IBM Almaden Research Center

Solid-State Storage: Technology, Design and Applications

Dr. Richard Freitas and Lawrence Chiu

Abstract

Most system designers dream of replacing slow, mechanical storage (disk drives) with fast, non-volatile memory. The advent of inexpensive solid-state disks (SSDs) based on flash memory technology and, eventually, on storage class memory technology is bringing this dream closer to reality.

This tutorial will briefly examine the leading solid-state memory technologies and then focus on the impact the introduction of such technologies will have on storage systems. It will include a discussion of SSD design, storage system architecture, applications, and performance assessment.

Author Biographies

- **Rich Freitas is a Research Staff Member at the IBM Almaden Research Center. Dr. Freitas received his PhD in EECS from the University of California at Berkeley in 1976. He then joined IBM at the IBM T.J. Watson Research Lab. He has held various management and research positions in architecture and design for storage systems, servers, workstations, and speech recognition hardware at the IBM Almaden Research Center and the IBM T.J. Watson Research Center. His current interest lies in exploring the use of emerging nonvolatile solid state memory technology in storage systems for commercial and scientific computing.**

- **Larry Chiu is Storage Research Manager and a Senior Technical Staff Member at the IBM Almaden Research Center. He co-founded the SAN Volume Controller product, a leading storage virtualization engine which has held the fastest SPC-1 benchmark record for several years. In 2008, he led a research team in the US and in the UK to demonstrate one million IOPS storage system using solid state disks. He is currently working on expanding solid state disk use cases in enterprise system and software. He has an MS in computer engineering from the University of Southern California and another MS in technology commercialization from the University of Texas at Austin.**

Acknowledgements

- **Winfried Wilcke**
- **Geoff Burr**
- **Bulent Kurdi**
- **Clem Dickey**
- **Paul Muench**
- **C. Mohan**
- **KK Rao**

Agenda

Introduction	10 min
Technology	40 min
System	30 min
Questions	10 min
Break	30 min
Applications	40 min
Performance	40 min
Questions	10 min

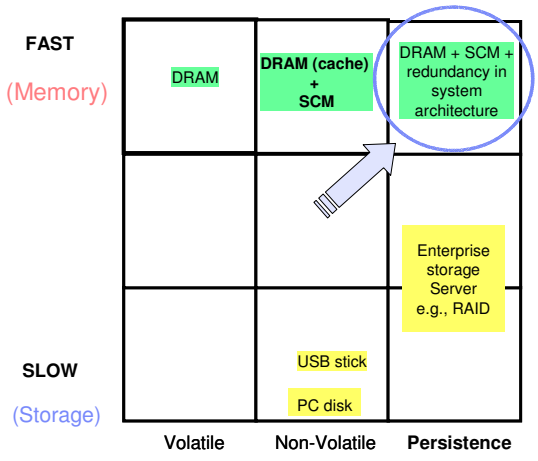
Introduction

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

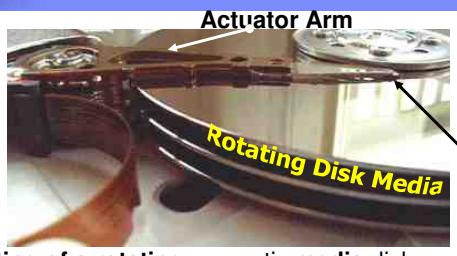
- **A new class of data storage/memory devices**
  - –many technologies compete to be the 'best' SCM
- **SCM features:**
  - –Non-volatile
  - –Short Access times (~ DRAM like)
  - –Low cost per bit (more DISK like – by 2020)
  - –Solid state, no moving parts
- **SCM *blurs the distinction* between**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

Speed/Volatility/Persistence Matrix

- **NVRAM = Non Volatile RAM**
  - –Data survives loss of power
  - –SCM is one example of NVRAM
  - –Other NVRAM types: DRAM+battery or DRAM+disk combos
- **Persistent Storage**
  - –Data survives despite component failure or loss of power
  - –Disk drives is not persistent but RAID array is



HDDs

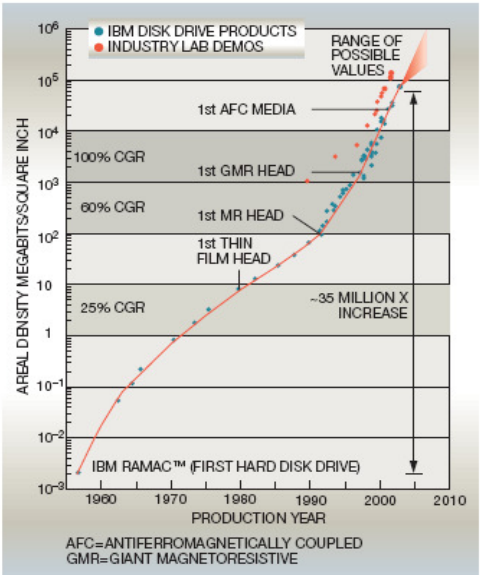


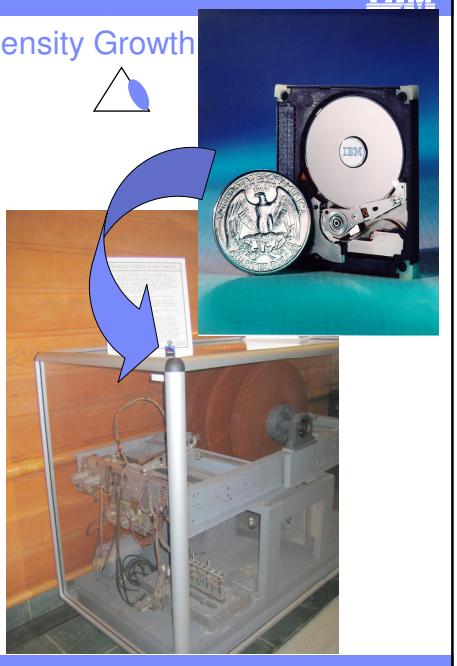
- **Invented in the 1950s**
- **Mechanical device consisting of a rotating** magnetic **media** disk **and** actuator **arm w/ magnetic** head

HUGE COST ADVANTAGES

- $ **High growth in** disk areal density **has driven the HDD success**
- $ **Magnetic thin-film head wafers have very few critical elements per chip (vs. billions of transistors per semiconductor chip)**
- $ **Thin-film head (GMR-head) has only one critical feature size controlled by optical lithography (determining track width)**
- $ **Areal density is control by track width times (X) linear density…**

History of HDD is based on Areal Density Growth





Future of HDD

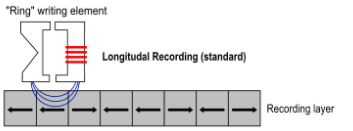
Higher densities through

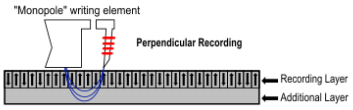
- perpendicular recording

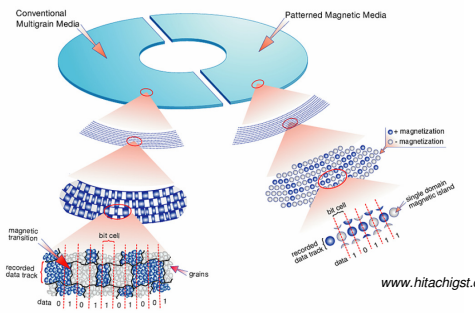
Jul 2008
610 Gb/in^2 → ~4 TB

- patterned media

Wikipedia

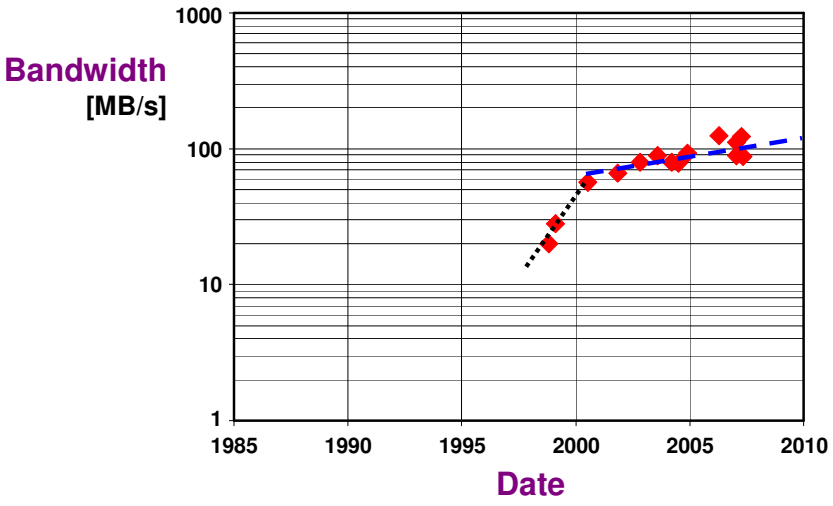




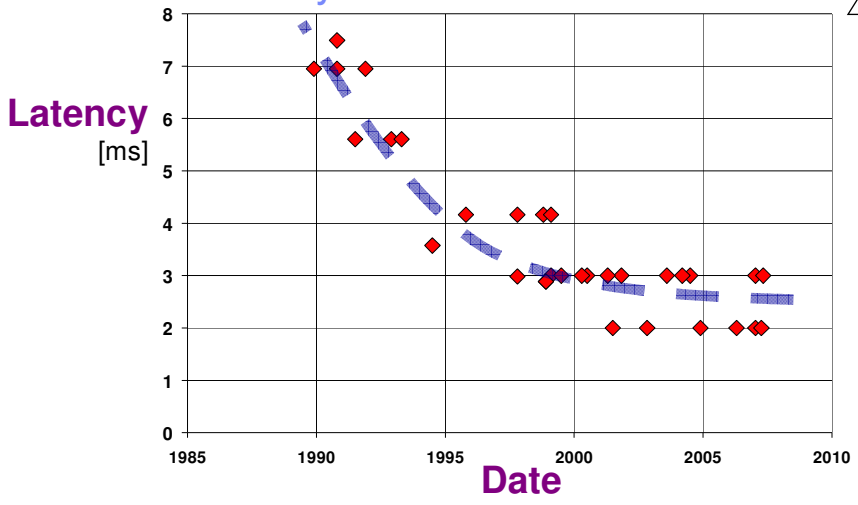


www.hitachigst.com/hdd/research/images/pm images/conventional_pattern_media.pdf

Disk Drive Maximum Sustained Data Rate



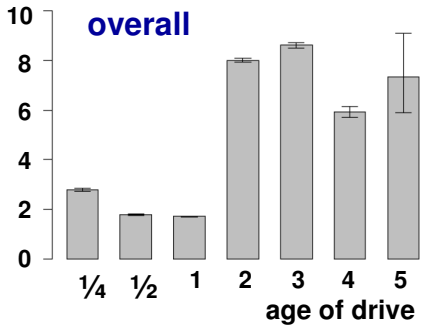
Disk Drive Latency

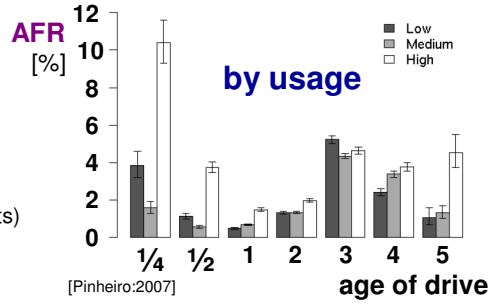


- ☒ **Bandwidth Problem** is getting much harder to hide with parallelism
- ☒ **Access Time Problem** is also not improving with caching tricks
- ☒ Power/Space/Performance Cost

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 during recovery from a drive failure…?





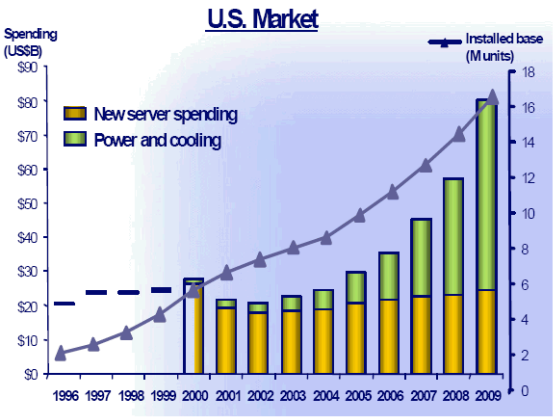
[Pinheiro:2007]

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

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 Mega**watts

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.

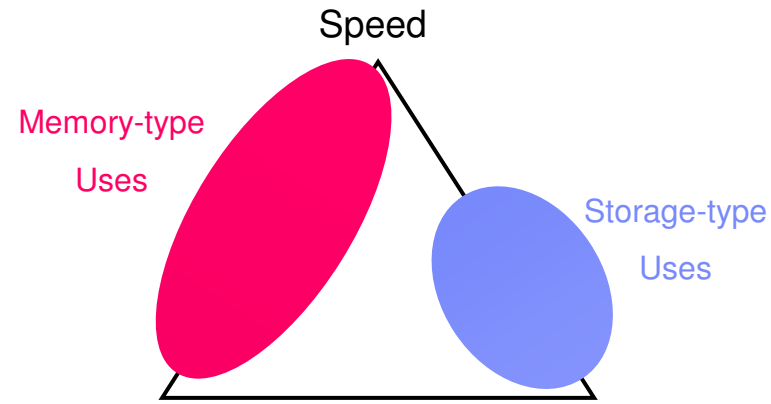
System Targets for SCM

Megacenters

Billions!

Mobile ✓ Desktop X Datacenter ✓

SCM Design Triangle



Power!

For more information

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).

Technology

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]

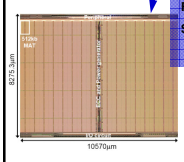
Even more Criteria

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

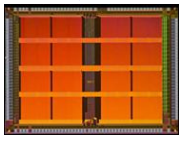
~ 16 criteria! This makes the SCM problem so hard

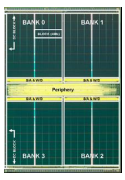
Emerging Memory Technologies

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	~~NEC~~			Nanolayer
Toshiba	NEC	Honeywell	~~Toshiba~~			
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					

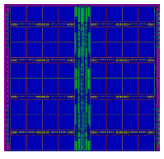


64Mb FRAM (Prototype)
0.13um 3.3V


4Mb MRAM (Product)
0.18um 3.3V



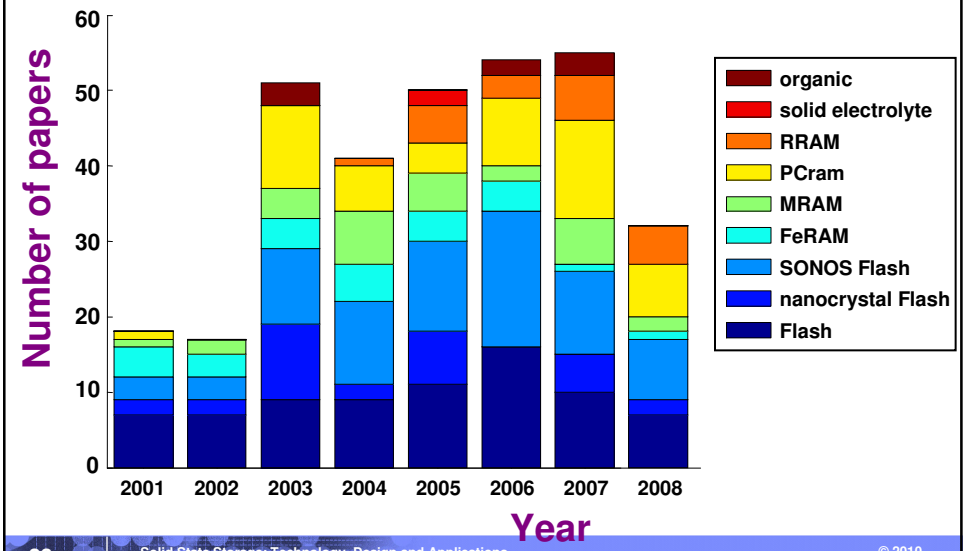
512Mb PRAM (Prototype)
0.1um 1.8V


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

Research interest

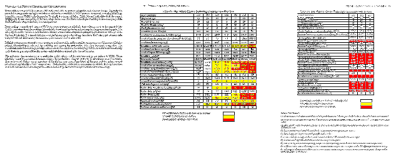
Papers presented at

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

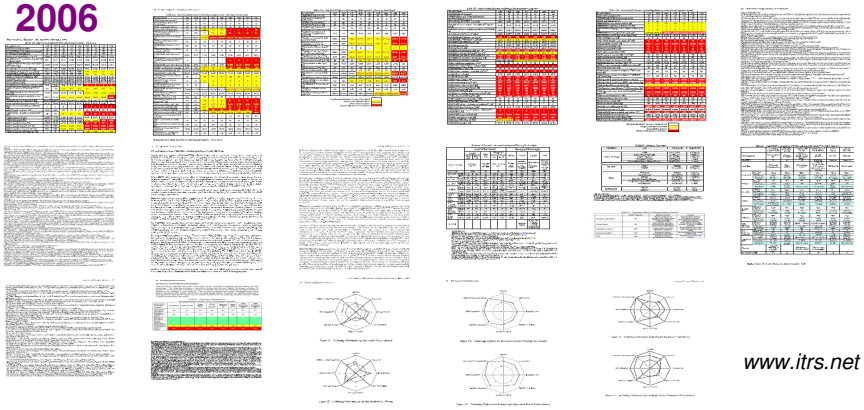


Industry interest in non-volatile memory **2001**

INTERNATIONAL TECHNOLOGY ROADMAP FOR SEMICONDUCTORS

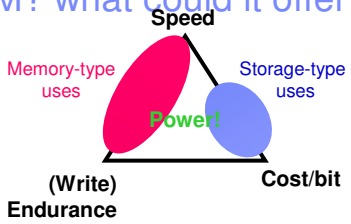


2006



www.itrs.net

What is SCM? what could it offer?

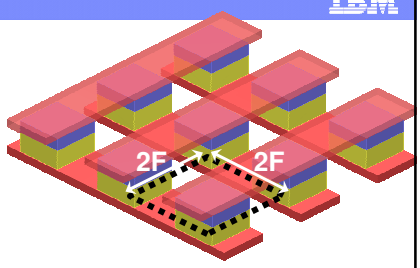


A solid-state memory that **blurs the boundaries** between storage and memory by being **low-cost**, **fast**, and **non-volatile**.

- **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

Density is key

Cost competition between IC, magnetic and optical devices comes down to effective areal density**.**

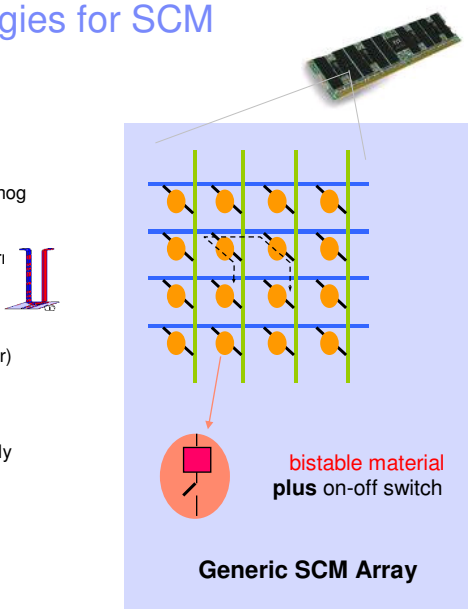


Device	Critical feature-size F	Area (F^2)	Density (Gbit /sq. in)
Hard Disk	**50 nm** (MR width)	**1.0**	**250**
DRAM	**45 nm** (half pitch)	**6.0**	**50**
NAND (2 bit)	**43 nm** (half pitch)	**2.0**	**175**
NAND (1 bit)	**43 nm** (half pitch)	**4.0**	**87**
Blue Ray	**210 nm** (λ /2)	**1.5**	**10**

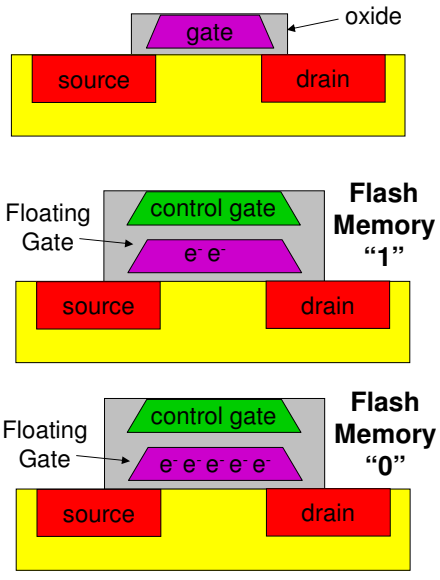
[Fontana:2004, web searches]

Many Competing Technologies 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 ter
- Ferroelectric RAM
  - used today, but poor scalability
- Solid Electrolyte and resistive RAM (Memristor)
  - early development, maybe?
- Organic, nano particle and polymeric RAM
  - many different devices in this class, unlikely
- **Improved FLASH**
  - still slow and poor write endurance

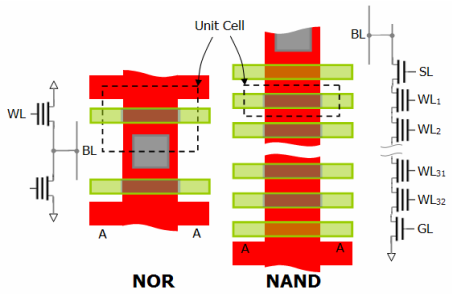


What is Flash?



- 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.

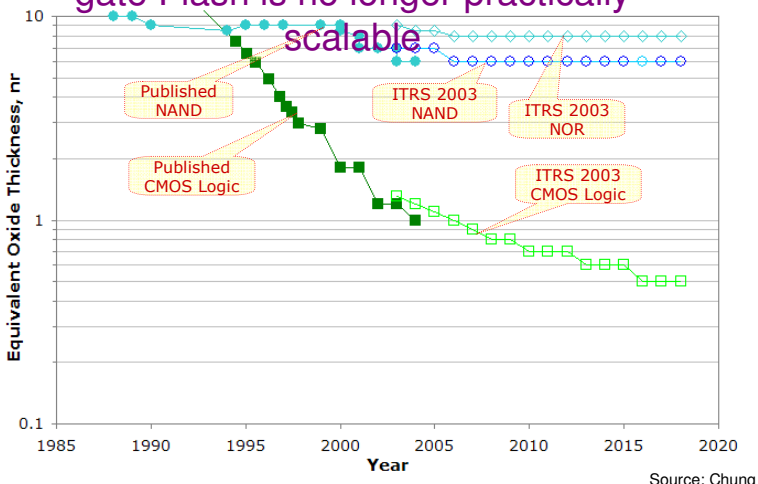
FLASH memory types and application



	NOR	NAND
Cell Size	9-11 F^2	2 F^2 ($4\ F^2$ physical x 2-bit MLC)
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

Flash – below the 100nm technology node

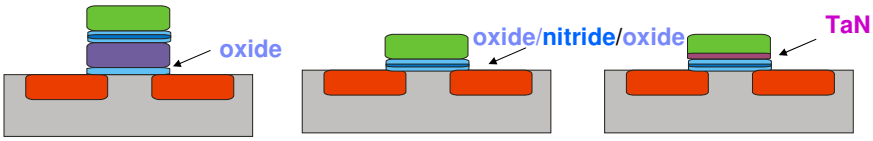
Tunnel oxide thickness in Floating-gate Flash is no longer practically scalable



Source: Chung Lam, IBM

Can Flash improve enough to help?

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



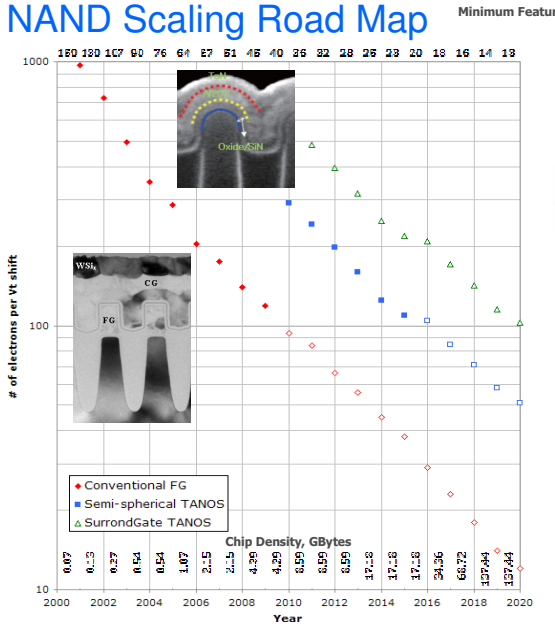
Floating Gate
<40nm ???

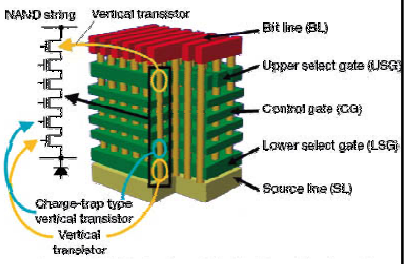
SONOS
Charge trapping in SiN trap layer

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…

NAND Scaling Road Map





Evolution??

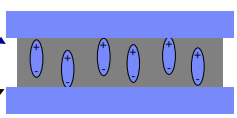
- **Migrating to Semi-spherical TANOS memory cell 2009**
- **Migrating to 3-bit cell in 2010**
- **Migrating to 4-bit cell in 2013**
- **Migrating to 450mm wafer size in 2015**
- **Migrating to 3D Surround-Gate Cell in 2017**

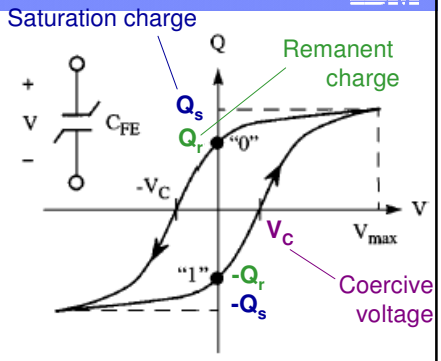
Source: Chung Lam, IBM

What is FeRAM?

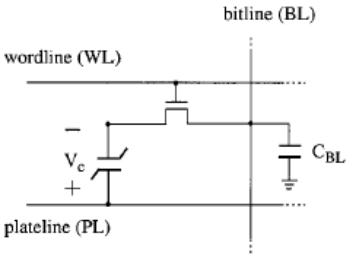
ferroelectric material
such as
lead zirconate titanate
(Pb(ZrxTi1-x)O) or PZT

metallic electrodes





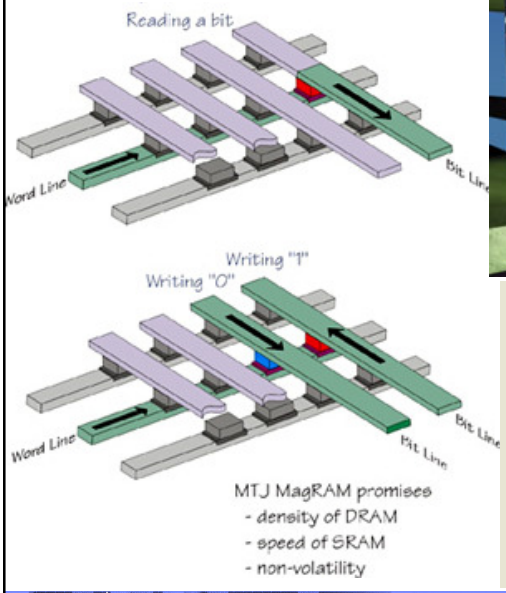
need select transistor –
"half-select" perturbs

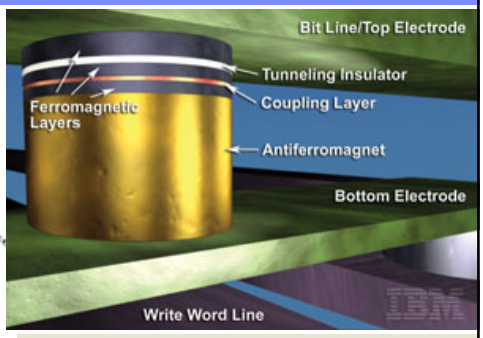


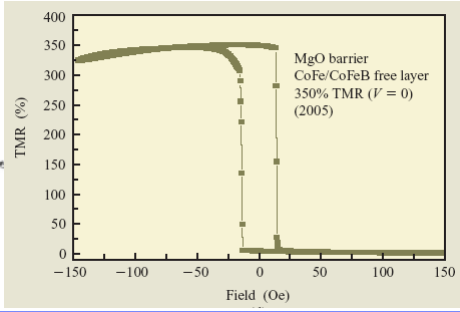
[Sheikholeslami:2000]

- 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

MRAM (Magnetic RAM)



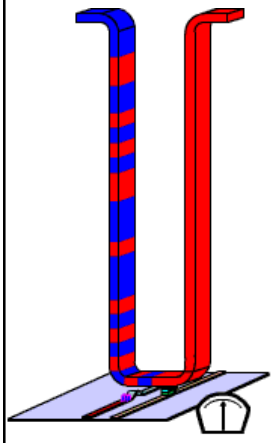




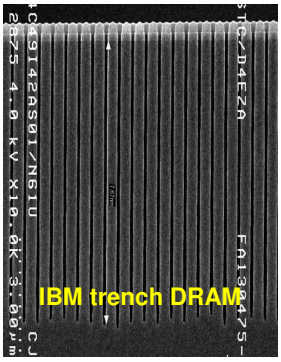
[Gallagher:2006]

Magnetic Racetrack Memory

MRAM alternatives **a 3-D shift register**

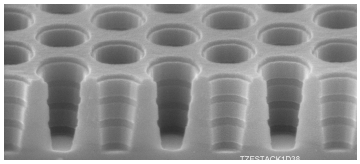


- Data stored as pattern of magnetic domains in long nanowire or “racetrack” of magnetic material.
- Current pulses move domains along racetrack
- Use deep trench to get many (**10-100**) bits per $4F^2$

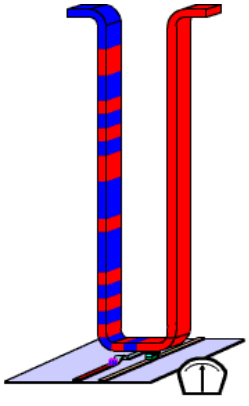


Magnetic Race Track Memory
S. Parkin (IBM), *US patents 6,834,005 (2004) & 6,898,132 (2005)*

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

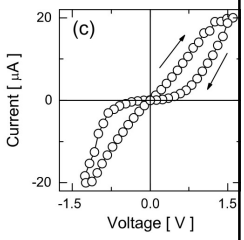


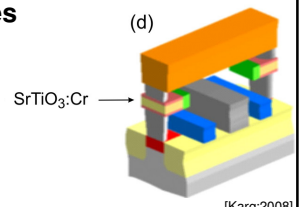
Promise (10-100 bits/F^2) is enormous…

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

RRAM (Resistive RAM)

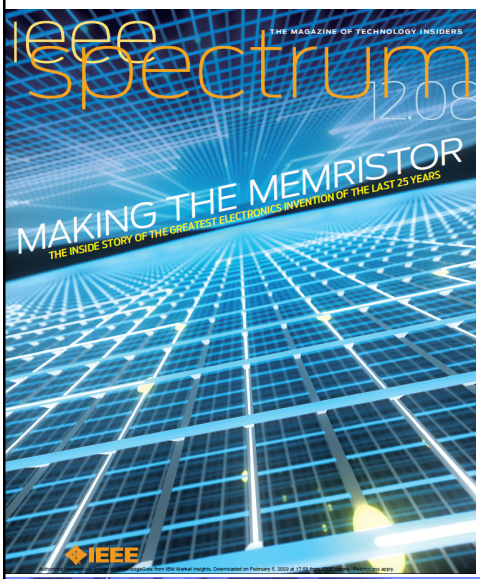
- Numerous examples of materials showing hysteretic behavior in their I-V curves
- Mechanisms not completely understood, but major materials classes include
  - metal nanoparticles(?) in **organics**
    - could they survive high processing temperatures?
  - oxygen vacancies(?) in **transition-metal 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**

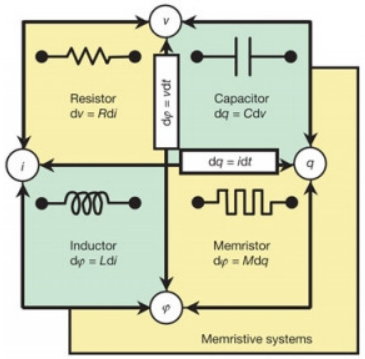




[Karg:2008]

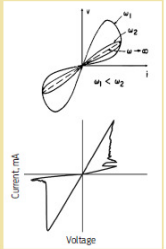
Memristor



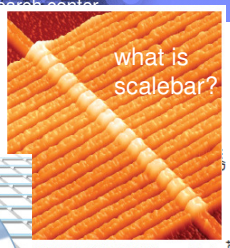


Bow Ties

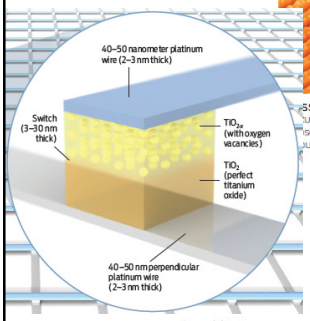
LEON CHUA'S original graph of the hypothetical memristor's behavior is shown at top right; the graph of R. Stanley Williams's experimental results in the *Nature* paper is shown below. The loops map the switching behavior of the device: it begins with a high resistance, and as the voltage increases, the current slowly increases. As charge flows through the device, the resistance drops, and the current increases more rapidly with increasing voltage until the maximum is reached. Then, as the voltage decreases, the current decreases but more slowly, because charge is flowing through the device and the resistance is still dropping. The result is an on-switching loop. When the voltage turns negative, the resistance of the device increases, resulting in an off-switching loop. —*R.S.W.*

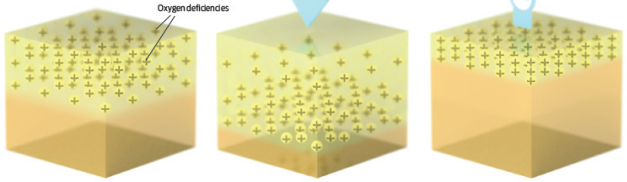


Memristor

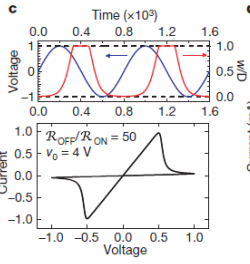


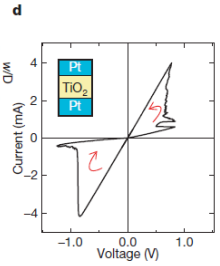
5SBAR ARCHITECTURE: A memristor's ure, shown here in a scanning tunneling scope image, will enable dense, stable uter memories. IMAGE P. STANLEY WILLIAMS/HP LABS





Note time-range chosen for simulations, and the required switching current (power)





Can nearly anything that involves a state variable *w* become a memristor...?

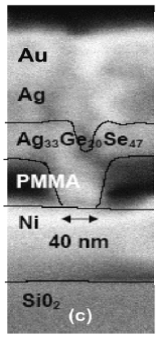
$$v = \mathcal{R}(w,i)i$$

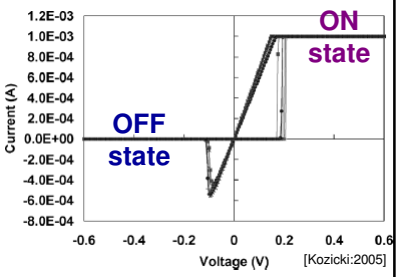
$$\frac{dw}{dt} = f(w,i)$$

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

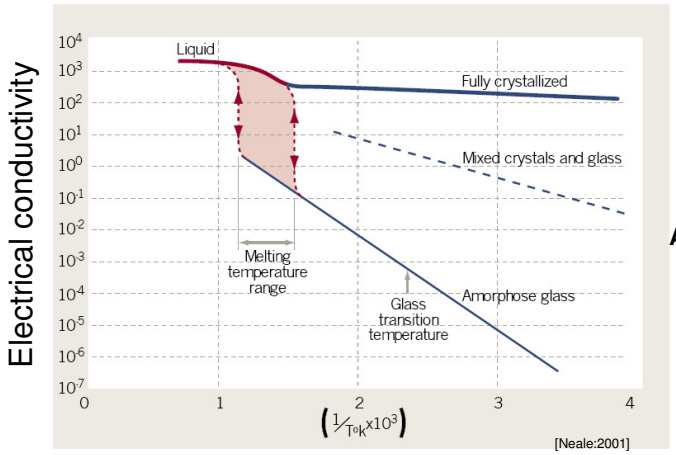
- Program and erase at very low voltages & currents
- High speed
- Large ON/OFF contrast
- Good endurance demonstrated
- Integrated cells demonstrated

Issues

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

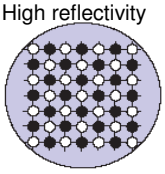
History of Phase-change memory

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

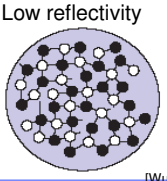


[Neale:2001]

Crystalline phase
Low resistance
High reflectivity

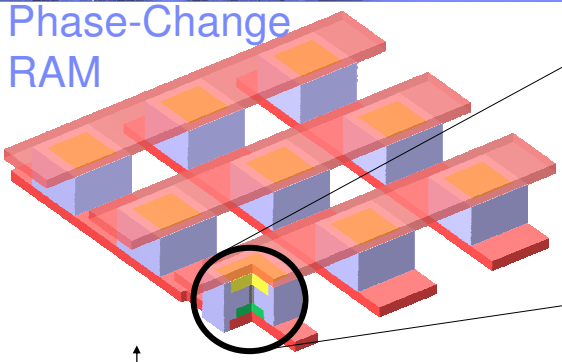


Amorphous phase
High resistance
Low reflectivity



[Wuttig:2007]

Phase-Change RAM

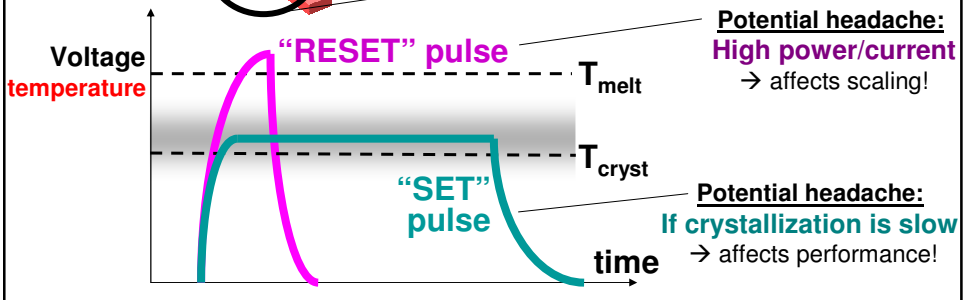


Bit-line

PCRAM
"programmable resistor"

Word -line

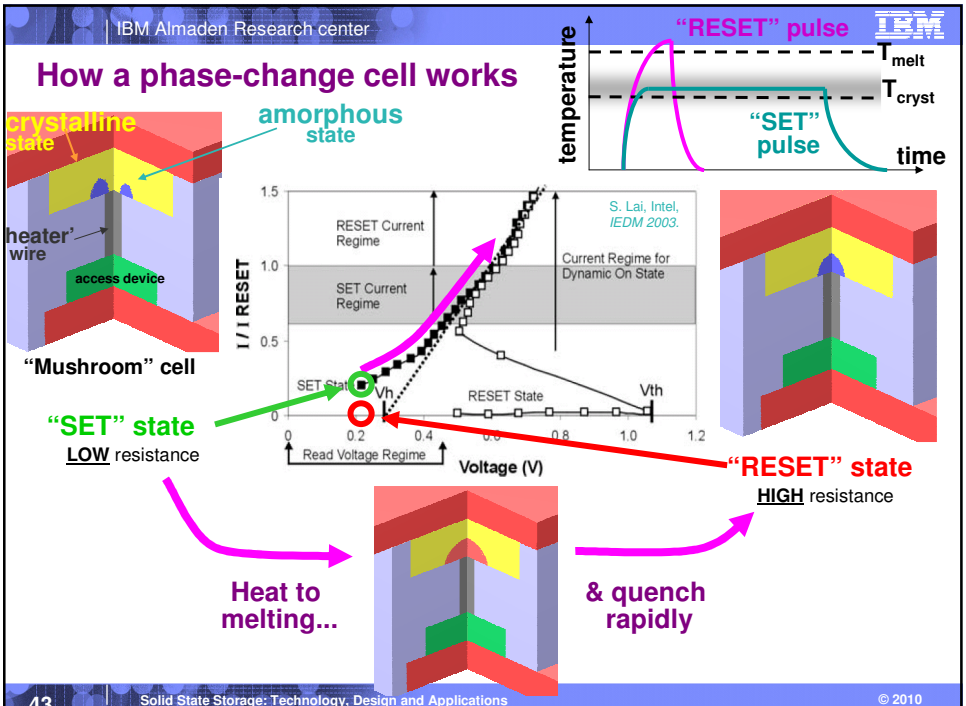
Access device
(transistor, diode)



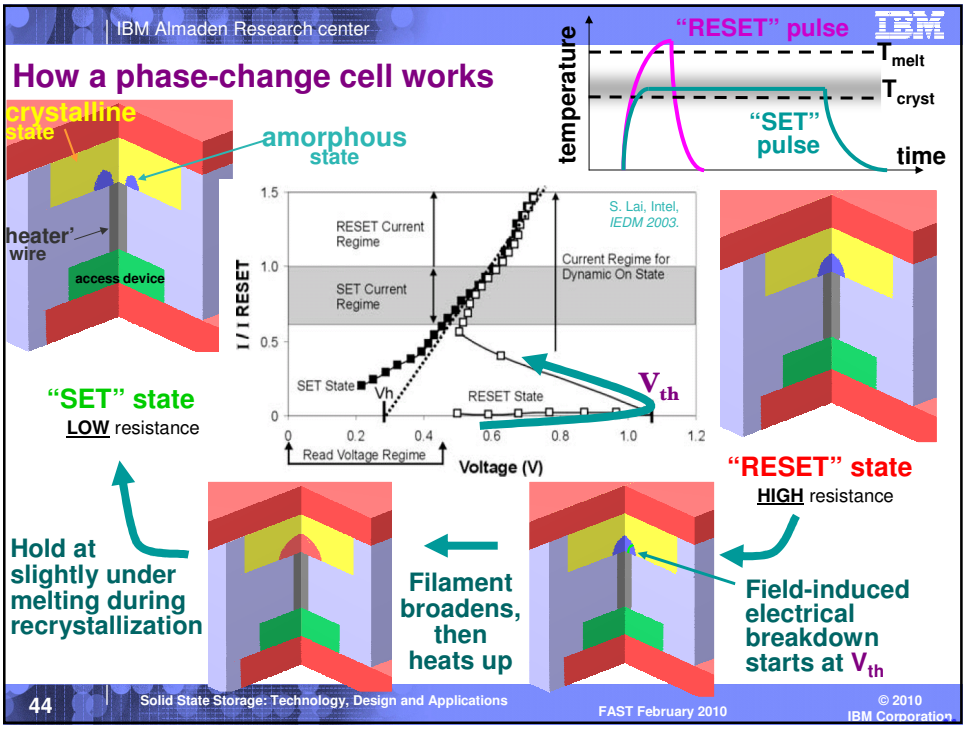
Potential headache:
High power/current
→ affects scaling!

Potential headache:
If crystallization is slow
→ affects performance!

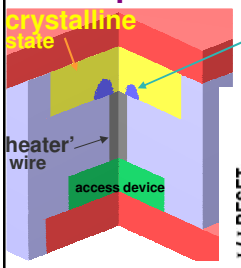
How a phase-change cell works



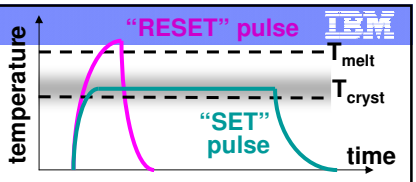
How a phase-change cell works

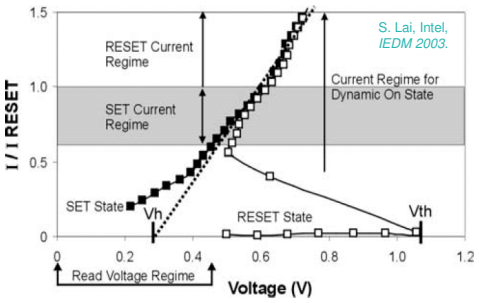


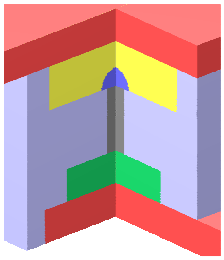
How a phase-change cell works



"SET" state
LOW resistance







"RESET" state
HIGH resistance

Issues for phase-change memory

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

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

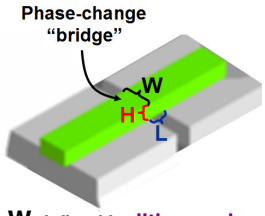
Any new non-volatile memory technology had better work for several device generations… ⇨ Will PC-RAM scale?

- ? 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?

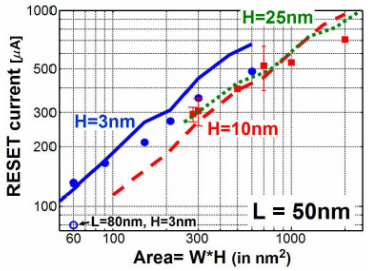
Phase-Change Nano-Bridge

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

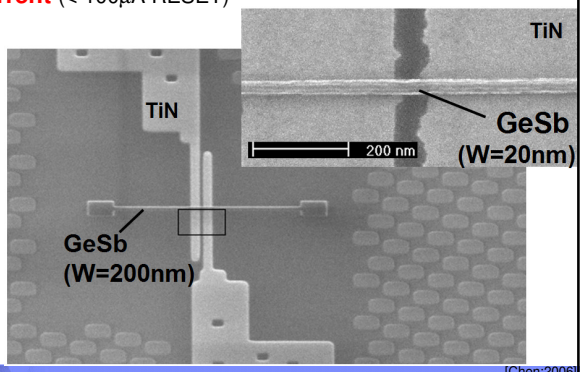
- 3nm * 20nm → 60nm^2
  ≈ Flash roadmap for **2013**
  → **phase-change *scales***
- **Fast** (<100ns SET)
- **Low current** (< 100μA RESET)



W defined by **lithography**
H by **thin-film deposition**

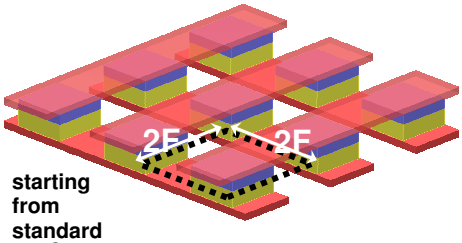


Current scales with area



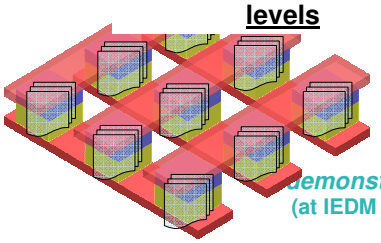
[Chen:2006]

Paths to ultra-high density memory

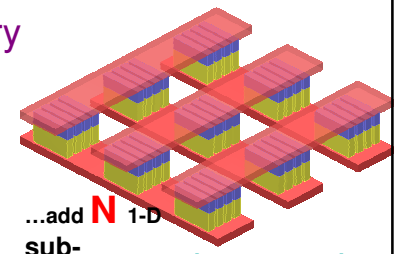


starting from standard $4F^2$ ...

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



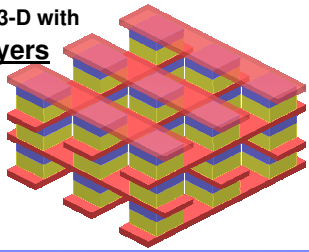
demonstrated (at IEDM 2007)



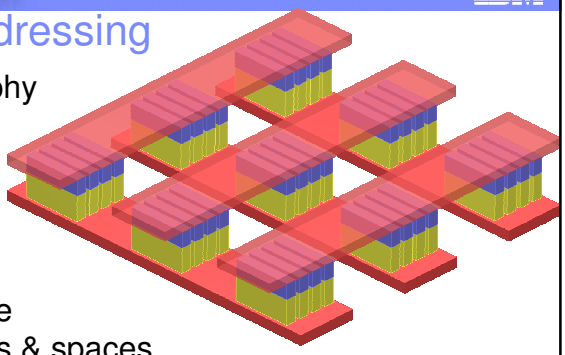
...add **N** 1-D **sub-lithographic** "fins" (N^2 with 2-D)

demonstrated (at IEDM 2005)

...go to 3-D with **L layers**



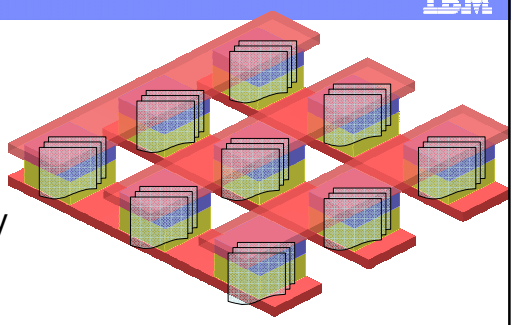
Sub-lithographic addressing



- 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

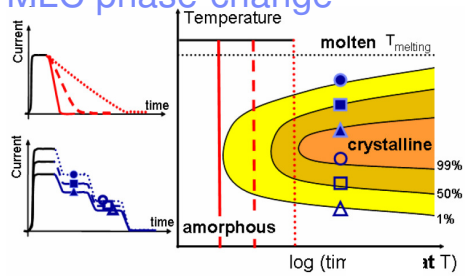
[Gopalakrishnan:2005 IEDM]

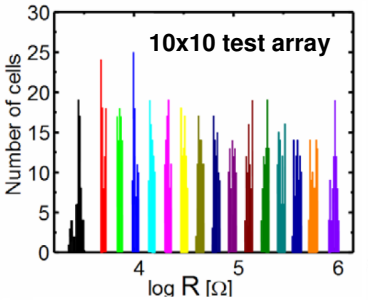
MLC (Multi-Level Cells)

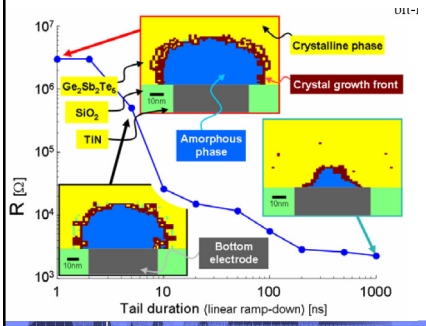


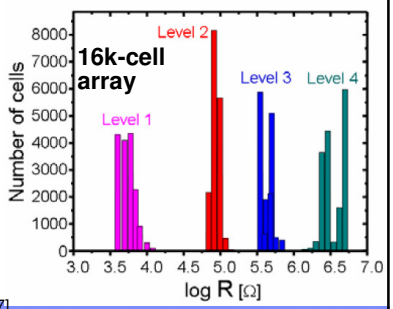
- Write and read multiple analog voltages
  → higher density at same fabrication difficulty
- Logarithm is not your friend:
  - 4 levels for 2 bits
  - 8 levels for 3 bits
  - 16 levels for 4 bits
- Coding & signal processing can help
- An iterative write scheme trades off performance for density → but useful to minimize resistance variability

MLC phase-change



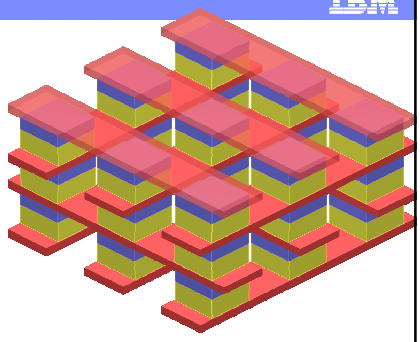






[Nirschl:2007]

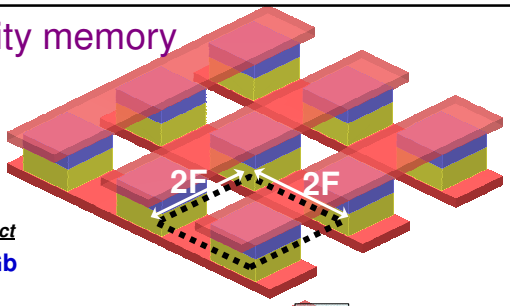
3-D stacking



- 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?)

Paths to ultra-high density memory

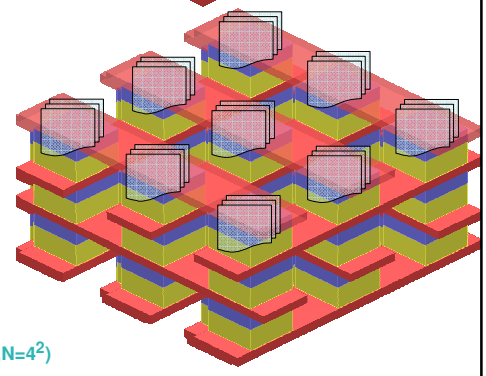
At the 32nm node in 2013, MLC NAND Flash (already M=2 → $2F^2$!) is projected* to be at…



	density	*product*
2x	43 Gb/cm² →	32Gb

if we could shrink $4F^2$ by…

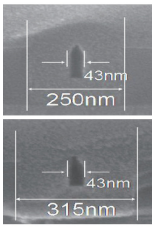
4x	86 Gb/cm² →	64Gb
	e.g., 4 layers of 3-D (L=4)	
16x	344 Gb/cm² →	256Gb
	e.g., 8 layers of 3-D, 2 bits/cell (L=8,M=2)	
64x	1376 Gb/cm² →	~1 Tb
	e.g., 4 layers of 3-D, 4x4 sublithographic (L=4,N=4^2)	



* 2006 ITRS Roadmap

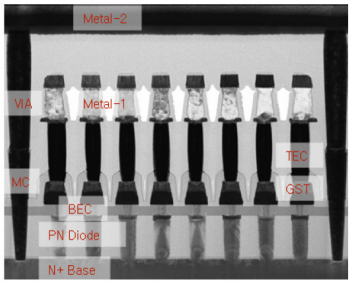
Industry SCM activities

- **Intel/ST-Microelectronics spun out Numonyx (FLASH & PCM)**
- **Samsung, Numonyx sample PCM chips**
  - –128Mb Numonyx chip (90nm) shipped in 12/08 to select customers
    Samsung started production of 512Mb (60nm) PCM in 9/09
  - –Working together on common PCM spec
- **Over 30 companies work on SCM**
  - –including all major IT players
  - –SCM research in IBM



IBM sub-litho PCM

Alverstone PCM



Samsung 512 Mbit PCM chip

For more information

Flash

- S. Lai, *IBM J. Res. Dev.*, 52(4/5), 529 (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).

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,"
IBM Journal of Research and Development, **52**(4/5), 449-464 (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., *IBM J. Res. Dev.*, **52**(4/5), 481-492 (2008).
- D. B. Strukov, et. al., *Nature*, **453**, 80(7191), 80-83 (2008).
- R. S. Williams, *IEEE Spectrum*, Dec 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).

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," *IBM Journal of Research and Development*, **52**(4/5), 465-480 (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).

FeRAM/MRAM/RRAM/SE References

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," *IBM Journal of Research and Development*, **52**(4/5), 449-464 (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).

In comparison…

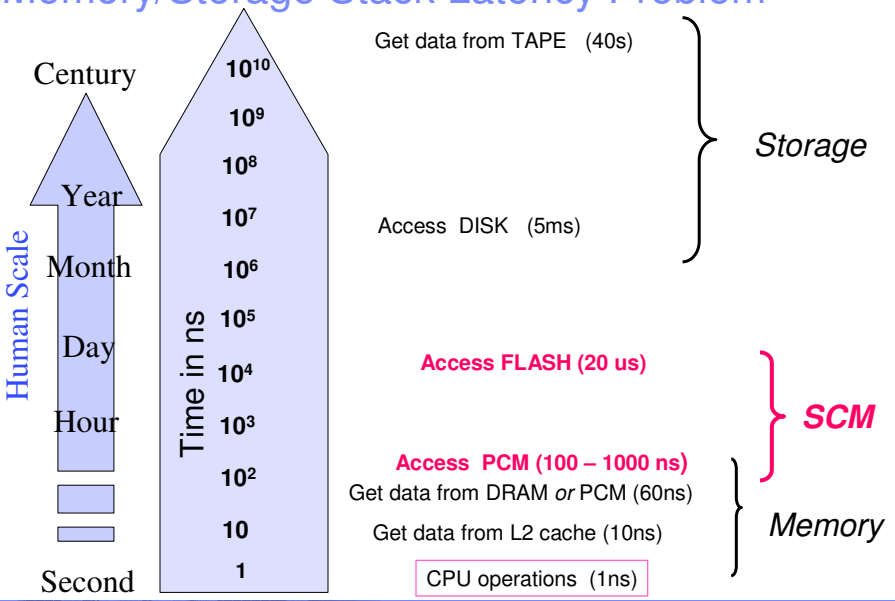
	Flash	SONOS Flash	Nanocrystal Flash	FeRAM	FeFET
Knowledge level	product	advanced development	development	product	basic research
Smallest demonstrated cell	4F^2 (2F^2 per bit)	4F^2 (1F^2 per bit)	16F^2 (@90nm)	15F^2 (@130nm)	—
Prospects for… …scalability	poor	maybe (enough stored charge?)	unclear (enough 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

Comparison continued

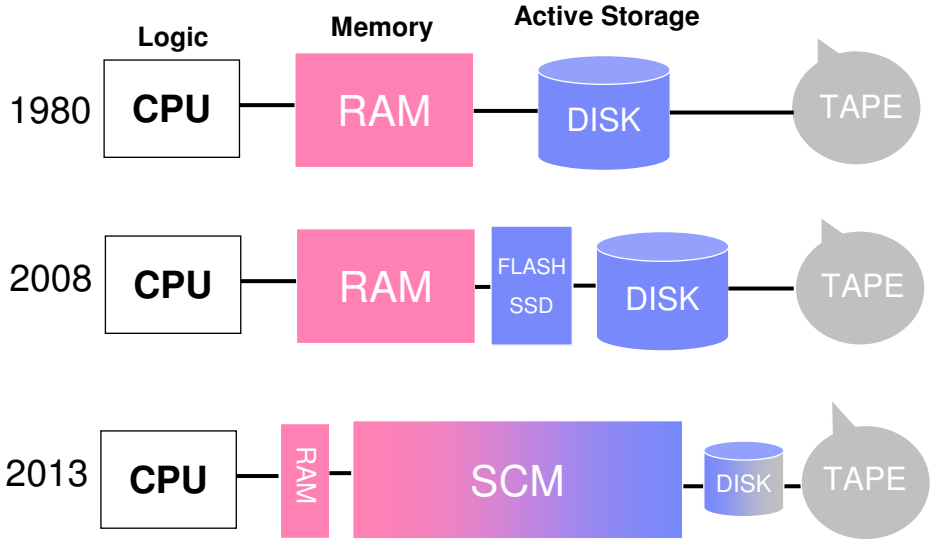
	MRAM	Racetrack	PCRAM	RRAM	solid electrolyte	organic memory
Knowledge level	product	basic research	advanced development	Early development	development	basic research
Smallest demonstrated cell	25F^2 @180nm	—	5.8F^2 (diode) 12F^2 (BJT) @90nm	—	8F^2 @90nm (4F^2 per bit)	—
Prospects for… …scalability	poor (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

Systems

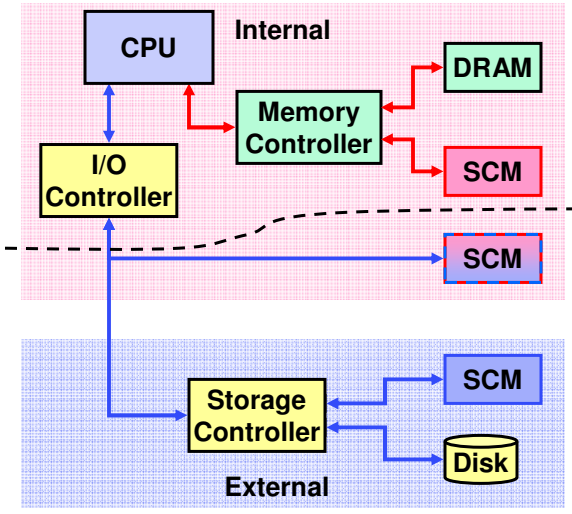
Memory/Storage Stack Latency Problem



SCM in a large System



Architecture



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

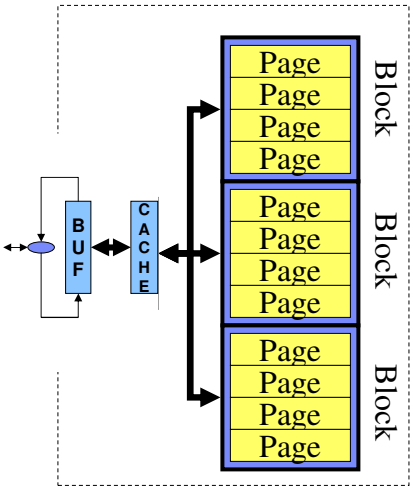
Asynchronous
- Software managed
- High overhead
- Processor doesn't wait
- Switch processes
- Flash and slow SCM
- Paging or storage

SCM Memory Classes

- **Storage device**
  - –NAND flash is current technology → eventually PCM
  - –Nonvolatile operation essential
  - –Erase vs write in place
  - –Medium speed --- Flash 20-50us, PCM for storage 1-5us est.
  - –Write endurance issues
- **Memory device**
  - –DRAM for most performance applications
  - – NOR flash for portable, etc. → PCM positioning here
    - nonvolatile operation may not be needed everywhere
    - Fast: DRAM 30-60ns, NOR 75ns, (wt very long), PCM ~75-1000ns est.
    - Write endurance issues, but not as severe
  - –Can PCM replace/augment DRAM is mainstream systems?

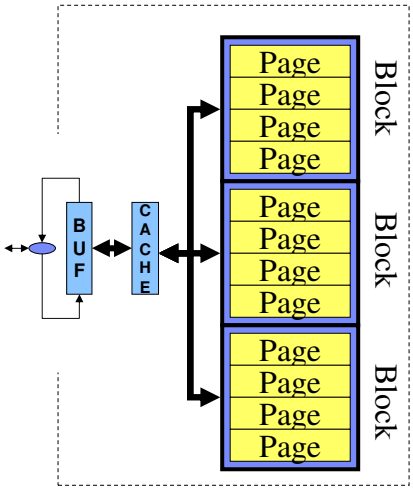
Representative NAND Flash Device

- **Power ≈ 100mW**
- **Interface: one or two bytes wide**
- **Data accessed in pages**
  - – **2112, 4224 or 8448 Bytes**
- **Data erased in blocks**
  - –Block = 64 - 128 Pages

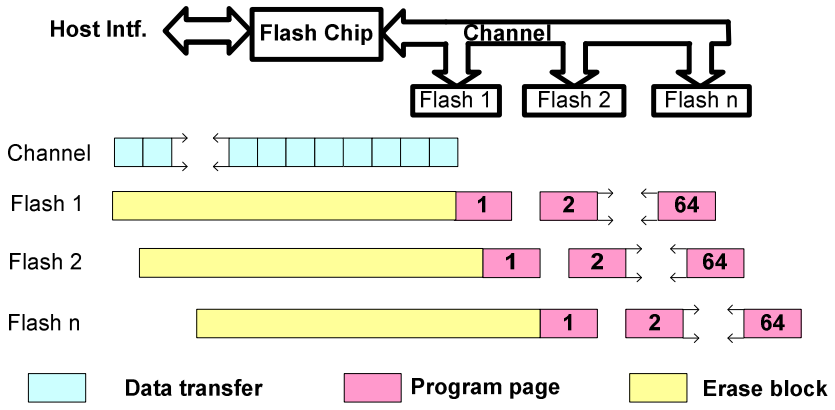


Representative NAND Flash Behavior

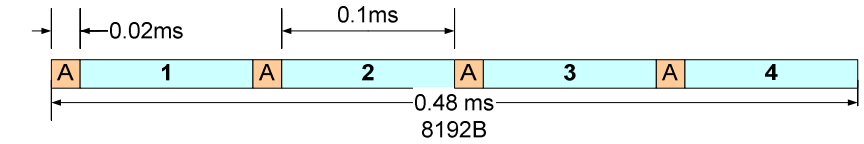
- **Read copies Page into BUF and streams data to host**
  - –Read 20 - 50 us access,
  - –20 MB/s transfer rate – sustained
  - –ONFI will take it to 200 MB/s
- **Write streams data from host into BUF**
  - –6 MB/s transfer rate sustained
  - –20 MB/s on standard bus
  - –ONFI increases this to
- **Program copies BUF into an erased 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
  - –A block must be erased before any of its pages may be programmed



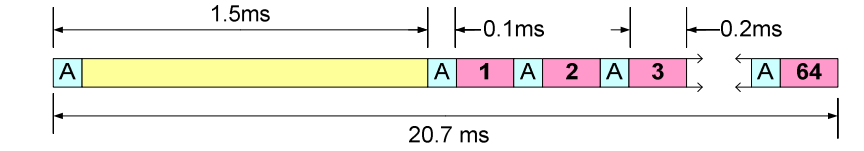
Flash Drive Channel



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

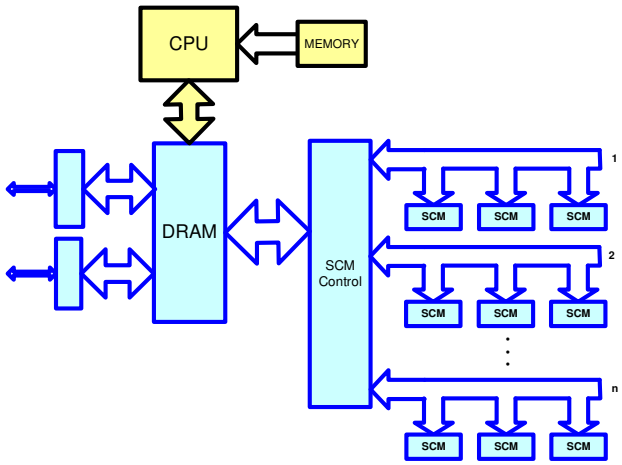
8KB Write: read 128KB, change 8KB, write 128KB → 35 IOP/s

Read Access **2 KB Data transfer** **Program page** **Erase block**

ONFI -- Open NAND Flash Interface

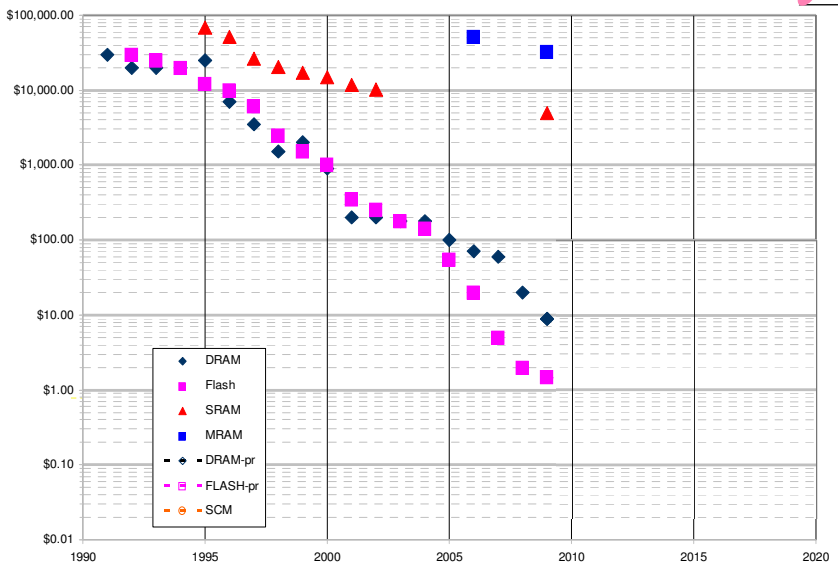
- **Industry workgroup of >80 companies who work with Flash**
- **Goal: devices interchangeable between vendors**
- **ONFI 2.1 supports up to 200MB/s R/W bandwidth**
- **Chip effectively contains multiple flash devices**
  - LUN and target addressability
  - Interleaving of commands
  - Cache register to improve read performance
- **Samsung not a member**

SCM: Generic Storage Design

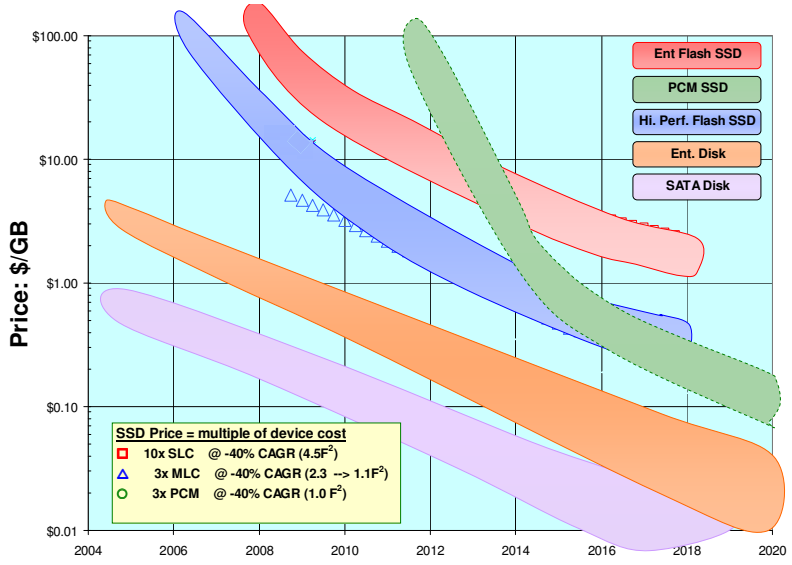


This can be the basis for a card design, an SSD design or a subsystem design.

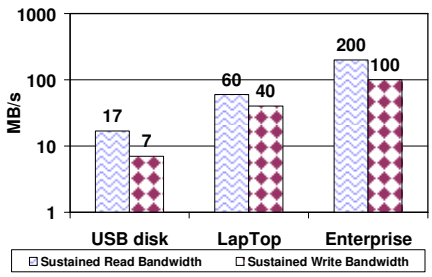
Input from the device cost crystal ball

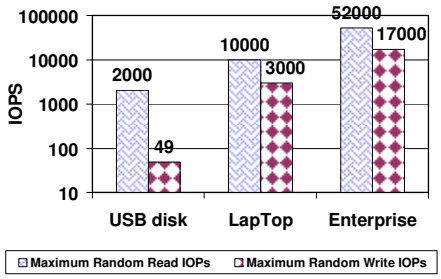


Input form the Subsystem Price Crystal Ball



Classes for Flash SSDs





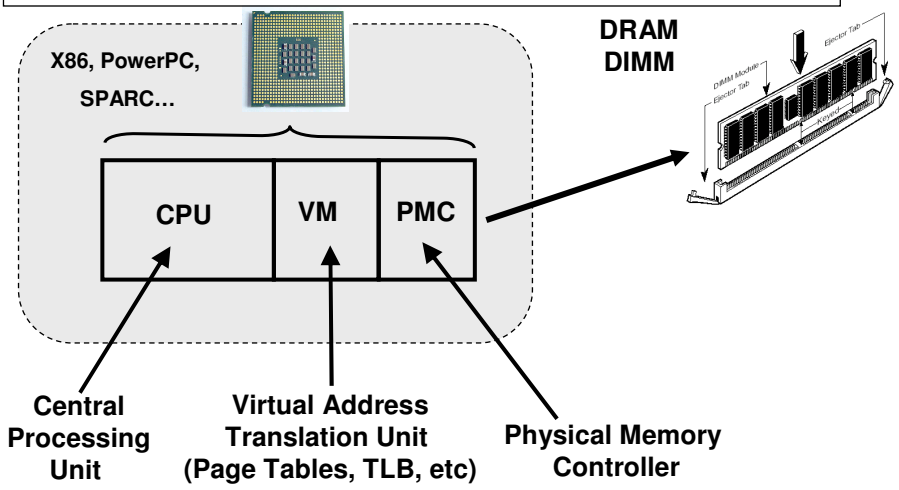
SCM-based Memory System

Logical Address > Translation > Wear Level > SCM Physical Add

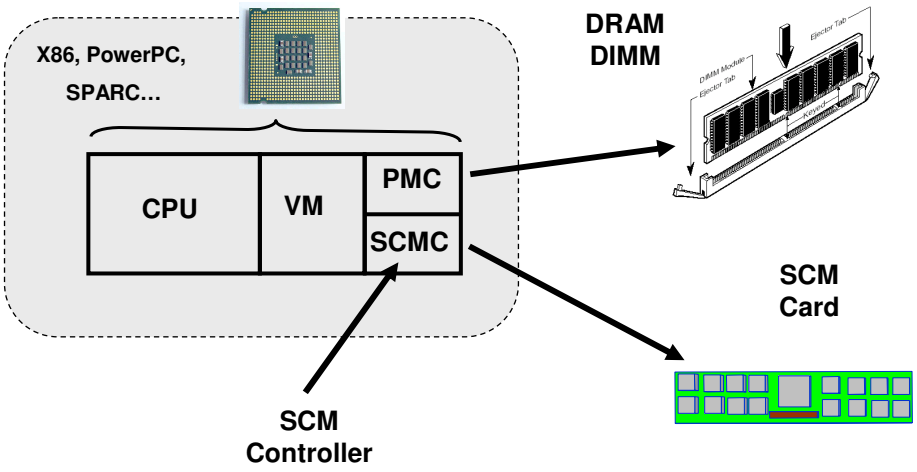
- **Treat WL as part of address translation flow**
  - –Option a – Separate WL/SCM controller
  - –Option b - Integrated VM/WL/SCM controller
  - –Option c - Software WL/Control
- **Also need physical controller for SCM**
  - –Different from DRAM physical controller

CPU & Memory System (Node) -- Today

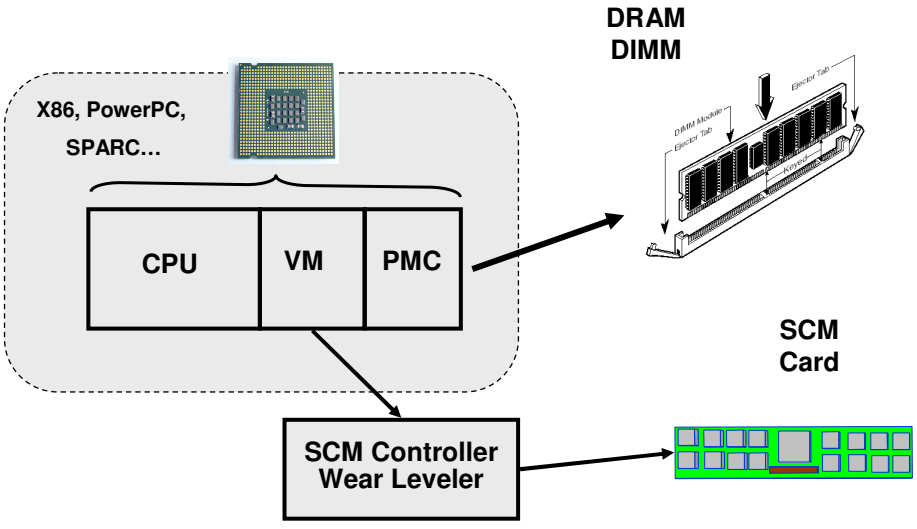
Logical Address > Address Translation > Physical Address



CPU & Memory System alternatives



CPU & Memory System alternatives



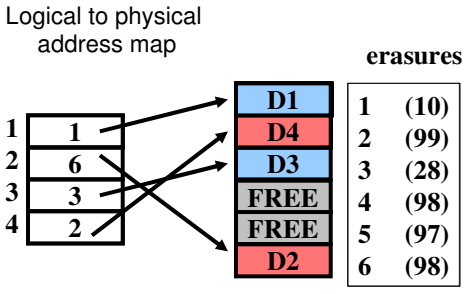
Challenges for SCM

- **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" for both memory and storage 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 writes/cell
    - Multi level cell (MLC) → 10^4 writes/cell
  - –Coping strategy → Wear-leveling, etc.

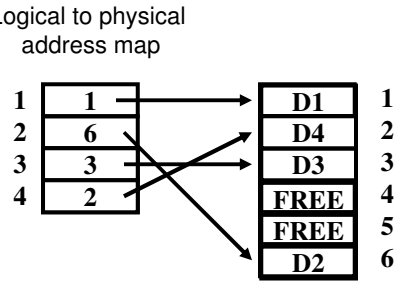
Static wear leveling

- **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**



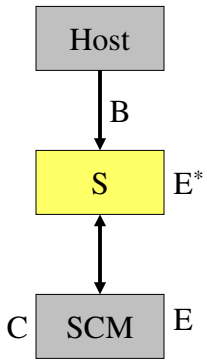
Dynamic wear leveling

- **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.**



Lifetime model (more details)

- **S are system level management 'tools' providing an effective endurance of $E^* = S(E)$**
  - 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
  - Over-provisioning
  - Compress, de-duplicate & write elimination…
  - $E^* = E * q * f(\text{error correction}) * g(\text{overprovisioning}) * h(\text{compress})\ldots$
  - With S included, $T_{life}(\text{System}) = T_{fill} * E^*$



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**
- **Simple wear leveling → each write is done to a new (empty) location**
  - Data unit is the smallest item that can be written/erased
  - Memory unit is the size of the largest item that can be wear-leveled

	DRAM	Disk	256GB Flash		8 GB SCM
Endurance	$>10^{16}$	$>10^{11}$	$10^5 \rightarrow 10^4$		10^8
Wear-eveled	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

Summary

- **There are a number of solid state memory technologies competing with DRAM and Disk**
  - Flash and PCM are the current leaders
- **An inexpensive nonvolatile memory with medium speed (1 – 50 us) will change the storage hierarchy**
- **An inexpensive memory with speed near DRAM will change the memory hierarchy**
  - Such a memory that is also nonvolatile will enable new areas
- **Write endurance is an issue for many of these technologies, but there are techniques to cope with it**

- Questions?

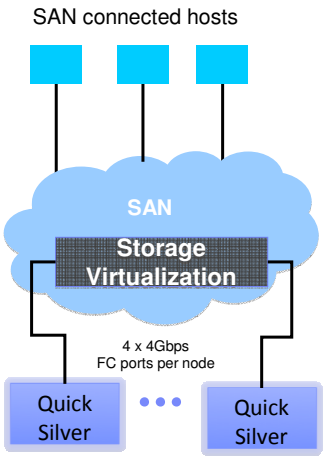
- Break Time

Performance

IBM QuickSilver Project 2008 → 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

87



SAN: Storage Area Network

SVC: San Volume Controller

QuickSilver Headlines in the Press (August 2008)

- Network World - *IBM flash memory breaks 1 million IOPS barrier*
  - "Flash storage is starting to catch on with enterprise customers as such vendors as EMC promise faster speeds and more efficient use of storage with solid-state disks. Speeds are typically orders-of-magnitude lower than what IBM is claiming to have achieved."
- Information Week - *IBM Plans Breakthrough Solid-State Storage System 'Quicksilver'*
  - "Compared to the fastest industry benchmarked disk system, the new technology had less than 1/20th the response time. In addition, the solid-state system took up 1/5th the floor space and required 55% of the power and cooling."
- Bloomberg - *IBM Breaks Performance Records through Systems Innovation*
  - "IBM has demonstrated, for the first time, the game-changing impact solid-state technologies can have on how businesses and individuals manage and access information."

Understanding Flash based SSD performance

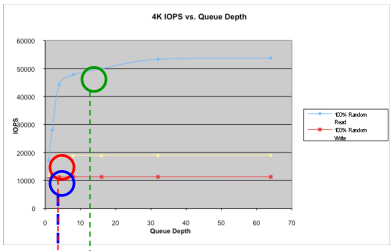
- Flash media can only do one the following three things: Read, Erase, Program
- IO Read -> Flash Read, IO Write -> Flash Erase and Flash Program
- Erase cycle is very time consuming (in msec)
- Major latency difference for IO Read operation (50usec) versus IO Write (100+usec) operation
- Flash based SSD device requires storage virtualization to deal with undesirable flash properties, erase latency and wear-leveling.
- Storage virtualization techniques typically used are : Relocate on write, batch write operation and , over provisioning.

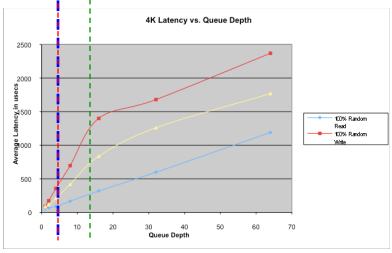
Vendor A SSD – IOPS and Latency

Sequential Precondition

Optimal IOPS			
R/W	IOPS	Latency - usecs	Queue Depth
100/0	47810	165	8
0/100	11316	85.8	1
50/50	17089	113	2

Minimal Latency			
R/W	IOPS	Latency - usecs	Queue Depth
100/0	17221	56.9	1
0/100	11316	85.8	1
50/50	11776	83	1



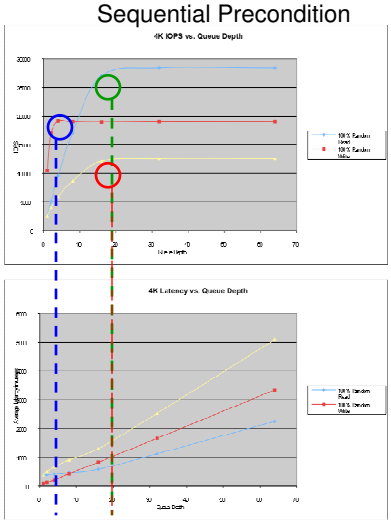


Vendor B SSD – IOPS and Latency

Optimal IOPS			
R/W	IOPS	Latency - usecs	Queue Depth
100/0	27048	583	16
0/100	19095	209	4
50/50	12125	1300	16

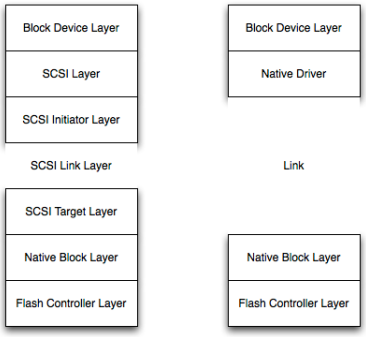
Minimum Latency			
R/W	IOPS	Latency - usecs	Queue Depth
100/0	2583	386	1
0/100	10525	92.7	1
50/50	2567	388	1

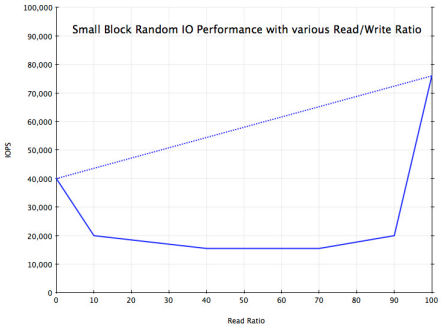
Sequential Precondition



Understanding Flash Based SSD Performance

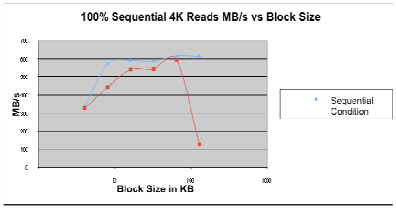
- **Latency model changes based on different storage hardware and software architecture.**
- **Read OPs are 2x+ comparing to Write Ops**



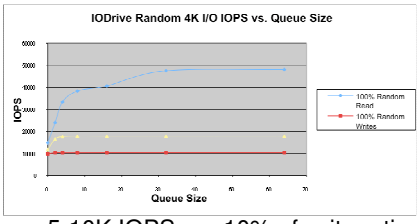


Read Write Bathtub Performance

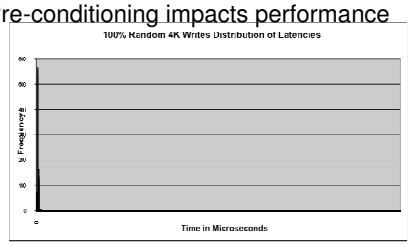
SSD Architecture is Sensitive to Writes



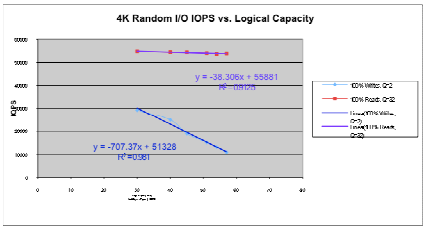
Pre-conditioning impacts performance



Lose 5-10K IOPS per 10% of write ratio



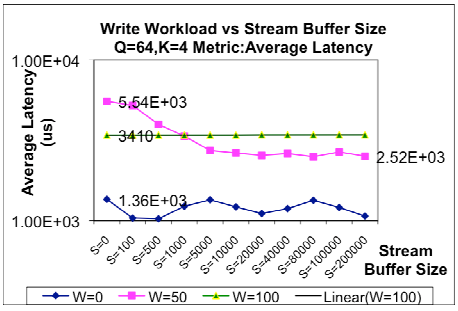
0.05% of write latencies > 45 ms

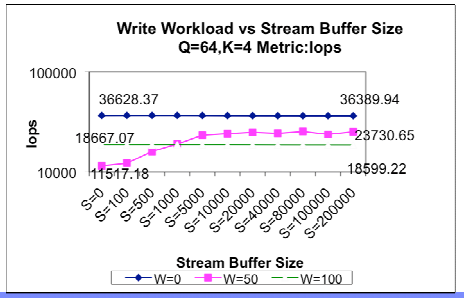


Decrease logical capacity for better write performance

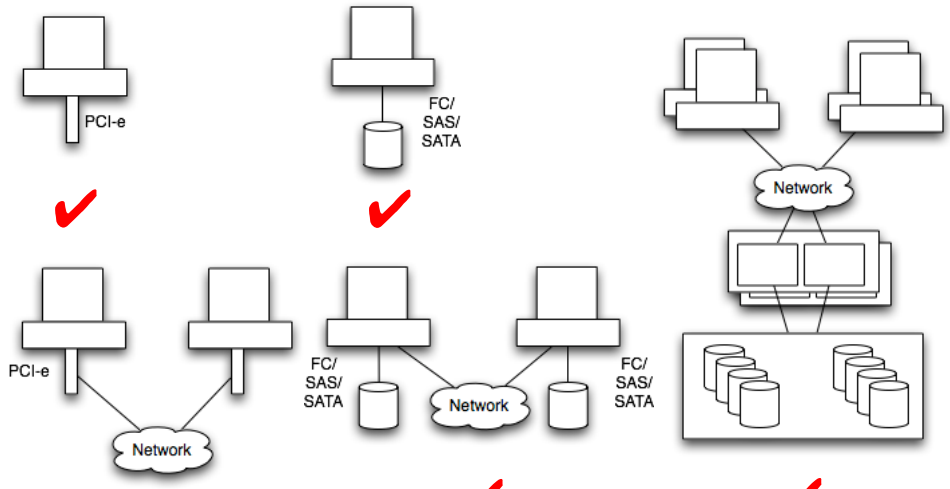
Improve Read/Write Bathtub Performance Characteristics

- **Developed IO stream shaping algorithm to improve SSD R/W Bathtub Performance behavior.**
- **Improve IOPS of Vendor B under stress in mixed R/W environment from 11500 IOPS to 23730 IOPS.**
- **Improve latency of Vendor B under stress in mixed R/W environment from 5.5 msec to 2.52 msec.**



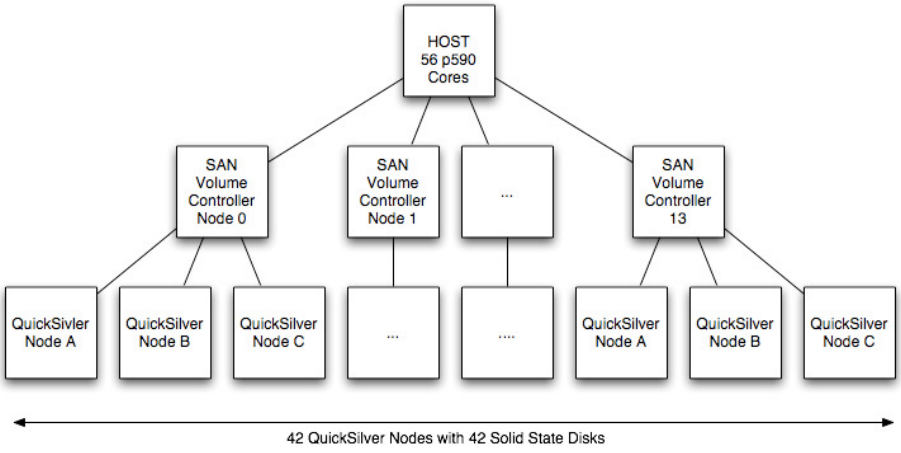


System Topology Using SSD

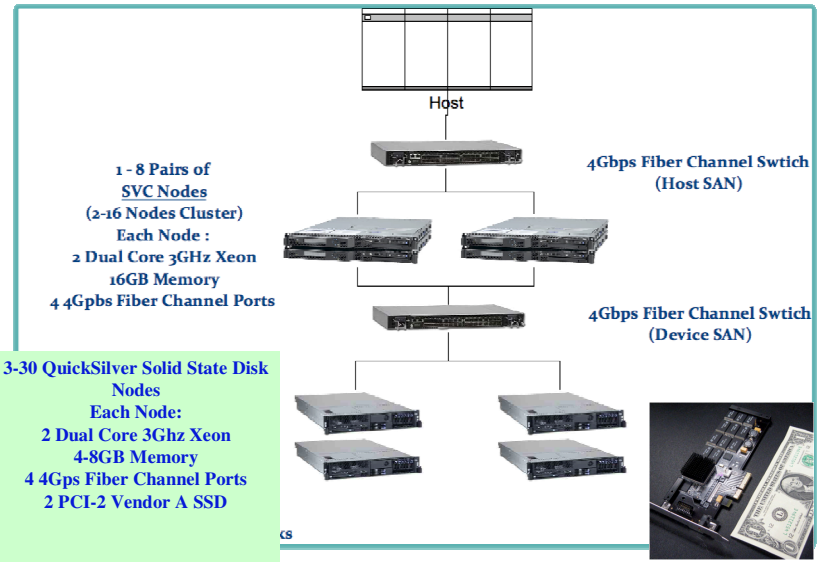


✔ = Used in Quicksilver Evaluation

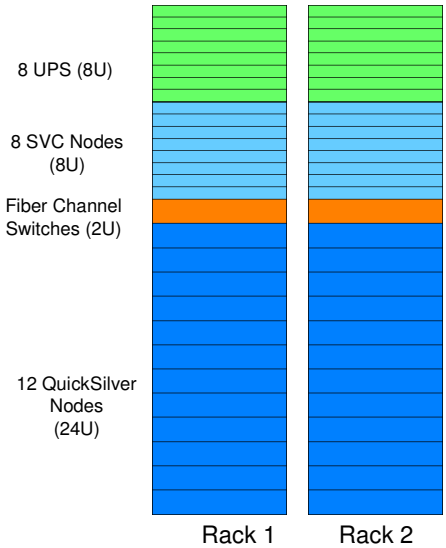
Logical Configuration View of QuickSilver 2008



Logical and Physical Configuration of QuickSilver SSD Node



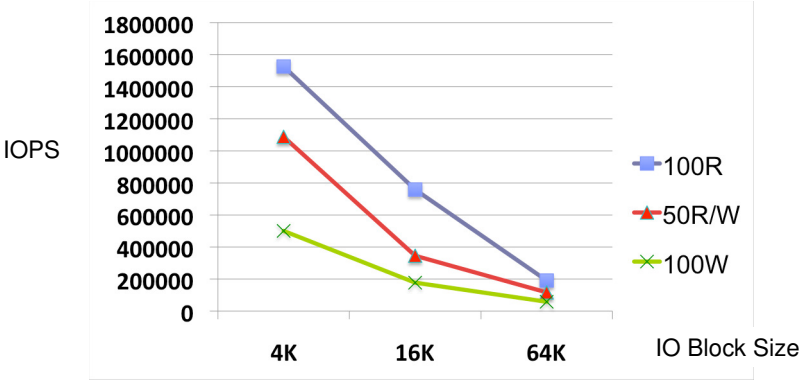
Rack Configuration as of QuickSilver 2008



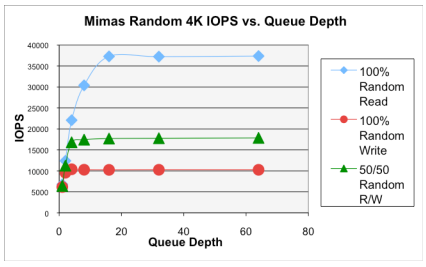
	Entry Configuration	Maximum Configuration
Description	2-nodes SVC 3-nodes QuickSilver 2 UPS	16-nodes SVC 24-nodes QuickSilver 16 UPS
Host Interface	4Gbps Fiber Channel x 8	4Gbps Fiber Channel x 56-64
Capacity	1.2TB (SLC) 2.4TB (MLC)	4.8TB (SLC) 9.6TB (MLC)
IOPS Performance	180KIOIPS 3.2GB/s	1.44 MIOPS+ 25.6GB/s
IO Latency	~500-700usec	
Physical attribute	10U (+4U FC Switch)	80U (+4U FC Switch)

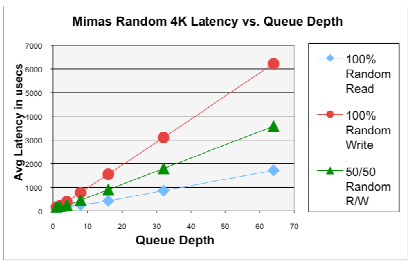
Microbenchmark on QuickSilver Cluster (46 Vendor A Cards)

Random IOPS	4K	16K	64K
100R	1526330	758593	190898
50/50 RW	1087675	345763	116967
100W	500636	177767	58606

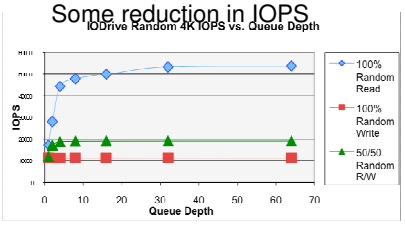


QuickSilver and Vendor A: QuickSilver Adds Overhead as Expected

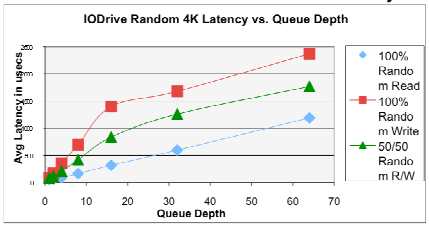




Some reduction in IOPS



Additional HW and SW add Latency



This IOPS is not equal to that IOPS

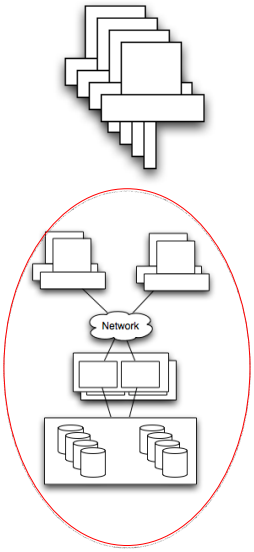
- **Low latency -> High IOPS**
  - You work faster -> You work more per unit time
- **Parallelism -> High IOPS**
  - More of you work -> More work is done per unit time

SSD Evaluation Service

- **Micro benchmarks**
  - Measure the performance of focused benchmarks:
    - Average IOPS for block size = 4k, queue depth=16, etc. e.g.
  - Metrics: latency → bandwidth and IOPS
  - Reports: hot spots, latency distribution, etc.
- **System benchmarks**
  - Measure performance storage system workloads: e.g., SPC-1
  - Metrics: sustained performance, etc.
- **Application benchmarks**
  - Measure performance of application workloads: TPC-C, etc.
  - Metrics: $/TPMC, etc.

Application

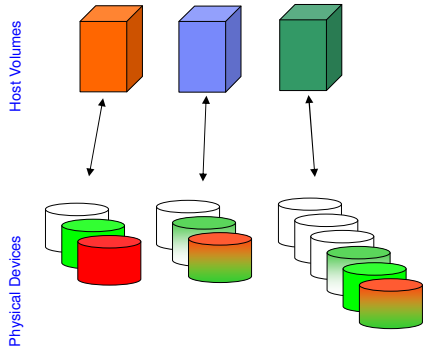
System Topology Using SSD



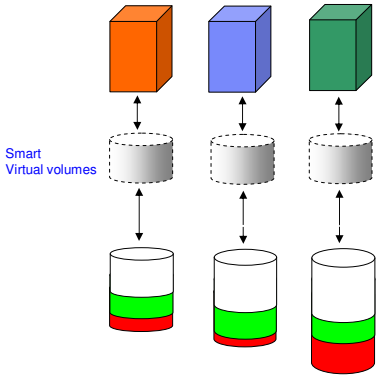
- **Purposed built Storage System using SSD**
  - Single purpose, well understood workload, high performance
  - Eg. Financial Advanced Trading Desk (Algorithmic Trading Desk)
  - Challenges : Balanced performance, cost, reliability and availability.
- **General Purpose Storage System**
  - Quickest time to market approach
  - Focus on consumability
  - Mixed SSD with HDD types in multiple tiered storage system.
  - Eg. Database workload, Batch processing
  - Challenges : Find the right balance between automation and policy based data placement.

Traditional Disk Mapping vs. *Smart* Storage Mapping

Traditional disk mapping



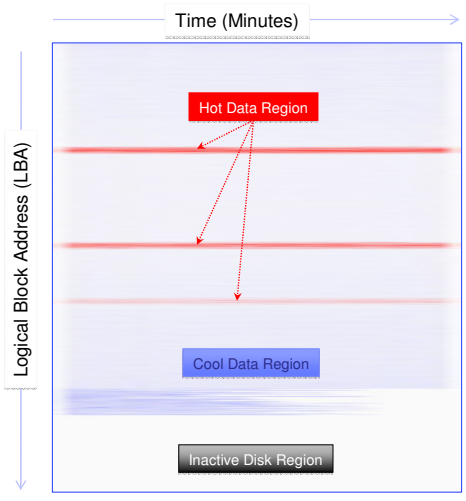
Smart storage mapping



Volumes have different characteristics. Applications need to place them on correct tiers of storage based on usage.

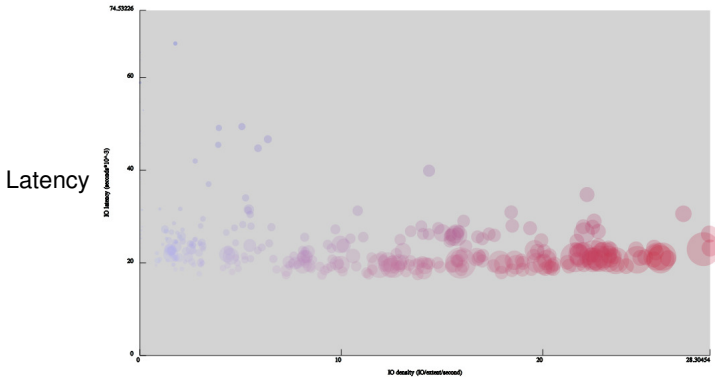
All volumes appear to be "logically" homogenous to apps. But data is placed at the right tier of storage based on its usage through data placement and migration.

Workload Learning



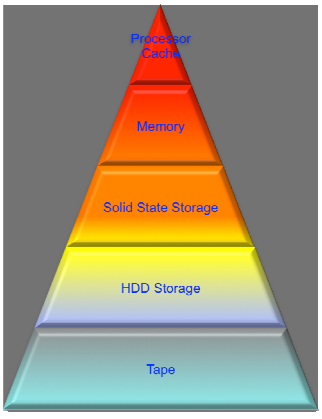
- **Each workload has its unique IO access patterns and characteristics over time.**
- **Heatmap will develop new insight to application optimization on storage infrastructure.**
- **Left diagram shows historical performance data for a LUN over 12 hours.**
  - Y-axis (Top to bottom) LBA ranges
  - X-axis (Left to right) time in minutes.
- **This workload shows performance skews in three LBA ranges.**

LUN Based Latency Density Chart



- **Each circle represents an IO (density,latency) point. The size of the circle is proportional to the number of IOs at that point. Color simply corresponds to location in the coordinate space; it has no additional meaning. Each circle has the same opacity.**

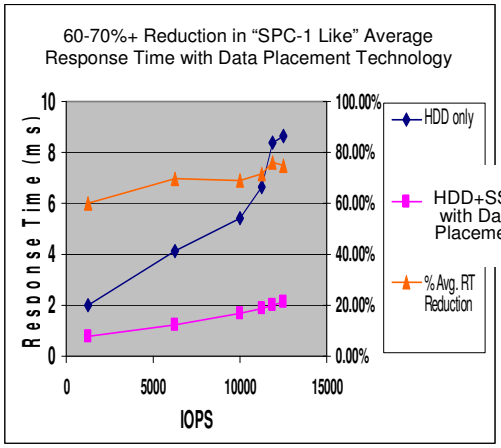
Data Placement



Memory Storage Hierarchy

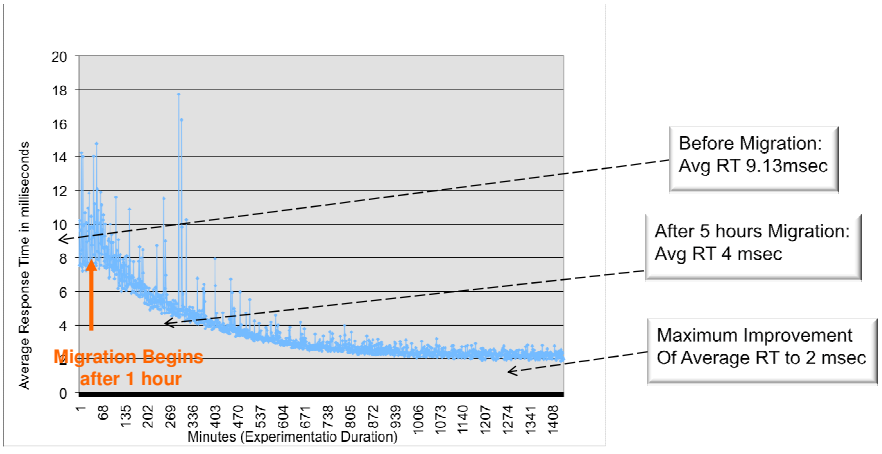
- **Data Placement : Key technology to improve consumability of SSD in a multiple tiered storage system.**
  - Automatically optimize storage performance within enterprise storage system to improve application demand or system configuration.
  - Maximize performance / cost benefit.
- **Highly dependent on workload characteristics**
  - Transaction Processing oriented
  - Analytics Workload
  - Daily Business Process
- **Leverage Memory Storage Hierarchy**
  - Assuming every layer, system hardware and software are optimized for best usage of resource.

Demonstration of Data Placement Technology on IBM Enterprise Storage System

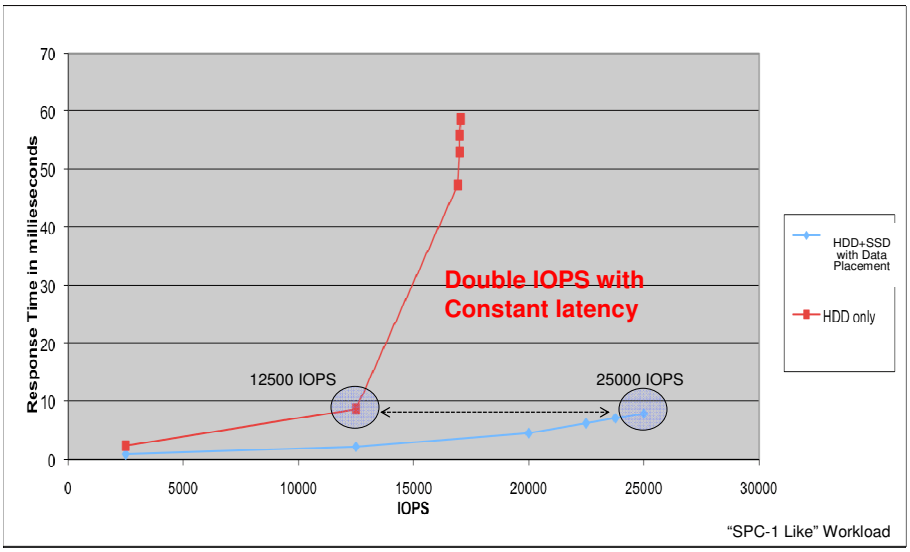


- Setup:
  - Single Enterprise Storage System with both HDD and SSD ranks. About 5-6% capacity is in SSD ranks.
- Demonstration of Data Placement:
  - Compare "SPC-1 like" workload on HDD versus "Data placement of HDD and SSD"
  - Data Placement Technology identifies and non-disruptively migrates "hot data" from HDD to SSD. About 4% of data is migrated from HDD to SSD.
- Result:
  - Response time reduction of 60-70+% at peak load
    - Sustainability test, 76%
    - Ramp test, 77%

Average Response Time Shows Significant Improvement with Data Placement and Migration Technology

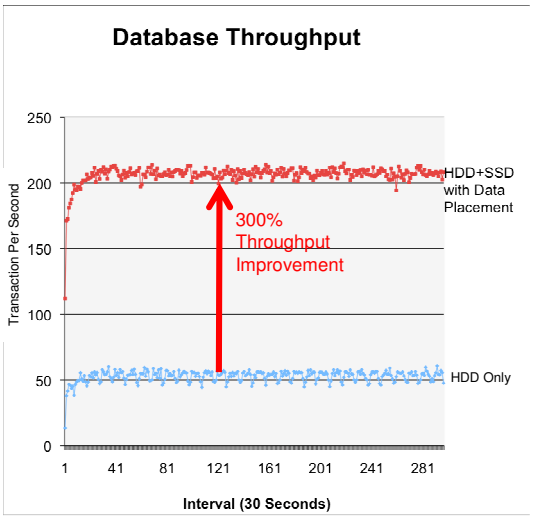


Workload Performance IOPS Doubled with Data Placement

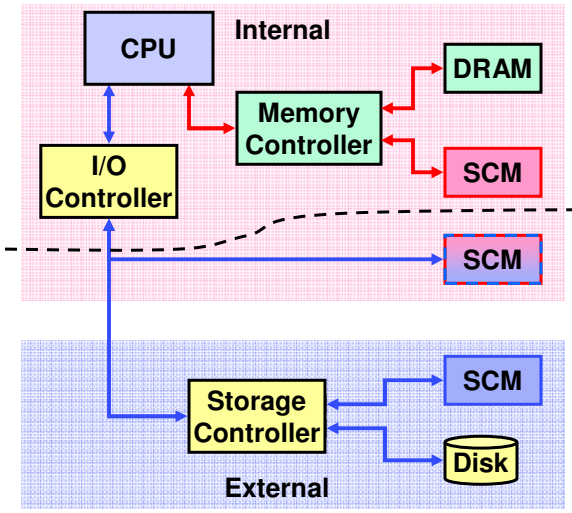


Brokerage Workload using DB2 and Data Placement

- **Identify hot "database objects" and placed in the right tier.**
- **Scalable Throughput**
  - 300% throughput improvement.
- **Substantial IO Bound Transaction Response time Improvement**
  - 45%-75% response time improvement.



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

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 – Fast boot and hibernate – New persistency model will emerge.
Storage:	– Active data on disk – Inactive data on tape – SANs in heavy use	– Active data on SCM – Inactive data on disk/tape – Direct Attached Storage Model and Shared Storage Model
Applications:	– Compute centric – Focus on hiding disk latency	– Data centric comes to fore – Focus on efficient memory use and exploiting persistence – Fast, persistent metadata

Paths Forward for SCM

- **Storage**
  - Direct disk replacement with an NAND Flash (SCM) packaged as a SSD
  - PCIe card that supports a high bandwidth local or direct attachment to a processor.
  - Design the storage system or the computer system around Flash or SCM from the start
- **Memory**
  - Possible positioning in the memory stack
  - Paging

SCM impact on software (Present to Future)

- **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

Issues with persistent memory

- **Shared state maintenance**
  - Storage difficult to corrupt, must set up a write operation
  - Directly mapped storage easily corrupted
  - Corrupted state is persistent
- **Memory pool management**
  - Complex management task
  - Fixed or “Virtually Fixed” allocation
  - Addressability
- **SCM Media Failure**
  - Bad block and Wearout
  - Complex recovery scenario in typical memory management model

Implications on Traditional Commercial Databases

- **Initial SCM in DB uses:**
  - Logging (for Durability)
  - Buffer pool

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

- **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)

PCM as Logging Store – Permits > Log Forces/sec?

- Obvious one but options exist even for this one!
- Should log records be written directly to PCM or

  first to DRAM log buffers and then be forced to PCM (rather than disk)
- In the latter case, is it really that beneficial if ultimately you still want to have log on disk since PCM capacity won't be as much as disk – also since disk is more reliable and is a better long term storage medium
- In former case, all writes will be way slowed down!

PCM replaces DRAM? - Buffer pool in PCM?

- This PCM BP access will be slower than DRAM BP access!
- Writes will suffer even more than reads!!
- Should we instead have DRAM BPs backed by PCM BPs?

  This is similar to DB2 z in parallel sysplex environment with BPs in coupling facility (CF)

  But the DB2 situation has well defined rules on when pages move from DRAM BP to CF BP
- Variation was used in SafeRAM work at MCC in 1989

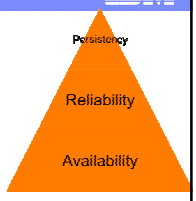
Assume whole DB fits in PCM?

- Apply old main memory DB design concepts directly?
- Shouldn't we leverage persistence specially?
- Every bit change persisting isn't always a good thing!
- Today's failure semantics lets fair amount of flexibility on tracking changes to DB pages – only some changes logged and inconsistent page states not made persistent!
- Memory overwrites will cause more damage!
- If every write assumed to be persistent as soon as write completes, then L1 & L2 caching can't be leveraged – need to do write through, further degrading performance.

Assume whole DB fits in PCM? …

- Even if whole DB fits in PCM and even though PCM is persistent, still need to externalize DB regularly since PCM won't have good endurance!
- If DB spans both DRAM and PCM, then
  - need to have logic to decide what goes where – hot and cold data distinction?
  - persistency isn't uniform and so need to bookkeep carefully

Data Availability and PCM



- **What about data availability model with PCM?**
  - Reliability, Recoverability and Availability
- **If PCM is used as permanent and persistent medium for data, what is the right kind of reliability model? Is memory failure detection and recovery sufficient?**
- **If PCM is used as memory and its persistence is taken advantage of, then such a memory should be dual ported (like for disks) so that its contents are accessible even if the host fails for backup to access**
- **Should locks also be maintained in PCM to speed up new transaction processing when host recovers**

What about Logging?

- If PCM is persistent and whole DB in PCM, do we need logging?
- Of course it is needed to provide at least partial rollback even if data is being versioned (at least need to track what versions to invalidate or eliminate); also for auditing, disaster recovery, …

Start from Scratch?

- **Maybe it is time for a fundamental rethink**
- **Design a DBMS from scratch keeping in mind the characteristics of PCM**
- **Reexamine data model, access methods, query optimizer, locking, logging, recovery, …**

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**

Q & A