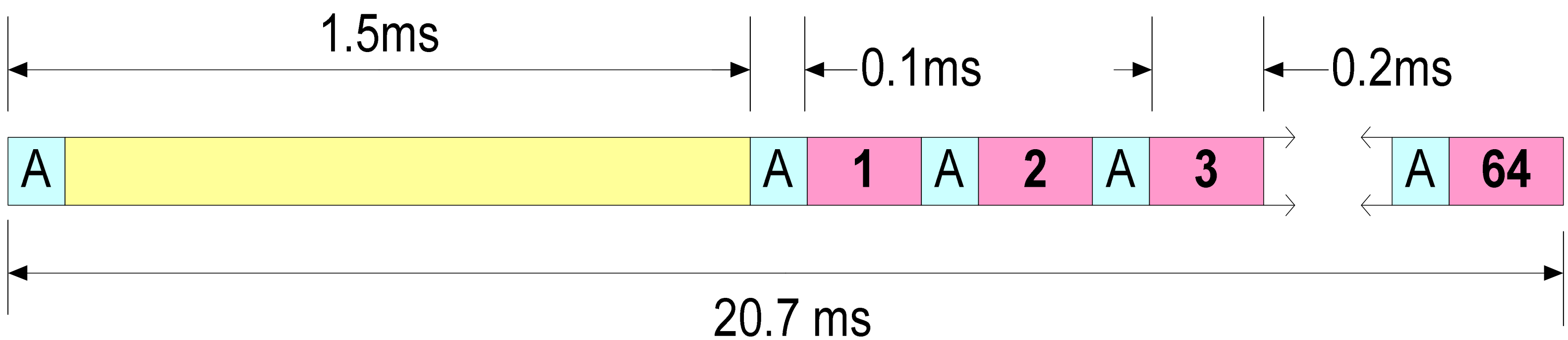


© 2008 IBM Corporation

NAND Flash Chip Read and Write timing



8 KB READ: sequential at 17MB/s sustained --- random at 2083 IOP/s

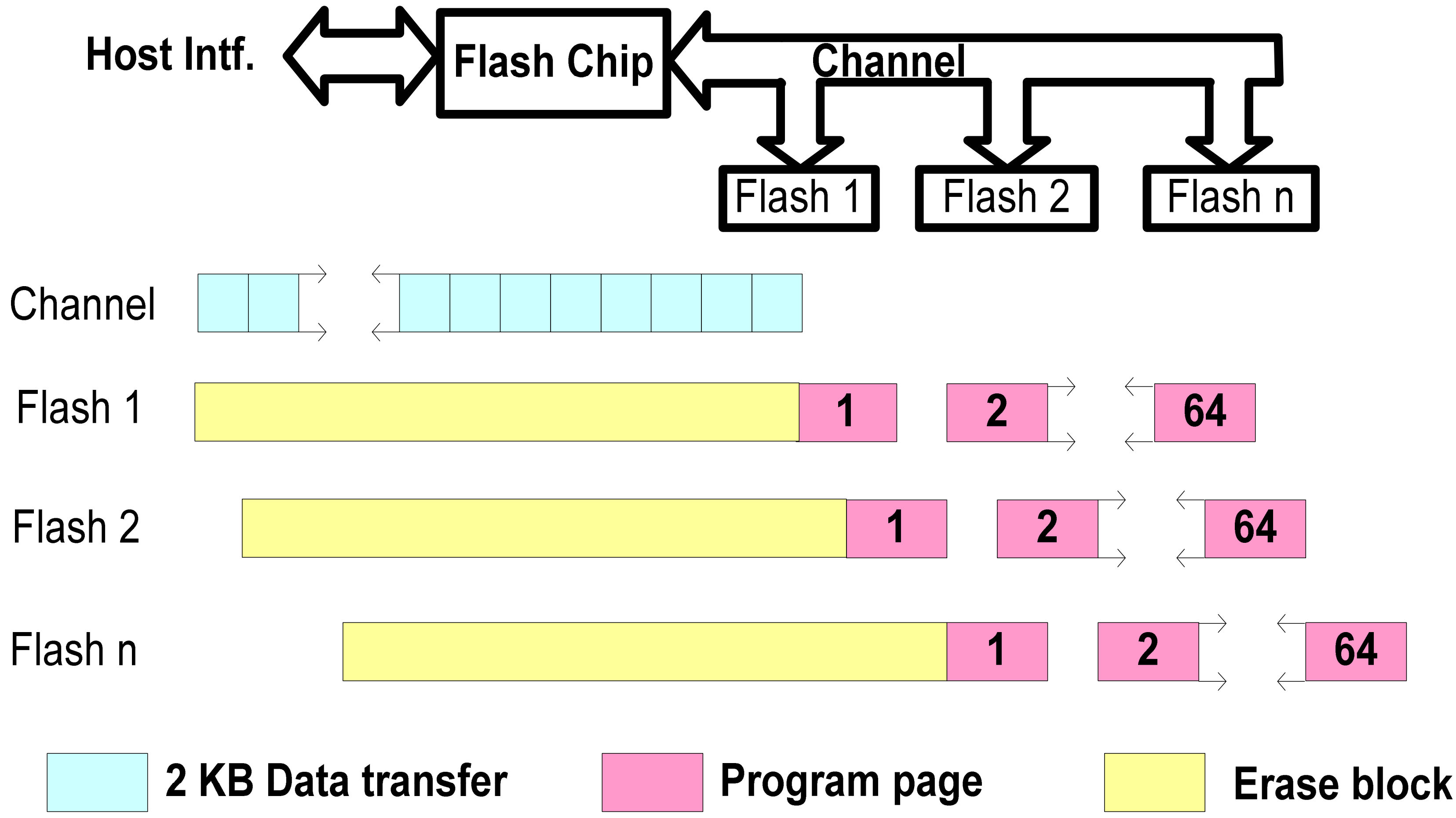


128KB Write: sequential at 6.55 MB/s sustained --- random at 49 IOP/s

8KB Write: read 128KB, change 8KB, write 128KB → 35 IOP/s

© 2008 IBM Corporation

Flash Drive Channel



© 2008 IBM Corporation

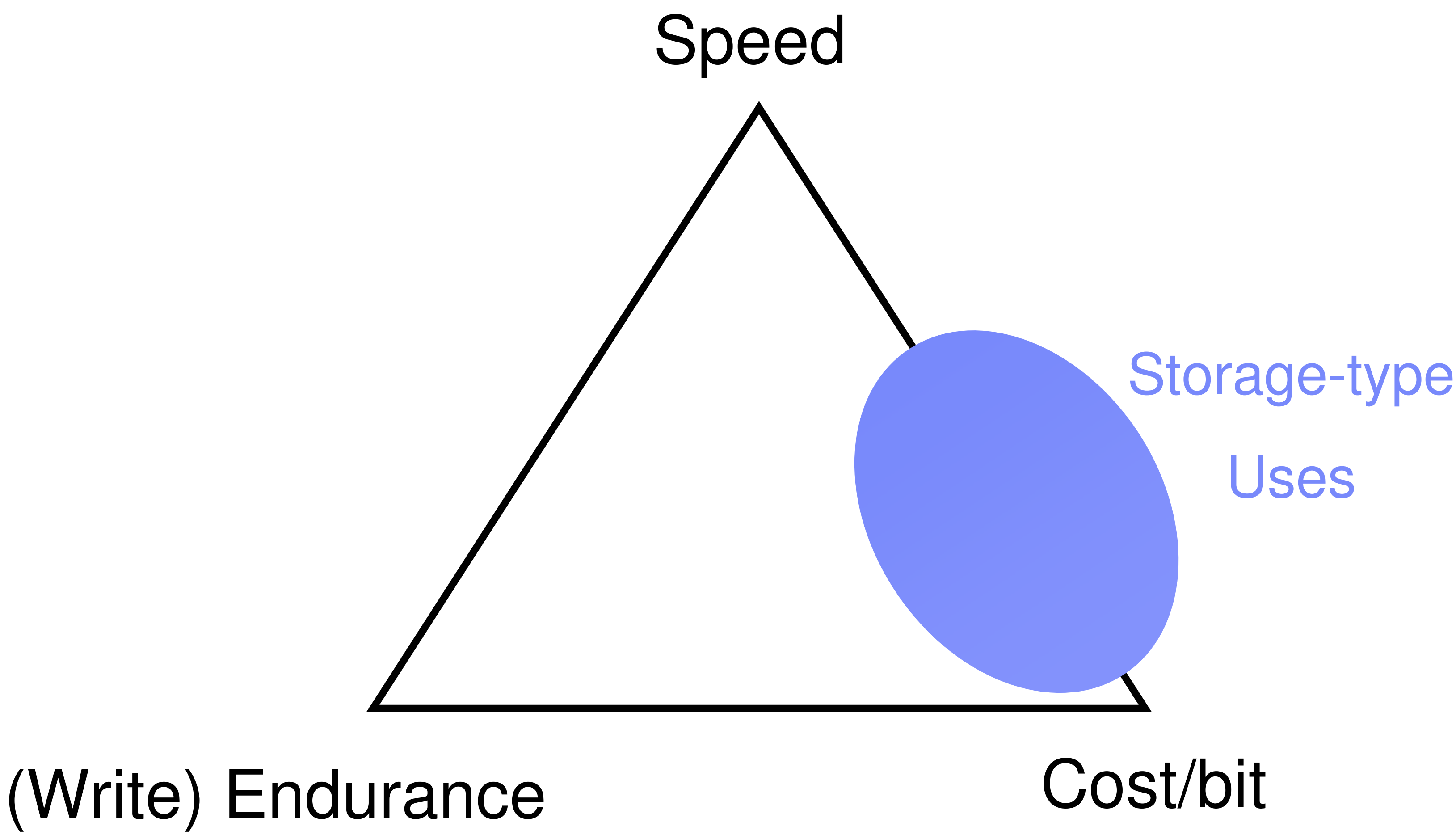
Can HDD & Flash improve enough to help?

- **Magnetic hard-disk drives (HDD)**
  - **bandwidth** issues (hidden with parallelism, but at power/space cost)
  - slow **access** time (not improving, hard to hide with caching tricks)
  - **reliability** (newest drives are *less reliable* → data losses inevitable)
  - **power** consumption (must keep drives spinning to avoid even longer access times)

- **Flash**
  - slow read/write **access time** (yet processors keep getting faster)
  - low write **endurance** ($<10^6$) (need $>10^8$ for continuously streaming data)
  - block architecture
  - **scalability** beyond the end of this decade?

© 2008 IBM Corporation

SCM Design Triangle



Power!

© 2008 IBM Corporation

Outline

▪ Motivation

- by 2020, server-room power & space demands will be too high
- evolution of hard-disk drive (HDD) storage and Flash cannot help
- need a new technology – **Storage Class Memory (SCM)** – that combines
  - ❖ the benefits of a solid-state memory (**high performance** and **robustness**)
  - ❖ the **archival capabilities** and **low cost** of conventional HDD

▪ How could we build an SCM?

- combine a scalable non-volatile memory (**Phase-change memory**)
- with **ultra-high density** integration, using
  - ❖ micro-to-nano addressing
  - ❖ multi-level cells
  - ❖ 3-D stacking

▪ Conclusion

- With its combination of **low-cost** and **high-performance**,
  SCM could impact much more than just the server-room...

© 2008 IBM Corporation

Can HDD & Flash improve enough to help?

Magnetic hard-disk drives (HDD)

- **bandwidth** issues (hidden with parallelism, but at power/space cost)
- slow **access** time (not improving, hard to hide with caching tricks)
- **reliability** (newest drives are *less reliable* → data losses inevitable)
- **power** consumption (must keep drives spinning to avoid even longer access times)

Flash

- slow read/write **access time** (yet processors keep getting faster)
- low write **endurance** ($<10^6$) (need $>10^9$ for continuously streaming data)
- block architecture
- **scalability** beyond the end of this decade?

© 2008 IBM Corporation

What is an HDD?

-
-

HUGE COST ADVANTAGES

$

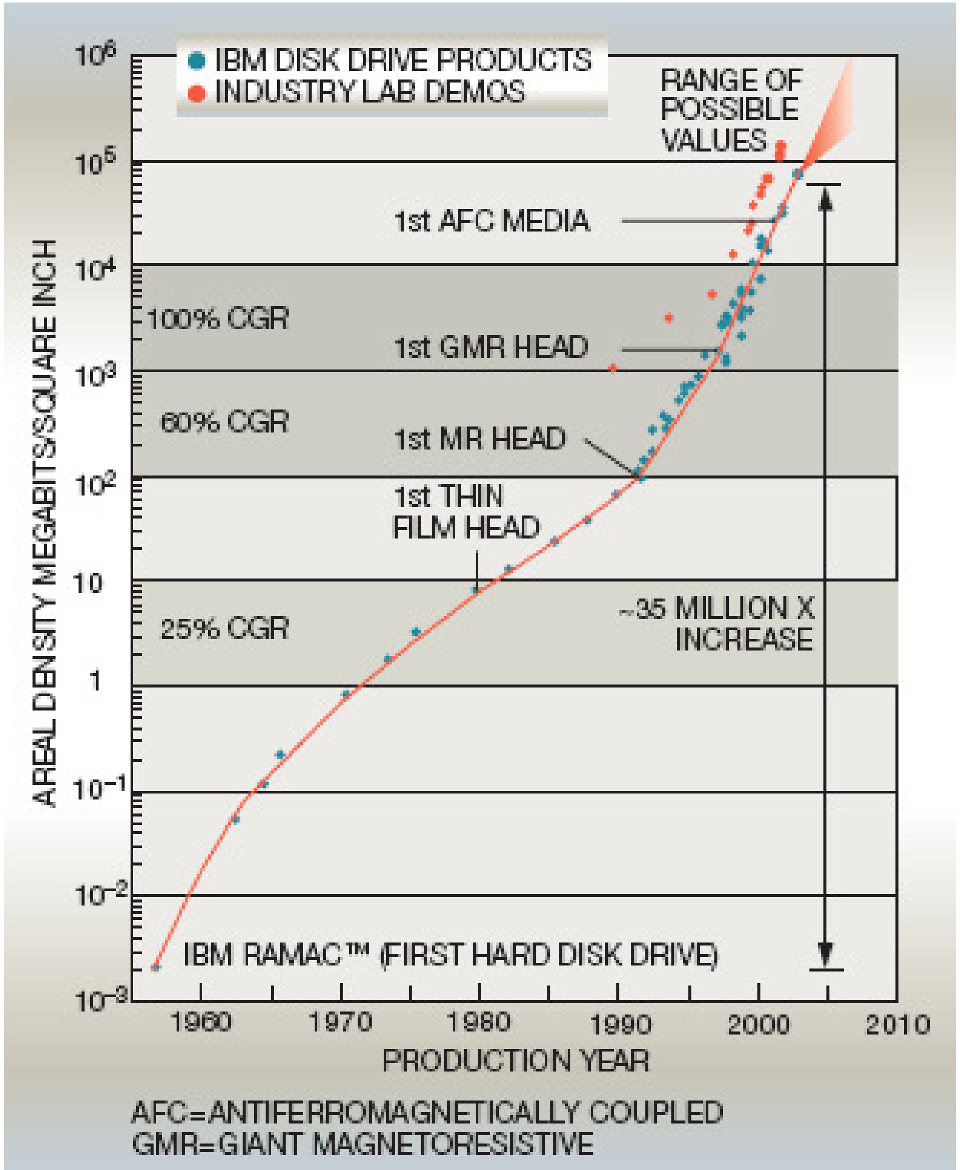
$

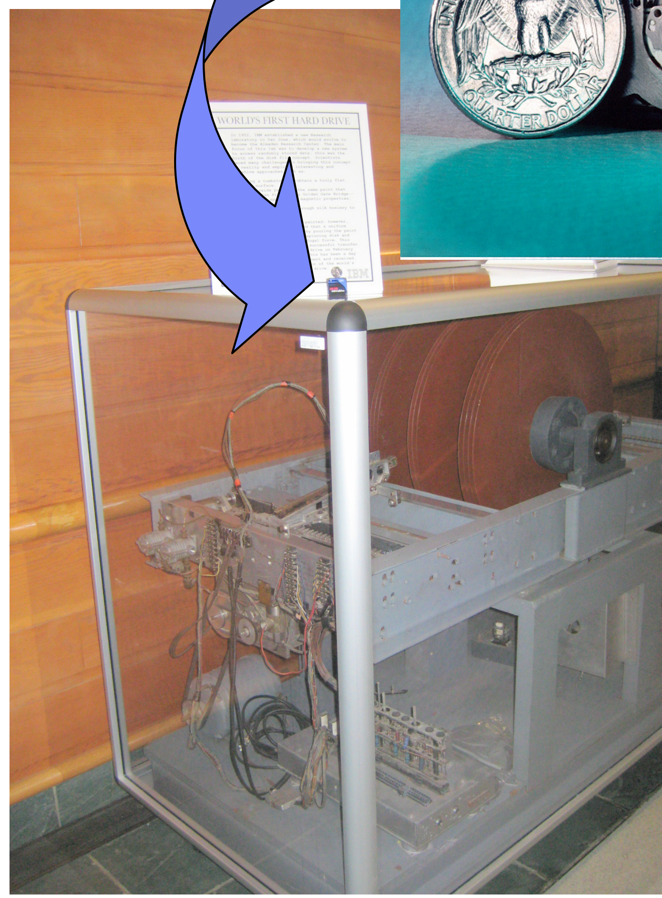
$

$

© 2008 IBM Corporation

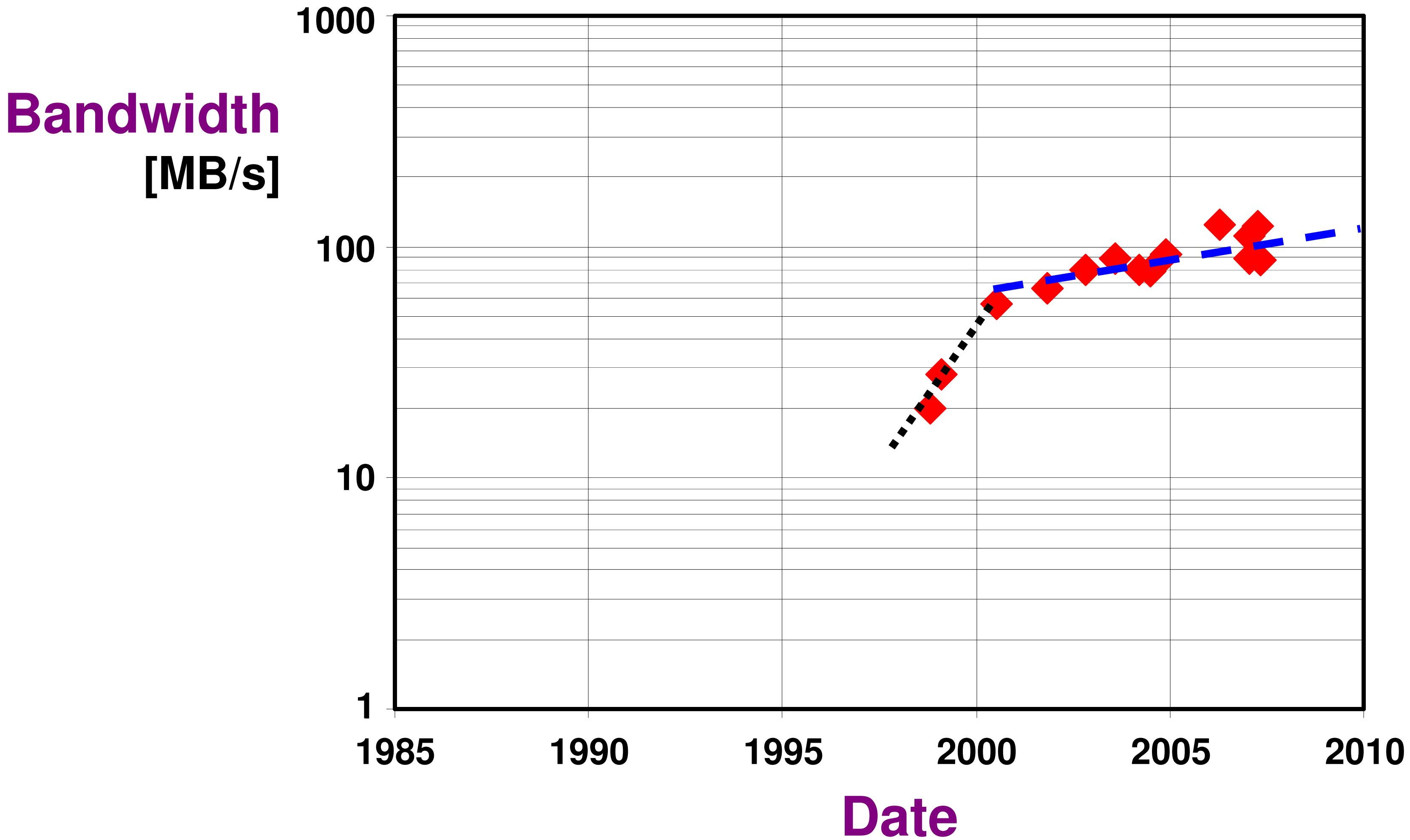
History of HDD is based on Areal Density Growth





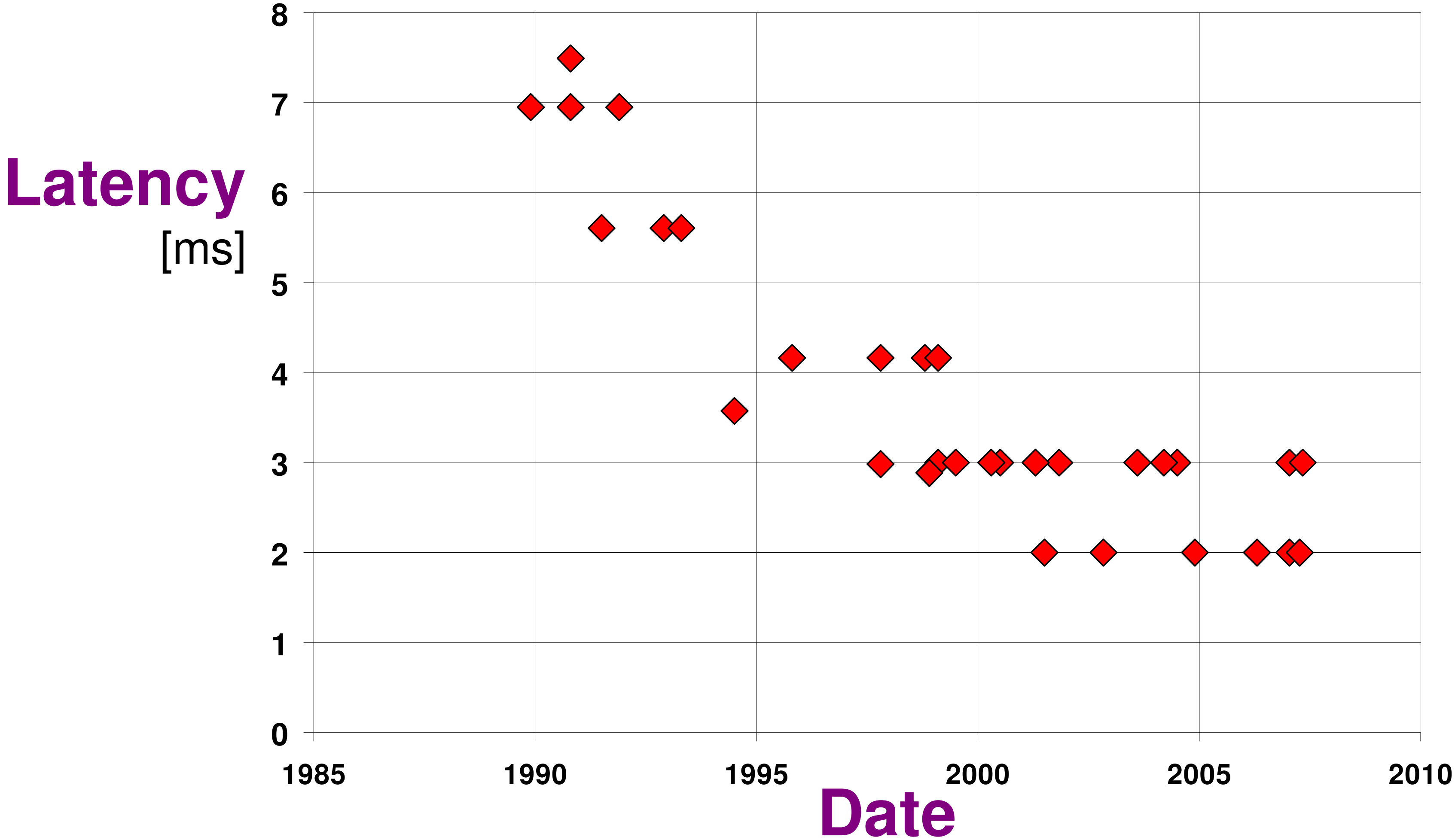
© 2008 IBM Corporation

Disk Drive Maximum Sustained Data Rate



 © 2008 IBM Corporation

Disk Drive Latency



- **Bandwidth Problem** is getting much harder to hide with parallelism
- **Access Time Problem** is also not improving with caching tricks
- Power/Space/Performance Cost

 © 2008 IBM Corporation

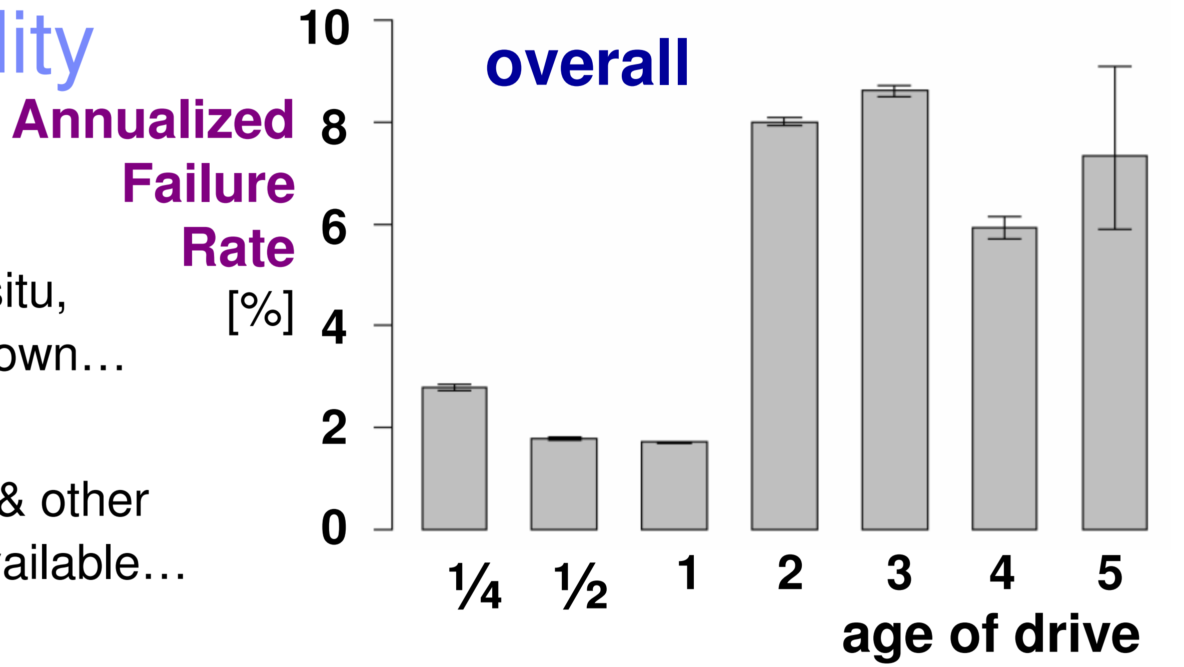
Disk Drive Reliability

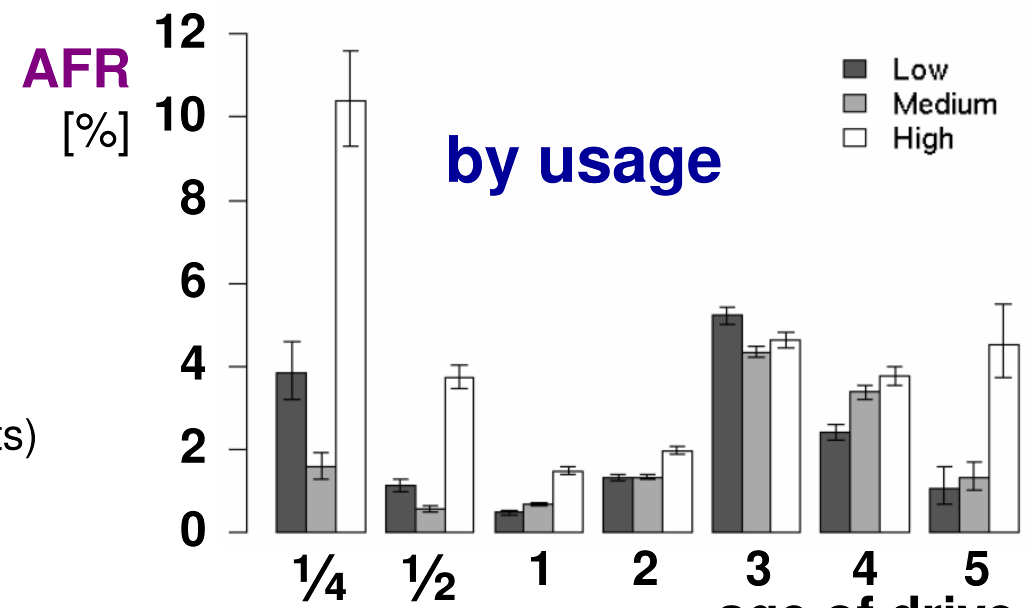
- with hundreds of thousands of server drives being used in-situ, **reliability problems** well known…

- similar understanding for Flash & other SCM technologies not yet available…

- Consider: drive failures <u>during</u> recovery from a drive failure…?

→ potential for improvement given
  - switch to solid-state (no moving parts)
  - faster time-to-fill (during recovery)





[Pinheiro:2007]

© 2008 IBM Corporation

Can HDD & Flash improve enough to help?

- **Magnetic hard-disk drives (HDD)**
  - **bandwidth** issues (hidden with parallelism, but at power/space cost)
  - slow **access** time (not improving, hard to hide with caching tricks)
  - **reliability** (newest drives are *less reliable* → data losses inevitable)
  - **power** consumption (must keep drives spinning to avoid even longer access times)

- **Flash**
  - slow read/write **access time** (yet processors keep getting faster)
  - low write **endurance** (<10^6) (need >10^9 for continuously streaming data)
  - block architecture
  - **scalability** beyond the end of this decade?

© 2008 IBM Corporation

Improved Flash

- An unpleasant tradeoff between **scaling**, **speed**, and **endurance,** designers are choosing to hold speed & endurance constant to keep the scaling going…

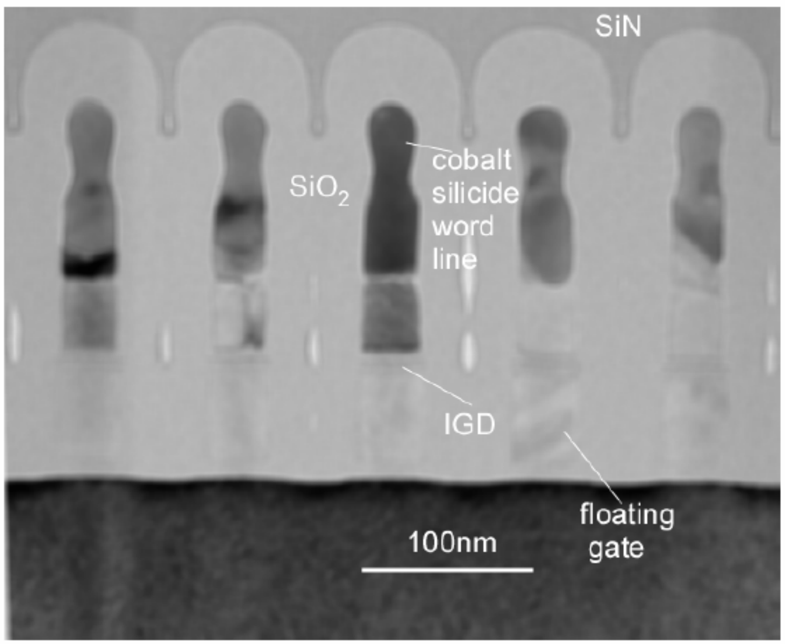
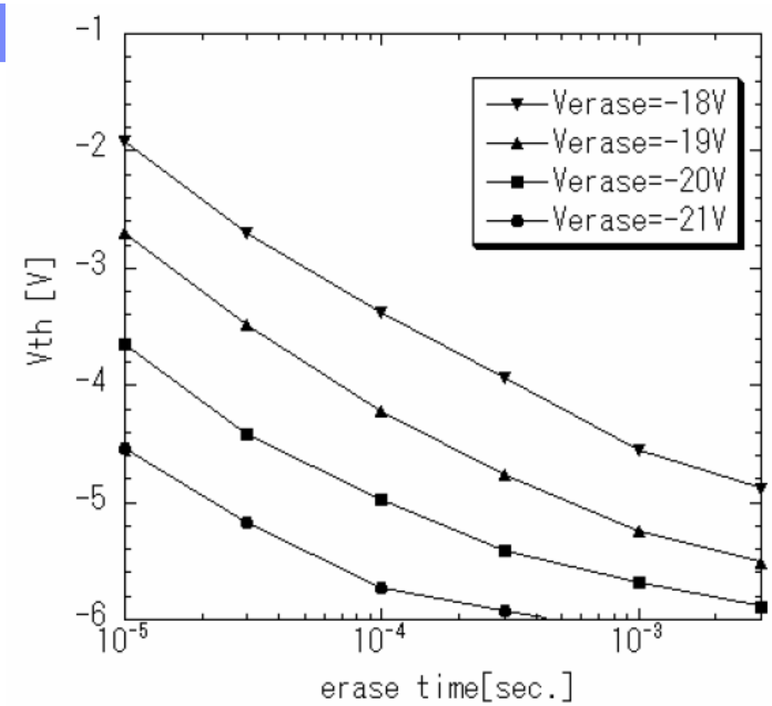
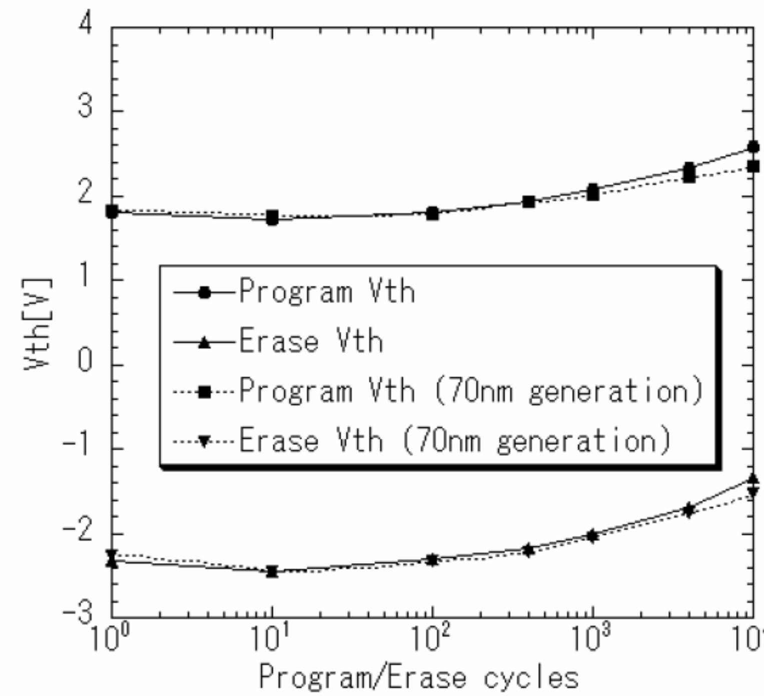


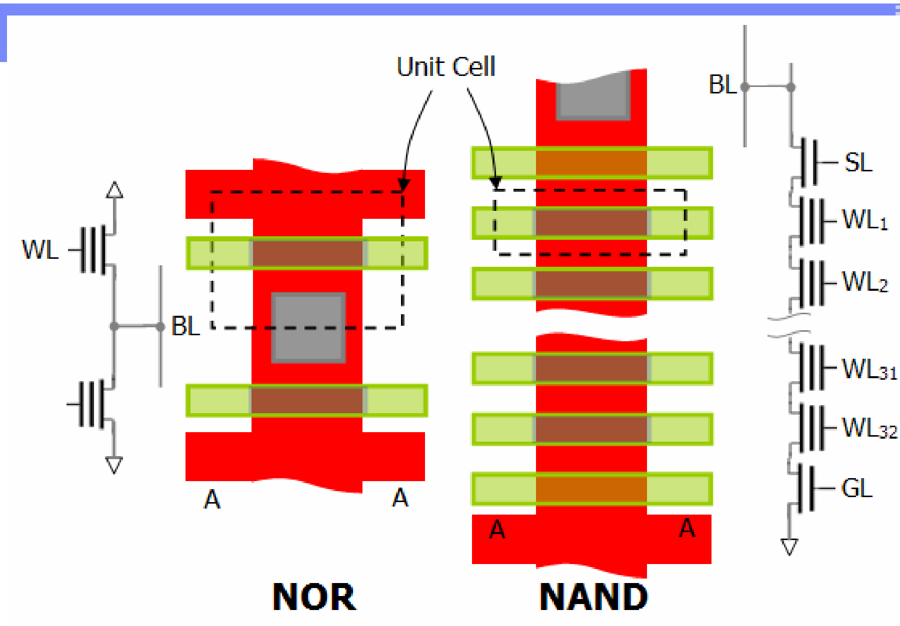
Fig. 1. Cross-sectional image of 43nm-node floating-gate memory cells in a shorter gate condition.

[Noguchi:2007]





FLASH memory types and application

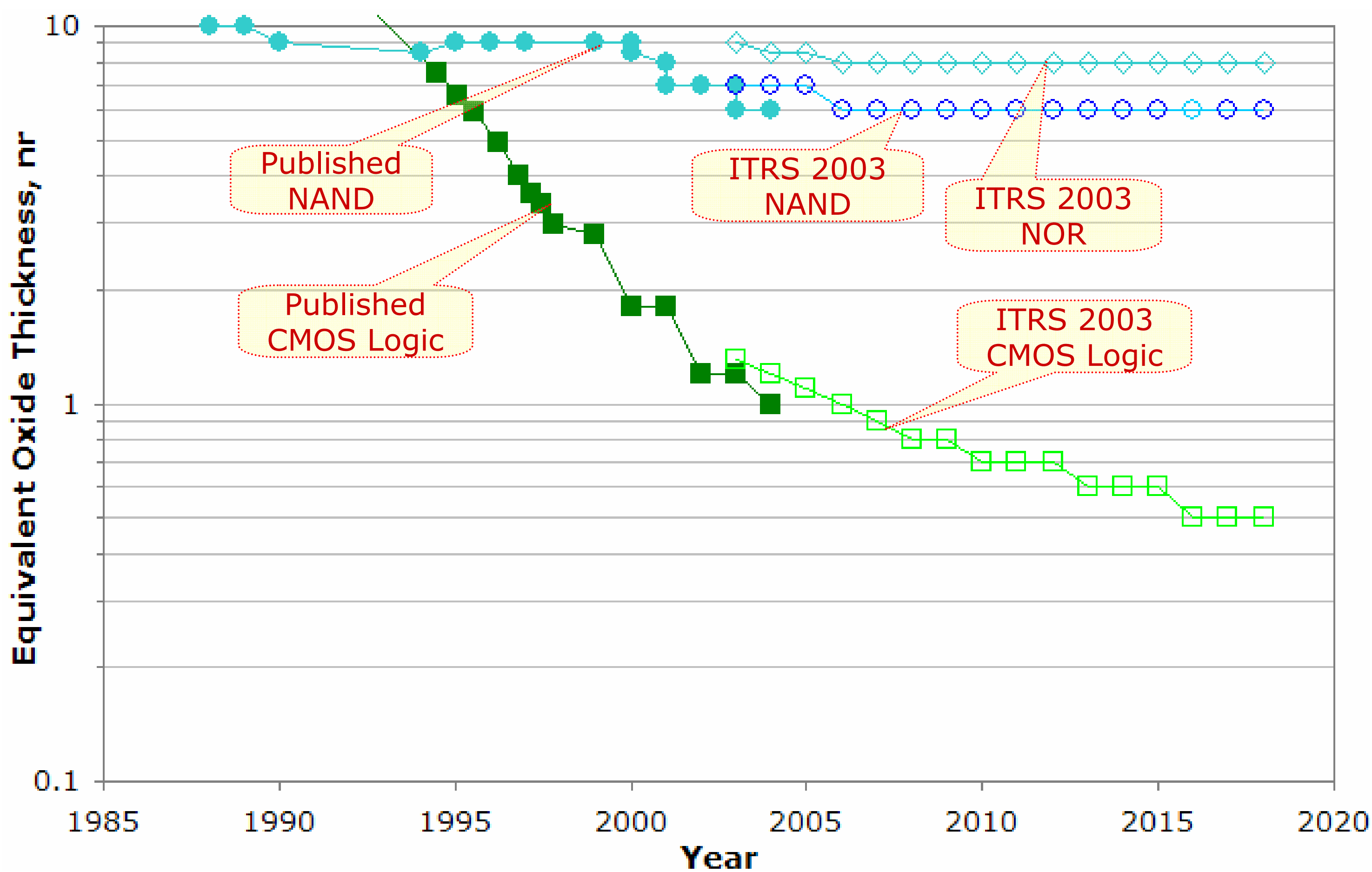


	NOR	NAND
Cell Size	9-11 F^2	2 F^2
Read	100 MB/s	18-25 MB/s
Write	<0.5MB/sec	8MB/sec
Erase	750msec	2ms
Market Size (2007)	$8B	$14.2B
Applications	Program code	Multimedia

© 2008 IBM Corporation

Flash – below the 100nm technology node

Tunnel oxide thickness in Floating-gate Flash is no longer practically scalable

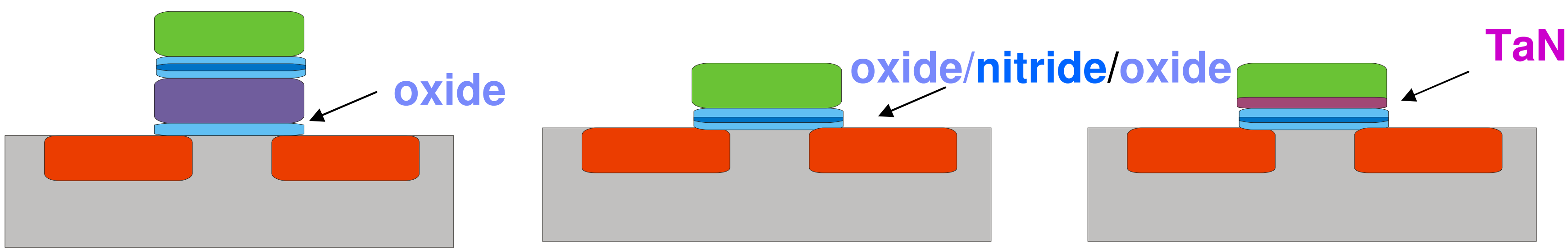


Source: Chung Lam, IBM

© 2008 IBM Corporation

Can Flash improve enough to help?

Technology Node: **40nm** → **30nm** → **20nm**



Floating Gate
<40nm ???

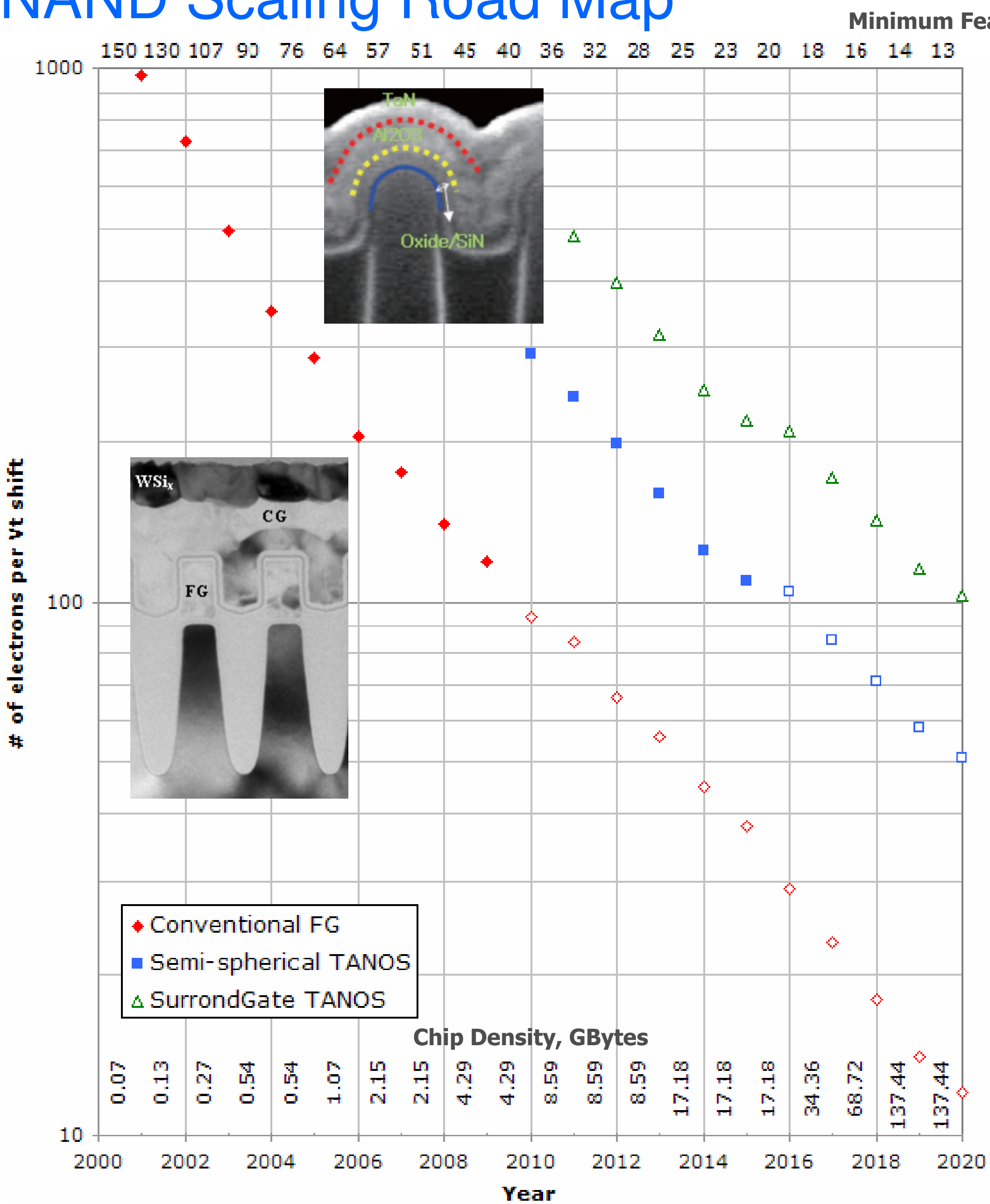
SONOS
Charge trapping in SiN trap layer

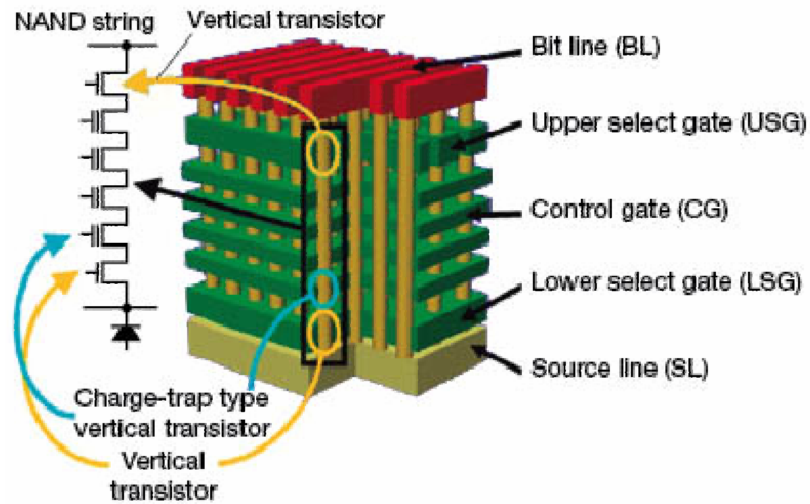
TaNOS
Charge trapping in novel trap layer coupled with a metal-gate (TaN)

Main thrust is to continue scaling yet maintain the **same** performance and write endurance specifications…

© 2008 IBM Corporation

NAND Scaling Road Map





Evolution??

- Migrating to Semi-spherical TANOS memory cell 2009
- Migrating to 3-bit cell in 2010
- Migrating to 4-bit cell in 2013
- Migrating to 450mm wafer size in 2015
- Migrating to 3D Surround-Gate Cell in 2017

Source: Chung Lam, IBM

© 2008 IBM Corporation

For more information (on HDD & Flash)

HDD

- E. Grochowski and R. D. Halem, *IBM Systems Journal*, **42**(2), 338-346 (2003)..
- R. J. T. Morris and B. J. Truskowski, *IBM Systems Journal*, **42**(2), 205-217 (2003).
- R. E. Fontana and S. R. Hetzler, *J. Appl. Phys.*, **99**(8), 08N902 (2006).
- E. Pinheiro, W.-D. Weber, and L. A. Barroso, *FAST'07* (2007).

Flash

- S. Lai, to appear in *IBM J. Res. Dev.*, (2008).
- R. Bez, E. Camerlenghi, et. al., *Proceedings of the IEEE*, **91**(4), 489-502 (2003).
- G. Campardo, M. Scotti, et. al., *Proceedings of the IEEE*, **91**(4), 523-536 (2003).
- P. Cappelletti, R. Bez, et. al., *IEDM Technical Digest*, 489-492 (2004).
- A. Fazio, *MRS Bulletin*, **29**(11), 814-817 (2004).
- K. Kim and J. Choi, *Proc. Non-Volatile Semiconductor Memory Workshop*, 9-11 (2006).
- M. Noguchi, T. Yaegashi, et. al., *IEDM Technical Digest*, 17.1 (2007).

 © 2008 IBM Corporation

Can HDD & Flash improve enough to help?

▪ Magnetic hard-disk drives (HDD)

- **bandwidth** issues (hidden with parallelism, but at power/space cost)
- slow **access** time (not improving, hard to hide with caching tricks)
- **reliability** (newest drives are *less reliable* → data losses inevitable)
- **power** consumption (must keep drives spinning to avoid even longer access times)

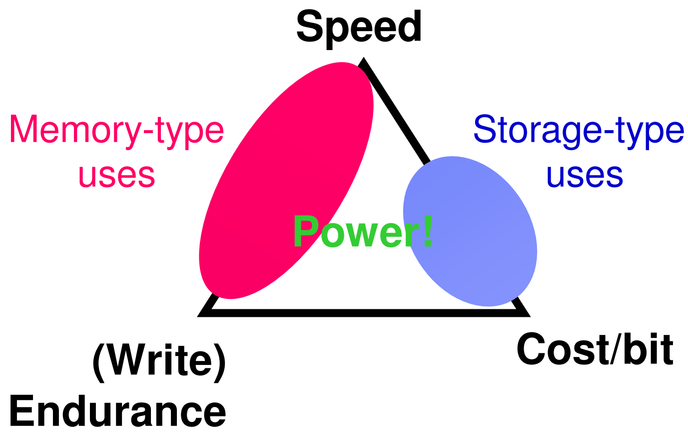
▪ Flash

- slow read/write **access time** (yet processors keep getting faster)
- low write **endurance** ($<10^6$) (need $>10^9$ for continuously streaming data)
- block architecture
- **scalability** beyond the end of this decade?

© 2008 IBM Corporation

Storage Class Memory

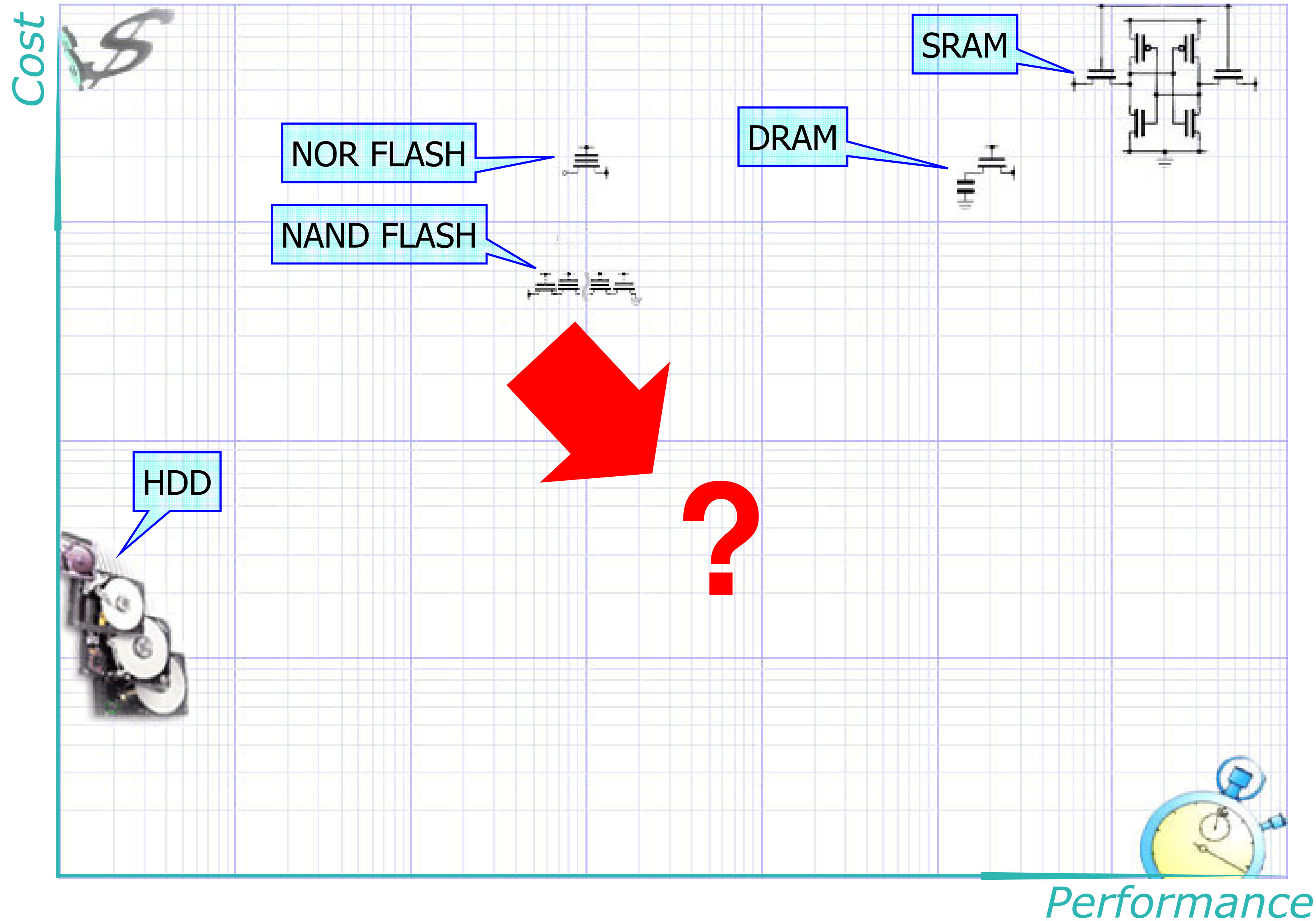
A solid-state memory that **blurs the boundaries** between storage and memory by being **low-cost**, **fast**, and **non-volatile**.



- **SCM system requirements for Memory (Storage) apps**
  - No more than 3-5x the **Cost** of enterprise HDD (< $1 per GB in 2012)
  - **<200nsec (<1 µsec) Read/Write/Erase time**
  - >100,000 **Read I/O operations** per second
  - **>1GB/sec (>100MB/sec)**
  - **Lifetime** of 10^9 – 10^{12} write/erase cycles
  - 10x lower **power** than enterprise HDD

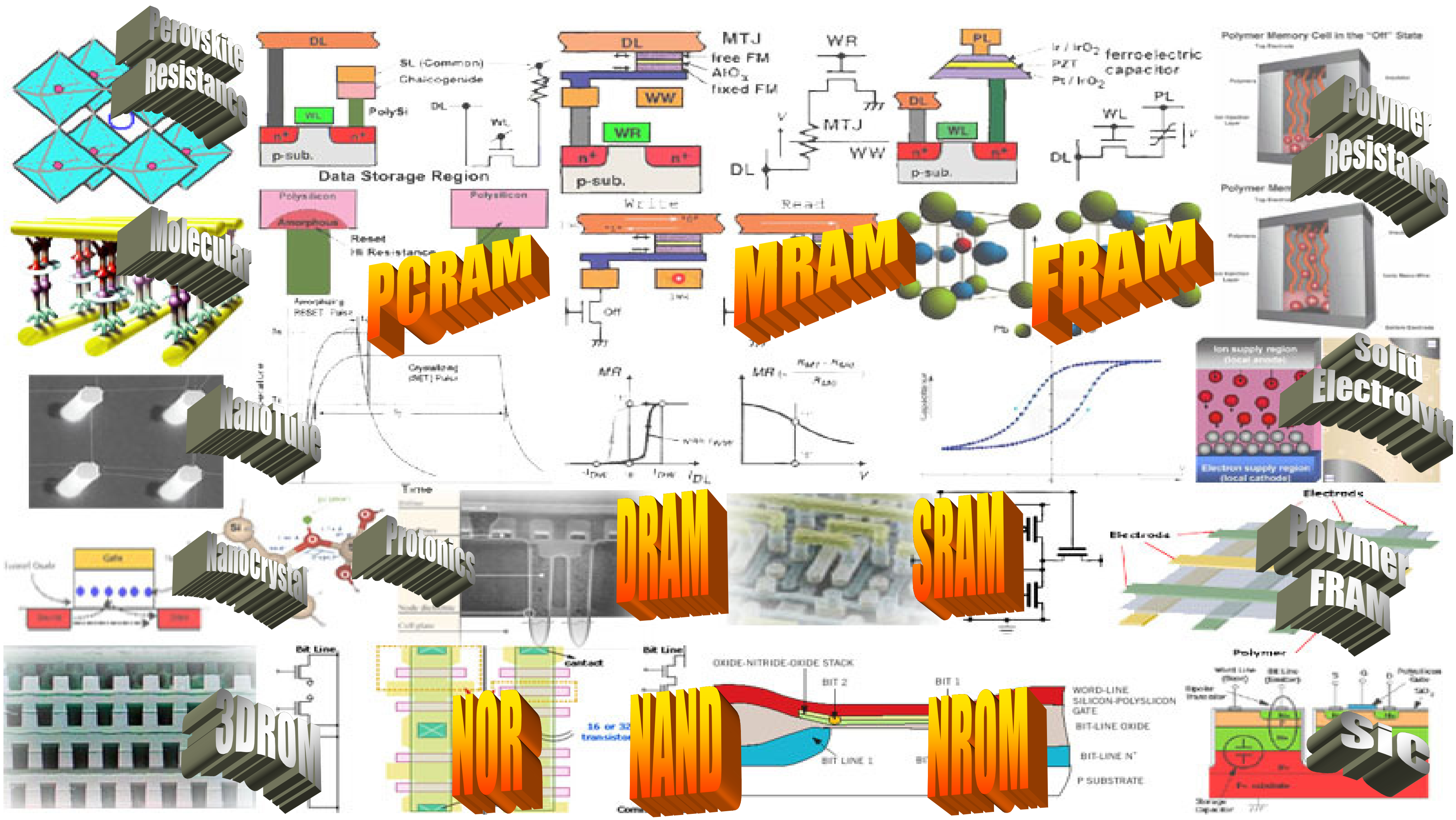
 © 2008 IBM Corporation

Landscape of existing technologies



 © 2008 IBM Corporation

Memory/storage landscape



 © 2008 IBM Corporation

Improved Flash

- An unpleasant tradeoff between **scaling**, **speed**, and **endurance,** designers are choosing to hold speed & endurance constant to keep the scaling going…

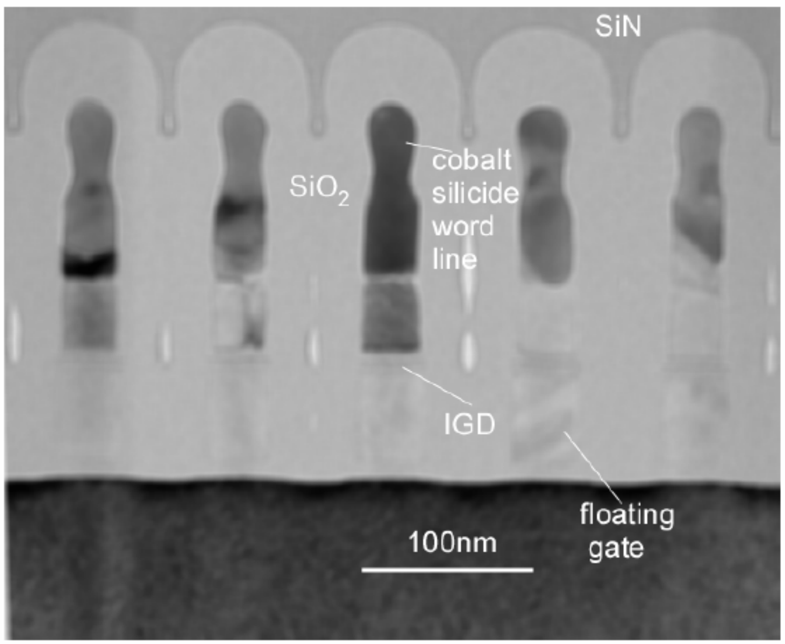
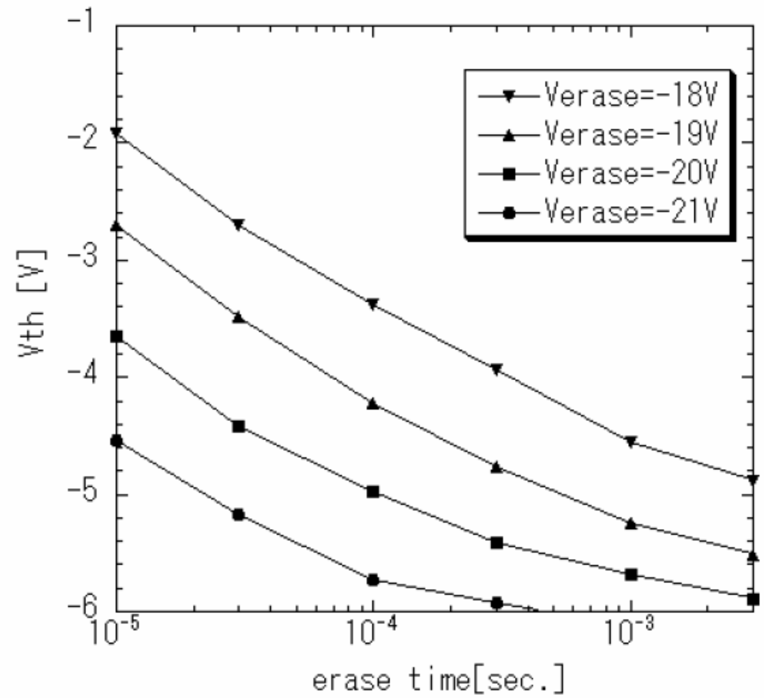
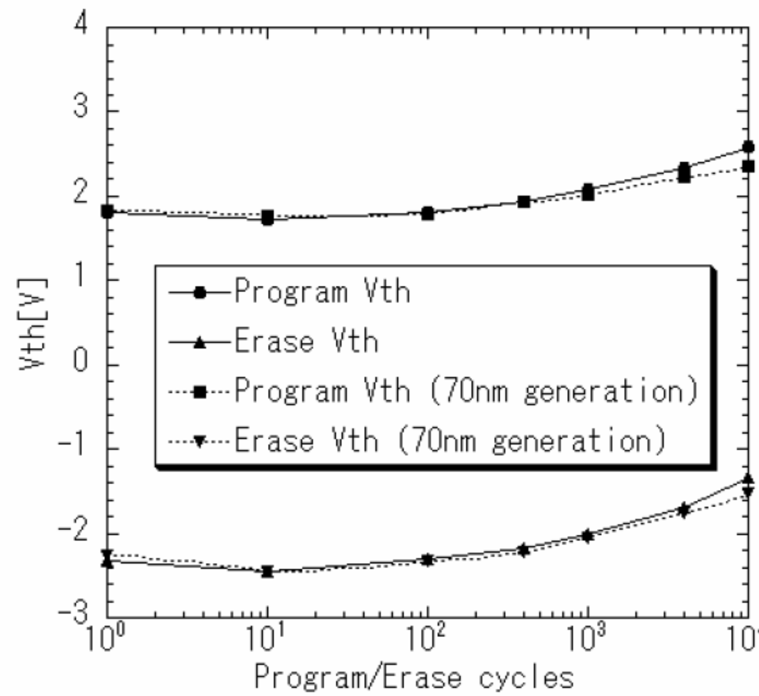


Fig. 1. Cross-sectional image of 43nm-node floating-gate memory cells in a shorter gate condition.

[Noguchi:2007]





Candidate device technologies

- **Improved Flash**
  - little change expected in write endurance or speed
- **FeRAM** (Ferroelectric RAM)
  - FeFET
- **MRAM** (Magnetic RAM)
  - Racetrack memory
- **RRAM** (Resistive RAM)
  - Organic & polymer memory
- **Solid Electrolyte**
- **PC-RAM** (Phase-change RAM)

© 2008 IBM Corporation

FeRAM progress

- Lots of attention in 1998-2003 timeframe
- Commercially available (Playstation 2), mostly as embedded memory

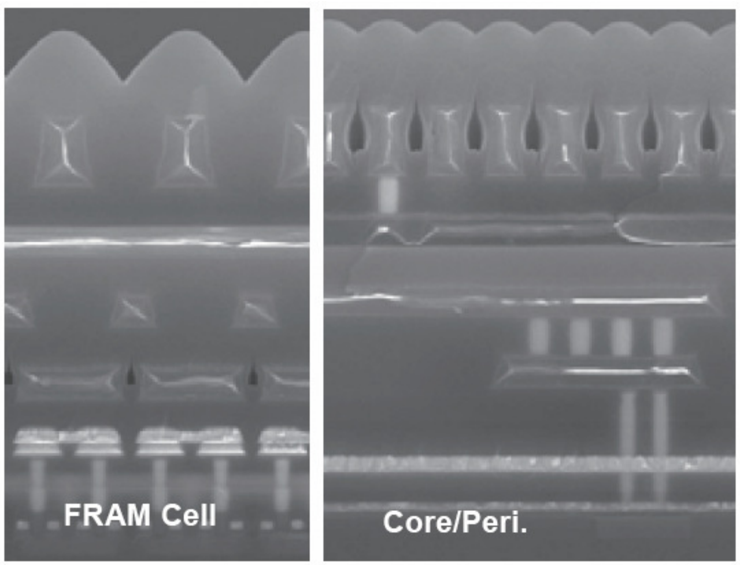
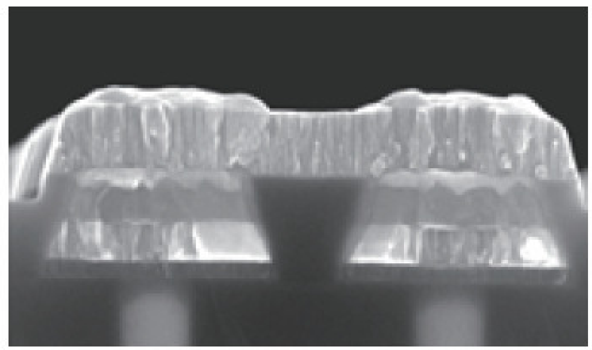
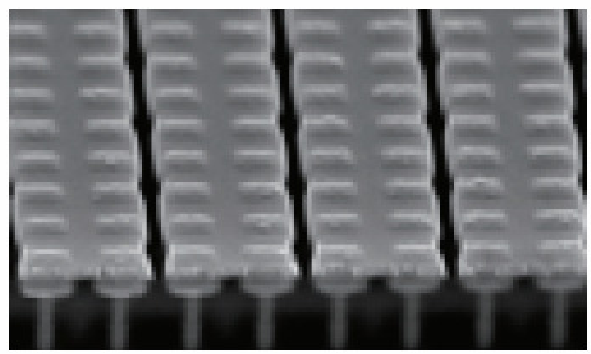


Fig. 1 A cross-sectional SEM image of 0.25 μm^2 , 64 Mb FRAM cells.





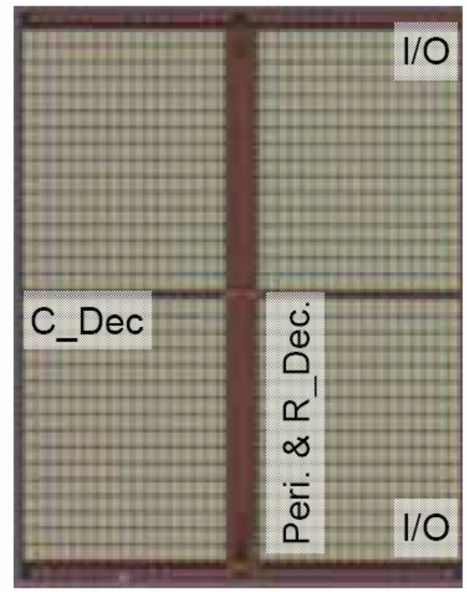


Fig. 9 An optical micrograph of 0.25 μm^2, 1T1C 64 Mb FRAM cell.

[Hong:2007]

© 2008 IBM Corporation

FeRAM difficulties

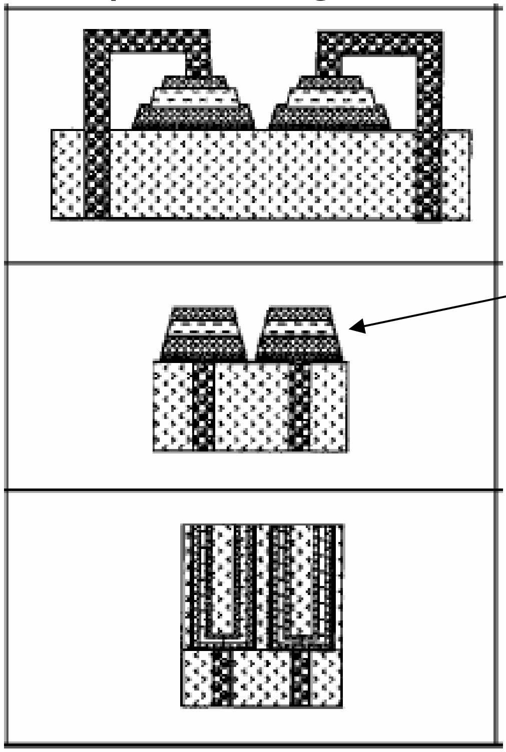
- Signal ΔV = transfer of charge $Q_r \sim 2\ P_r$ Area onto bitline capacitance C_b
  - scaling to smaller devices means lower signal !!
  - need material with large remanent polarization P_r
  - tradeoff speed for signal with C_b

- Forces more complex integration schemes to keep effective area large

"Strapped" — > 30 F^2

"Stacked" — 10 - 30 F^2

"3-D" — < 10 F^2



Materials difficult to etch vertically – forces guard bands and thus less area

[Kim:2006]

© 2008 IBM Corporation

FeRAM difficulties

SBT = $SrBi_2Ta_2O_9$
strontium bismuth tantalate

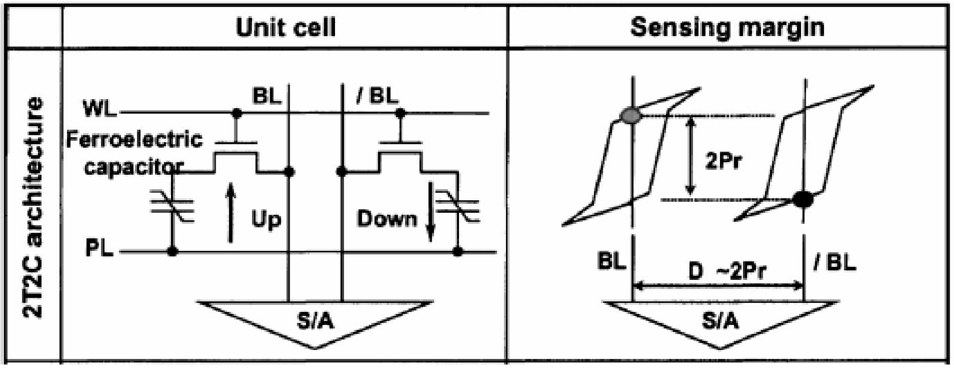
- Many reliability & processing difficulties to overcome…

fatigue	remanent polarization P_r decreases with cycling	• Change electrodes from metals to metal-oxides • Change FE material (PZT → SBT)
imprint	a device left in one state tends to favor that polarization, causing hysteresis loop to shift	• Eliminate defects introduced during fabrication by hydrogen • Change FE material (PZT → SBT)
retention	Stored polarization is lost over time	• Change FE material (PZT → SBT)
High temperature processing	For crystalline FE material	• Change FE material (→ PZT)
insufficient P_r	$\propto$ voltage signal	• Change FE material (→ PZT)

© 2008 IBM Corporation

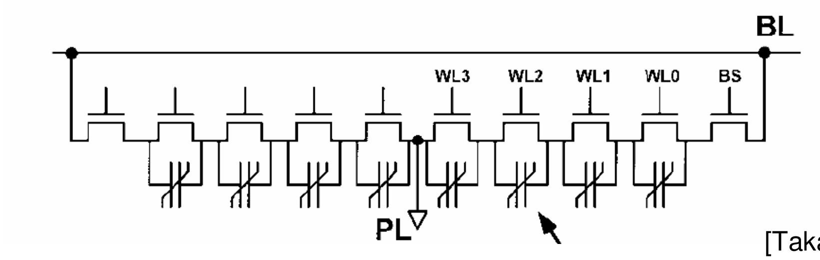
Alternative FeRAM concepts

- **2T-2C** concept – twice the signal but also twice the area



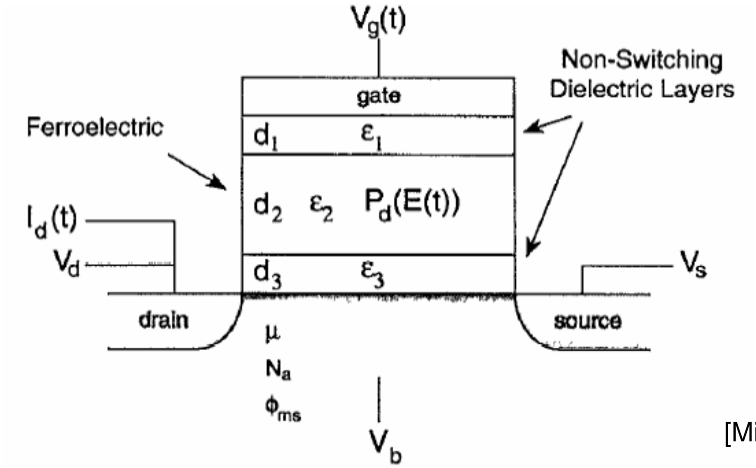
[Kim:2006]

- **Chain-FeRAM** – improves signal but decreases speed, only minor density improvement



[Takashima:1998]

- **FeFET** – perhaps more scalable but requires integration onto silicon and tends to sacrifice the non-volatility



[Miller:1992]

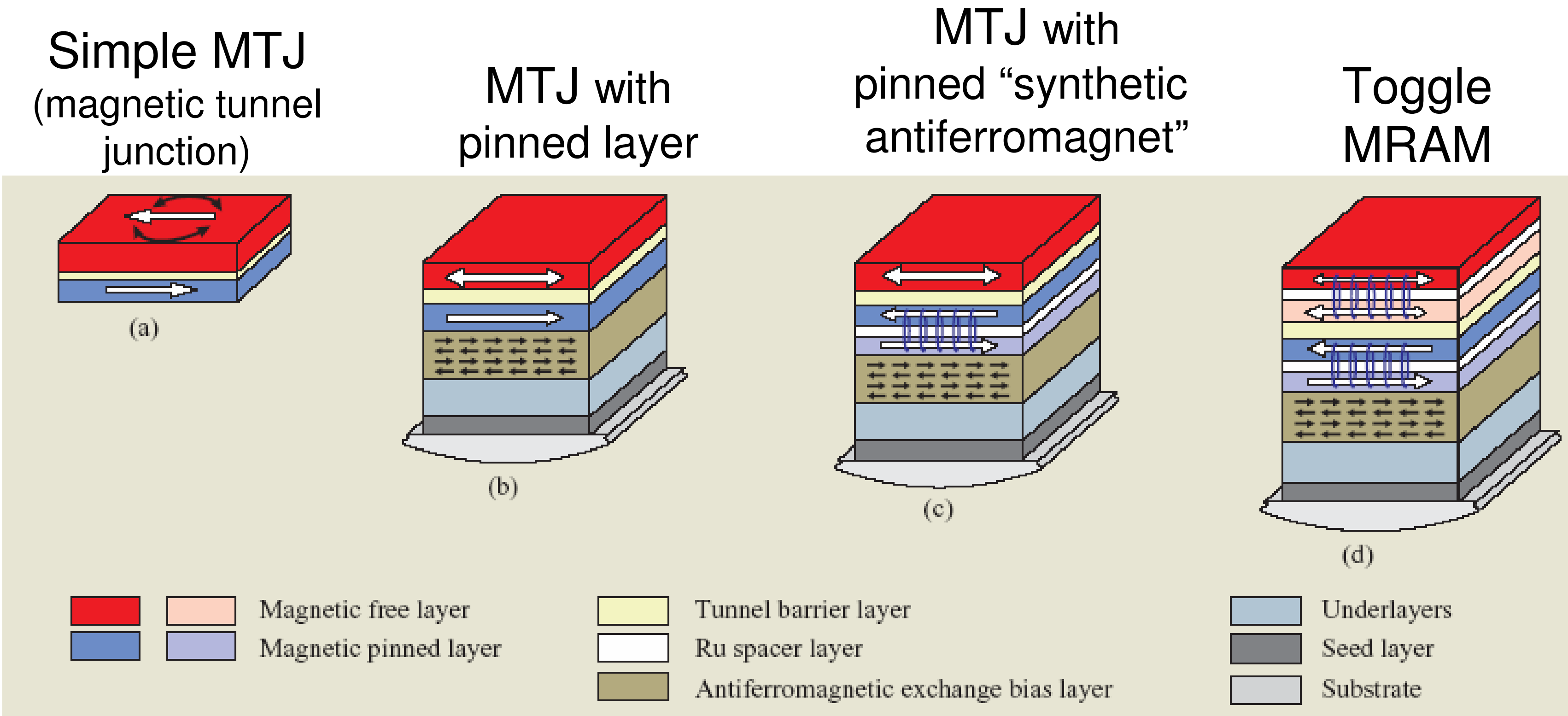
© 2008 IBM Corporation

Candidate device technologies

- **Improved Flash**
  - little change expected in write endurance or speed
- **FeRAM** – commercial product but difficult to scale!
  - FeFET – old concept, with many roadblocks
- **MRAM** (Magnetic RAM)
  - Racetrack memory
- **RRAM** (Resistive RAM)
  - Organic & polymer memory
- **Solid Electrolyte**
- **PC-RAM** (Phase-change RAM)

© 2008 IBM Corporation

MRAM (Magnetic RAM)



[Gallagher:2006]

- inherently **fast write speed**
- straightforward placement in the **CMOS back-end**
- **very high endurance** (no known wear-out mechanism)
- write by simply passing current through two nearby wires
  (superimposed magnetic field exceeds a write threshold)
  (need transistor upon reading for good SNR)

© 2008 IBM Corporation

Progress in MRAM

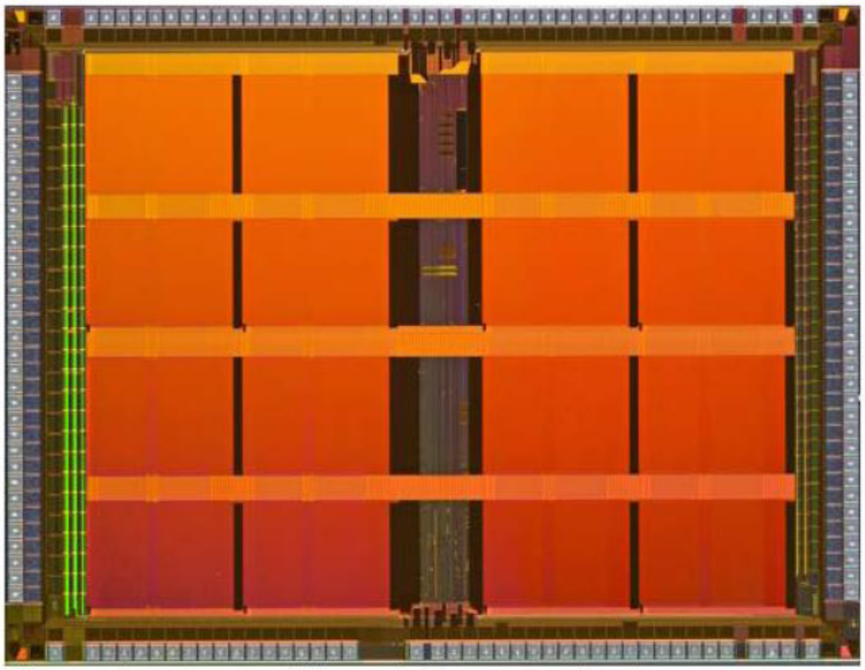
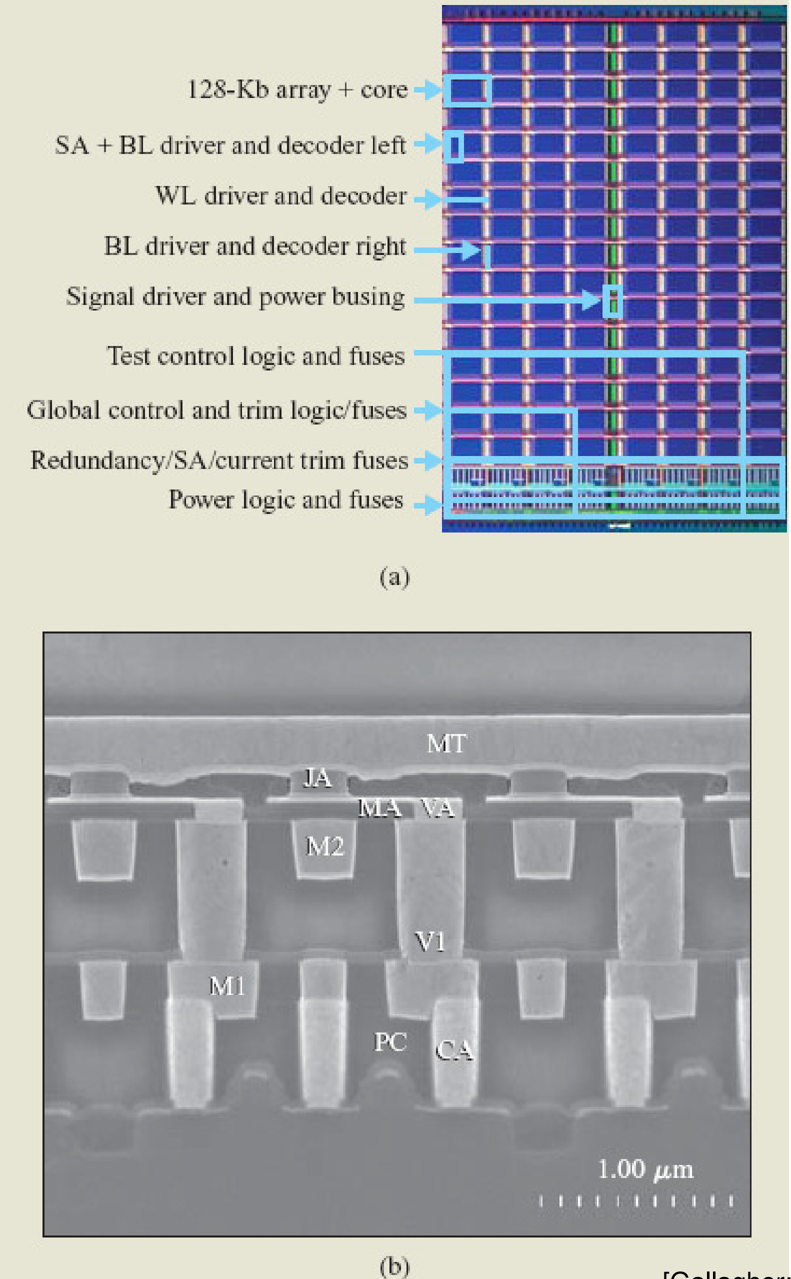


Fig. 2. First Commercially available MRAM circuit MR2A16A

[Durlam:2007]

- lots of progress 2001-2004
- commercially available
  - focus on embedded memory

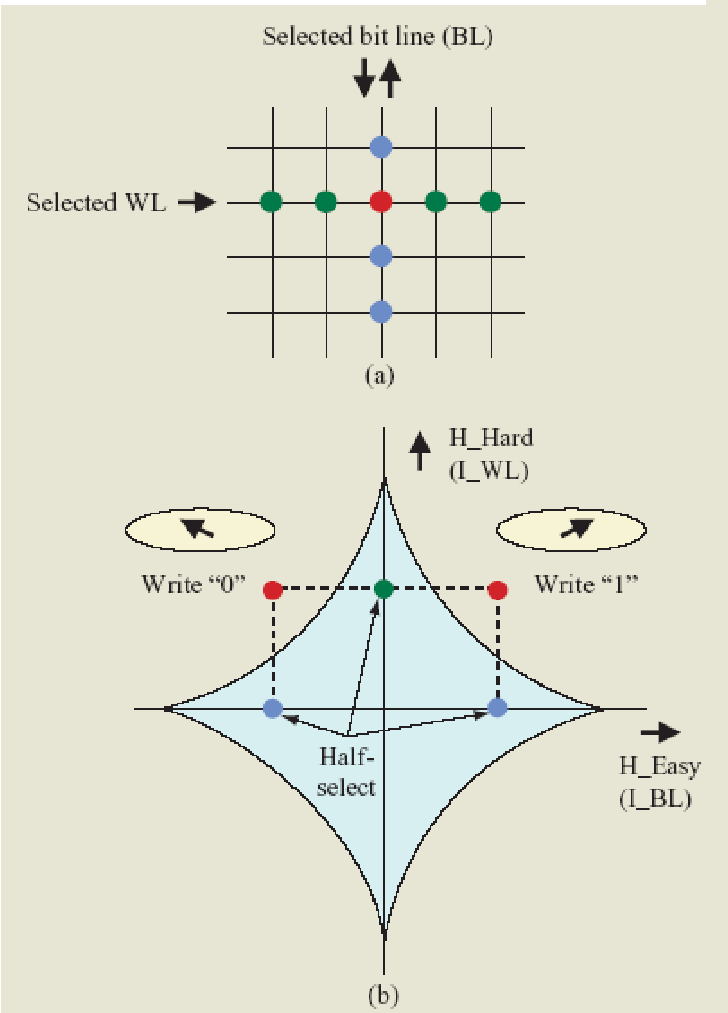


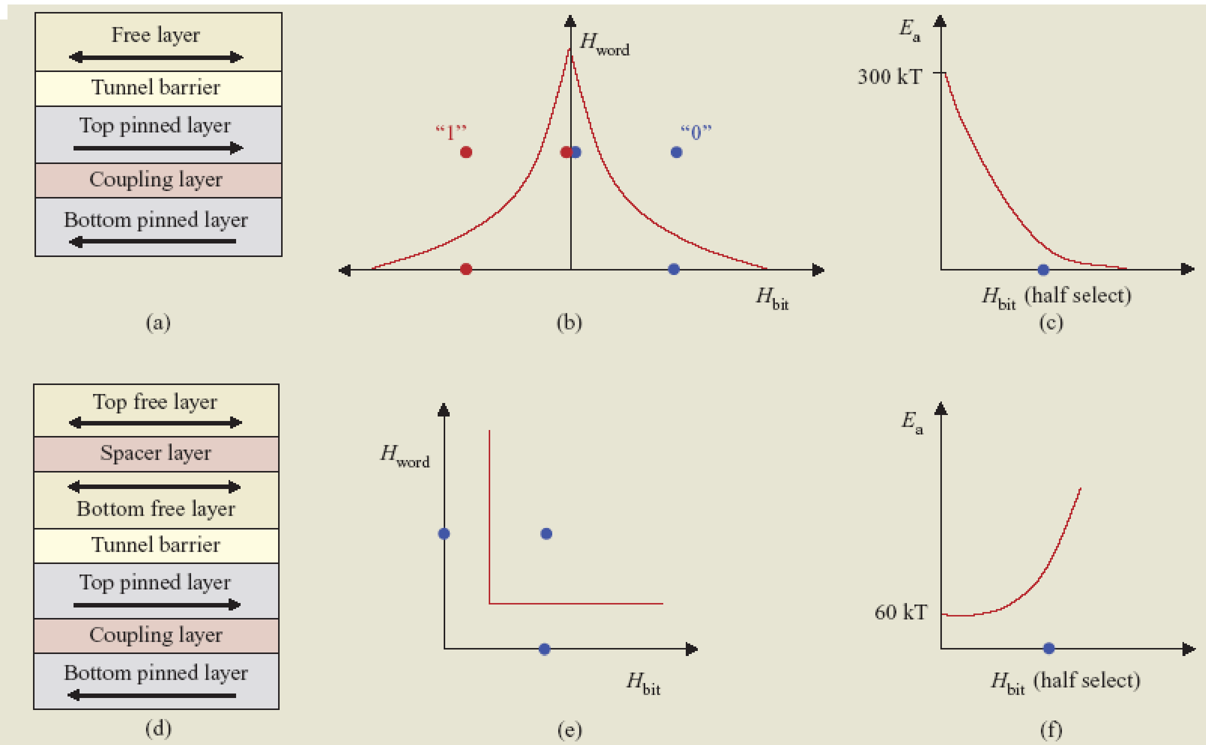
[Gallagher:2006]

© 2008 IBM Corporation

Problems with MRAM

- “Half-select problem”
  - → solved by Toggle-MRAM, but introduces a read-before-write





[Worledge:2006]

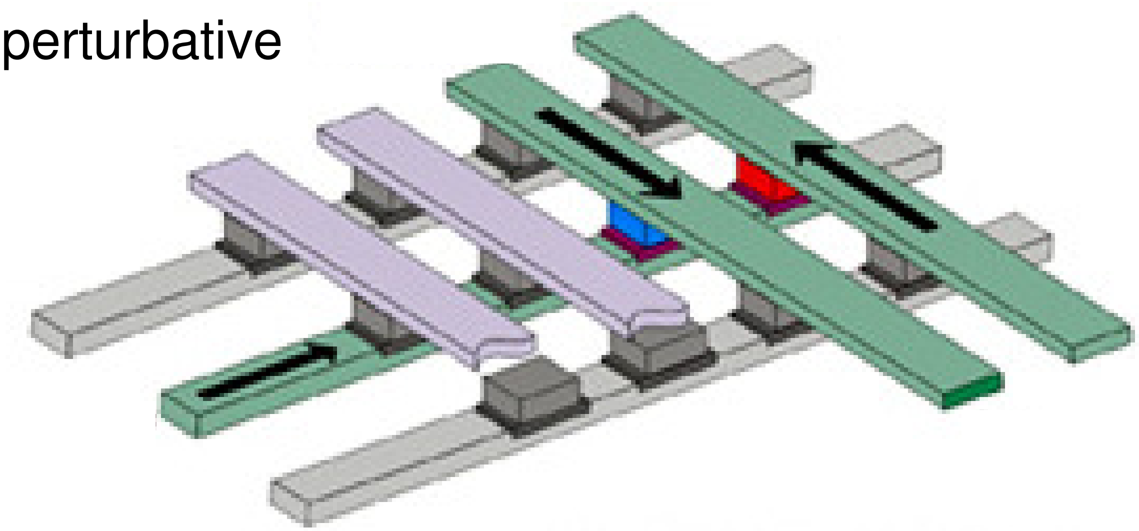
[Gallagher:2006]

© 2008 IBM Corporation

Problems with MRAM

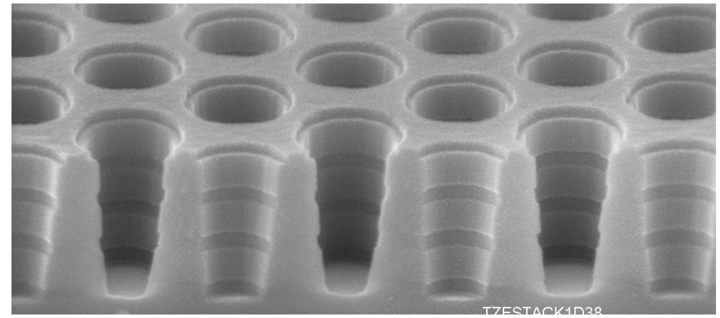
- Write currents very high – do not appear to scale well
  → electromigration even at 180nm node

- **Possible solutions**
  - **heat MTJ to reduce required current**
  - **use “spin-torque” effect**
    → rotate magnetization by passing current through the cell
    → now can have a wear-out mechanism (thin tunneling layers)
    → must insure read is non-perturbative

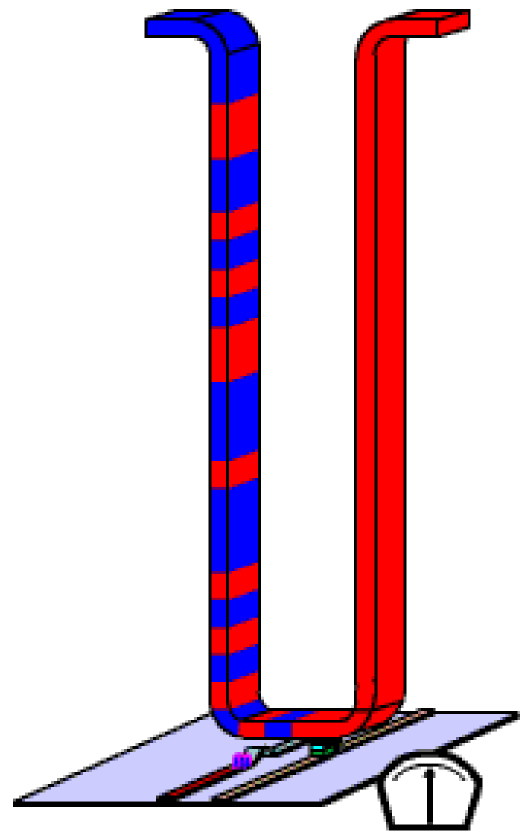


© 2008 IBM Corporation

Magnetic Racetrack Memory



- Need deep trench with notches to “pin” domains
- Need sensitive sensors to “read” presence of domains
- Must insure a moderate current pulse moves every domain one and only one notch
- Basic physics of current-induced domain motion being investigated



Promise (10-100 bits/F^2) is enormous…

but we’re still working on our basic understanding of the physical phenomena…

© 2008 IBM Corporation

Candidate device technologies

- **Improved Flash**
  - little change expected in write endurance or speed
- **FeRAM** – commercial product but difficult to scale!
  - FeFET – old concept, with many roadblocks
- **MRAM** – commercial product, also difficult to scale!
  - Racetrack memory – new concept w/ promise, still at point of early basic physics research
- **RRAM** (Resistive RAM)
  - Organic & polymer memory
- **Solid Electrolyte**
- **PC-RAM** (Phase-change RAM)

© 2008 IBM Corporation

For more information (on FeRAM, MRAM, RRAM & SE)

G. W. Burr, B. N. Kurdi, J. C. Scott, C. H. Lam, K. Gopalakrishnan, and R. S. Shenoy, "An overview of candidate device technologies for Storage-Class Memory," to appear in *IBM Journal of Research and Development*, (2008).

FeRAM

- A. Sheikholeslami and P. G. Gulak, *Proc. IEEE,* **88**, No. 5, 667-689 (2000).
- Y.K. Hong, D.J. Jung, et. al., *Symp. VLSI Technology*, 230-231 (2007).
- K. Kim and S. Lee, *J. Appl. Phys.,* **100**, No. 5, 051604 (2006).
- N. Setter, D. Damjanovic, et. al., *J. Appl. Phys.*, **100**(5), 051606 (2006).
- D. Takashima and I. Kunishima, *IEEE J. Solid-State Circ.,* **33**, No. 5, 787-792 (1998).
- S. L. Miller and P. J. McWhorter, *J. Appl. Phys.*, **72**(12), 5999-6010 (1992).
- T. P. Ma and J. P. Han, *IEEE Elect. Dev. Lett.,* **23**, No. 7, 386-388 (2002).

MRAM

- R. E. Fontana and S. R. Hetzler, *J. Appl. Phys.*, **99**(8), 08N902, (2006).
- W. J. Gallagher and S. S. P. Parkin, *IBM J. Res. Dev.* **50**(1), 5-23, (2006).
- M. Durlam, Y. Chung, et. al., *ICICDT Tech. Dig.*, 1-4, (2007).
- D. C. Worledge, *IBM J. Res. Dev.* **50**(1), 69-79, (2006).
- S.S.P. Parkin, *IEDM Tech. Dig.*, 903-906 (2004).
- L. Thomas, M. Hayashi, et. al., *Science*, **315**(5818), 1553-1556 (2007).

RRAM

- J. C. Scott and L. D. Bozano, *Adv. Mat.*, **19**, 1452-1463 (2007).
- Y. Hosoi, Y. Tamai, et. al., *IEDM Tech. Dig.*, 30.7.1-4 (2006).
- D. Lee, D.-J. Seong, et. al., *IEDM Tech. Dig.*, 30.8.1-4 (2006).
- S. F. Karg, G. I. Meijer, et. al., to appear in *IBM J. Res. Dev.*, (2008).

SE

- M. N. Kozicki, M. Park, and M. Mitkova, *IEEE Trans. Nanotech.*, **4**(3), 331-338 (2005).
- M.N. Kozicki, M. Balakrishnan, et. al., *Proc. IEEE NVSM Workshop*, 83-89 (2005).
- M. Kund, G. Beitel, et. al., *IEDM Tech. Dig.*, 754-757 (2005).
- P. Schrögmeier, M. Angerbauer, et. al., *Symp. VLSI Circ.*, 186-187 (2007).

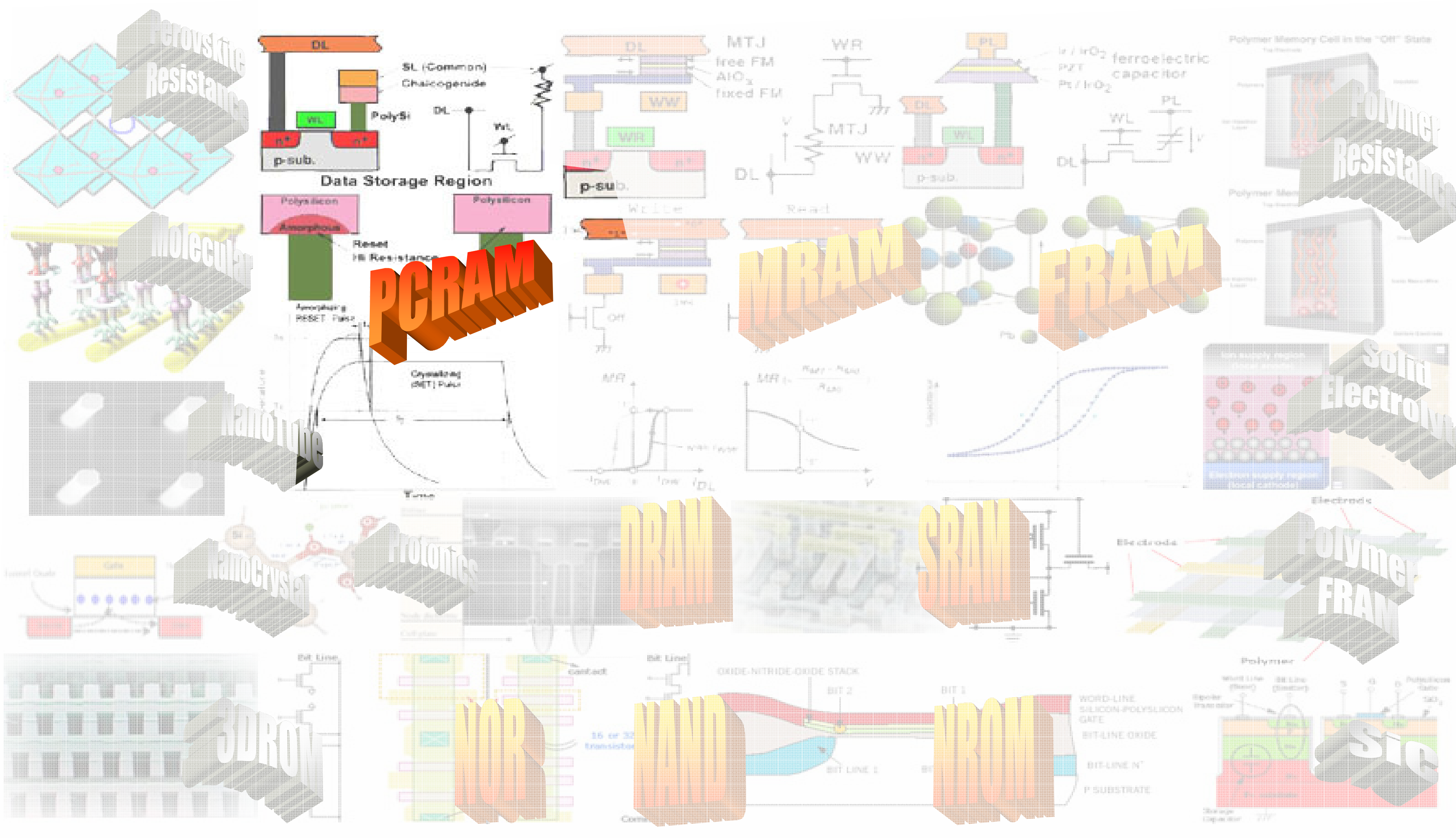
 © 2008 IBM Corporation

Candidate device technologies

- **Improved Flash**
  - little change expected in write endurance or speed
- **FeRAM** – commercial product but difficult to scale!
  - FeFET – old concept, with many roadblocks
- **MRAM** – commercial product, also difficult to scale!
  - Racetrack memory – new concept w/ promise, still at point of early basic physics research
- **RRAM** – few demos showing real CMOS integration
  - Organic & polymer memory – temperature compatibility?
- **Solid Electrolyte** – shows real promise if tradeoff between retention & overprogramming can be solved…
- **PC-RAM** (Phase-change RAM)

© 2008 IBM Corporation

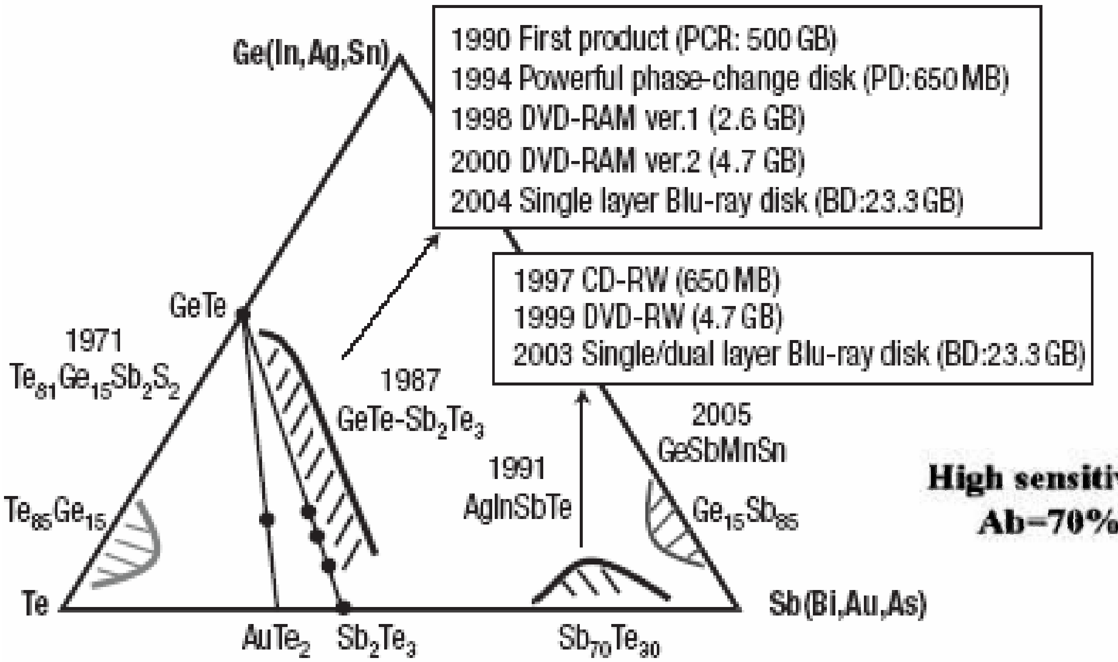
Memory/storage landscape



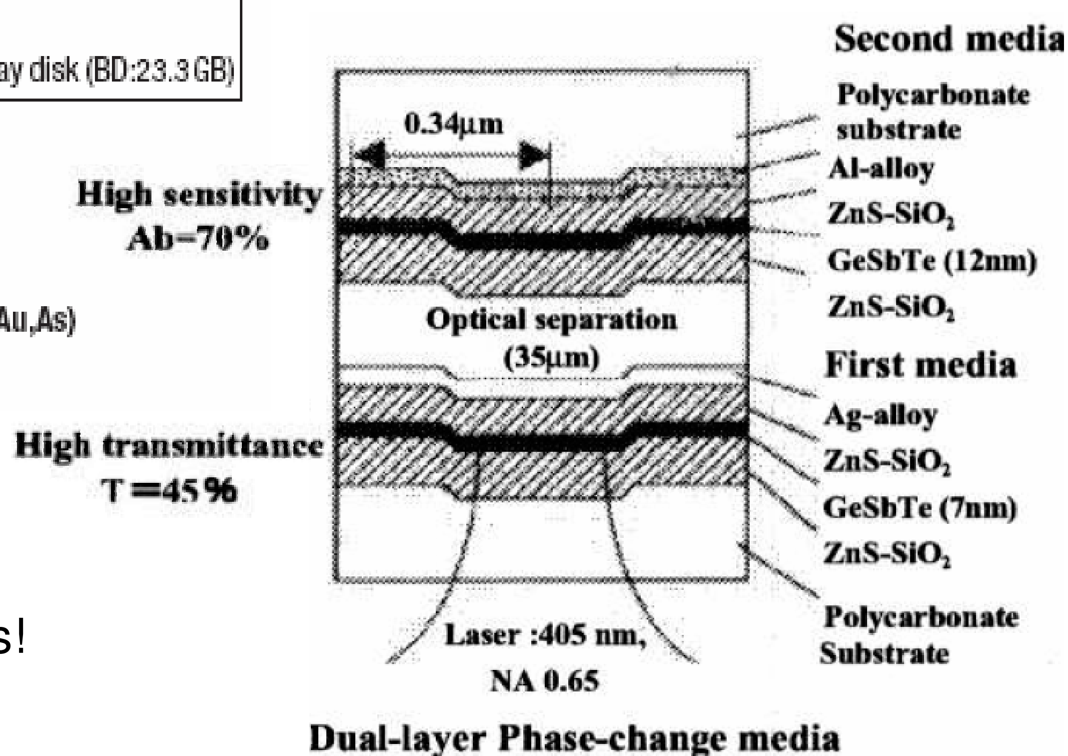
 © 2008 IBM Corporation

History of Phase-change memory

- late 80's – 90's – **Fast** phase-change materials discovered & optimized for re-writeable optical storage



[Wuttig:2007]

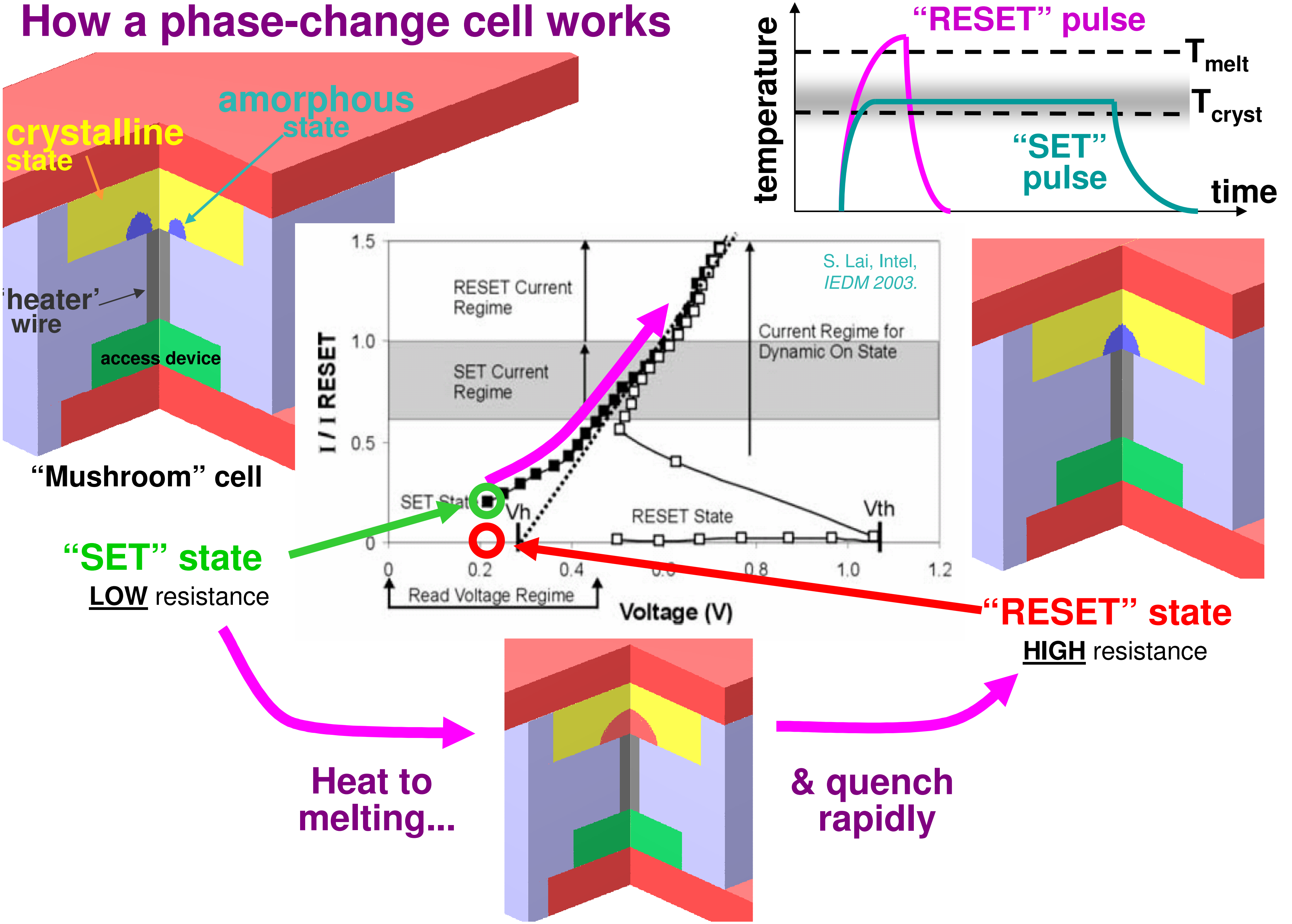


[Ohta:2001]

- late-1990's and on – return to PC-RAM with fast materials!

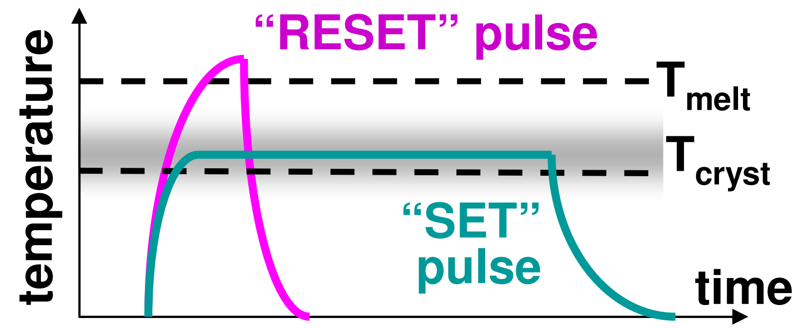
© 2008 IBM Corporation

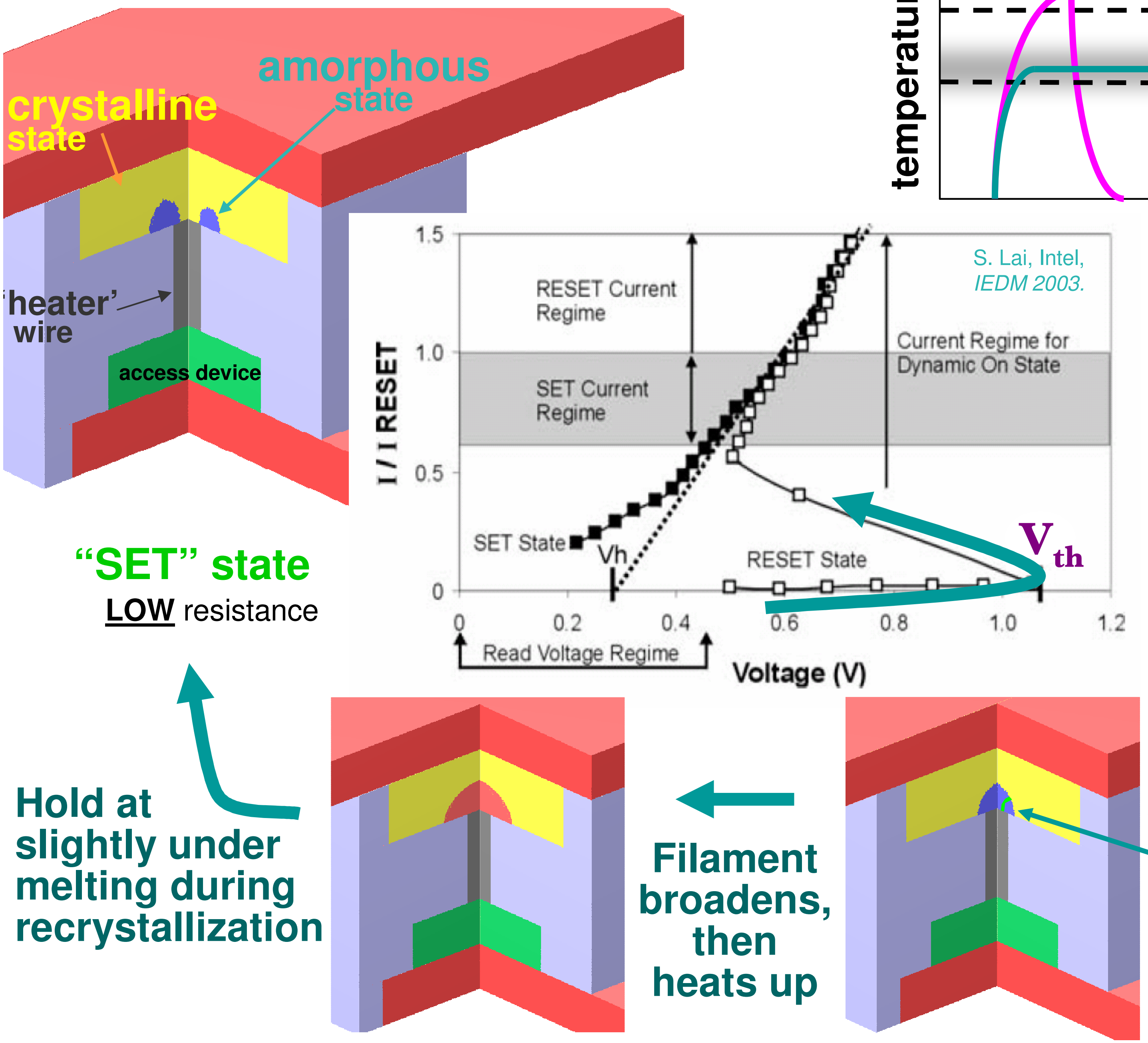
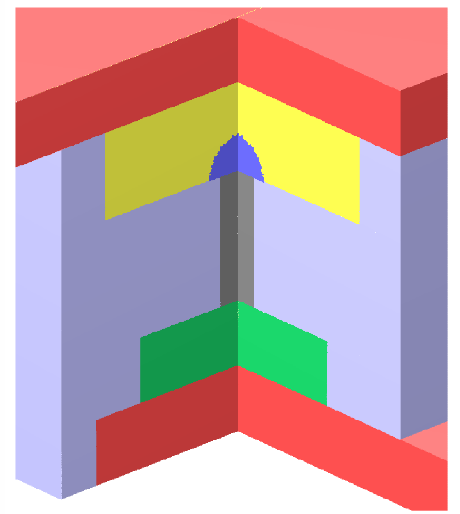
How a phase-change cell works



© 2008 IBM Corporation

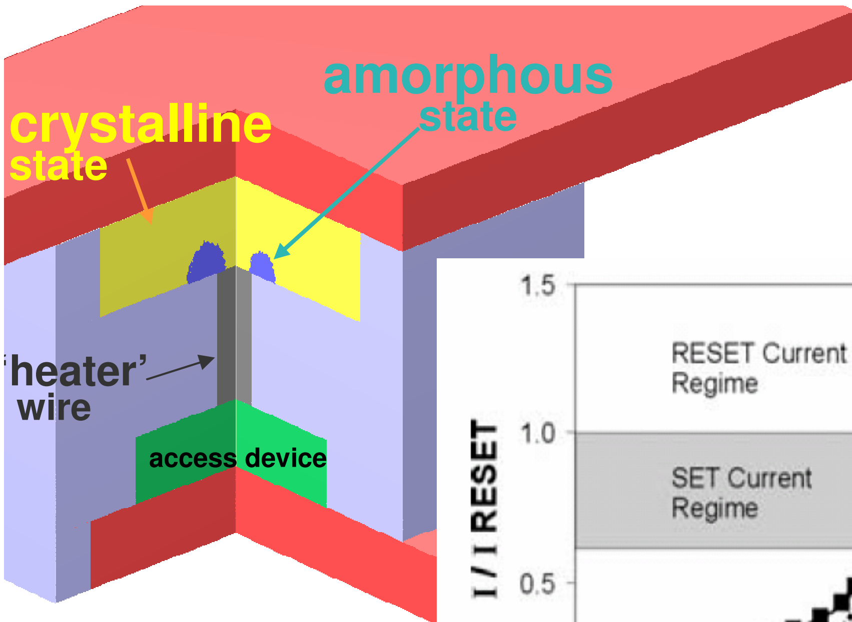
How a phase-change cell works

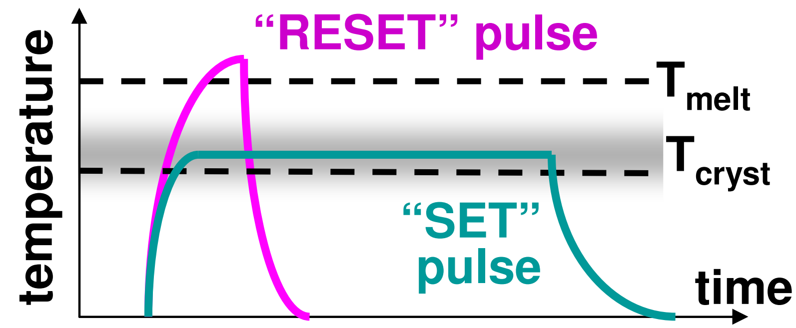


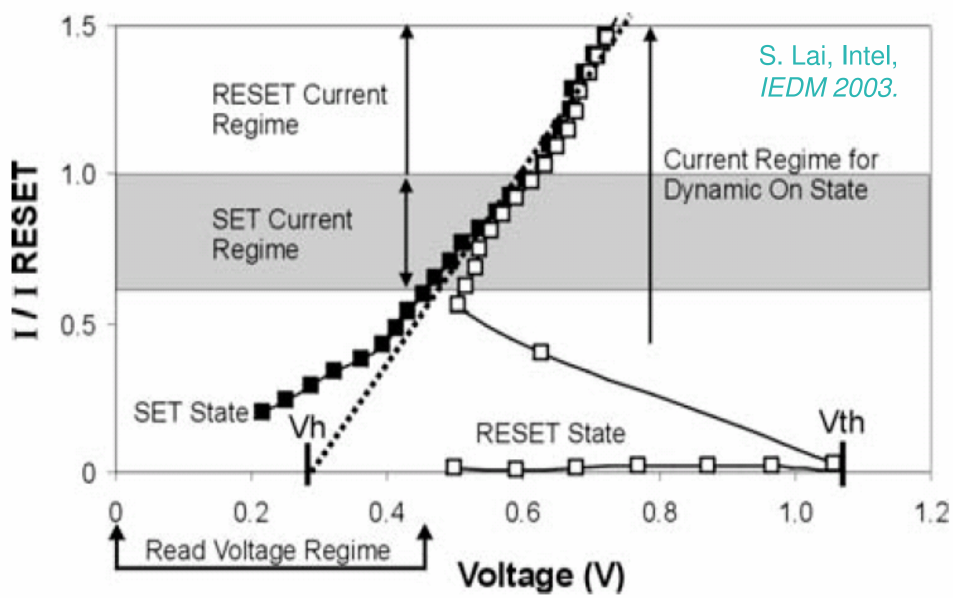



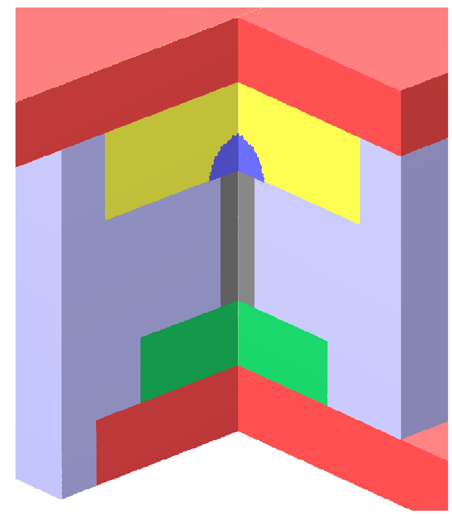
© 2008 IBM Corporation

How a phase-change cell works









"SET" state
<u>**LOW**</u> resistance

"RESET" state
<u>**HIGH**</u> resistance

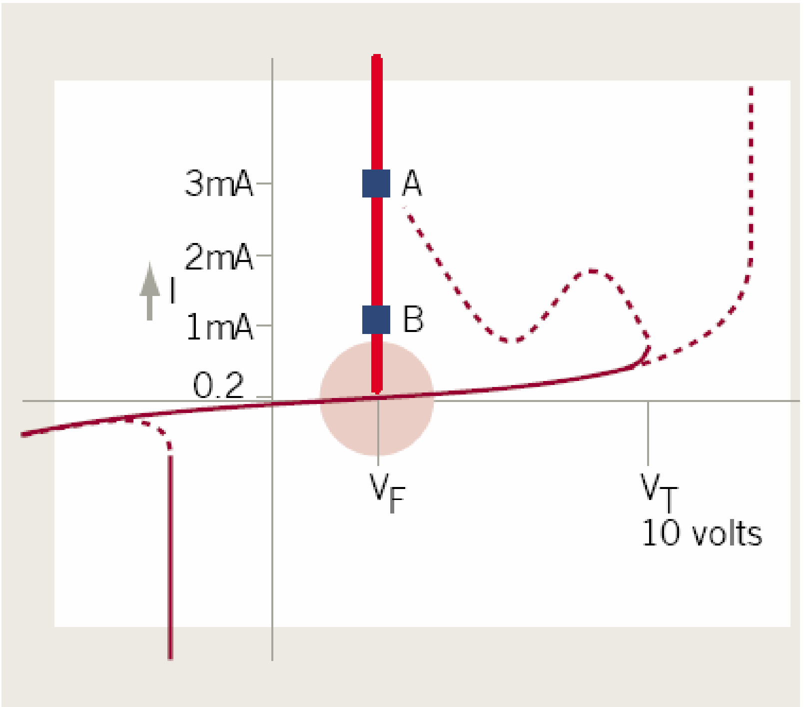
<u>Issues for phase-change memory</u>

- Keeping the **RESET current** low
- Multi-level cells (for **>1bit** / **cell**)
- Is the technology **scalable**?

© 2008 IBM Corporation

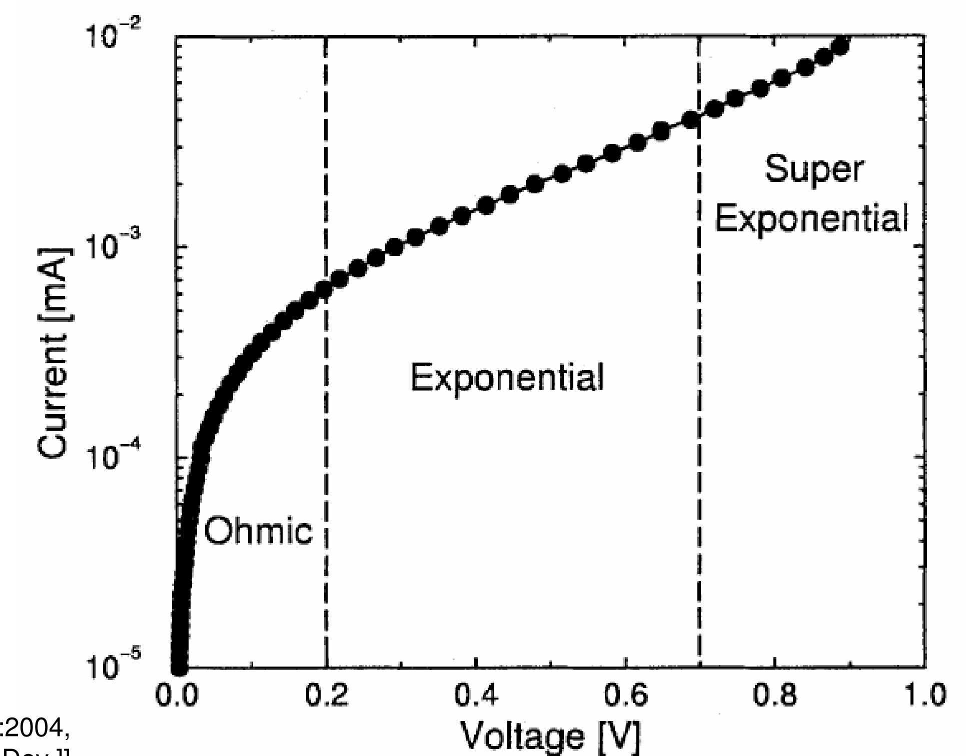
Electrical "breakdown" in PCM devices

- 70's – Study of **electrical breakdown** – "memory switching' vs. "threshold switching"



[Neale:2001]

- Recent studies – electrical resistivity drops rapidly with electric field…



[Pirovano:2004, IEEE Tr. Electr. Dev.]]

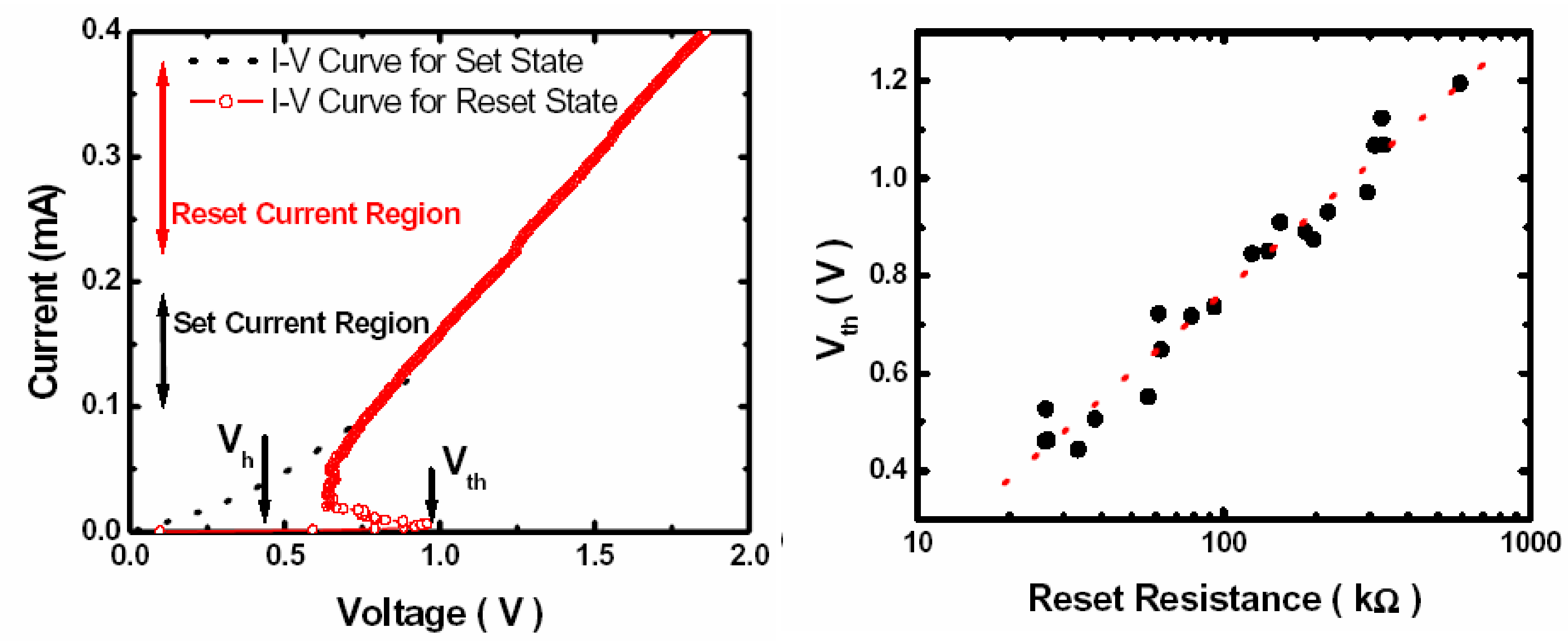
© 2008 IBM Corporation

Electrical "breakdown" in PC-RAM devices

- 70's – Study of **electrical breakdown** – "memory switching' vs. "threshold switching"

- Recent studies – electrical resistivity drops rapidly with electric field…

- "Threshold voltage" observed to be a function of the "size" of the amorphous plug…



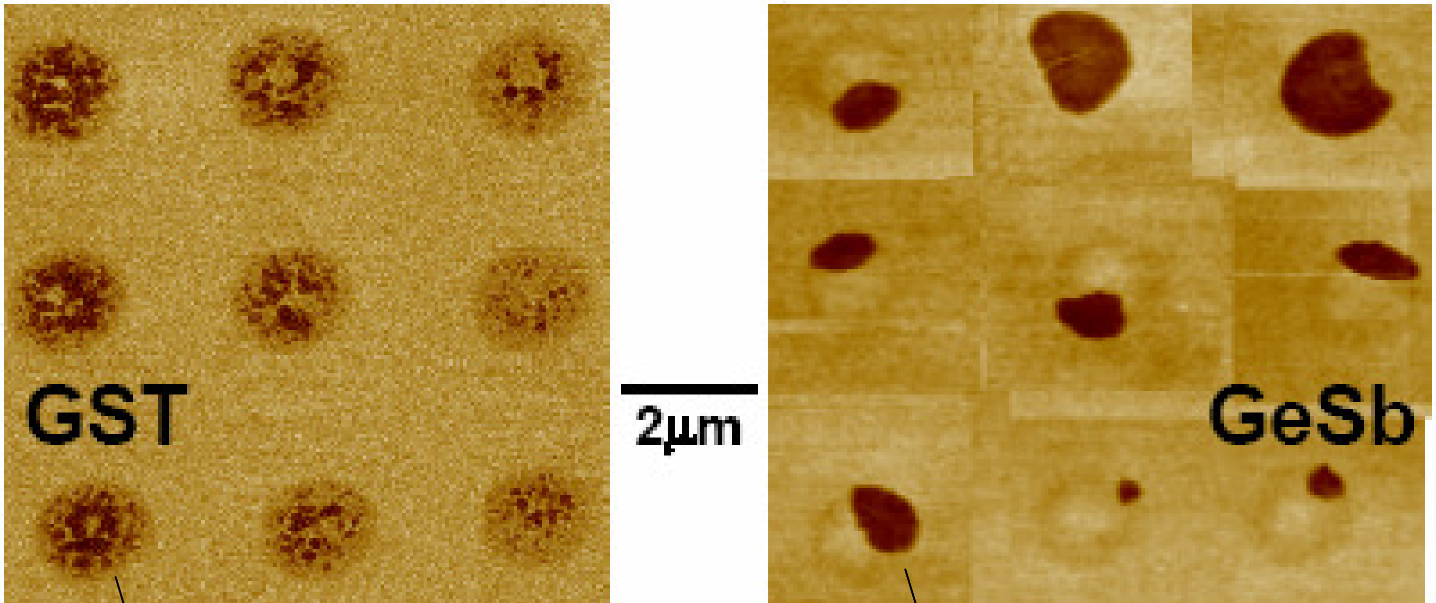
[Ha:2003]

© 2008 IBM Corporation

Phase-change materials

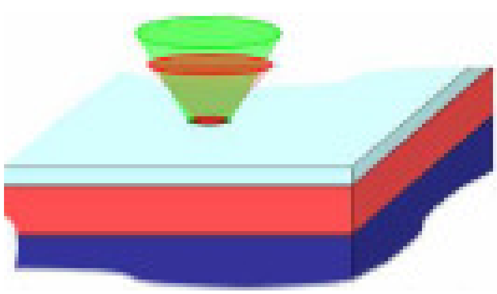
- Two types of materials: "**nucleation-dominated**" vs. "**growth-dominated**"

AFM taken after optical experiments on "as-deposited" amorphous material…



Nucleation-dominated
Many crystalline nuclei start growing inside each optical spot

Growth-dominated
After a long incubation time where nothing happens, one nuclei then gets started and rapidly grows to cover the entire optical spot



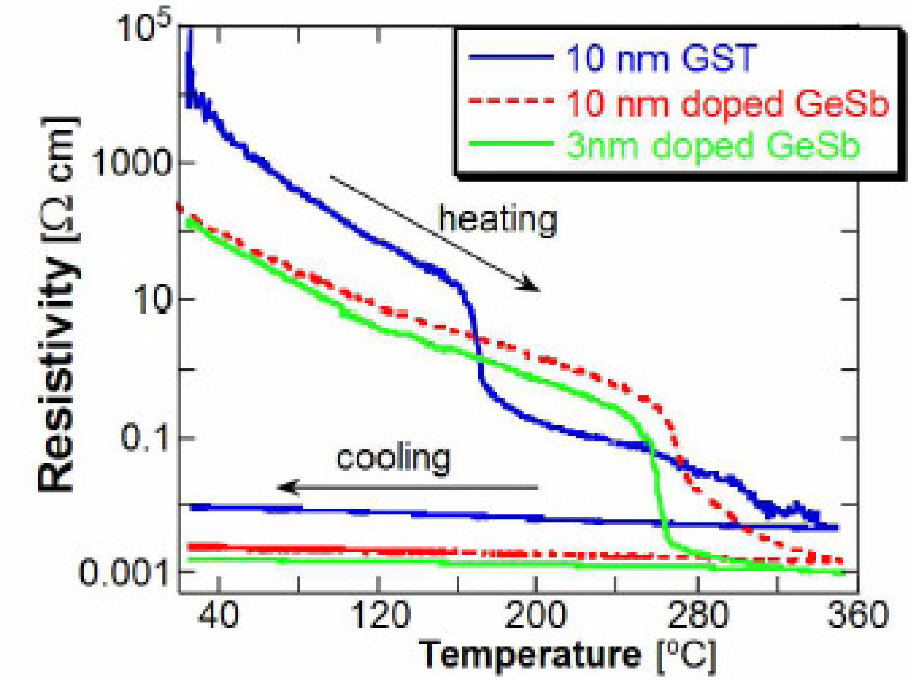
[Chen:2006]

© 2008 IBM Corporation

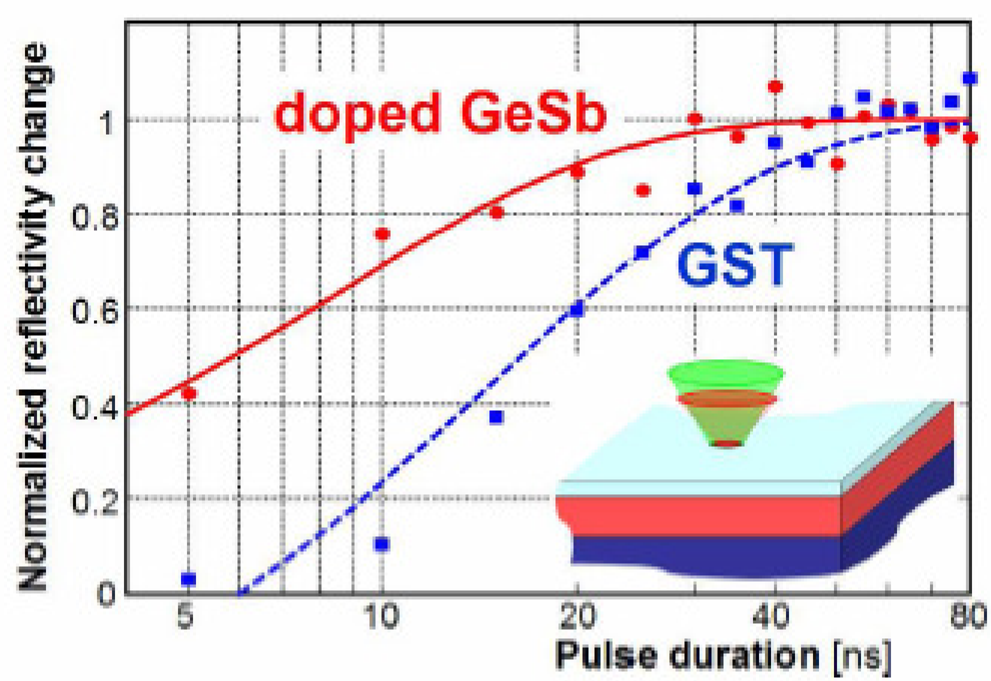
Phase-change materials

We want a material that…

…**retains data** at moderate temperature…



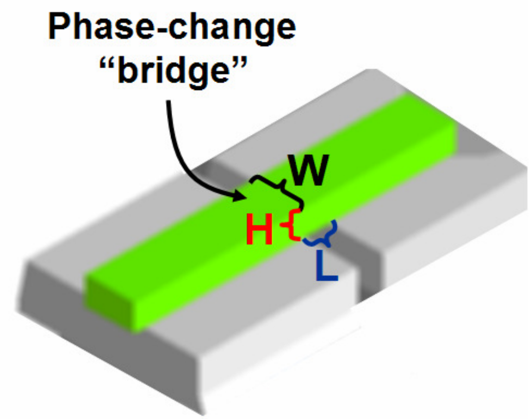
yet **switches rapidly** at high temperature.



[Chen:2006]

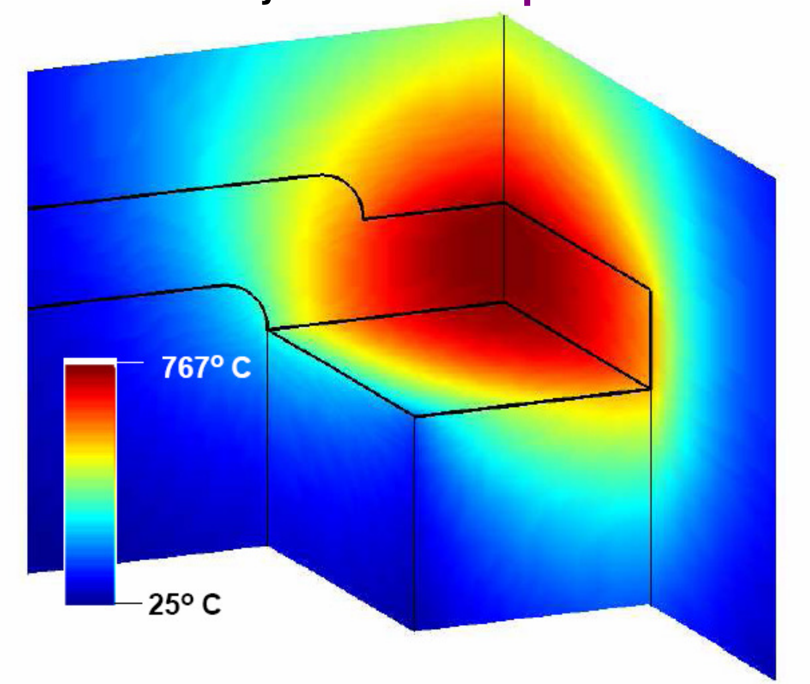
© 2008 IBM Corporation

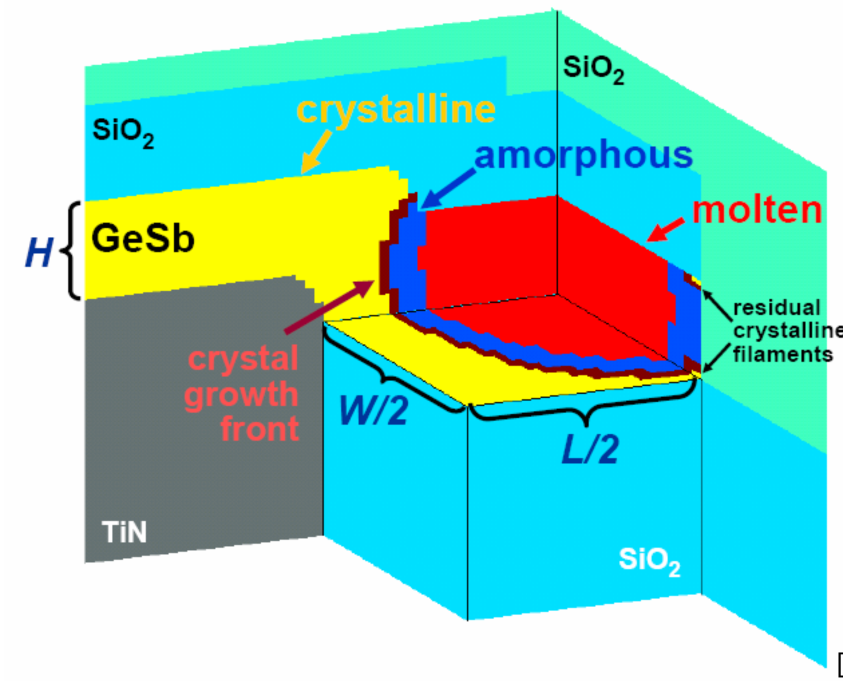
Designing for lower RESET current



W defined by lithography
H by thin-film deposition

- We use **modeling** to help understand how the phase-change cell works
- In particular, design choices that can **reduce RESET current/power** are particularly important





[Chen:2006]

© 2008 IBM Corporation

Scalability of PCM

Basic requirements

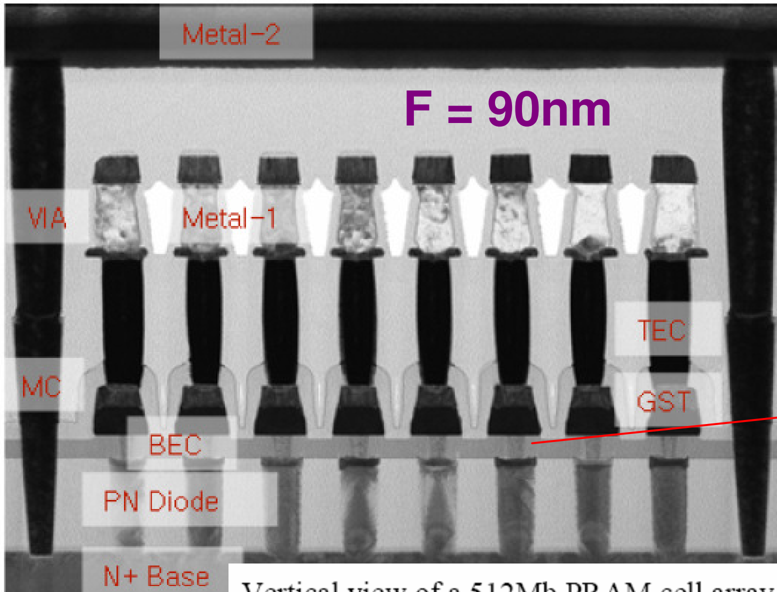
- ✓ widely separated SET and RESET resistance distributions
- ✓ switching with accessible electrical pulses
- ✓ the ability to read/sense the resistance states without perturbing them
- ✓ high write **endurance** (many switching cycles between SET and RESET)
- ✓ long data **retention** ("10-year data lifetime" at some elevated temperature)
  - → avoid unintended re-crystallization
- ✓ **fast** SET speed
- ✓ **MLC** capability – more than one bit per cell

Any new non-volatile memory technology had better work for several device generations… ⇨ Will PC-RAM scale?

- ? will the phase-change process even work at the 22nm node?
- ? can we fabricate tiny, high-aspect devices?
- ? can we make them all have the same Critical Dimension (CD)?
- ? what happens when the # of atoms becomes countable?

© 2008 IBM Corporation

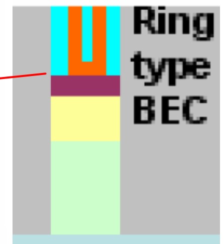
PCM state-of-the-art



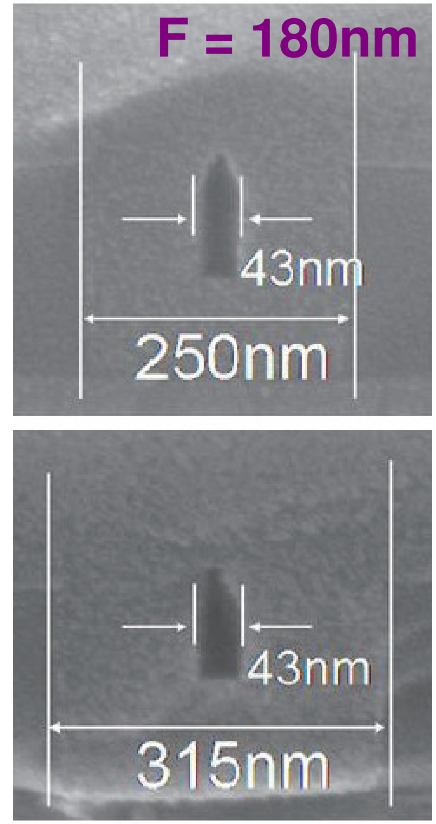
Vertical view of a 512Mb PRAM cell array.

[Oh:2006]

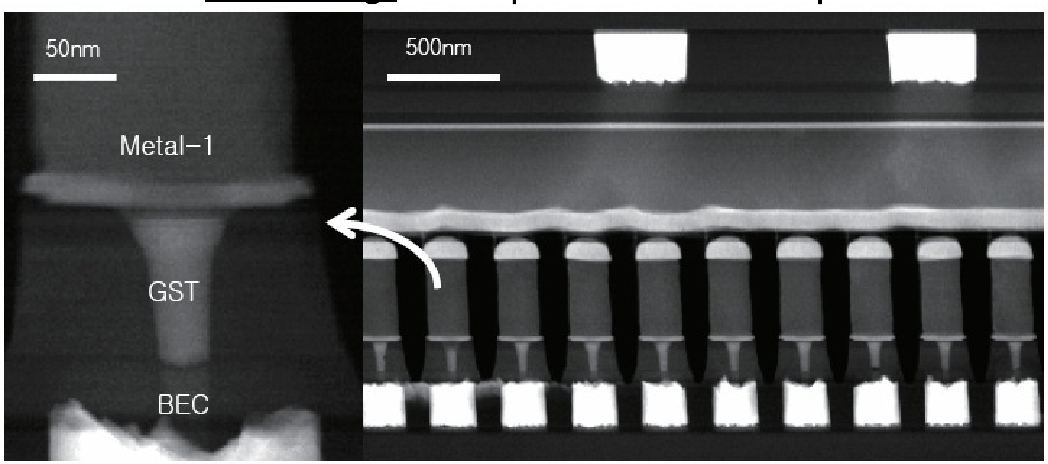
Samsung:
- ring bottom electrode (BEC) **reduces CD variations**
- diode → more current
- 90nm process

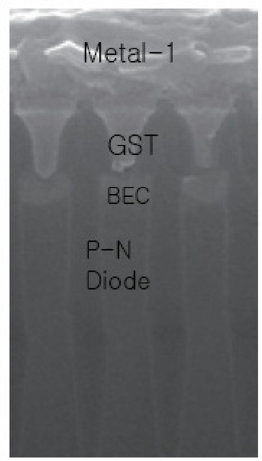


[Breitwisch:2007]



Samsung: CVD process fills deep holes





[Lee:2007 VLSI]

IBM/Macronix/Qimonda: make features only F/4 in size yet **reduce CD variations**

© 2008 IBM Corporation

Outlook of PCM

- ✓ will the phase-change process even work at the 22nm node?
- ✓ can we fabricate tiny, high-aspect devices?
- ✓ can we make them all have the same Critical Dimension (CD)?
- ? what happens when the # of atoms becomes countable?

Scaling outlook appears to be “good” for PC-RAM

By adding two bits per cell, Intel and ST Microelectronics have put phase-change memory on par with today's flash technology, says H.-S. Philip Wong, professor of electrical engineering at Stanford University. Intel has already mastered a similar trick with flash

Phase-change memory has made a lot of progress in the past few years, Wong adds. "A few years ago it looked promising," he says. "But now it's going to happen. There's no doubt about it."

February 4, 2008

[http://www.technologyreview.com/Infotech/20148/]

→ Focus now on novel IP, implementation, and cost reduction.

© 2008 IBM Corporation

For more information (on PCRAM)

S. Raoux, G. W. Burr, M. J. Breitwisch, C. T. Rettner, Y. Chen, R. M. Shelby, M. Salinga, D. Krebs, S. Chen, H. Lung, and C. H. Lam, "Phase-change random access memory — a scalable technology," to appear in *IBM Journal of Research and Development*, (2008).

PCRAM

- S. R. Ovshinsky, *Phys. Rev. Lett.*, **21**(20), 1450 (1968).
- D. Adler, M. S. Shur, et. al., *J. Appl. Phys.*, **51**(6), 3289-3309 (1980).
- R. Neale, *Electronic Engineering*, **73**(891), 67-, (2001).
- T. Ohta, K. Nagata, et. al., *IEEE Trans. Magn.*, **34**(2), 426-431 (1998).
- T. Ohta, J. Optoelectr. Adv. Mat., **3**(3), 609-626 (2001).
- S. Lai, *IEDM Technical Digest*, 10.1.1-10.1.4, (2003).
- A. Pirovano, A. L. Lacaita, et. al., *IEDM Tech. Dig.*, 29.6.1-29.6.4, (2003).
- A. Pirovano, A. Redaelli, et. al., *IEEE Trans. Dev. Mat. Reliability*, **4**(3), 422-427, (2004).
- A. Pirovano, A. L. Lacaita, et. al., *IEEE Trans. Electr. Dev.*, **51**(3), 452-459 (2004).
- Y. C. Chen, C. T. Rettner, et. al., *IEDM Tech. Dig.*, S30P3, (2006).
- J.H. Oh, J.H. Park, et. al., *IEDM Tech. Dig.*, 2.6, (2006).
- S. Raoux, C. T. Rettner, et. al., *EPCOS 2006*, (2006).
- M. Breitwisch, T. Nirschl, et. al., *Symp. VLSI Tech.*, 100-101, (2007).
- T. Nirschl, J. B. Philipp, et. al., *IEDM Technical Digest*, 17.5, (2007).
- J.I. Lee, H. Park, *Symp. VLSI Tech.*, 102-103 (2007).
- S.-H. Lee, Y. Jung, and R. Agarwal, *Nature Nanotech.*, **2**(10), 626-630 (2007).
- D. H. Kim, F. Merget, et. al., *J. Appl. Phys.*, **101**(6), 064512 (2007).
- M. Wuttig and N. Yamada, *Nature Materials*, **6**(11), 824-832 (2007).

 © 2008 IBM Corporation

Outline

- **Motivation**
  - ✓ by 2020, server-room power & space demands will be <u>too high</u>
  - ✓ evolution of hard-disk drive (HDD) storage and Flash cannot help
  - ✓ need a new technology – **Storage Class Memory (SCM)** – that combines
    - ❖ the benefits of a solid-state memory (**high performance** and **robustness**)
    - ❖ with the **archival capabilities** and **low cost** of conventional HDD

- **How could we build an SCM?**
  - ✓ combine a scalable non-volatile memory (**Phase-change memory**)
  - • with **ultra-high density** integration, using
    - ❖ micro-to-nano addressing
    - ❖ multi-level cells
    - ❖ 3-D stacking

© 2008 IBM Corporation

Cost structure of silicon-based technology

Co$t determined by

- cost per wafer
  - # of dies/wafer
    - memory area per die [sq. µm]
      - **memory density** [bits per $4F^2$]
        - **patterning density** [sq. µm per $4F^2$]

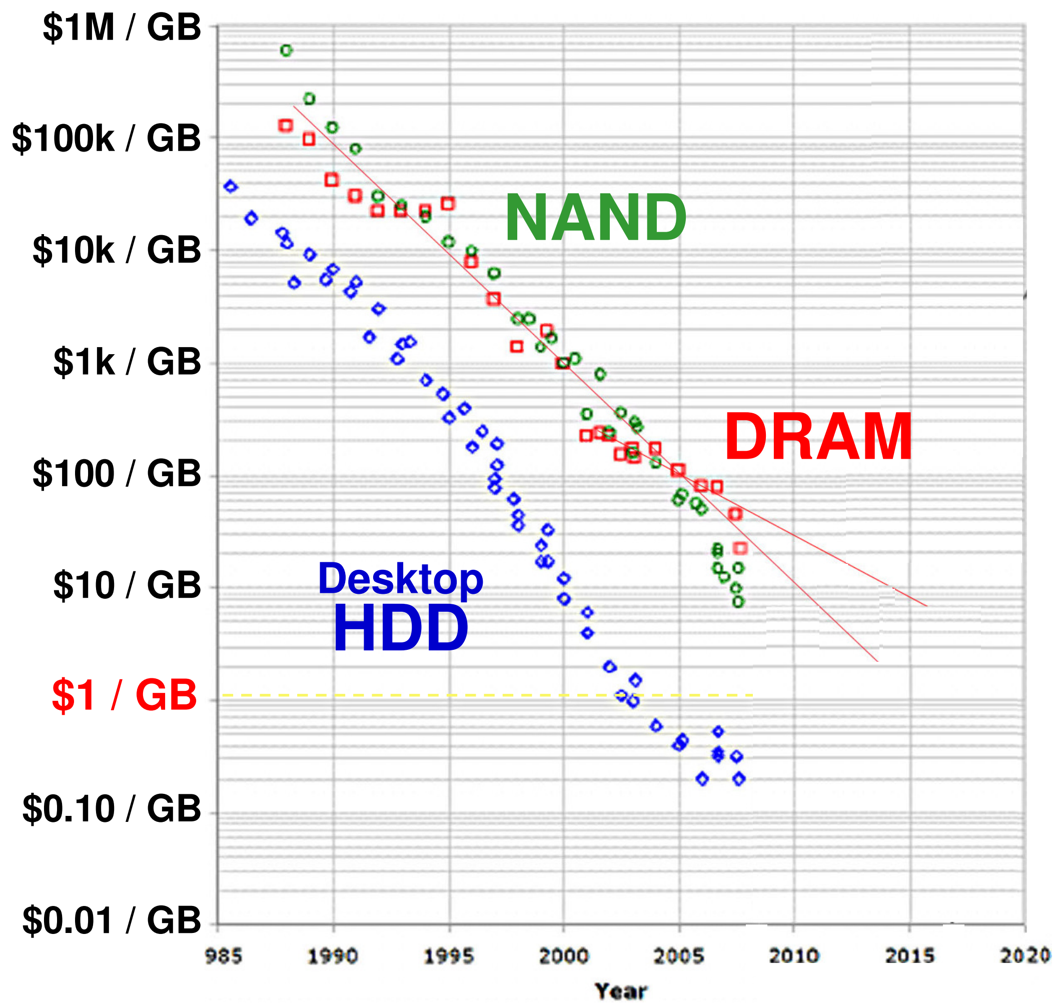
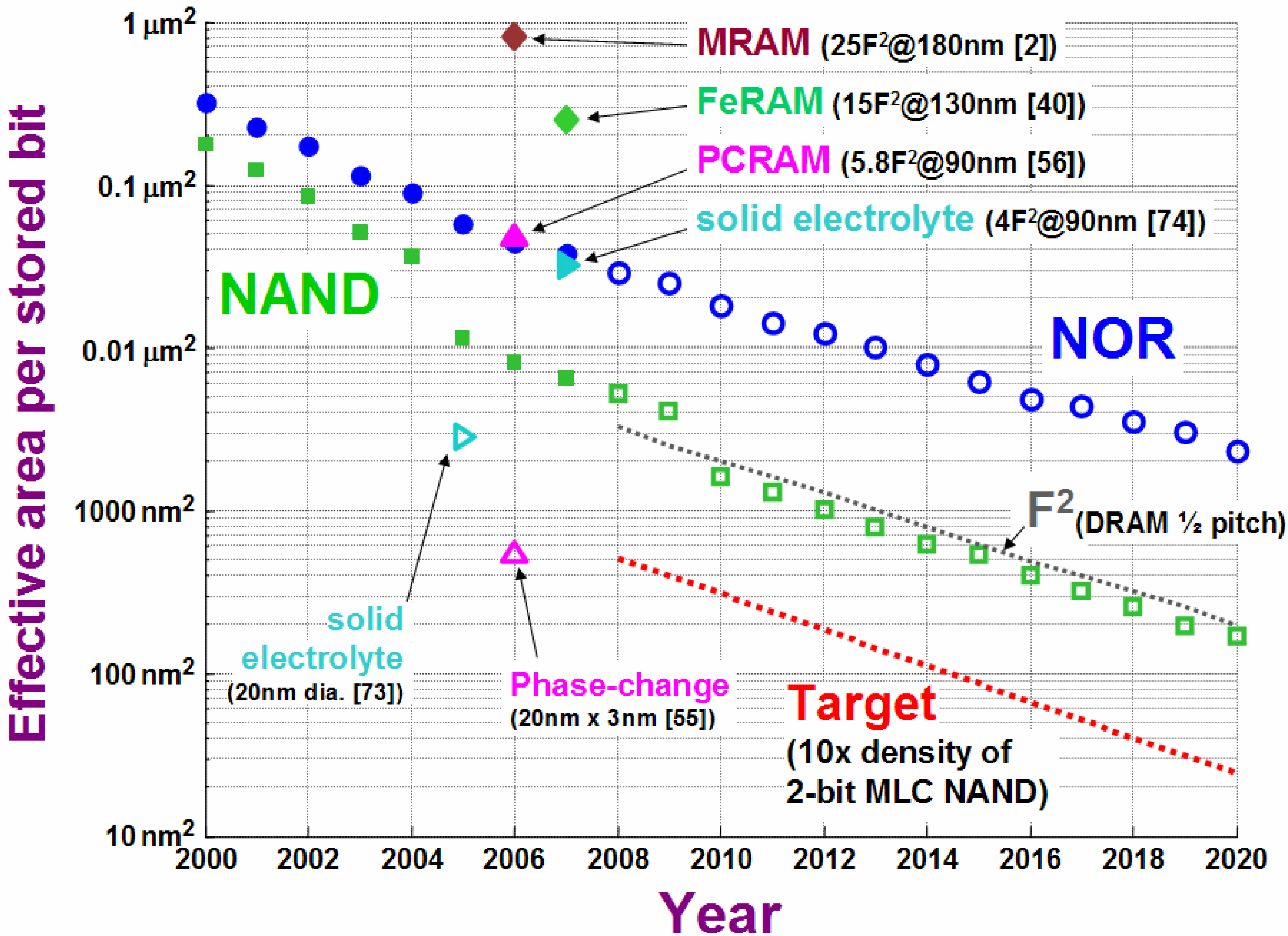


Chart courtesy of Dr. Chung Lam, IBM Research
To be published in *IBM Journal R&D*

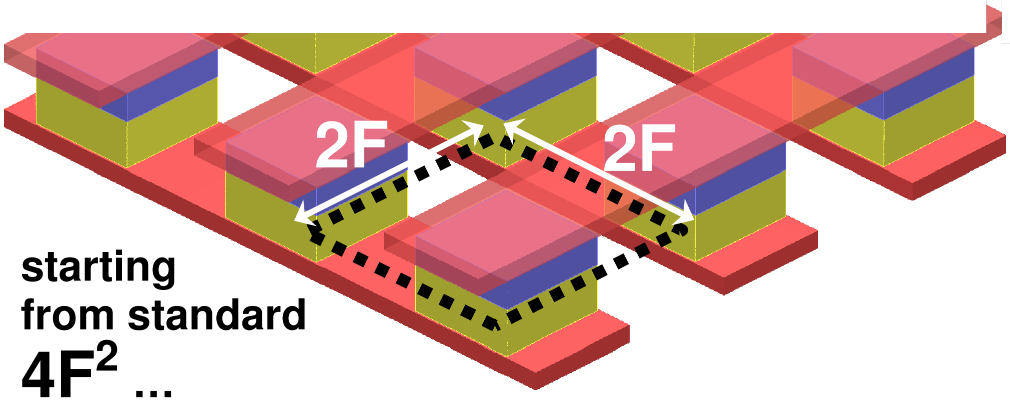
© 2008 IBM Corporation

Need a 10x boost in density BEYOND Flash!

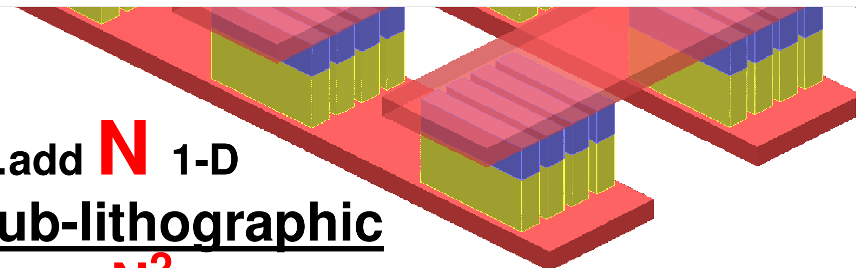


© 2008 IBM Corporation

Paths to ultra-high density memory



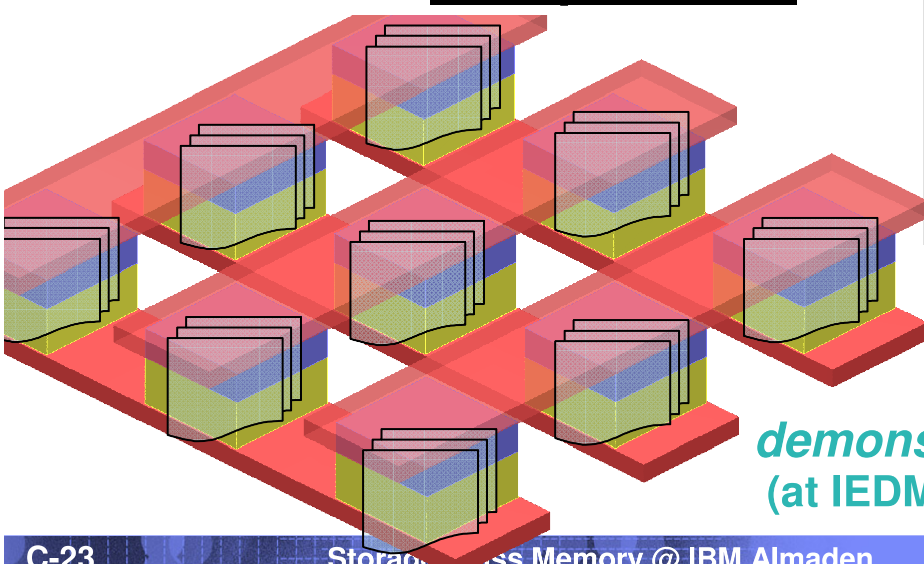
starting from standard $4F^2$ …



…add **N** 1-D <u>sub-lithographic</u> "fins" (**N^2** with 2-D)

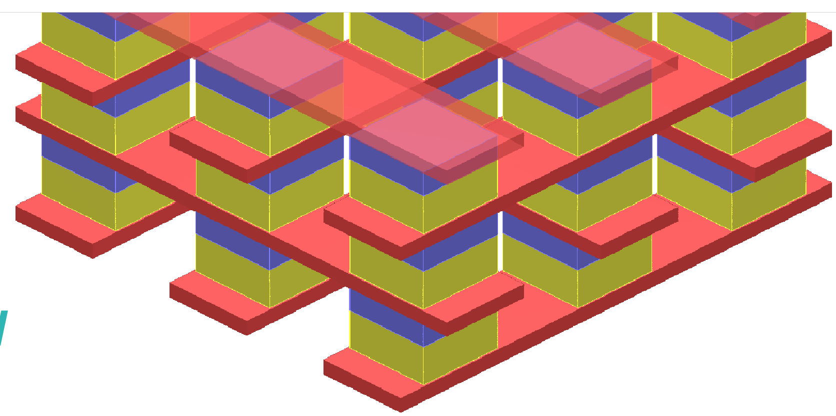
demonstrated (at IEDM 2005)

…store **M** bits/cell with **2^M** <u>multiple levels</u>



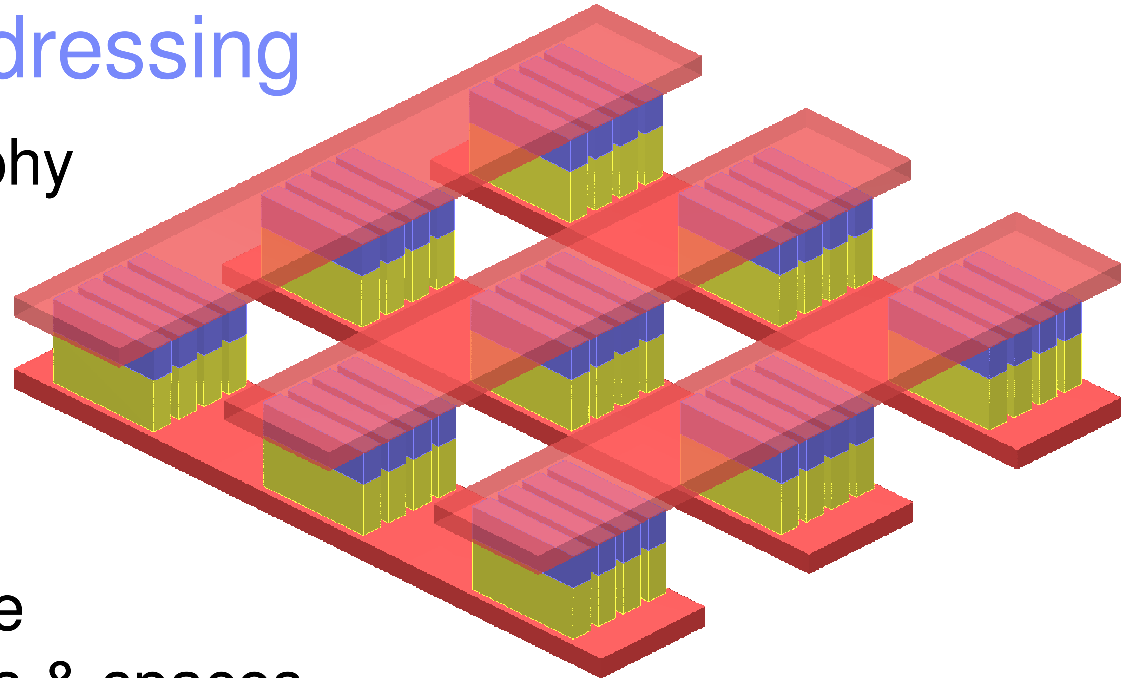
demonstrated (at IEDM 2007)

…go to 3-D with **L** <u>layers</u>



© 2008 IBM Corporation

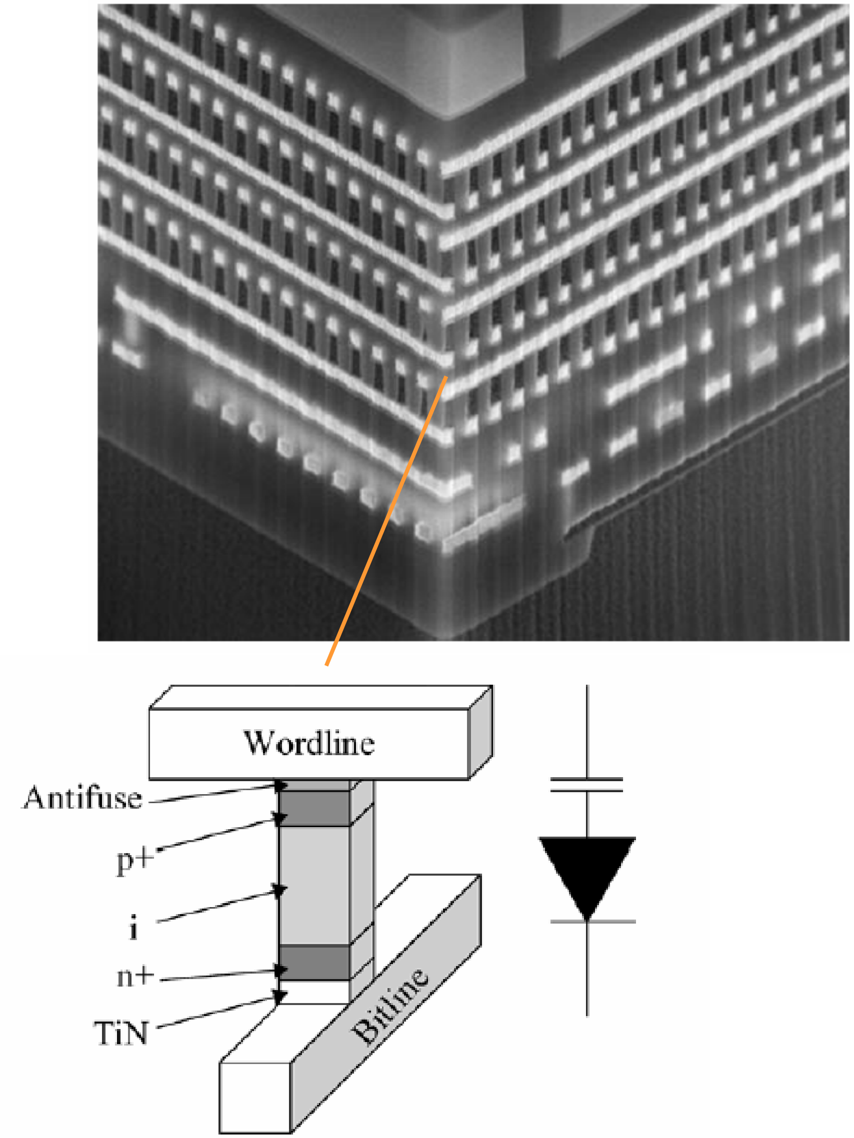
Sub-lithographic addressing



- Push beyond the lithography roadmap to pattern a dense memory
- But nano-pattern has more complexity than just lines & spaces
- Must find a scheme to connect the surrounding micro-circuitry to the dense nano-array

© 2008 IBM Corporation

3-D stacking

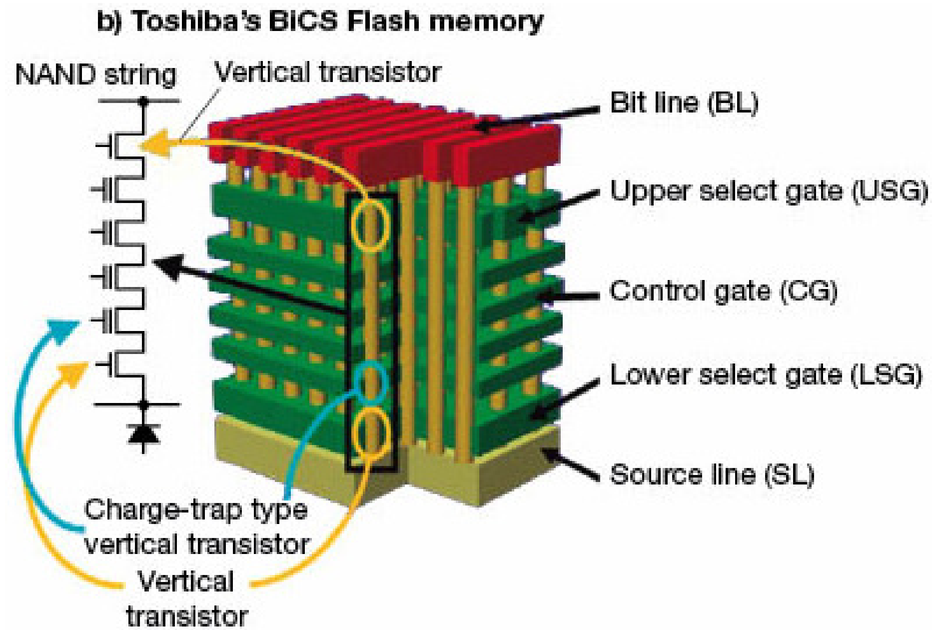


[Li:2004]

- 3-D anti-fuse
  (Matrix semiconductor)

[Tanaka:2007]

- 3-D Flash (Toshiba)



 © 2008 IBM Corporation

For more information (on ultra-high density)

G. W. Burr, B. N. Kurdi, J. C. Scott, C. H. Lam, K. Gopalakrishnan, and R. S. Shenoy, "An overview of candidate device technologies for Storage-Class Memory," to appear in *IBM Journal of Research and Development*, (2008).

- ITRS roadmap, *www.itrs.net*
- T. Nirschl, J. B. Philipp, et. al., *IEDM Technical Digest*, 17.5 (2007).
- K. Gopalakrishnan, R. S. Shenoy, et. al., *IEDM Technical Digest*, 471-474 (2005).
- F. Li, X. Y. Yang, et. al. *IEEE Trans. Dev. Materials Reliability*, **4**(3), 416-421 (2004).
- H. Tanaka, M. Kido, et. al., *Symp. VLSI Technology*, 14-15 (2007).

 © 2008 IBM Corporation

Technology conclusions

▪ Motivation

- by 2020, server-room power & space demands will be too high
- evolution of hard-disk drive (HDD) storage and Flash cannot help
- need a new technology – **Storage Class Memory (SCM)** – that combines
  - ❖ the benefits of a solid-state memory (**high performance** and **robustness**)
  - ❖ with the **archival capabilities** and **low cost** of conventional HDD

▪ How to build SCM

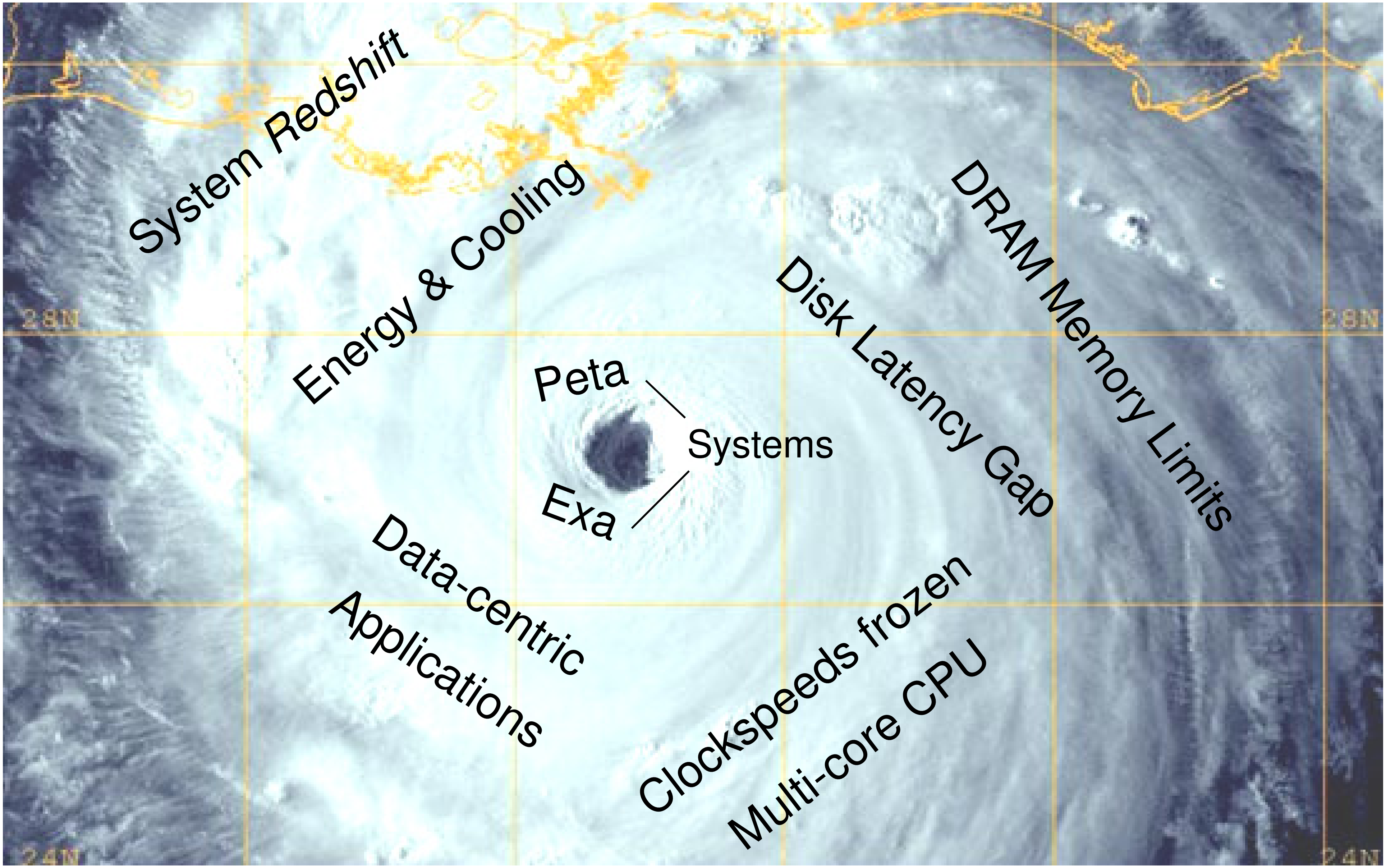
- combine a scalable non-volatile memory (**Phase-change memory**)
- with **ultra-high density** integration, using
  - ❖ micro-to-nano addressing
  - ❖ multi-level cells
  - ❖ 3-D stacking

▪ If you build it, they will come

- With its combination of **low-cost** and **high-performance**,
  SCM could impact much more than just the server-room...

© 2008 IBM Corporation

A Perfect Storm for Large Systems Architectures



Typical System Time Scale

Time [ns]

Time [ns]	Human scale		
10^{10}	century	Get data from TAPE (40s)	*Storage*
10^9	decade		
10^8	year		
10^7	month	Get data from DISK (5ms)	
10^6	week	Write to FLASH, random (1 ms)	
10^5	day		
10^4		Read a FLASH device (20 us)	
10^3	hour		*SCM*
10^2	minute	Get data from DRAM (60ns)	*Memory*
10		Get data from L2 cache (10ns)	
1	second	CPU operations (1ns)	

© 2007 IBM Corporation

Price/MB for DRAM-NAND FLASH- SCM - HDD

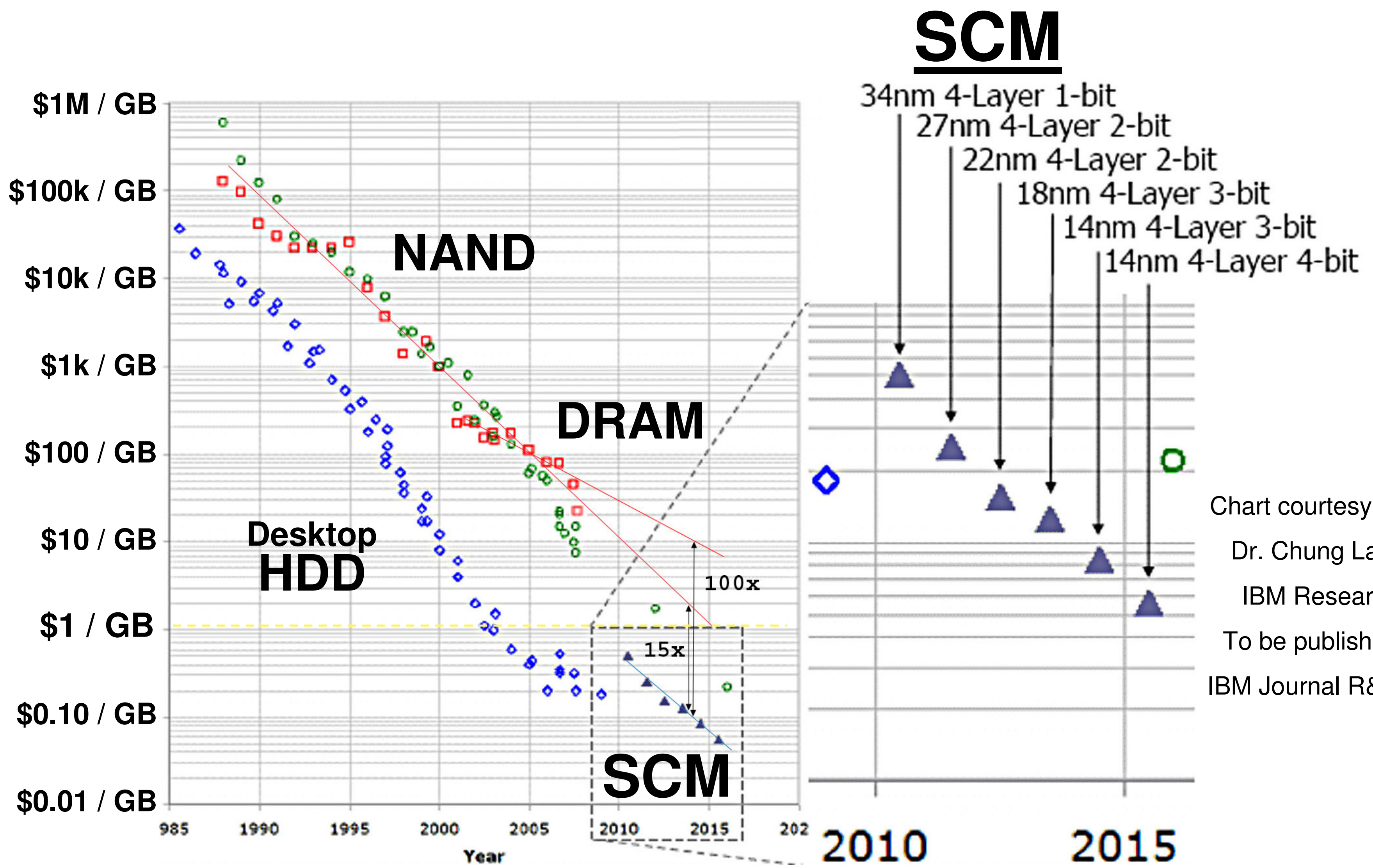


Chart courtesy of

Dr. Chung Lam

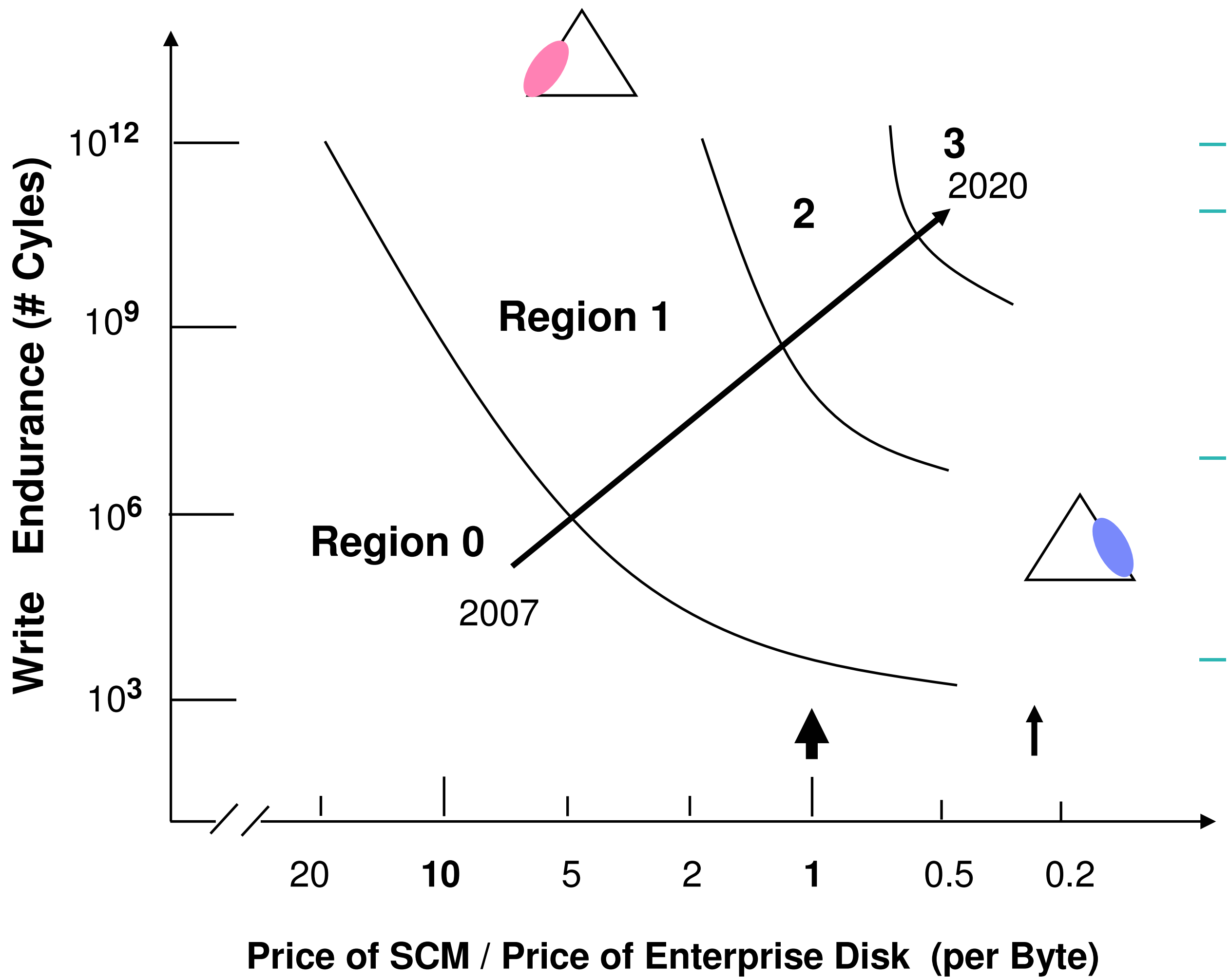
IBM Research

To be published

IBM Journal R&D

© 2007 IBM Corporation

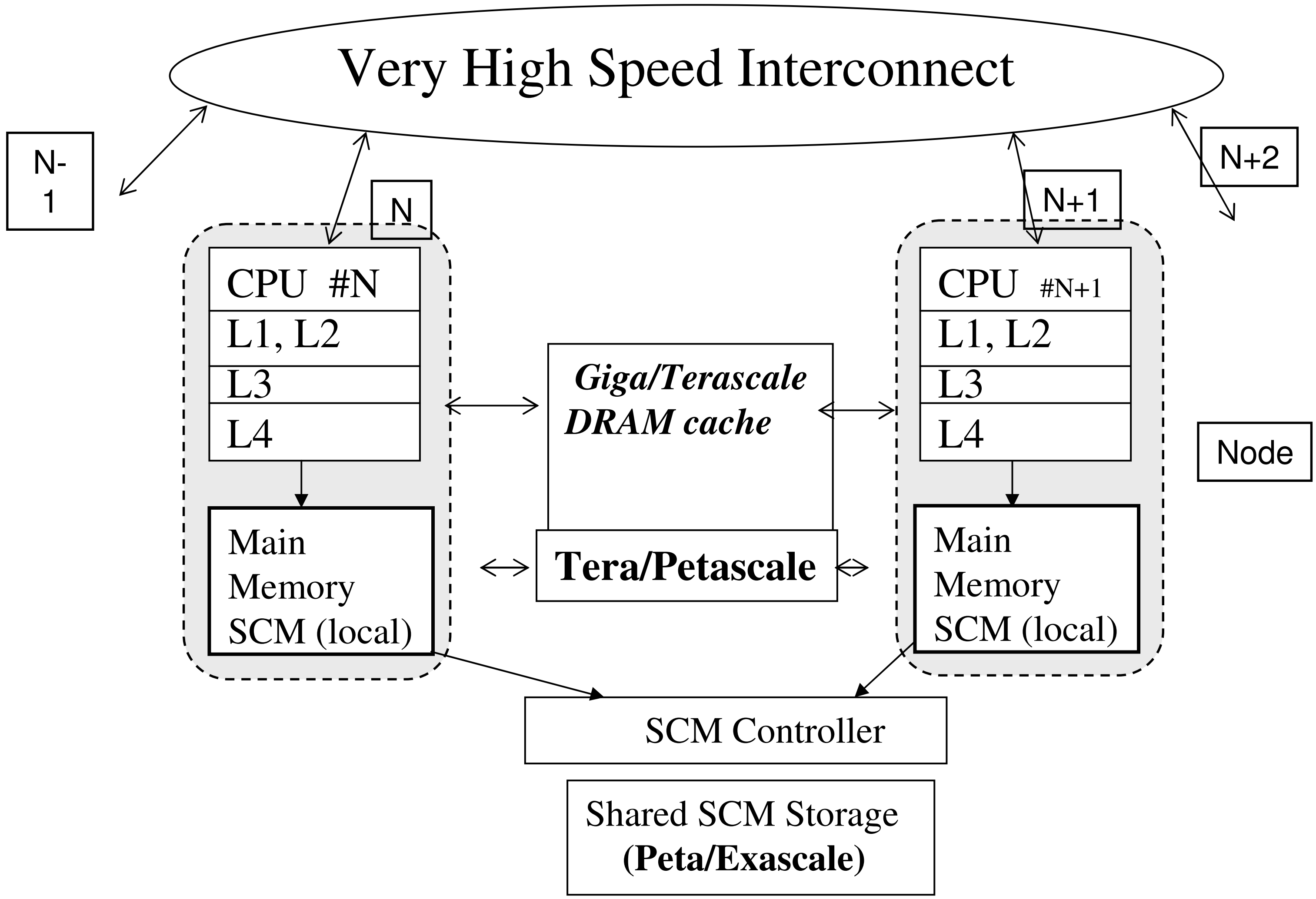
Opportunity Bands for SCM (very approximate)



# of Region
- 0: Rad-hard uses
- 1: Use as additional layers in memory hierarchy (for performance)
- 2: Replace **enterprise disks** and supplement FLASH
- 3: Replace commodity disks and FLASH, too

© 2007 IBM Corporation

Peta-scale System Diagram (to Exa-scale by 2015)

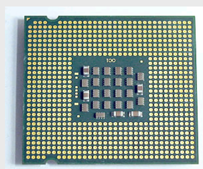


© 2007 IBM Corporation

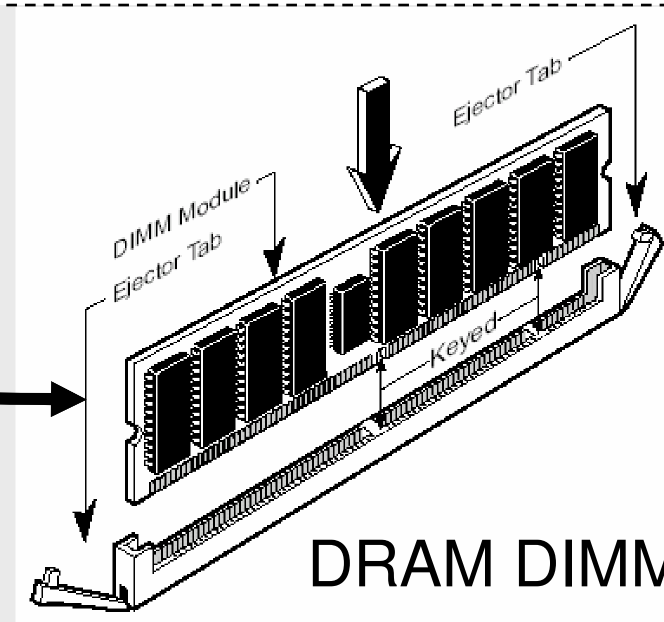
CPU & Memory System (Node) in 2008

Logical Address > VM Translation > Physical Address

X86, PowerPC, SPARC…



CPU	VM	PMC



DRAM DIMM

Central Processing Unit

Virtual Memory Translation Unit (Page Tables, TLB, etc)

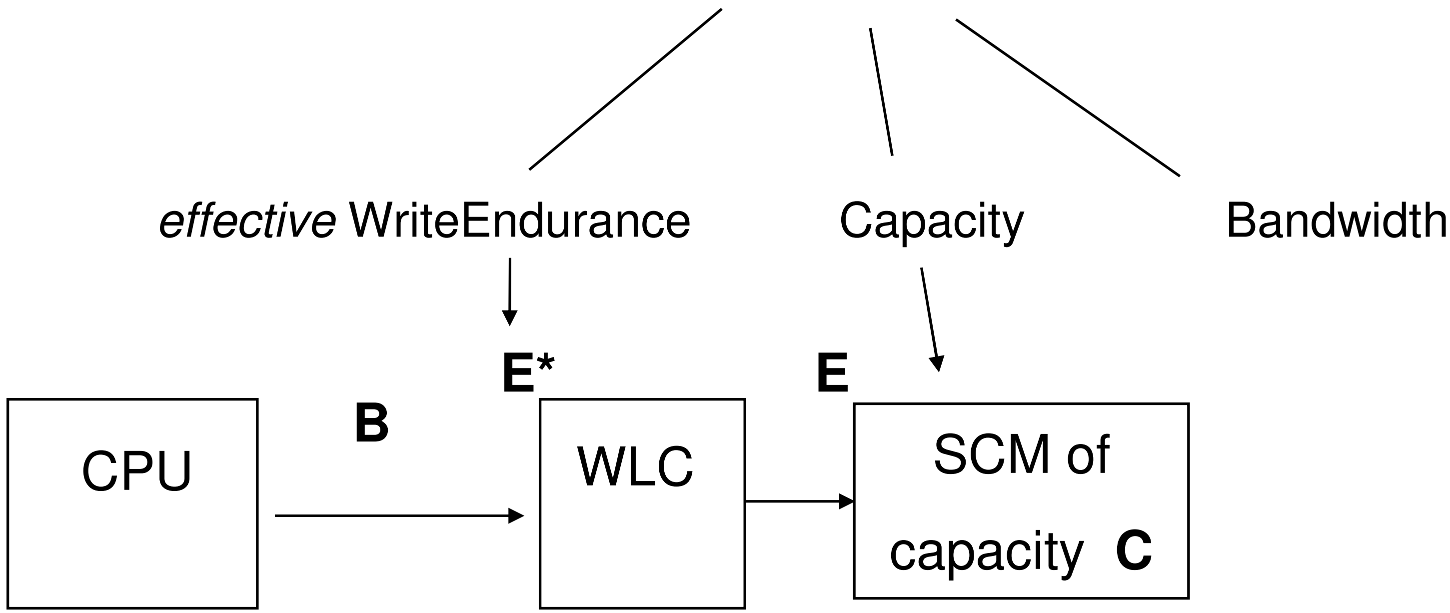
Physical Memory Controller

© 2007 IBM Corporation

Wear Level Control (WLC) is essential

Without it, system can die in seconds. With it, it lives for years

$$T_{life} = E^{*} \bullet T_{fill} = E^{*} \bullet C/B$$



© 2007 IBM Corporation

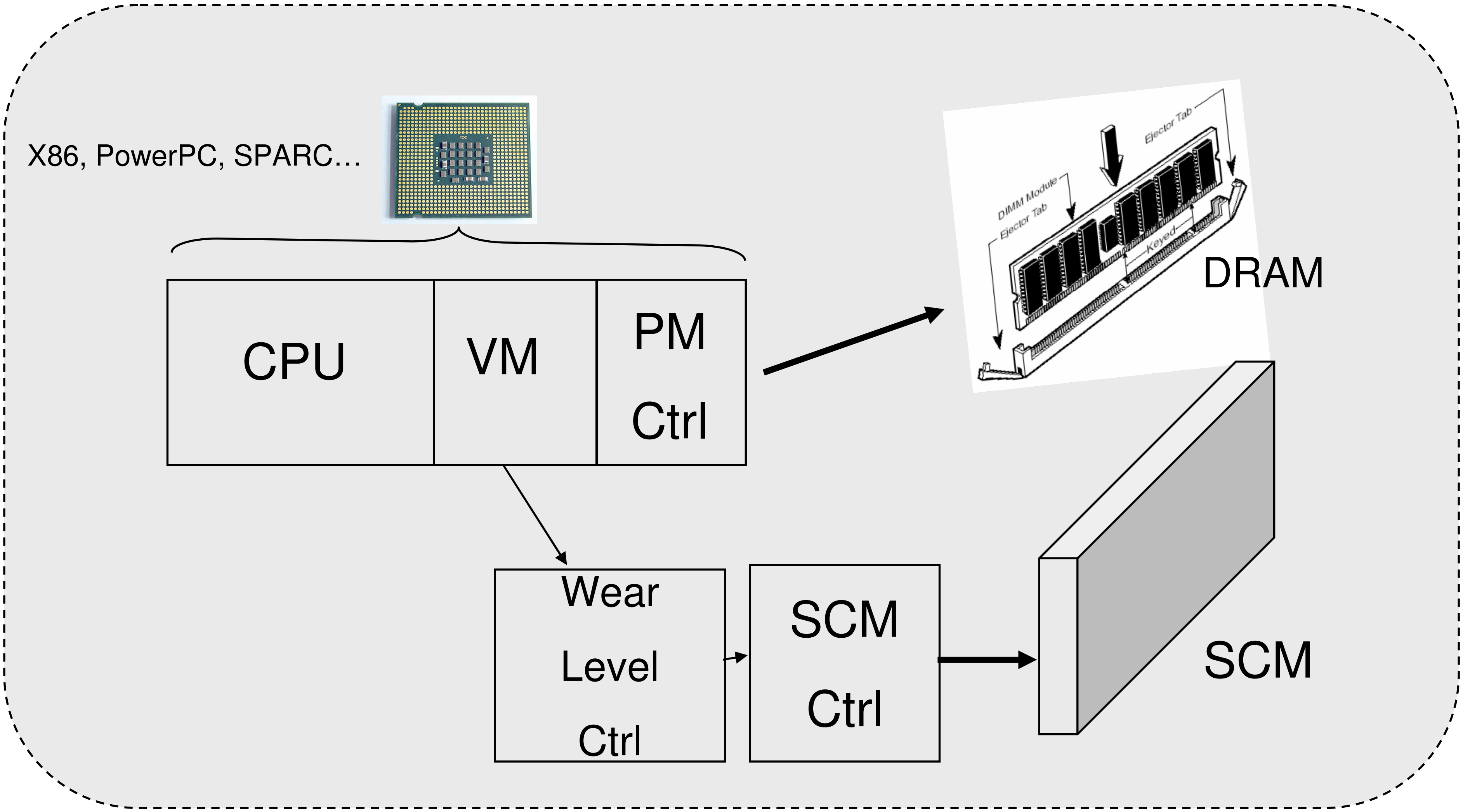
SCM-based Memory System

Logical Address > VM-Translation > WL-Translation > SCM Physical Add

- **Treat WL as part of address translation flow**
  - Option a – Separate WL/SCM controller
  - Option b - Integrated VM/WL/SCM controller
  - Option c - Software WL/Control
- **Also need physical controller for SCM**
  - Different from DRAM physical controller

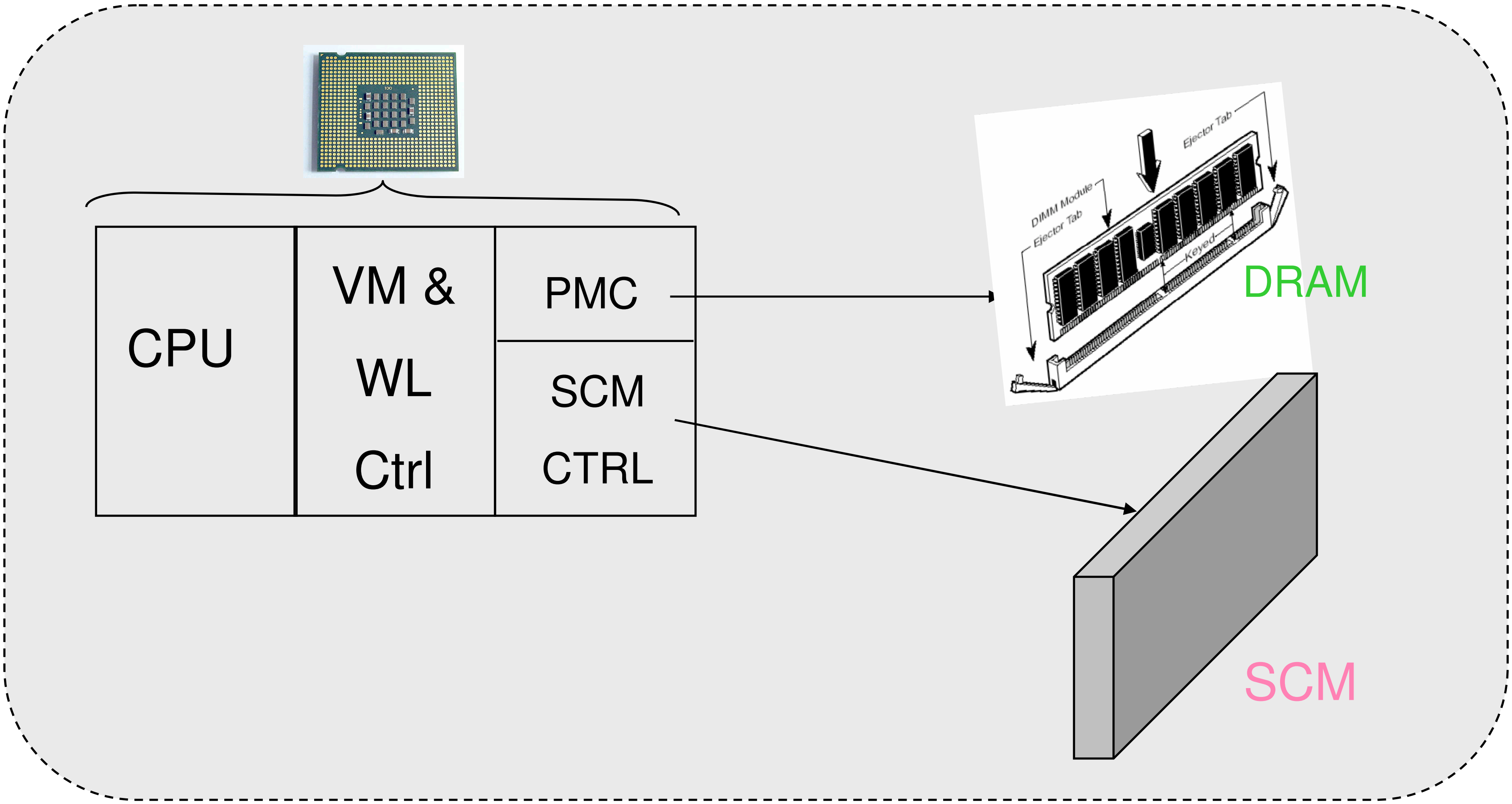
© 2007 IBM Corporation

Separate WL/SCM Controller



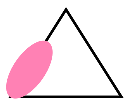
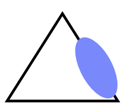
© 2007 IBM Corporation

Integrated VM/WLC System



© 2007 IBM Corporation

Uses of SCM in overall memory/storage stack

Access Mode	Use Mode	Comments
Address oriented (Memory-like) 	Cache (e.g. Level 4)	Wear level too high?
	Main memory - version (a)	Separate WL/SCM controller
	Main memory - version (b)	Integrated WL/SCM/RAM controller
	Main memory - version (c)	SCM Wear level managed by software & VM manager <u>(dangerous)</u>
Block oriented (Storage-like) 	Via legacy I/O busses	Easy, but wastes SCM performance
	Via new interfaces	Good for memory mapping use model
	Paging Device	Very promising use
	I/O Cache and/or meta-date storage for a disk controller	Act as NVRAM, good use

© 2007 IBM Corporation

Implications on Traditional Commercial Databases

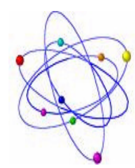
JOHN	DOE	49	NYC
FRANK	DOHERTY	67	NYC
JAMES	DUNDEE	36	SYDNEY

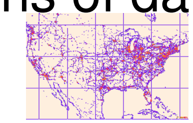
- **Initial SCM in DB uses:**
  - Logging (for Durability)
  - Buffer pool
- **Long term, deep Impact: Random access replaces paging**
  - DB performance depends heavily on good guesses what to page in
  - Random access eliminates column/row access tradeoffs
  - Reduces energy consumption (big effect)
- **Existing trend is to replace 'update in place' with 'appends'**
  - that's good – helps with write endurance issue
- **Reduce *variability* of data mining response times**
  - from hours and days (today) to seconds (SCM)

© 2007 IBM Corporation

‘Data-centric’ High Performance Applications

	▪ ***Compute-centric paradigm***		▪ ***Data-centric paradigm***
Focus on:	– solving diff. equations	→	– analyzing mountains of data
Bottleneck:	– CPU/Memory	→	– Storage & I/O
Examples:	Comp Fluid Dynamics, Finite Element Analysis Multibody Simulations Protein Folding ….		Search & text analysis (Google…) Graph Analysis (human networks) Video/Image Processing & Analysis Environmental & economic modeling Genetics Climate Modeling





Problem *Disks can’t keep up w/data centric applications*

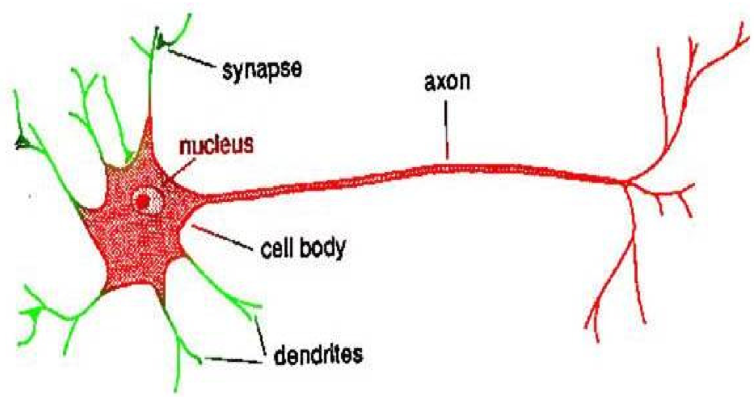
- *@ current trends: Million disks per HPC system in 2020!*

Solution *Need new technology for memory/Storage => SCM*

© 2007 IBM Corporation

Human-Scale Brain Simulation with SCM

- **Challenges are storing of synaptic weights (and network)**
- **Mouse & rat scale models today ~ 55 Million Neurons**
  - Almaden C2 Simulator (Supercomputing 2007)
- **Human Brain ~ 20 Billion Neurons @ 8000 Synapses each**
  - $1.6 * 10^{14}$ Synapses @ 16 Bytes / Synapse (C2 Sim.)
  - 2.5 PetaBytes of synaptic state
  - Ideal application for SCM
    - Random addressing, slowly varying
      - <1 Hz> => no write endurance issues



- **My prediction: Human Scale Brain Simulation by ~ 2017**

© 2007 IBM Corporation