IBM Almaden Research Center

Storage Class Memory: Technology, Systems and Applications

Rich Freitas

freitas@almaden.ibm.com

Nonvolatile Memory Seminar
Hot Chips Conference
August 22, 2010
Memorial Auditorium, Stanford University

Agenda

- **Technology**
  - Disks
  - Flash
  - Phase Change Memory
- **Systems**
  - Memory Systems
  - Storage Systems
- **Applications**

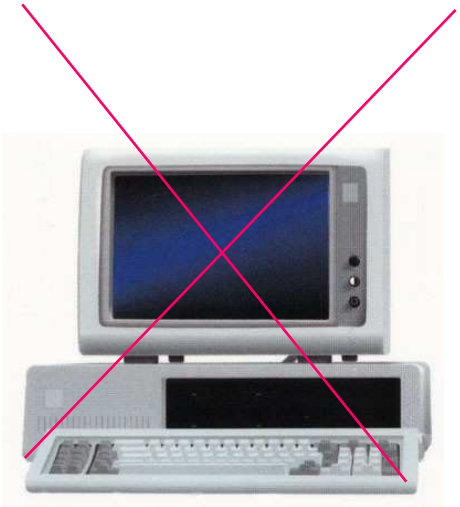
Definition of **S**torage **C**lass **M**emory **SCM**

- **A new class of data storage/memory devices**
  - many technologies compete to be the ‘best’ SCM
- **SCM features:**
  - Non-volatile
  - Short Access times (~ DRAM like)
  - Low cost per bit (more DISK like – by 2020)
  - Solid state, no moving parts
- **SCM *blurs the distinction* between**
  - MEMORY (*fast, expensive, volatile*) and
  - STORAGE (*slow, cheap, non-volatile*)

System Targets for SCM

Billions!

Mobile ✓

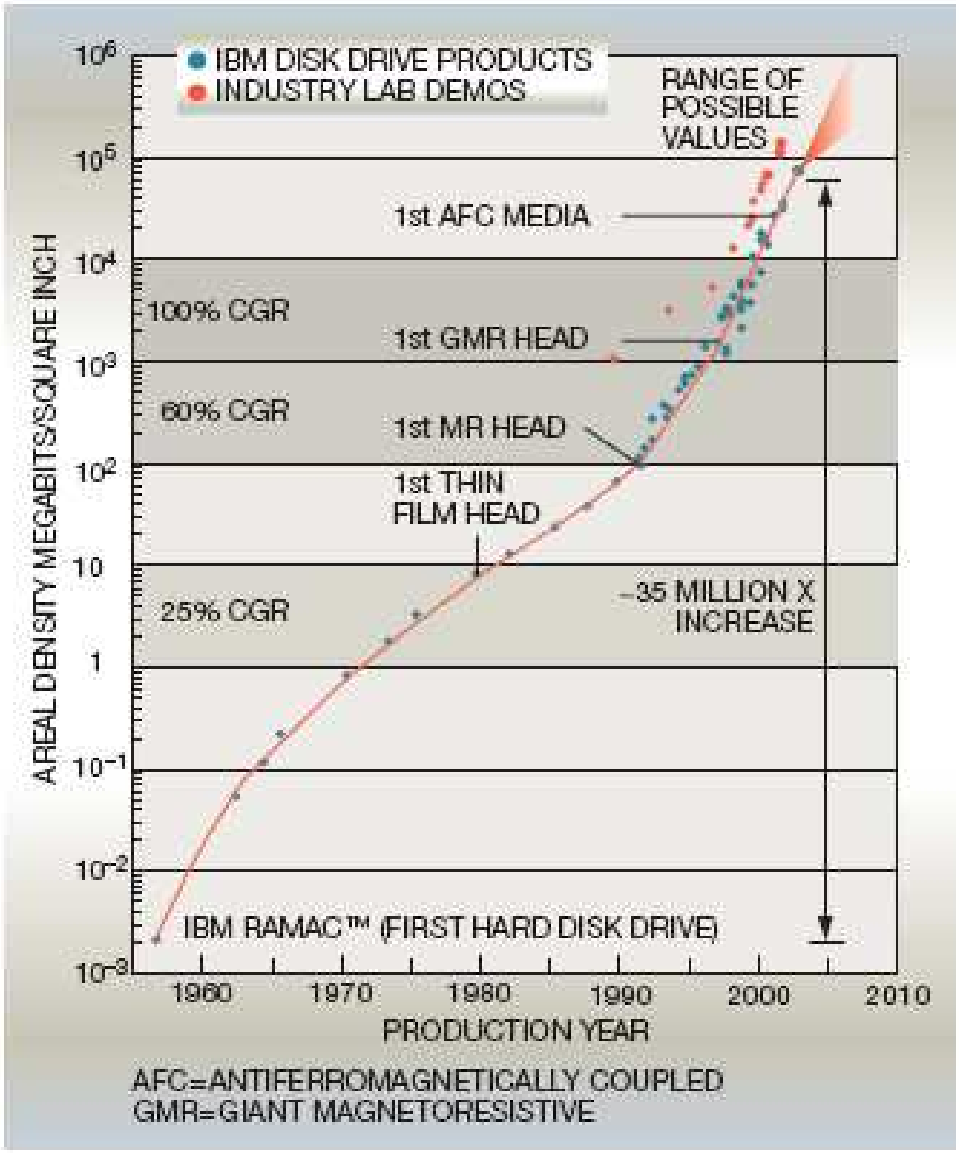


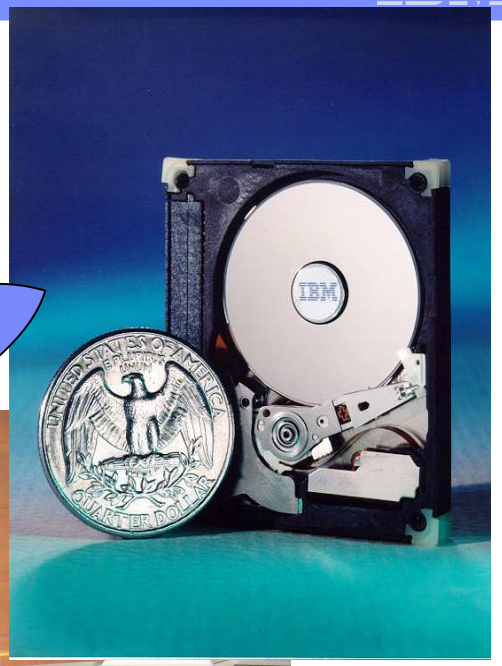
Desktop X

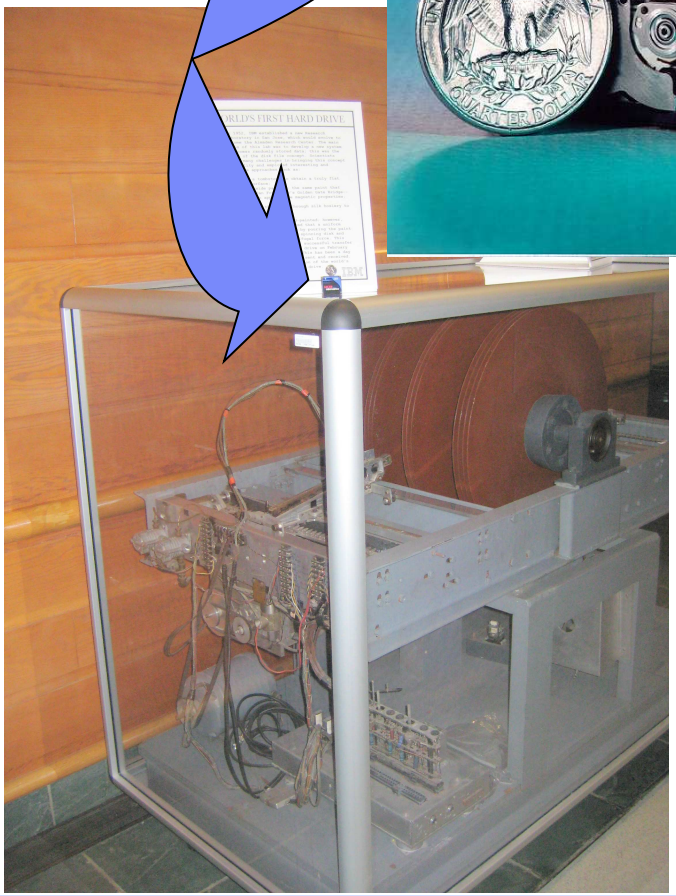
Megacenters

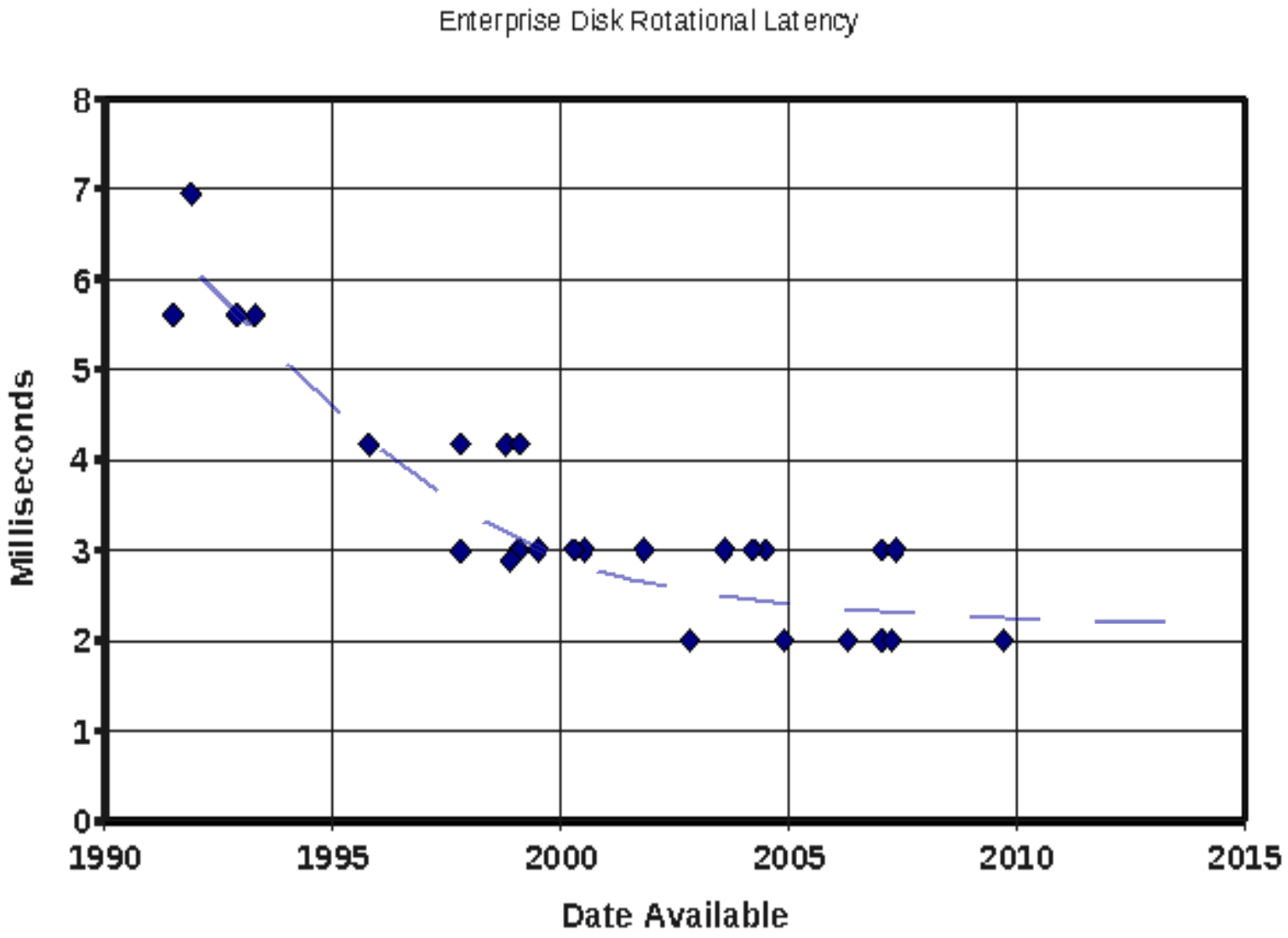
Datacenter ✓

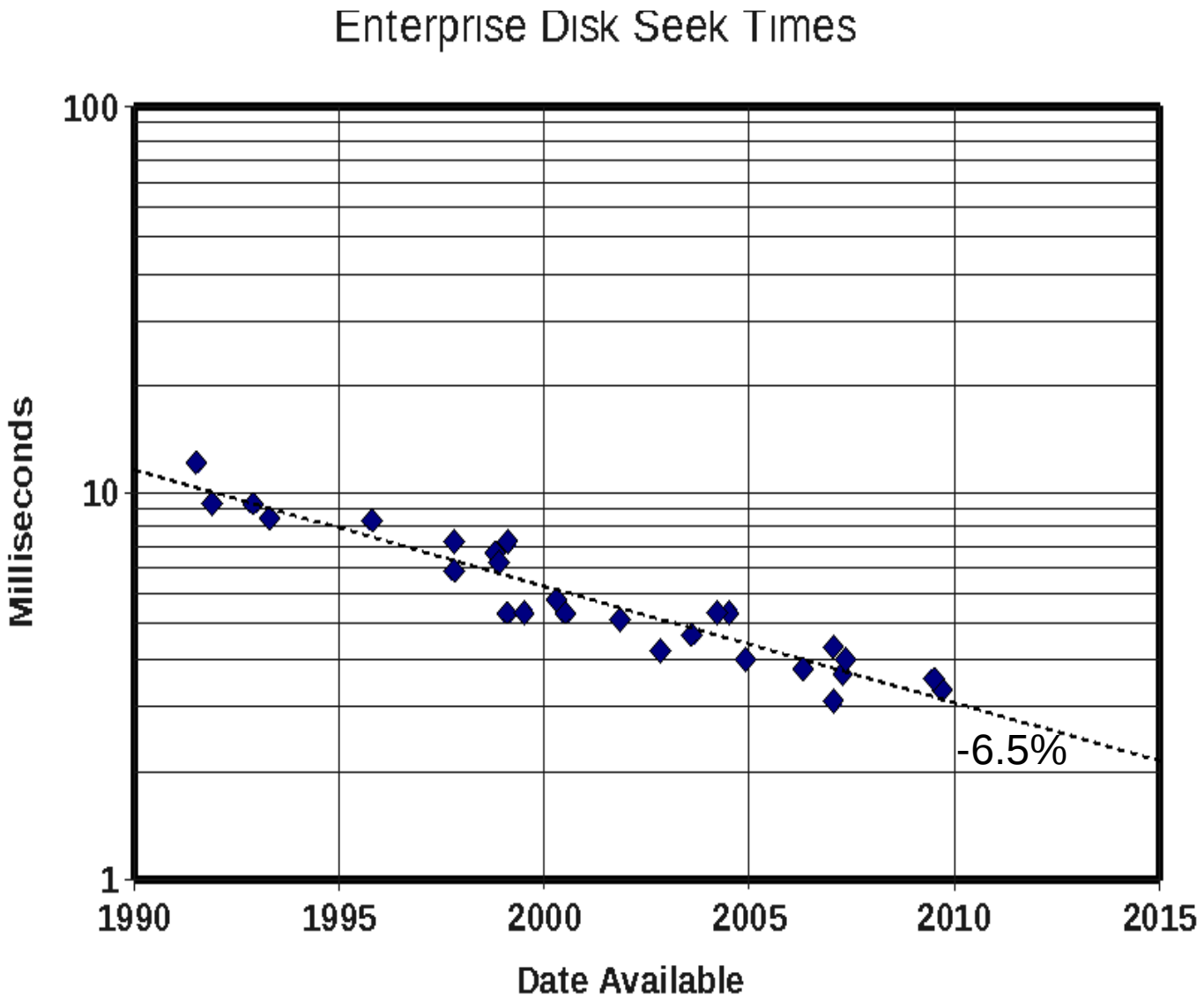
History of HDD is based on Areal Density Growth

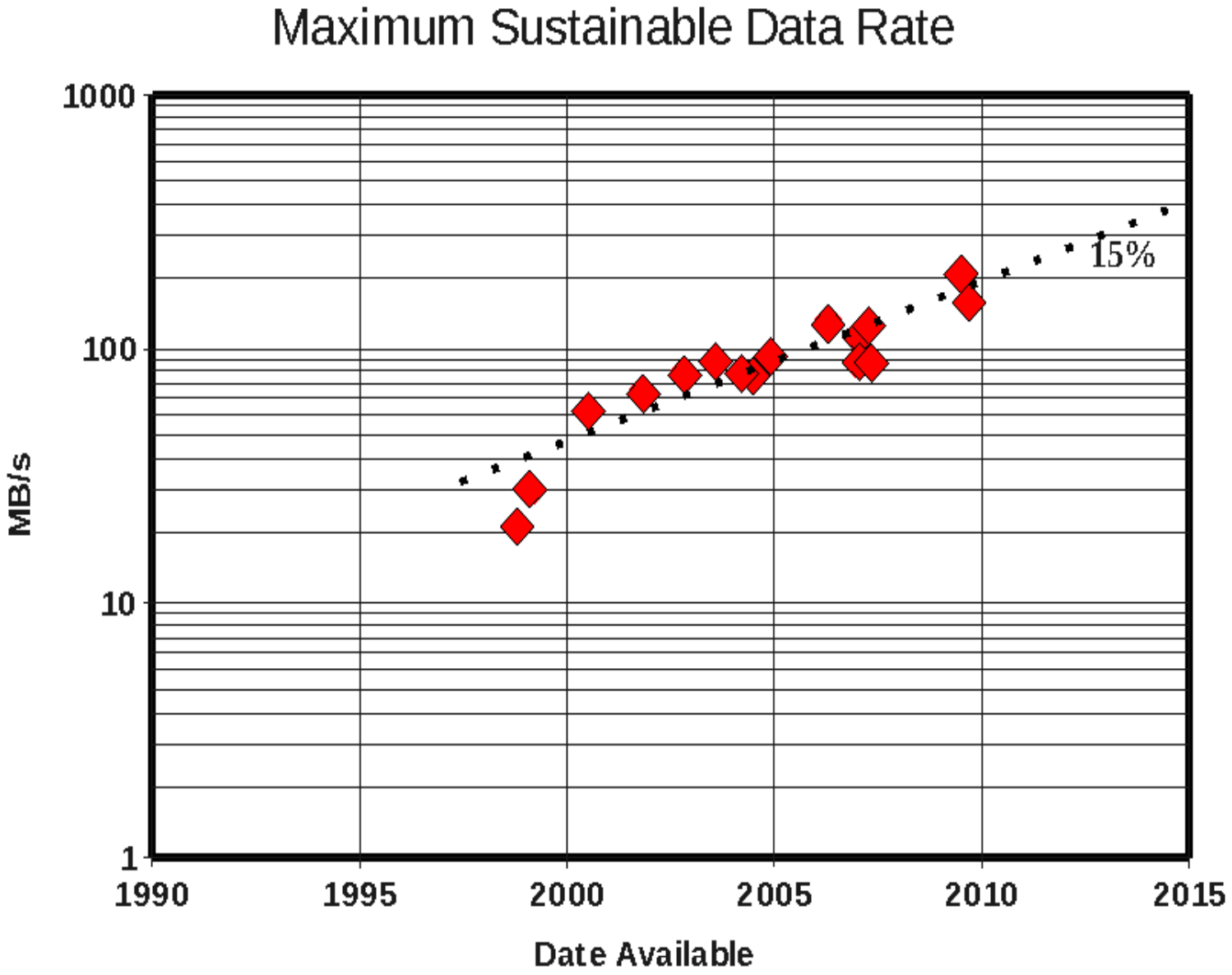






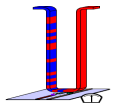


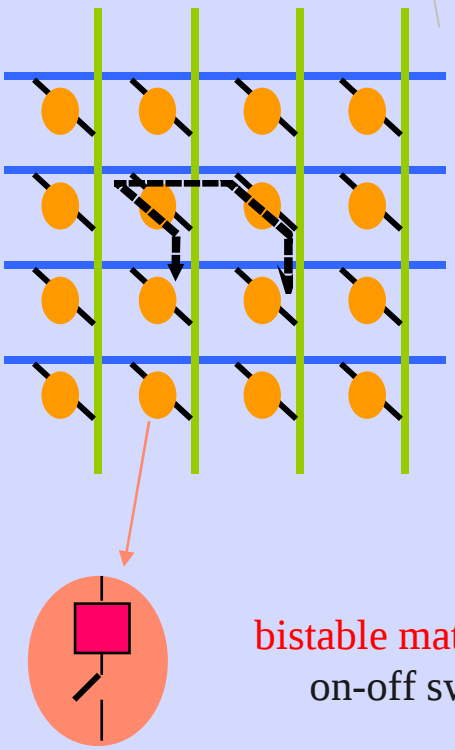




Many device technologies considered for SCM

- Phase Change RAM
  - most promising now (scaling)
- Magnetic RAM
  - used today, but poor scaling and a space hog
- Magnetic Racetrack
  - basic research, but very promising long term
- Ferroelectric RAM
  - used today, but poor scaleability
- Solid Electrolyte and resistive RAM (Memristor)
  - early development, promising
- Organic, nano particle and polymeric RAM
  - many different devices in this class, unlikely
- Improved FLASH
  - still slow and poor write endurance

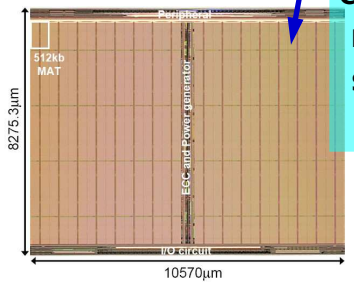




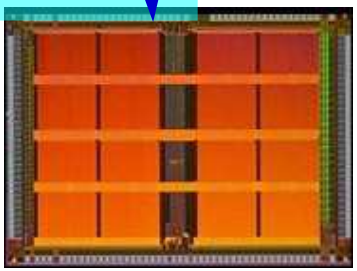
Generic SCM Array

Emerging Memory Technologies

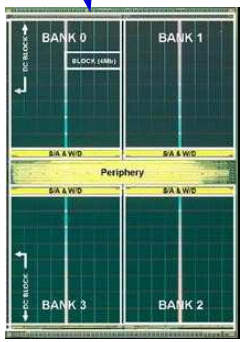
FLASH Extension	FRAM	MRAM	PCRAM	RRAM	Solid Electrolyte	Polymer/ Organic
Trap Storage	Ramtron	IBM	Ovonyx	IBM	Axon	Spansion
Saifun *NROM*	Fujitsu	Infineon	BAE	Sharp	Infineon	Samsung
Tower	STMicro	Freescale	Intel	Unity		TFE
Spansion	TI	Philips	STMicro	Spansion		MEC
Infineon	Toshiba	STMicro	Samsung	Samsung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



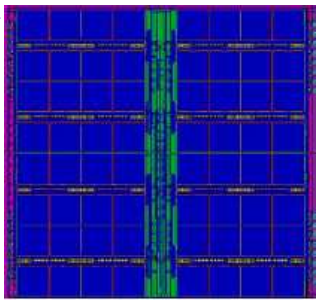
64Mb FRAM (Prototype)
0.13um 3.3V



4Mb MRAM (Product)
0.18um 3.3V



512Mb PRAM (Prototype)
0.1um 1.8V

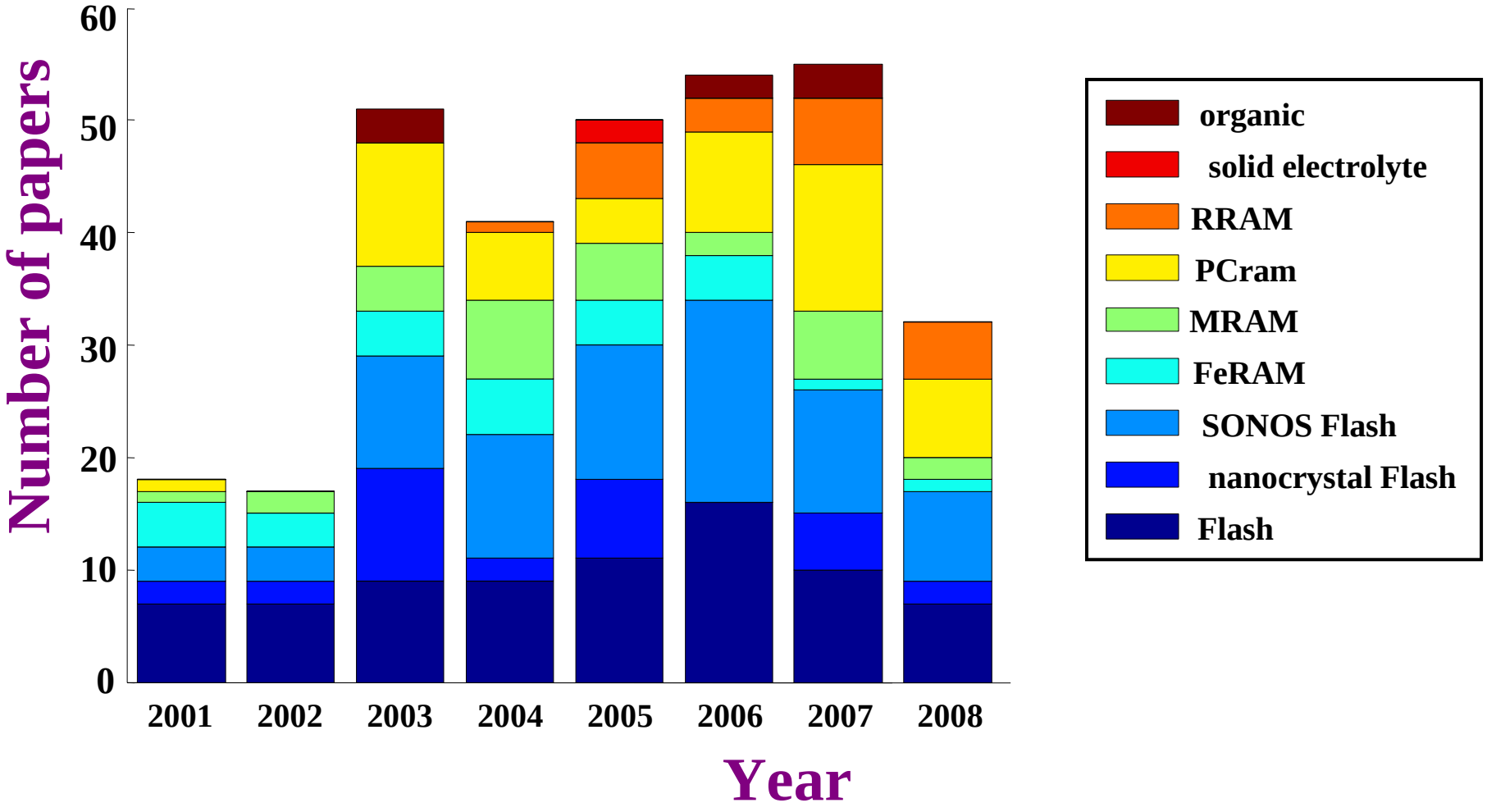


4Mb C-RAM (Product)
0.25um 3.3V

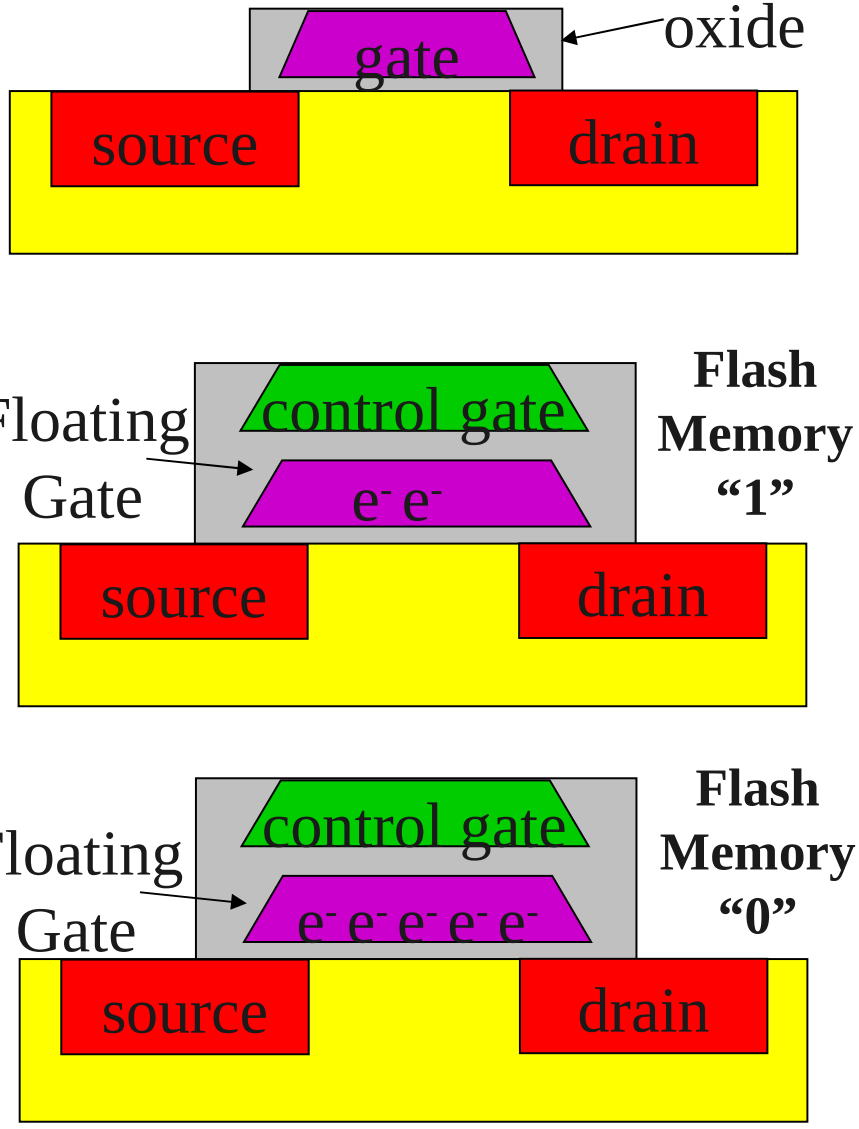
Research interest

Papers presented at

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



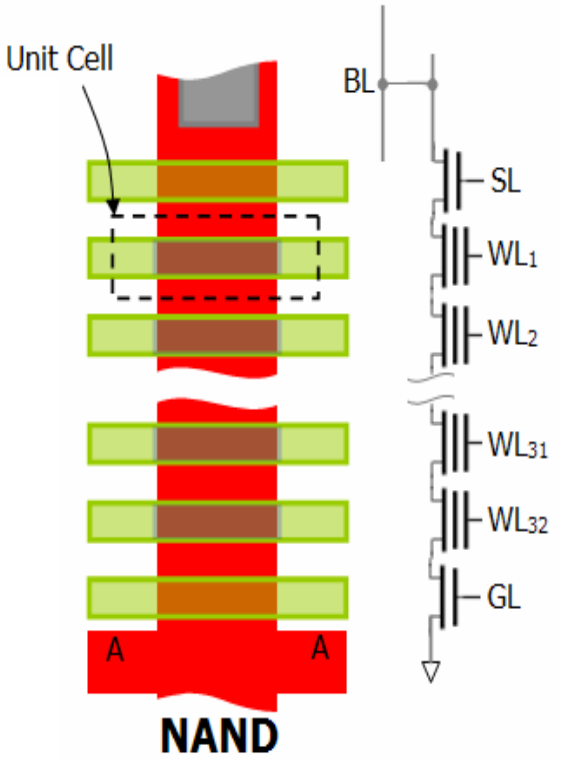
What is Flash?



- 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.

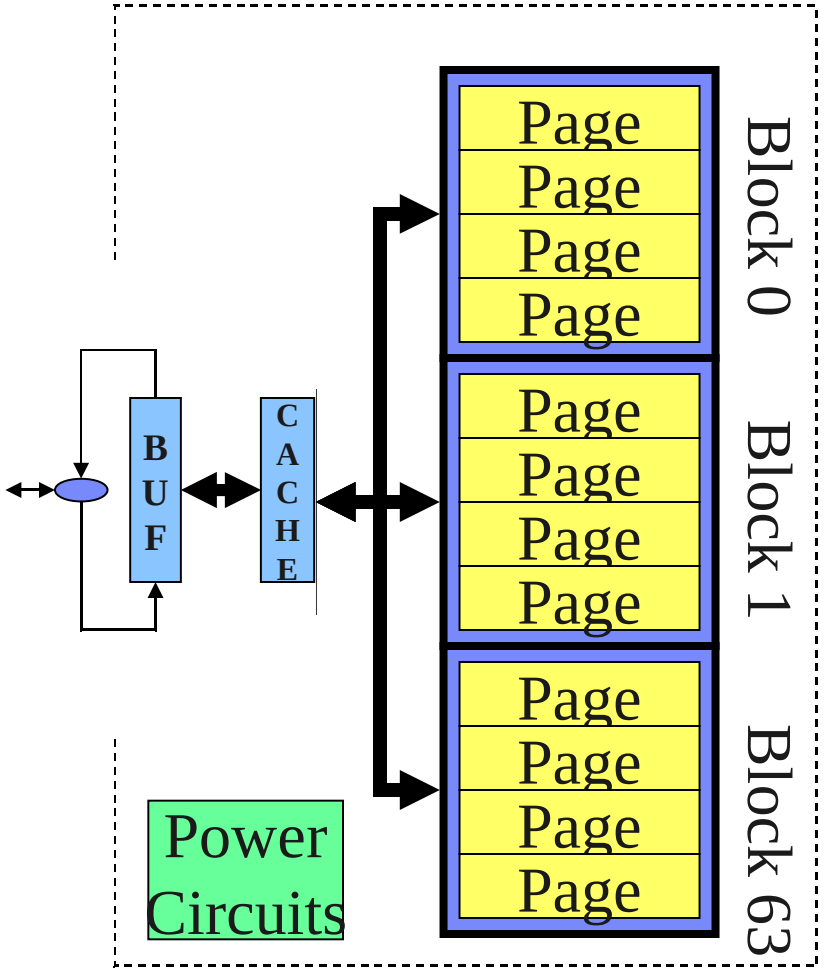
Feeds and Speeds for typical NAND Flash

	NAND
Cell Size	4 F^2 (2 F^2 virtual x 2-bit MLC)
Read Access Time	20-50 us
Read	15-25 MB/s
Write	5-8MB/sec
Erase	2ms
Start Up Time	50-100 us
Market Size (2007)	$14.2B
Applications	Multimedia



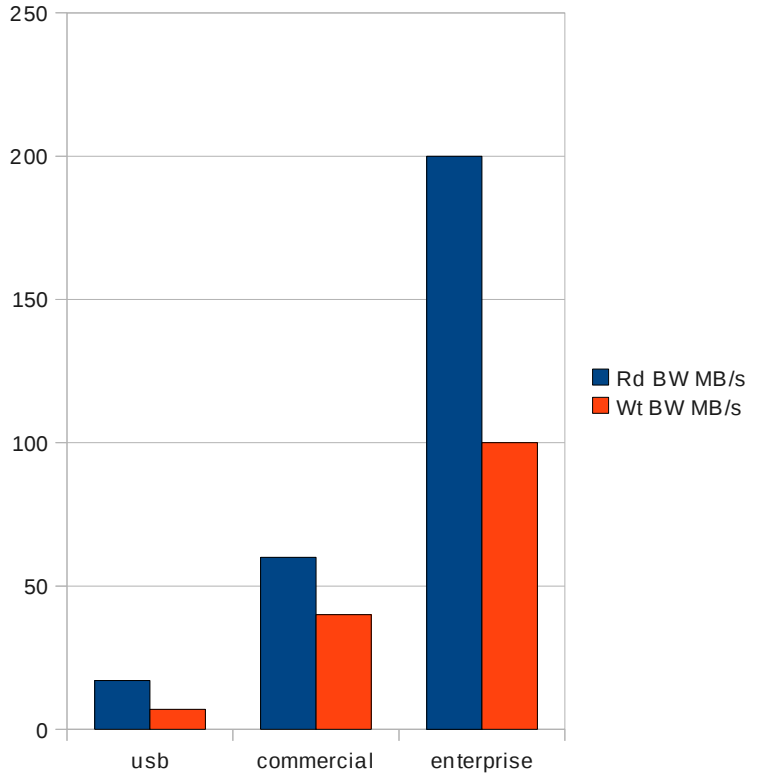
Representative NAND Flash Device

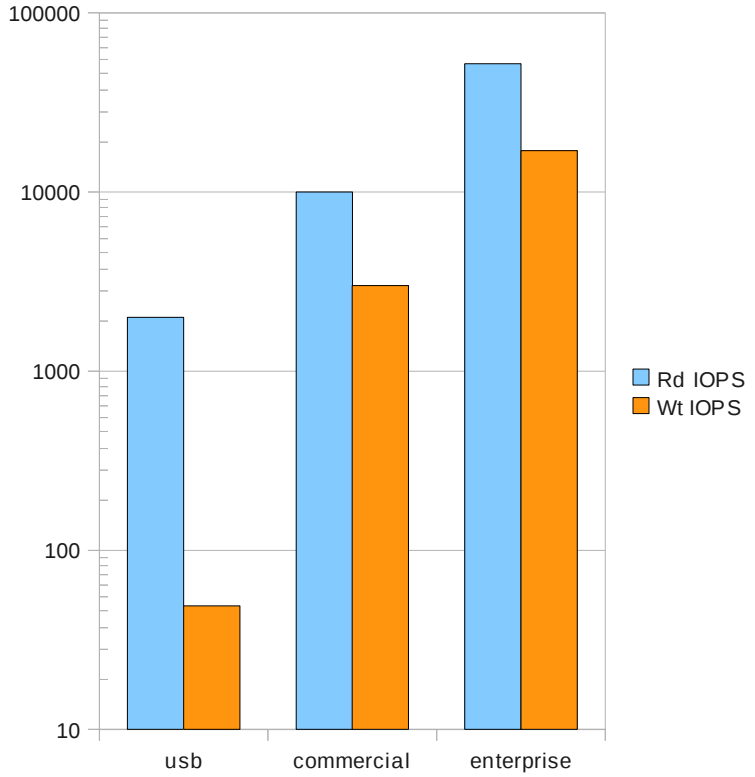
- **Interface: one or two bytes wide**
  - Transition to ONFI for some vendors
- **Data accessed in pages**
  - **2112, 4224 or 8448 Bytes**
- **Data erased in blocks**
  - Block = 64 - 128 Pages
- **Power circuits**
  - Charge Pumps
  - Clock drivers
  - Etc.



ONFI → Open NAND Flash Interface

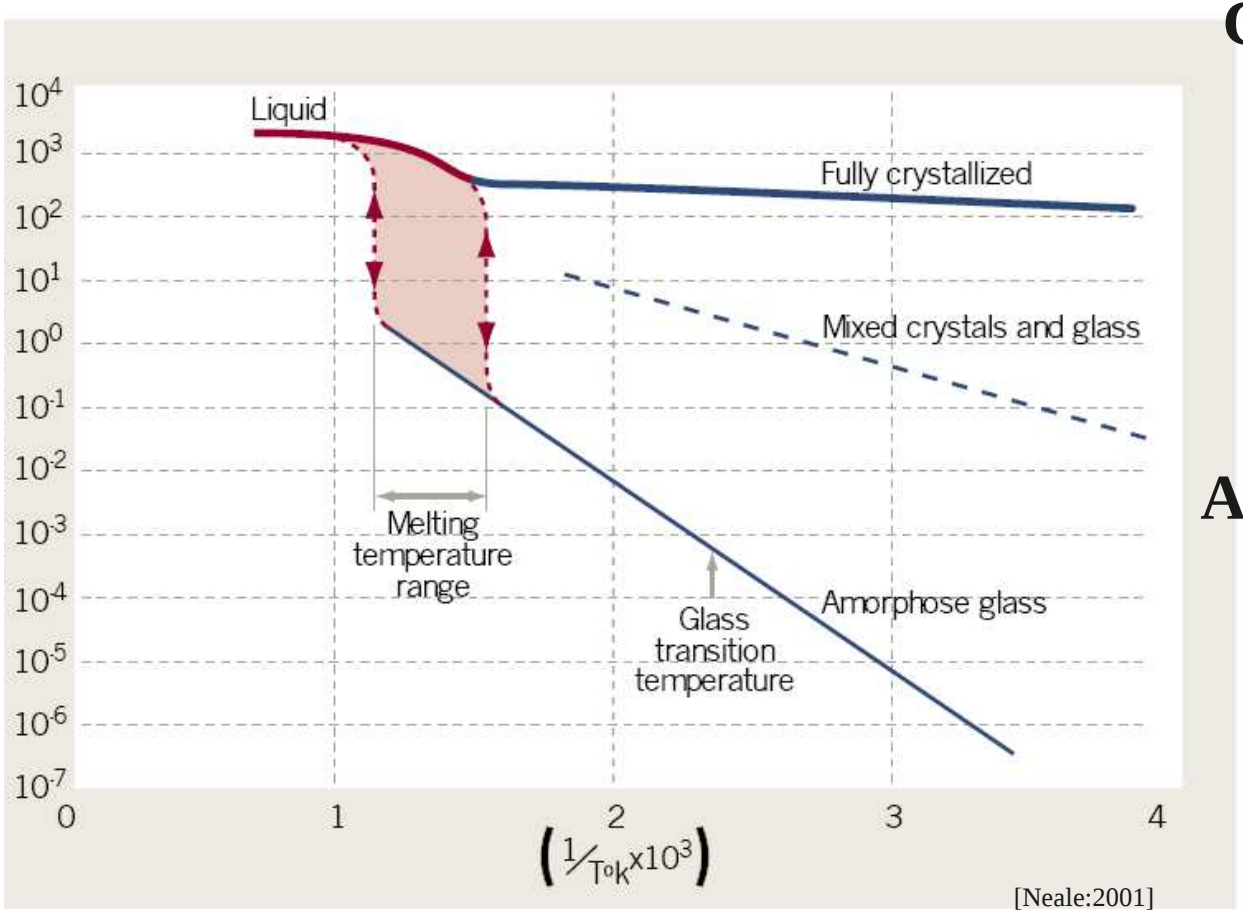
Representative Flash SSD Classes





History of Phase-change memory

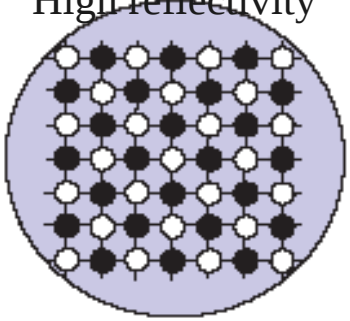
- late 1960's – Ovshinsky shows reversible electrical switching in disordered semiconductors
- early 1970's – much research on mechanisms, but everything was too slow!



[Neale:2001]

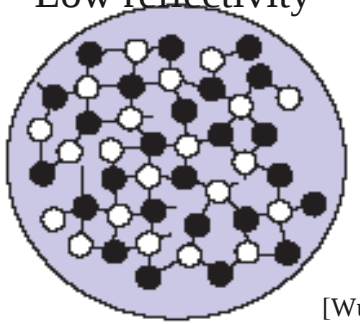
Crystalline phase

Low resistance
High reflectivity



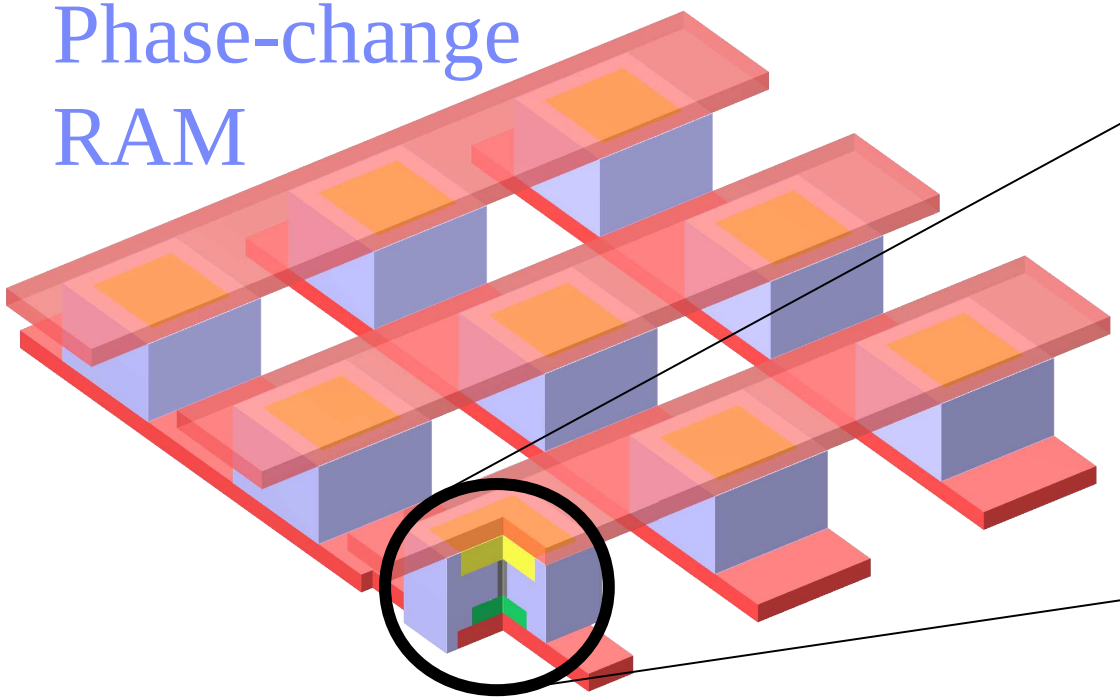
Amorphous phase

High resistance
Low reflectivity



[Wuttig:2007]

Phase-change RAM

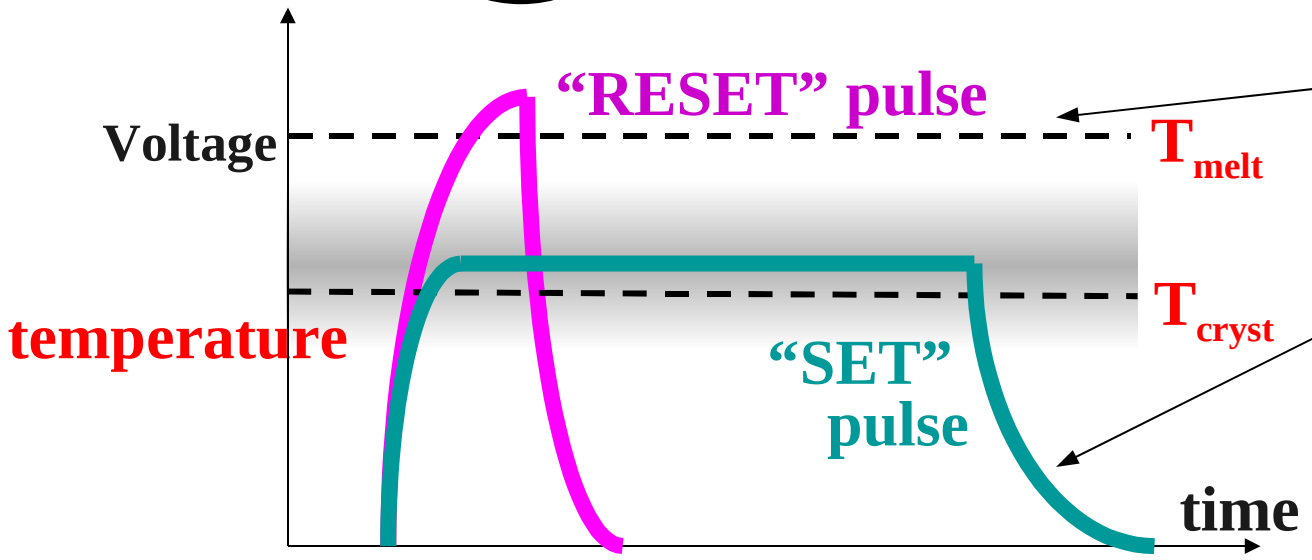


Bit-line

PCRAM
"programmable resistor"

Word-line

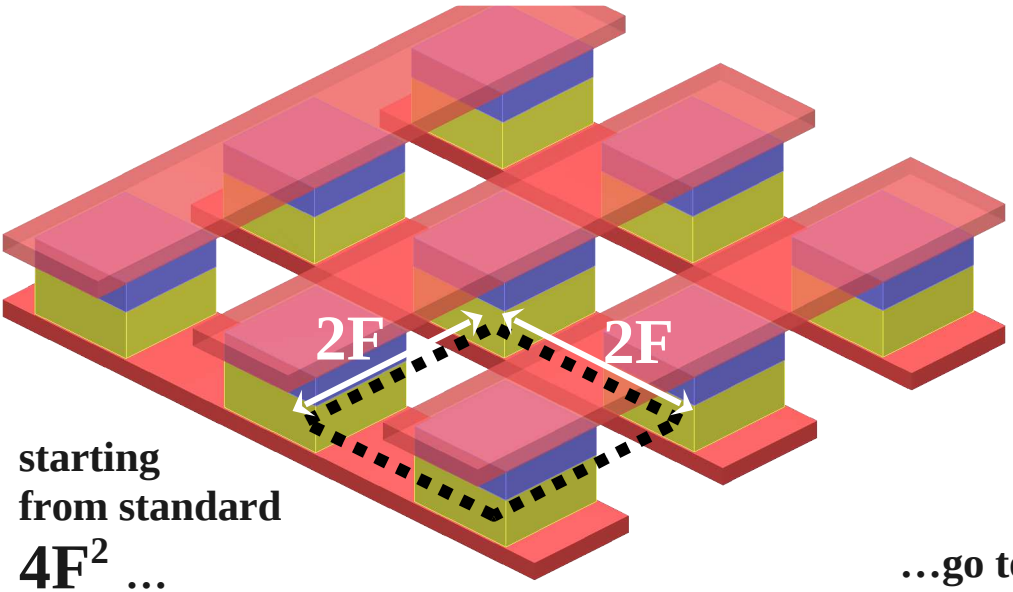
Access device
(transistor, diode)



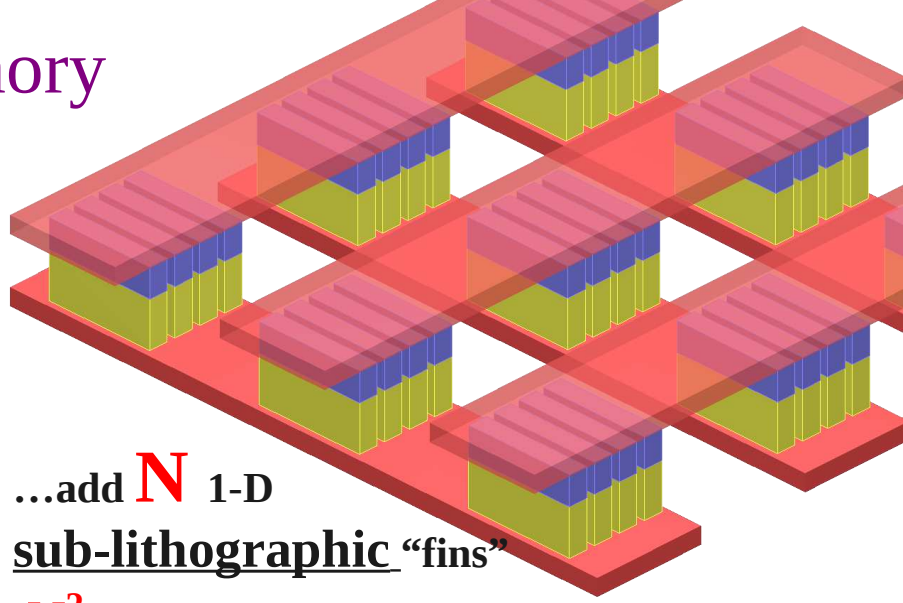
Potential headache:
High power/current
→ affects scaling!

Potential headache:
If crystallization is slow
→ affects performance!

Paths to ultra-high density memory



starting from standard $4F^2$ …

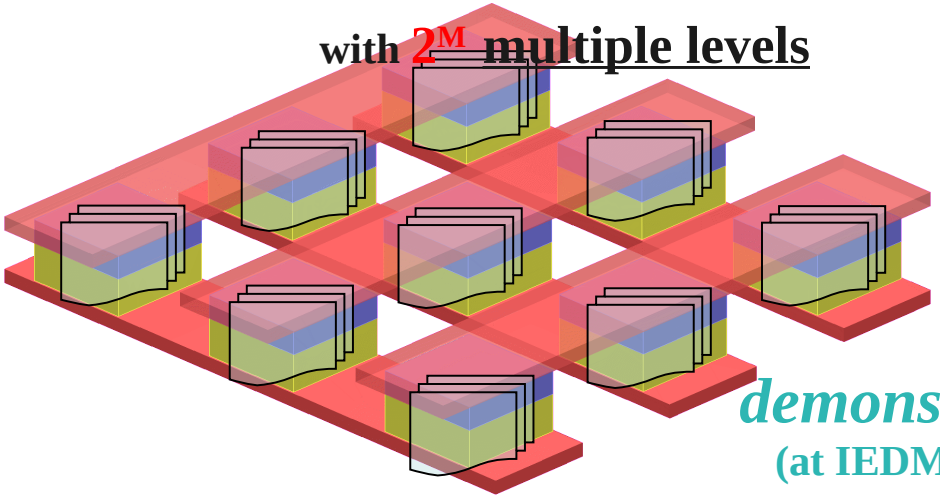


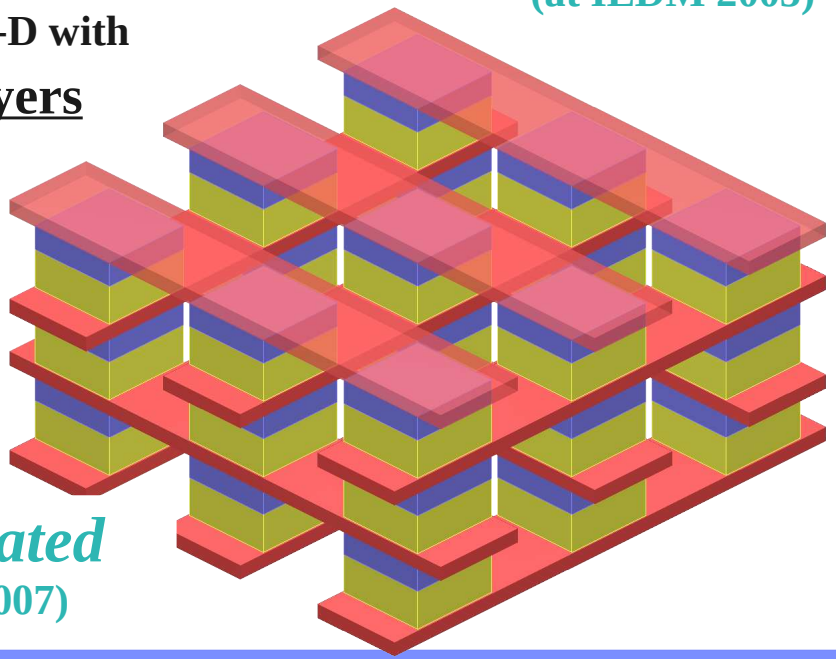
…add **N** 1-D **sub-lithographic** "fins" (N^2 with 2-D)

demonstrated
(at IEDM 2005)

…go to 3-D with **L layers**

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

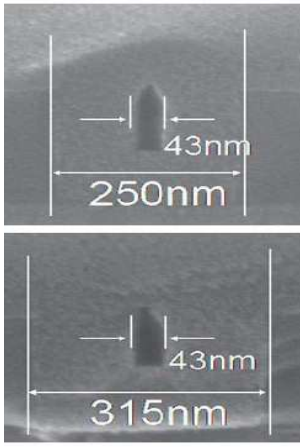




demonstrated
(at IEDM 2007)

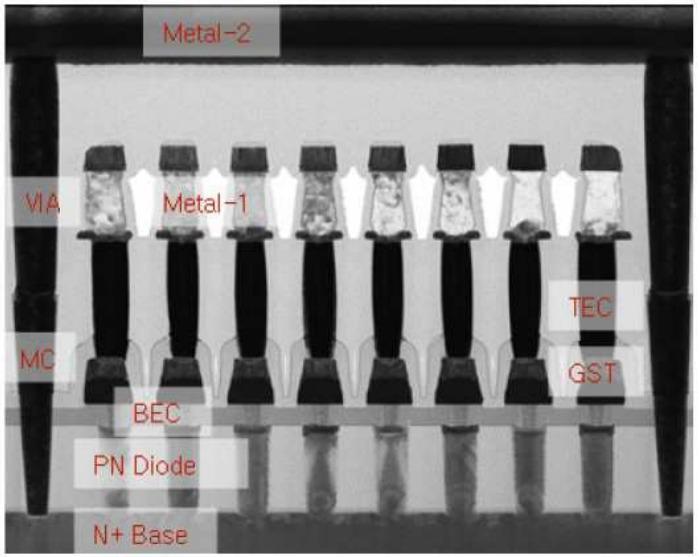
Industry SCM activities

- **SCM research in IBM**
- **Intel/ST-Microelectronics spun out Numonyx (FLASH & PCM)**
- **Samsung, Numonyx sample PCM chips**
- **Over 30 companies work on SCM**
  - including all major IT players



IBM sub-litho PCM

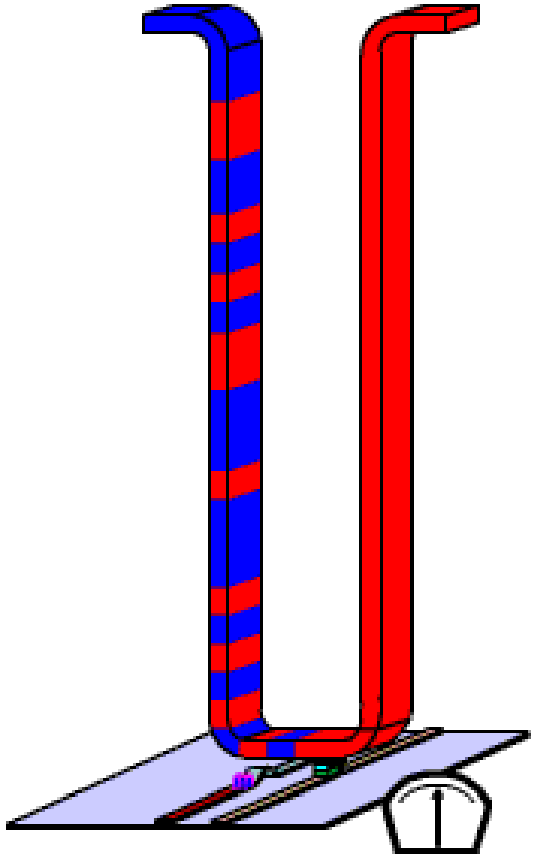
Alverstone PCM



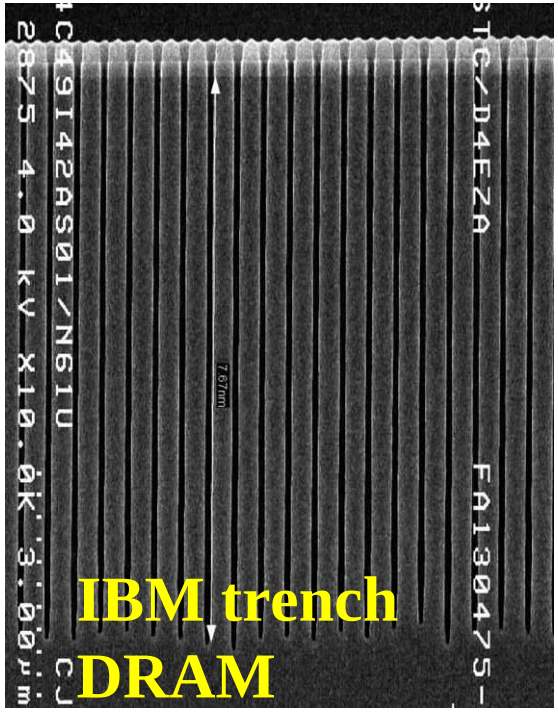
Samsung 512 Mbit PCM chip

Magnetic Racetrack Memory

MRAM alternatives
a 3-D shift register

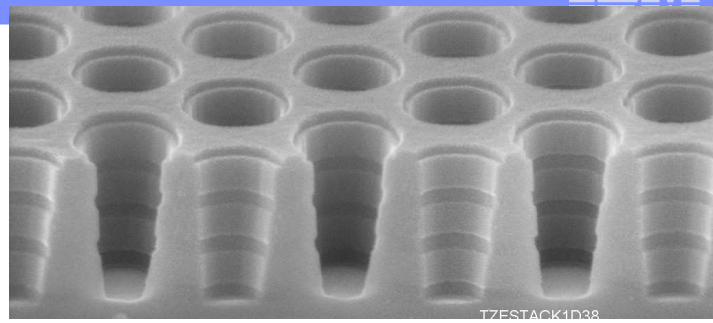


- Data stored as pattern of magnetic domains in long nanowire or “racetrack” of magnetic material.
- Current pulses move domains along racetrack
- Use deep trench to get many (**10-100**) bits per $4F^2$

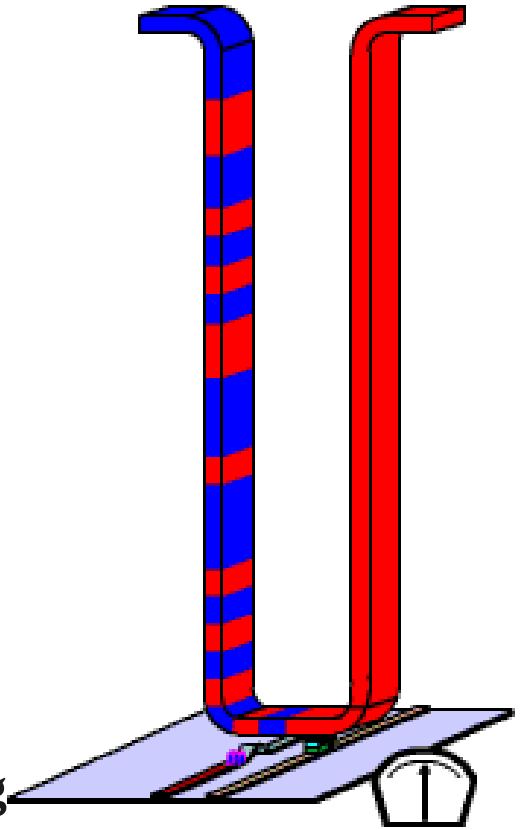


Magnetic Race Track Memory
S. Parkin (IBM), *US patents 6,834,005 (2004) & 6,898,132 (2005)*

Magnetic Racetrack Memory



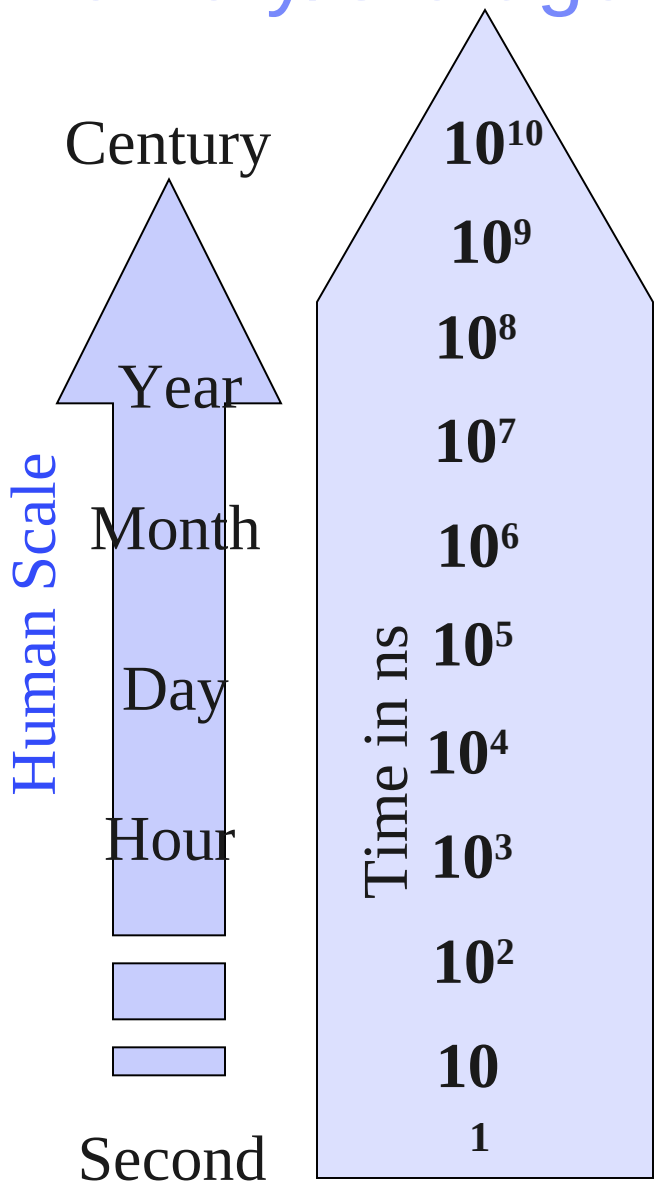
- Need deep trench with notches to “pin” domains
- Need sensitive sensors to “read” presence of domains
- Must insure a moderate current pulse moves every domain one and only one notch
- Basic physics of current-induced domain motion being investigated



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

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

Memory/Storage Stack Latency Problem



Get data from TAPE (40s)

Access DISK (5ms)

Storage

Access FLASH (20 us)

SCM

Access PCM (100 – 1000 ns)

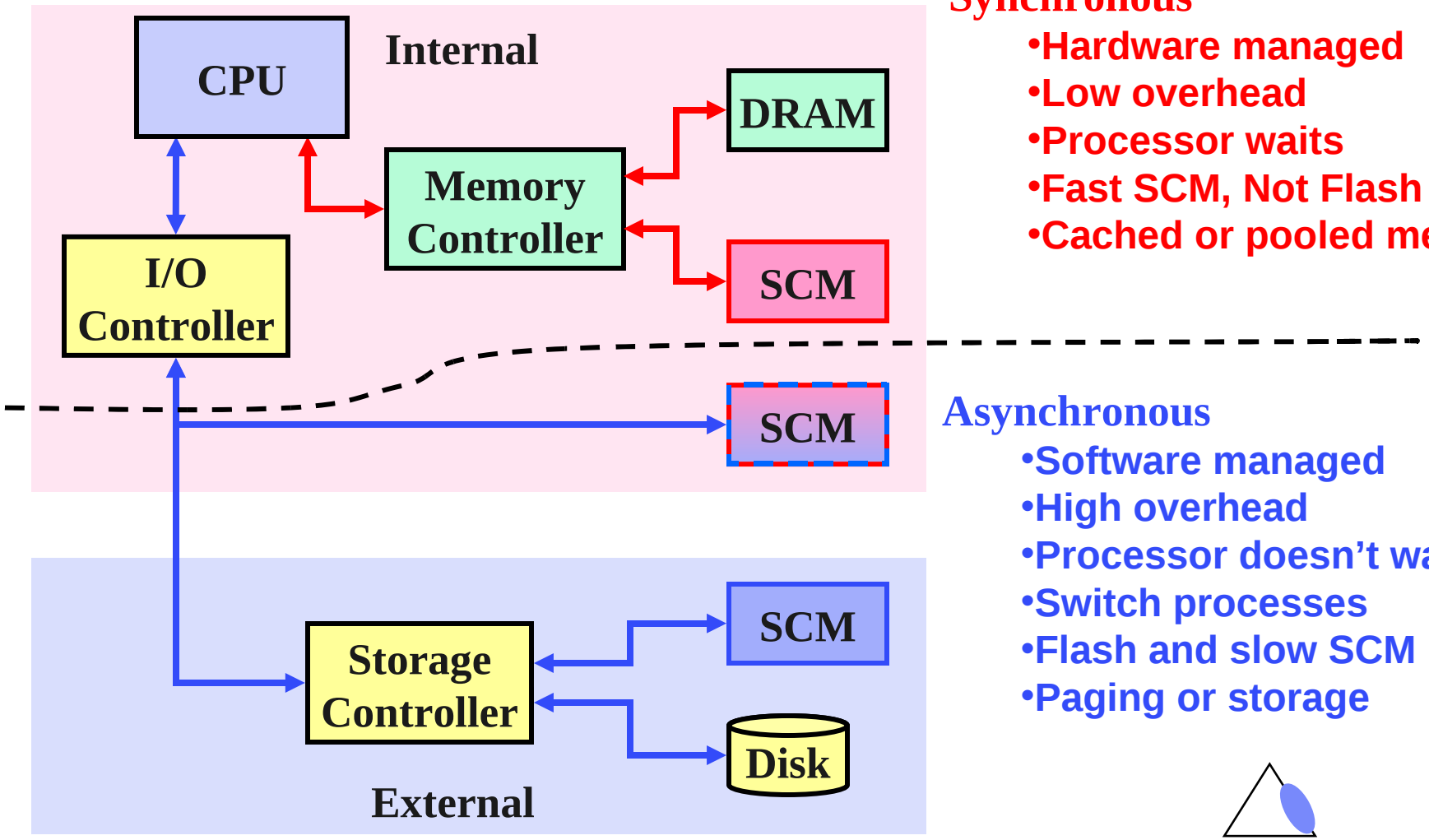
Get data from DRAM *or* PCM (60ns)

Get data from L2 cache (10ns)

Memory

CPU operations (1ns)

Architecture



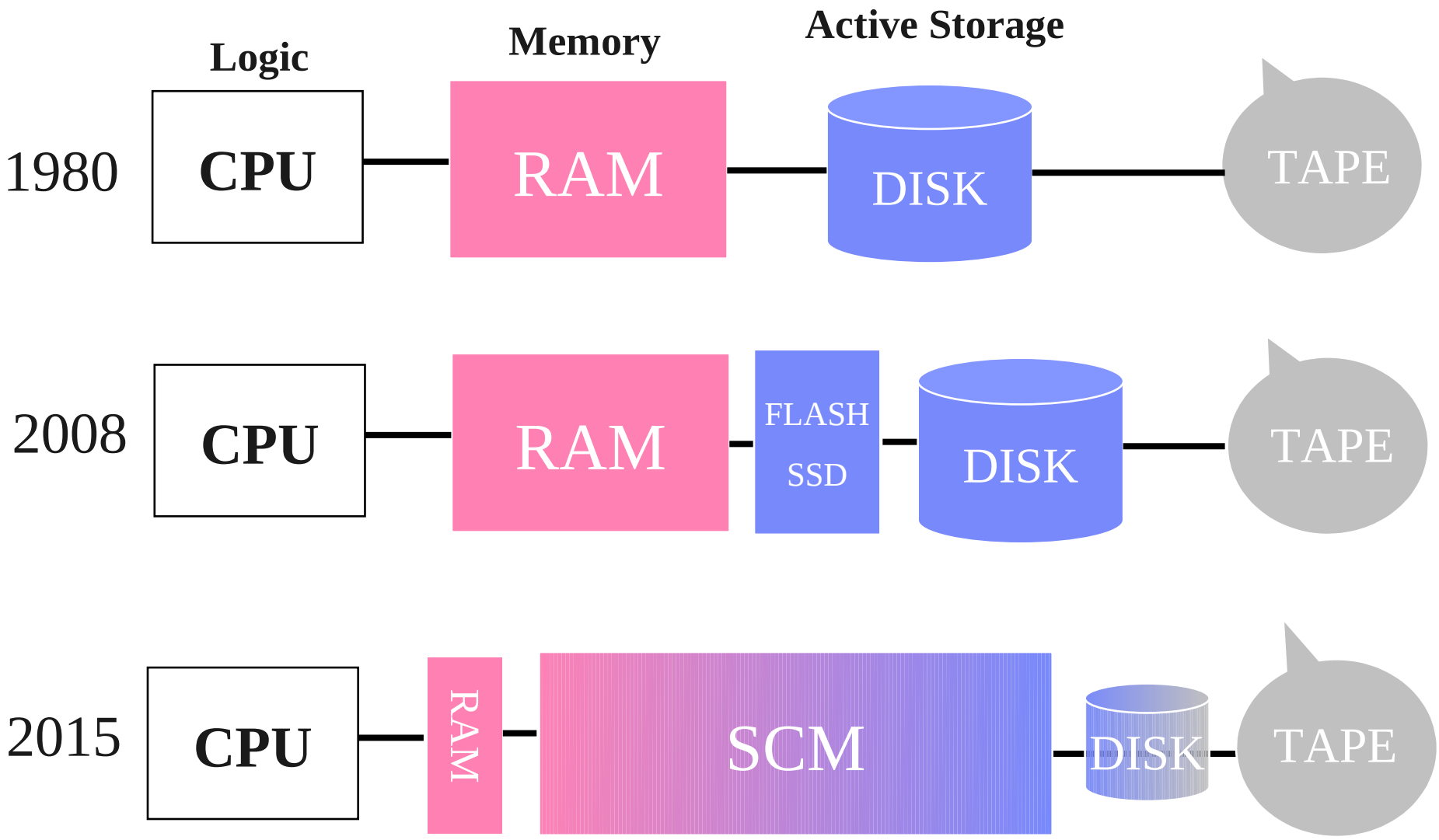
Synchronous

- Hardware managed
- Low overhead
- Processor waits
- Fast SCM, Not Flash
- Cached or pooled memory

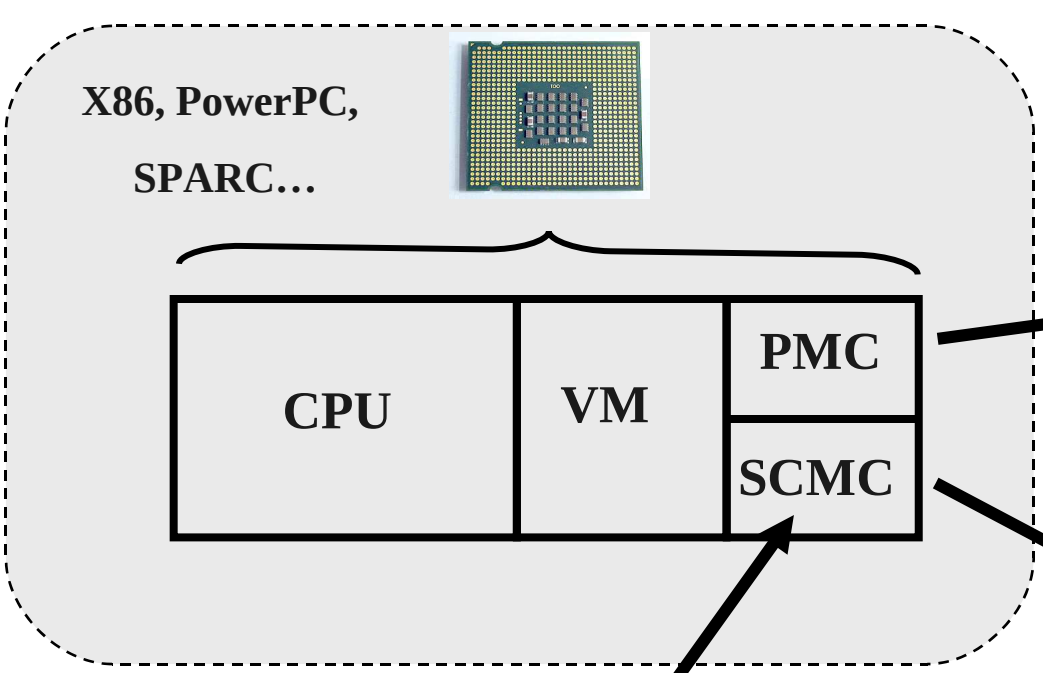
Asynchronous

- Software managed
- High overhead
- Processor doesn't wait
- Switch processes
- Flash and slow SCM
- Paging or storage

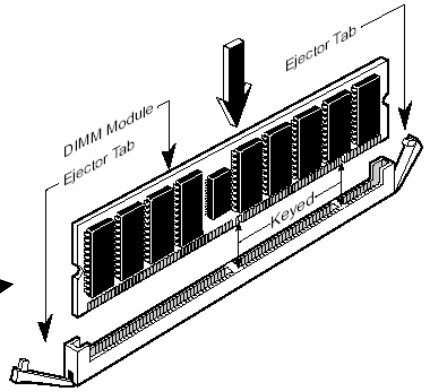
SCM in a large System



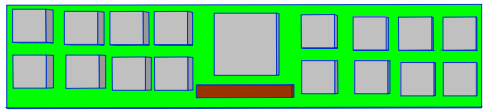
CPU & Memory System Conceptual Alternatives



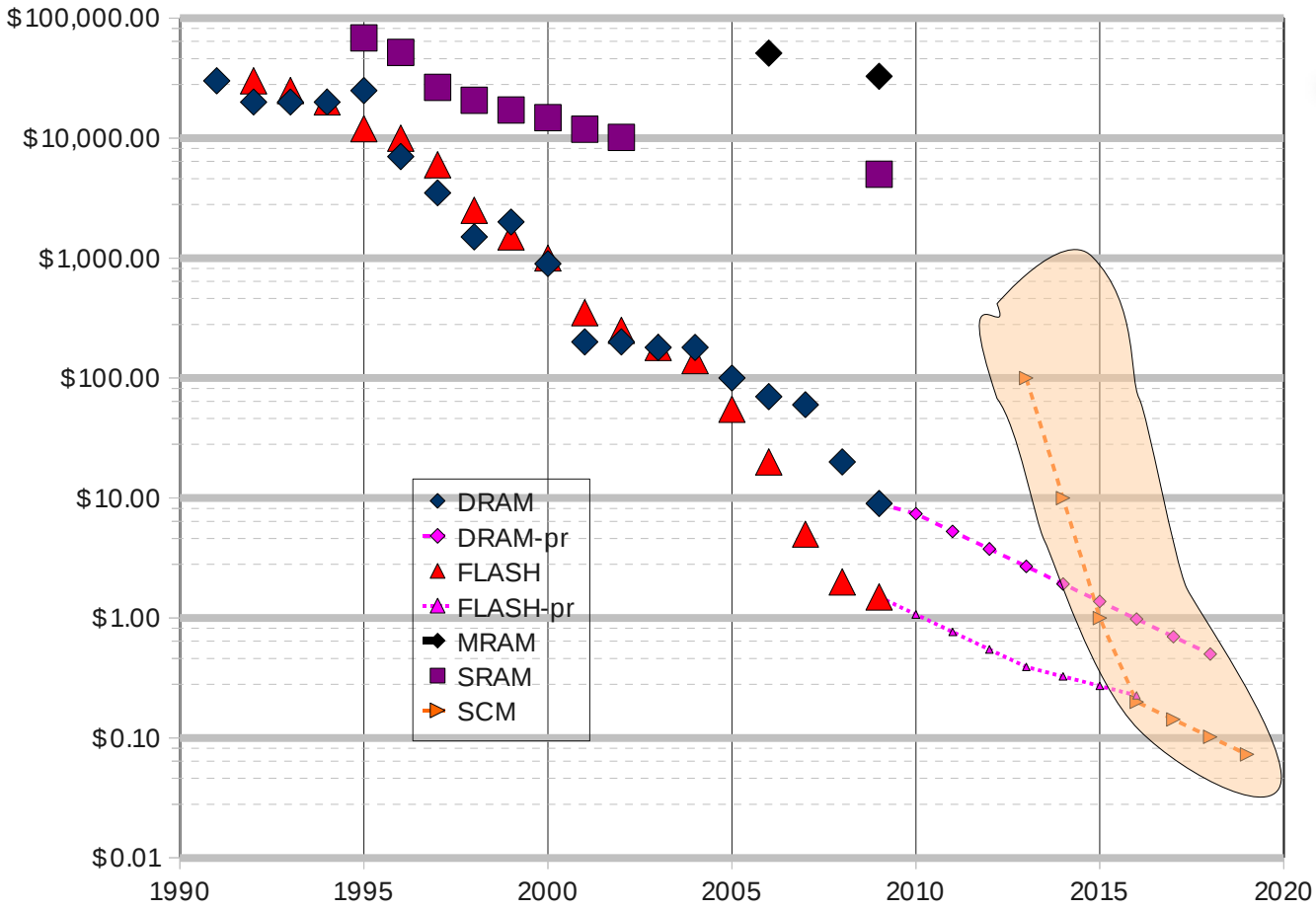
DRAM DIMM



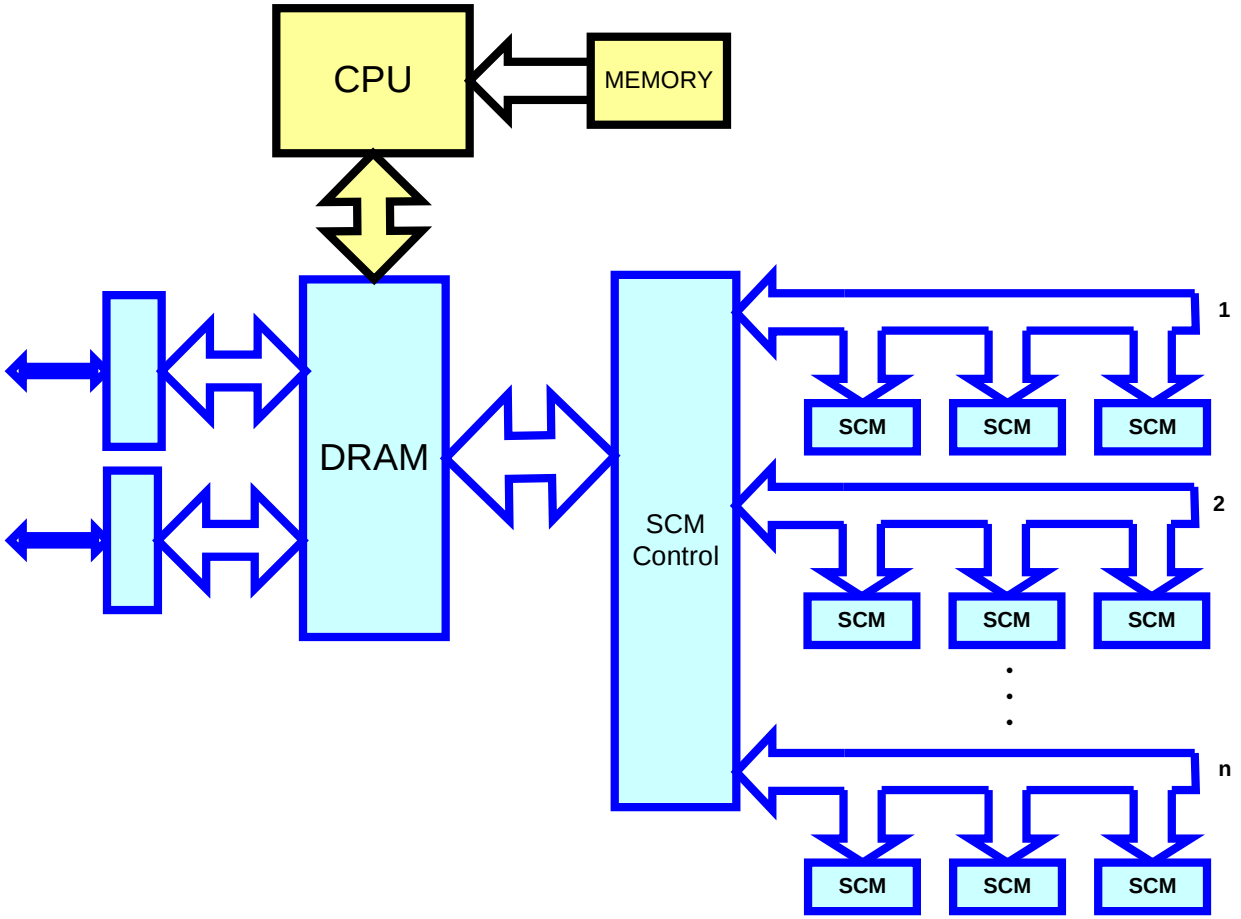
SCM Card



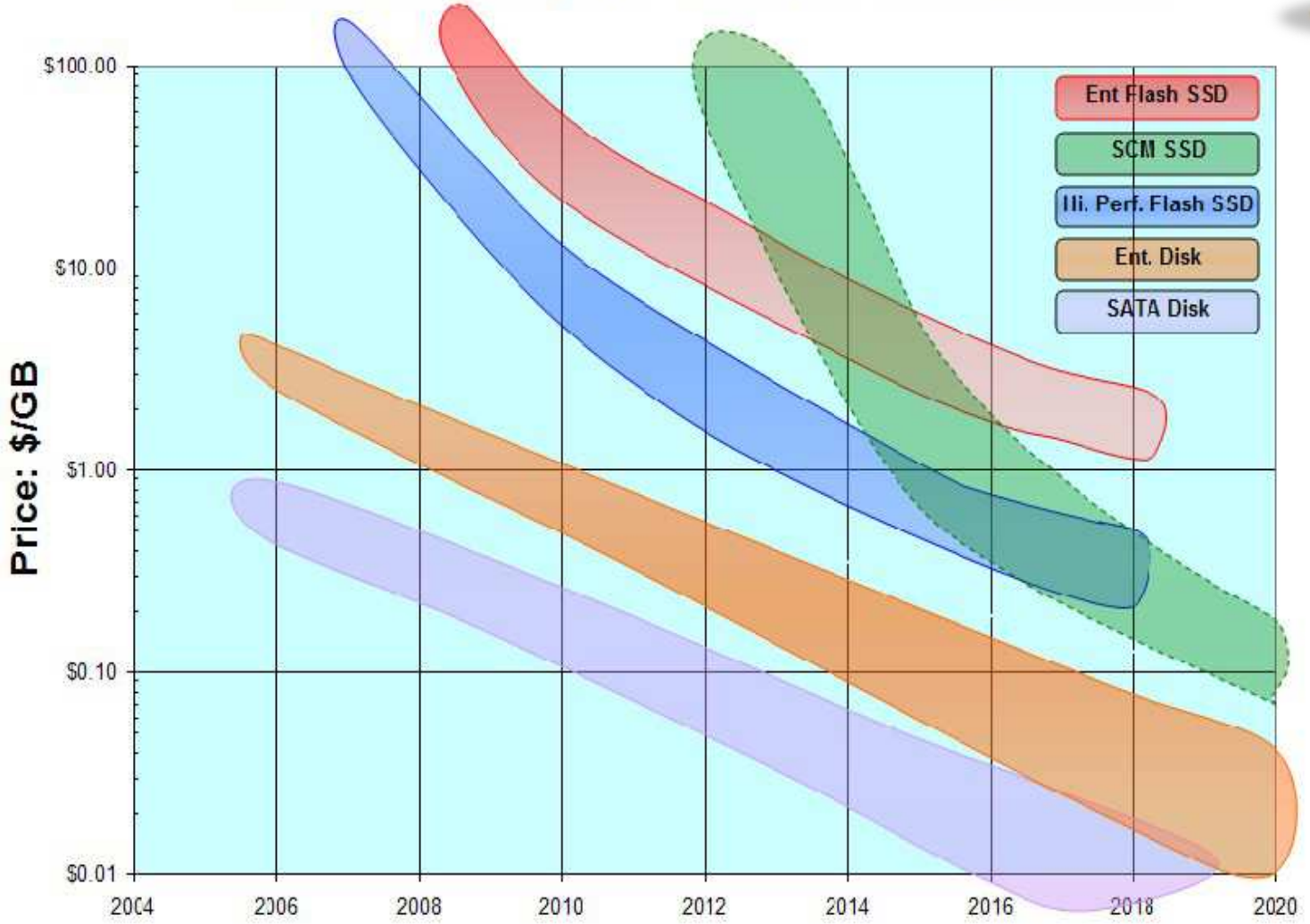
Input from the device cost crystal ball



SCM: Generic Storage Design



Price Trends: Magnetic disks and Solid State Disks



Challenges with SCM

- **Asymmetric performance**
  - Flash: writes much slower than reads
  - Not as pronounced in other technologies
- **Program/erase cycle**
  - Issue for flash
  - Most are write-in-place
- **Data retention and Non-volatility**
  - It's all relative
  - Use case dependent
- **Bad blocks**
  - Devices are shipped with bad blocks
  - Blocks wear out, etc.

- **The "fly in the ointment" is write endurance**
  - In many SCM technologies writes are cumulatively destructive
  - For Flash it is the program/erase cycle
  - Current commercial flash varieties
    - Single level cell (SLC) 10^5
    - Multi-level cell (MLC) 10^4
  - Coping strategy --> wear leveling
  - Typically hidden from applications by infrastructure

Write and/or read endurance and life-time of SCM devices

- **In DRAM and disks (magnetic) there is no known wear out mechanism**
- **In flash and many SCM technologies there are known wear out mechanisms**

  T_{life} = Endurance • Fill-Time

  Fill-Time: time to write a memory unit (what's a data unit?)

- **Simple wear leveling → each write is done to a new (empty) location**

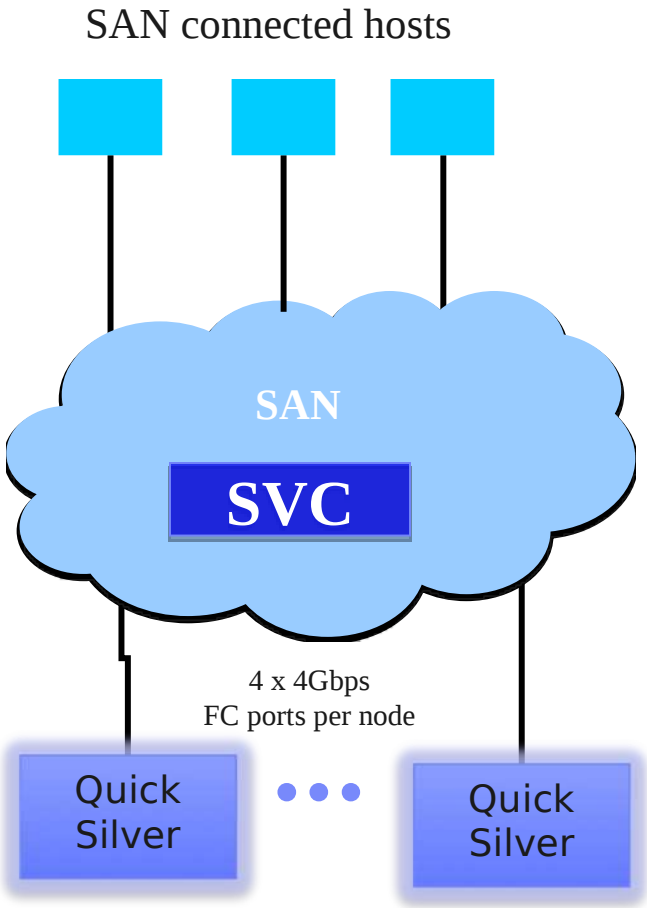
	DRAM	Disk	256GB Flash		8 GB SCM
Endurance	$>10^{16}$	$>10^{11}$	$10^5 \rightarrow 10^4$		10^8
Wear leveled	N	N	N	Y	Y
Memory unit	1 B	512 B	128 KB	256 GB	8 GB
Data unit	1 B	512 B	128 KB	128 KB	128 B
Fill Time	100 ns	4 ms	2 ms	4000 s	500 s
Life Time	>31 yrs	>12 yrs	<4 min	>12 yrs	>190 yrs

SCM impact on software

- **Operating systems**
  - Extend state information kept about memory pages
  - New mechanisms to manage new resource
  - Enhanced to provide hints to other layers of software
  - Potential for direct involvement in managing caches and pools
- **Middle ware and applications → evolutionary**
  - Improved performance impact immediate – full exploitation will occur gradually
  - Little near term demand for non-volatility
  - Cost improvements will drive memory size
  - Memory size will drive larger and more complex data structures.
  - Reload time on a crash will be exacerbated
  - User's need for non-volatility, persistence, etc. will be driven by these effects – blurring of memory and storage

IBM QuickSilver Project → SSD proof of concept

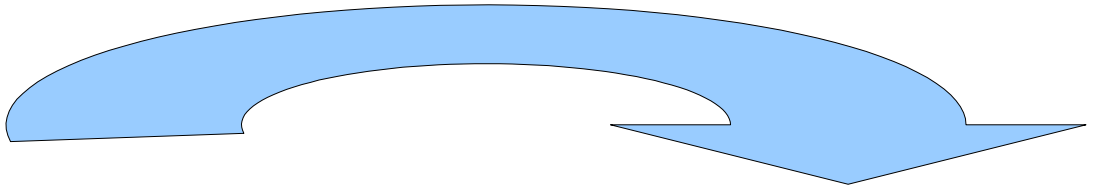
- **Ultra-fast storage performance without managing 1000's of disks.**
  - **Demonstrated performance of over 1 million IOPS using 40 SSDs.**
  - **Reduced $/IOPS, significantly lower than traditional disk storage farm.**
  - **Reduced floor space per IOPS**
  - **Improved energy efficiency for high performance workloads.**
  - **Reduced number of storage elements to manage**



SAN: Storage Area Network

SVC: San Volume Controller

Shift in Systems and Applications

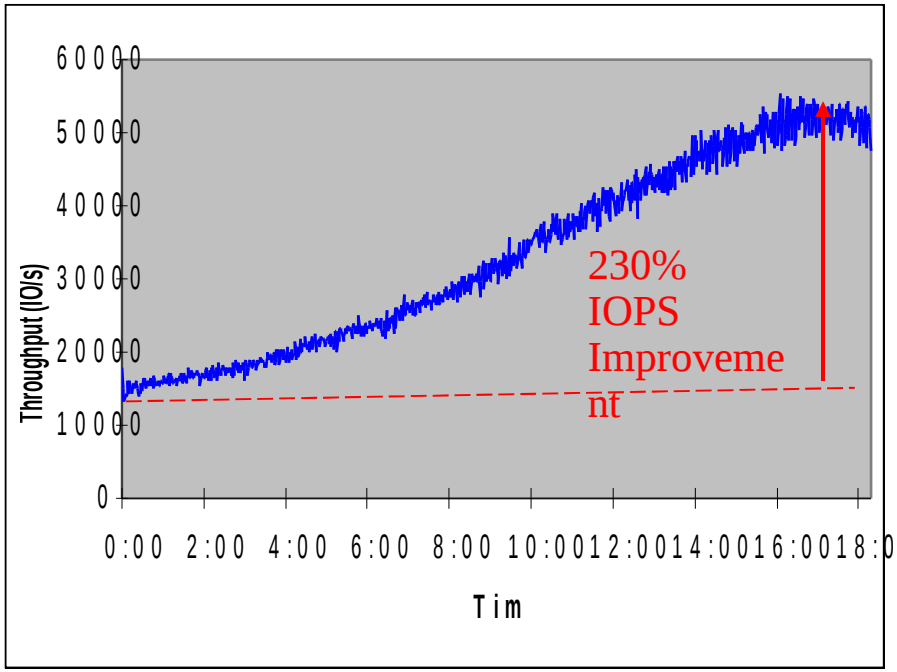


Main Memory:	▪ ***DRAM – Disk – Tape*** –Cost & power constrained –Paging not used –Only one type of memory: volatile	▪ ***DRAM – SCM – Disk – Tape*** –Much larger memory space for same power and cost –Paging viable –Memory pools: different speeds, some persistent –Fast boot and hibernate
Storage:	–Active 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 –Fast, persistent metadata

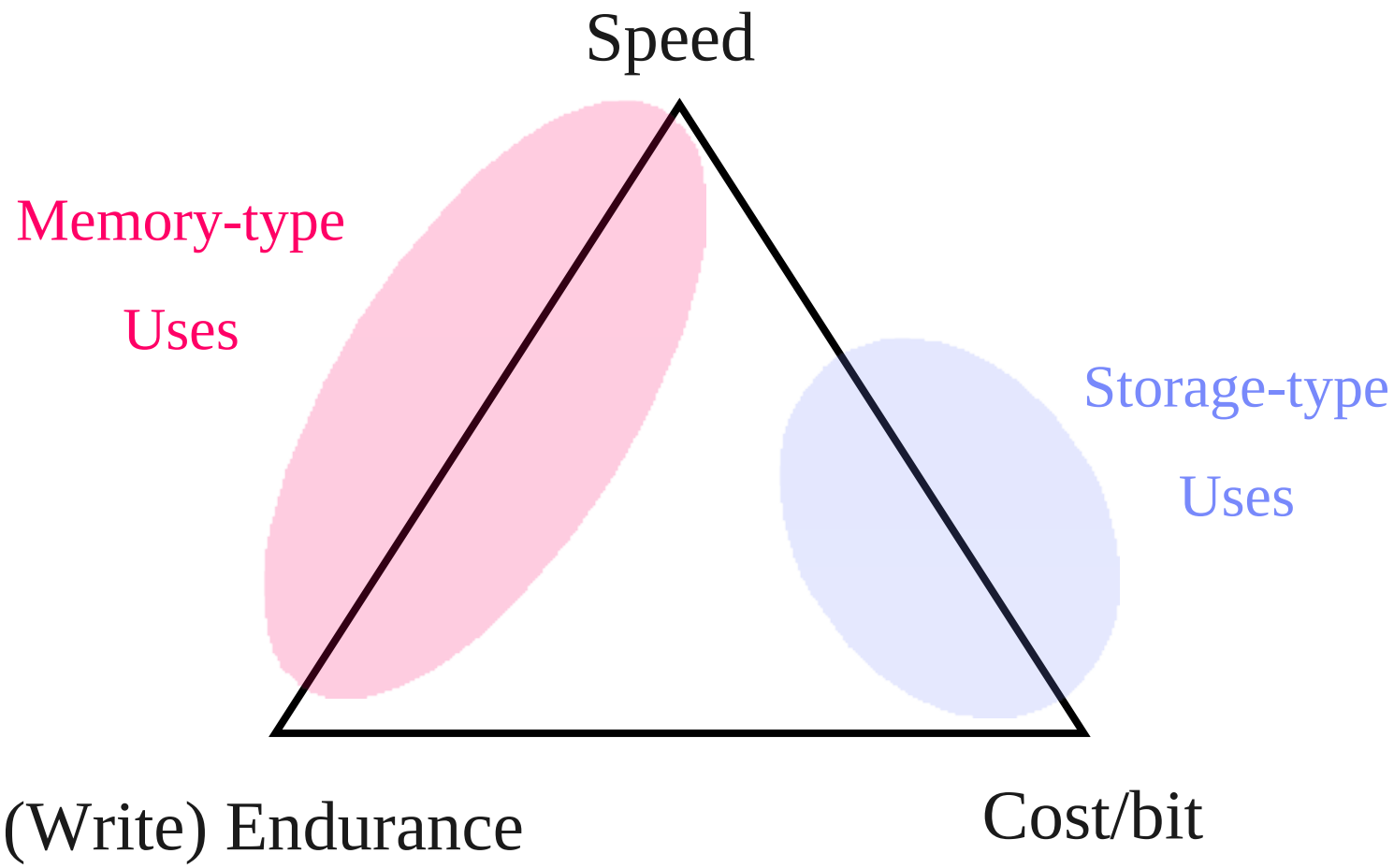
EASY TIER KEY MESSAGES

- SMART STORAGE
- EASY AND SIMPLE
- WORKLOAD OPTIMIZED
- GREEN

Performance increase of 230% by automatic movement of 3% of the application's data to SSD



SCM Design Triangle



Summary

- **Storage Class Memory is a new class of data storage/memory technology → many technologies are competing to be the 'best' SCM**
- **SCM *blurs the distinction* between memory and storage**
- **SCM will impact the design of computer systems and applications**
- **Flash, which has may SCM characteristics, is available now and various SCMs are in the wings.**
- **EasyTier like software will foster exploitation of Flash and SCM**

References

- **FAST2010 Tutorial**
  - T2: Freitas and Chiu, Solid State Storage: Technology, Design and Applications
  - **http://www.usenix.org/events/fast10/tutorials**
- IBM Journal of Research and Development
  - **Special issue on storage**
  - **http://www.research.ibm.com/journal/rd52-45.html**
  - **Four papers related to SCM**

Questions?