Agenda

- **A: Introduction** (Winfried Wilcke) : 20 min
- **B: Technology of Storage Class Memory** (Bülent Kurdi) :60
  - Question period: 10 min
  - Break: 30 min
- **C: Phase Change Memory** (Geoffrey Burr) : 30 min
- **D: Systems and Applications** (Rich Freitas) : 50 min
  - Question period: 10 min
Definition of **Storage Class Memory** (SCM)

- A new class of data storage/memory devices
  - many technologies compete to be the ‘best’ SCM
- **SCM blurs the distinction between**
  - MEMORY (fast, expensive, volatile) and
  - STORAGE (slow, cheap, non-volatile)

**SCM features:**

- Non-volatile
- Short Access times (~ DRAM like)
- Low cost per bit more (DISK like – by 2020)
- Solid state, no moving parts

Some Terminology Clarification

- **SCM = Storage Class Memory**
  - SCM describes a *technology*, not a *use*
  - FLASH isn’t quite SCM (can’t be used as memory)
- **NVRAM = Non Volatile RAM**
  - SCM is one example of NVRAM
  - Other NVRAM types: DRAM+battery or DRAM+disk combos
- **SSD = Solid State Disk**
  - Use of NVRAM for *block oriented* storage applications
Industry SCM activities

- SCM research in IBM
  - approximately 50 persons
- Intel/ST-Microelectronics spun out Numonyx (FLASH & PCM)
- Samsung, Numonyx sample PCM chips
- Over 30 companies work on SCM
  - including all major IT players

System Targets for SCM

- Mobile
- Desktop
- Datacenter

Over 30 companies work on SCM, including all major IT players. Samsung, Numonyx sample PCM chips. Intel/ST-Microelectronics spun out Numonyx (FLASH & PCM). SCM research in IBM, approximately 50 persons.
SCM in mobile devices: straightforward uses

- Functional replacement for FLASH
- Solid State Disks replace magnetic, rotating disks
- Cost, Power & Ruggedness are most critical

The Memory/Storage Bottleneck

<table>
<thead>
<tr>
<th>SLOW</th>
<th>T [ns]</th>
<th>Read or Write from TAPE (40s)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$10^{10}$ century</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$10^{9}$ decade</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$10^{8}$ year</td>
<td>Read or write to DISK (5ms)</td>
</tr>
<tr>
<td></td>
<td>$10^{7}$ month</td>
<td>Write to FLASH, random (1 ms)</td>
</tr>
<tr>
<td></td>
<td>$10^{6}$ week</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$10^{5}$ day</td>
<td>Read a FLASH device (20 us)</td>
</tr>
<tr>
<td></td>
<td>$10^{4}$ hour</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$10^{3}$ minute</td>
<td>Read/Write PCM (100 – 1000 ns)</td>
</tr>
<tr>
<td></td>
<td>$10^{2}$ second</td>
<td>Get data from DRAM (60ns)</td>
</tr>
<tr>
<td></td>
<td>1 second</td>
<td>CPU operations – e.g. ADD (1ns)</td>
</tr>
</tbody>
</table>

FAST

$T \times 10^9$

Storage

SCM

Memory
System Evolution

1980

Logic

CPU → RAM → DISK → TAPE

Memory

fast, synch

1980

Active Storage

slow, asynch

1980

Archival

2008

CPU → RAM → DISK → TAPE

2008

Logic

CPU → RAM → FLASH SSD → DISK → TAPE

Logic

2013+

CPU → RAM → SCM → DISK → TAPE

2013+

Logic

Speed/Volatility/Persistency Matrix

FAST

(Memory)

DRAM

DRAM + SCM

DRAM + SCM & System Architecture

SLOW

(Storage)

Enterprise storage server

Persistent storage will not lose data

Volatile

Non-Volatile

Persistent

USB stick

PC disk

Enterprise storage server

Persistent storage will not lose data
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 scalability
- Solid Electrolyte and resistive RAM (Memristor)
  - early development, maybe promising
- Organic, nano particle and polymeric RAM
  - many different devices in this class, unlikely
- Improved FLASH
  - still slow and poor write endurance

SCM Design Triangle:

- Speed
- Memory-type Uses
- (Write) Endurance
- Storage-type Uses
- Cost/bit
- Power!
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** [°C]
- **Radiation resistance** [Rad]

~ 16 criteria! This makes the SCM problem so hard
Which device technology will be King of SCM?
The Technology of Storage Class Memory

Towards a disruptively low-cost solid-state non-volatile memory

Geoffrey W. Burr
Bülent Kurdi
IBM Almaden Research Center

February 24, 2009
Tutorial T3
parts B & C

Outline

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

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

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

![Graph showing spending on new server, power and cooling over years](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


Storage Class Memory

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

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

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

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

More about HDD

- 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/in² → ~4 TB

- patterned media

Wikipedia

610 Gb/in²

~4 TB
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...?

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

Can HDD & Flash improve enough to help?

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

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

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

<table>
<thead>
<tr>
<th>Device</th>
<th>Critical feature-size F</th>
<th>Area ($F^2$)</th>
<th>Density (Gbit/sq. in)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hard Disk</td>
<td>50 nm (MR width)</td>
<td>1.0</td>
<td>250</td>
</tr>
<tr>
<td>DRAM</td>
<td>45 nm (half pitch)</td>
<td>6.0</td>
<td>50</td>
</tr>
<tr>
<td>NAND (2 bit)</td>
<td>43 nm (half pitch)</td>
<td>2.0</td>
<td>175</td>
</tr>
<tr>
<td>NAND (1 bit)</td>
<td>43 nm (half pitch)</td>
<td>4.0</td>
<td>87</td>
</tr>
<tr>
<td>Blue Ray</td>
<td>210 nm ($\lambda/2$)</td>
<td>1.5</td>
<td>10</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th></th>
<th>NOR</th>
<th>NAND</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell Size</td>
<td>9-11 $F^2$</td>
<td>2 $F^2$</td>
</tr>
<tr>
<td></td>
<td>(4 $F^2$ physical x 2-bit MLC)</td>
<td></td>
</tr>
<tr>
<td>Read</td>
<td>100 MB/s</td>
<td>18-25 MB/s</td>
</tr>
<tr>
<td>Write</td>
<td>&lt;0.5MB/sec</td>
<td>8MB/sec</td>
</tr>
<tr>
<td>Erase</td>
<td>750msec</td>
<td>2ms</td>
</tr>
<tr>
<td>Market Size (2007)</td>
<td>$8B$</td>
<td>$14.2B$</td>
</tr>
<tr>
<td>Applications</td>
<td>Program code</td>
<td>Multimedia</td>
</tr>
</tbody>
</table>

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**

Charge trapping in SiN trap layer

**SONOS**

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

**TaNOS**

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

---

**NAND Scaling Road Map**

- 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
For more information (on HDD & Flash)

**HDD**

**Flash**

Can HDD & Flash improve enough to help?

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

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

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

SCM system requirements for Memory (Storage) apps

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

SCM device requirements

- Desired attributes
  - high performance (>1 GB/sec data rate, < 200nsec access time)
  - low active & standby power (100mW ON power, 1mW standby)
  - high read/write endurance ($10^9$ – $10^{12}$ cycles)
  - non-volatility
  - compatible with existing technologies
  - continuously scalable
  - lowest cost per bit (target: cost of Enterprise HDD)
Landscape of existing technologies

- NOR FLASH
- NAND FLASH
- DRAM
- HDD
- Cost
- Performance

Memory/storage landscape

- PCRAM
- MRAM
- FRAM
- NOR
- NAND
- ROM
- SRAM
### Emerging Memory Technologies

**Memory technology remains an active focus area for the industry**

<table>
<thead>
<tr>
<th>FLASH Extension</th>
<th>FRAM</th>
<th>MRAM</th>
<th>PCRAM</th>
<th>RRAM</th>
<th>Solid Electrolyte</th>
<th>Polymer/Organic</th>
</tr>
</thead>
<tbody>
<tr>
<td>64Mb FRAM (Prototype)</td>
<td>0.13um 3.3V</td>
<td>4Mb MRAM (Product)</td>
<td>0.18um 3.3V</td>
<td>512Mb PRAM (Prototype)</td>
<td>0.1um 1.8V</td>
<td>4Mb C-RAM (Product)</td>
</tr>
<tr>
<td>64Mb FRAM (Prototype)</td>
<td>0.13um 3.3V</td>
<td>4Mb MRAM (Product)</td>
<td>0.18um 3.3V</td>
<td>512Mb PRAM (Prototype)</td>
<td>0.1um 1.8V</td>
<td>4Mb C-RAM (Product)</td>
</tr>
</tbody>
</table>

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

![Bar chart showing the number of papers presented per year from 2001 to 2007.](image)

**Organic:**
- 10 papers

**Solid Electrolyte:**
- 15 papers

**RRAM:**
- 20 papers

**PCram:**
- 25 papers

**MRAM:**
- 30 papers

**FeRAM:**
- 35 papers

**SONOS Flash:**
- 40 papers

**nanocrystal Flash:**
- 45 papers

**Flash:**
- 50 papers

---

**Research interest**

- IBM
- Infineon
- Freescale
- Mitsubishi
- Samsung
- Toshiba
- NEC
- Fujitsu
- Philips
- Renesas
- Spansion
- Macronix
- BAE
- Intel
- STMicro
- Elpida
- IBM
- Hitachi
- Philips
- Samsung
- Hynix
- Matsushita
- HP
- Cypress
- Oki
- Rohm
- Cypress
- Fujitsu
- Infineon
- Hynix
- Samsung
- TSMC
- Infineon
- TI
- NEC
- Toshiba
- Spansion
- Matsushita
- Fujitsu
- Seiko Epson

---

**Emerging Memory Technologies**

- Trap Storage
- Saifun NROM
- Tower
- Spansion
- Infineon
- Macronix
- Samsung
- Toshiba
- Spansion
- Macronix
- NEC
- Nano-CM
- Freescale
- Matsushita

---

**Technology**

- 4Mb C-RAM (Product)
- 0.25um 3.3V
- 512Mb PRAM (Prototype)
- 0.1um 1.8V
- 4Mb MRAM (Product)
- 0.18um 3.3V
- 64Mb FRAM (Prototype)
- 0.13um 3.3V
Industry interest in non-volatile memory

2001

2006

www.itrs.net

Candidate device technologies

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

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

Candidate device technologies

- **Improved Flash**
  - little change expected in write endurance or speed

- **FeRAM** (Ferroelectric RAM)
  - FeFET

- **MRAM** (Magnetic RAM)
  - Racetrack memory

- **RRAM** (Resistive RAM)
  - Organic & polymer memory
  - Memristor

- **Solid Electrolyte**

- **PC-RAM** (Phase-change RAM)
FeRAM (Ferroelectric RAM)

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

metallic electrodes

need select transistor – “half-select” perturbs

• perovskites (ABO₃) = 1 family of FE materials
• destructive read → forces need for high write endurance
• inherently fast, low-power, low-voltage
• first demonstrations ~1988

Saturation charge

Remanent charge

Coercive voltage

FeRAM progress

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

Fig. 1 A cross-sectional SEM image of 0.25 μm², 64 Mb FRAM cells.

Fig. 9 An optical micrograph of 0.25 μm², ITIC 64 Mb FRAM cell.
FeRAM difficulties

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

- Forces more complex integration schemes to keep effective area large

![Diagram showing different integration schemes]

<table>
<thead>
<tr>
<th>Integration Scheme</th>
<th>Area ($F^2$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Strapped”</td>
<td>&gt; 30</td>
</tr>
<tr>
<td>“Stacked”</td>
<td>10 - 30</td>
</tr>
<tr>
<td>“3-D”</td>
<td>&lt; 10</td>
</tr>
</tbody>
</table>

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

---

FeRAM difficulties

- Many reliability & processing difficulties to overcome…

<table>
<thead>
<tr>
<th>Difficulty</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>fatigue</td>
<td>remanent polarization $P_r$ decreases with cycling</td>
</tr>
<tr>
<td>imprint</td>
<td>a device left in one state tends to favor that polarization, causing hysteresis loop to shift</td>
</tr>
<tr>
<td>retention</td>
<td>Stored polarization is lost over time</td>
</tr>
<tr>
<td>High temperature processing</td>
<td>For crystalline FE material</td>
</tr>
<tr>
<td>insufficient $P_r$</td>
<td>$\propto$ voltage signal</td>
</tr>
</tbody>
</table>

- Change electrodes from metals to metal-oxides
- Change FE material ($PZT \rightarrow SBT$)
- Eliminate defects introduced during fabrication by hydrogen
- Change FE material ($PZT \rightarrow SBT$)
- Change FE material ($\rightarrow PZT$)
- Change FE material ($\rightarrow PZT$)

SBT = SrBi$_2$Ta$_2$O$_9$
Strontium bismuth tantalate
Alternative FeRAM concepts

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

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

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

Candidate device technologies

- **Improved Flash**
  - little change expected in write endurance or speed

- **FeRAM** – commercial product but difficult to scale!
  - **FeFET** – old concept, with many roadblocks

- **MRAM (Magnetic RAM)**
  - Racetrack memory

- **RRAM (Resistive RAM)**
  - Organic & polymer memory
  - Memristor

- **Solid Electrolyte**

- **PC-RAM (Phase-change RAM)**
**MRAM (Magnetic RAM)**

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

---

**Simple MTJ (magnetic tunnel junction)**

**MTJ with pinned layer**

**MTJ with pinned “synthetic antiferromagnet”**

**Toggle MRAM**

---

**Simple MTJ**

**MTJ with pinned layer**

**MTJ with pinned “synthetic antiferromagnet”**

---

**Toggle MRAM**
Progress in MRAM

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

Problems with MRAM

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

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

**Possible solutions**

- Heat MTJ to reduce required current

- Use “spin-torque” effect
  → rotate magnetization by passing current through the cell
  → now can have a wear-out mechanism (thin tunneling layers)
  → must insure read is non-perturbative

Magnetic Racetrack Memory

- 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 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/F²) is enormous…

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

Candidate device technologies

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

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

Can nearly anything that involves a state variable \( w \) become a memristor...?

\[
v = R(w, i) i \\
\frac{d}{dt} w = f(w, i)
\]

Solid Electrolyte

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

- Ag and/or Cu-doped \( \text{Ge}_x\text{Se}_{1-x} \), \( \text{Ge}_x\text{S}_{1-x} \), or \( \text{Ge}_x\text{Te}_{1-x} \)
- Cu-doped \( \text{MoO}_x \)
- Cu-doped \( \text{WO}_x \)
- RbAg\(_4\)I\(_5\) system

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)
Candidate device technologies

- Improved Flash
  - little change expected in write endurance or speed

- FeRAM – commercial product but difficult to scale!
  - FeFET – old concept, with many roadblocks

- MRAM – commercial product, also difficult to scale!
  - Racetrack memory – new concept w/ promise, still at point of early basic physics research

- RRAM – few demos showing real CMOS integration
  - Organic & polymer memory – temperature compatibility?
  - Memristor – hype may have outraced performance specifications

- Solid Electrolyte – shows real promise if tradeoff between retention & overprogramming can be solved...

- PC-RAM (Phase-change RAM)
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!

Crystalline phase
- Low resistance
- High reflectivity

Amorphous phase
- High resistance
- Low reflectivity
History of Phase-change memory

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

- 1990 First product (PCR: 500 GB)
- 1994 Powerful phase-change disk (PD: 650 MB)
- 1990 DVD-RAM ver.1 (2.6 GB)
- 2000 DVD-RAM ver.2 (4.7 GB)
- 2004 Single layer Blu-ray disk (BD: 23.3 GB)

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

---

Phase-change RAM

- **Access device** (transistor, diode)
- **PCRAM** “programmable resistor”

- **“RESET” pulse**
- **“SET” pulse**

- **Potential headache:**
  - High power/current affects scaling!
  - If crystallization is slow affects performance!
How a phase-change cell works

“SET” state
LOW resistance

“RESET” state
HIGH resistance

Hold at slightly under melting during recrystallization

Filament broadens, then heats up

Field-induced electrical breakdown starts at $V_{th}$
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**?

Electrical “breakdown” in PCM devices

- **70’s** – Study of *electrical breakdown* – “memory switching’ vs. “threshold switching”
- Recent studies – electrical resistivity drops rapidly with electric field…
Electrical “breakdown” in PC-RAM devices

- 70's – Study of electrical breakdown – “memory switching’ vs. “threshold switching”
  - Recent studies – electrical resistivity drops rapidly with electric field…

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

![I-V Curve for Set State](image1)

![I-V Curve for Reset State](image2)

Phase-change materials

- Two types of materials: “nucleation-dominated” vs. “growth-dominated”

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

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

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

![AFM images of GST and GeSb](image3)
Phase-change materials

We want a material that...

...retains data at moderate temperature...

yet switches rapidly at high temperature.

Designing for lower RESET current

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

W defined by lithography
H by thin-film deposition
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 → 60nm² ≈ Flash roadmap for 2013 → phase-change scales

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

Current scales with area
**PCM state-of-the-art**

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

Vertical view of a 512Mb PRAM cell array.

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

**Samsung:** CVD process fills deep holes

*C-15* Storage Class Memory @ IBM Almaden

**PCM state-of-the-art**

*Samsung:* "dash-shaped" holes filled with CVD

*Nemonyx:* bipolar-selected MLC chips

Fig. 5 TEM images of dash-type confined cell structure. The width of PCM in the contact is approximately 7.5nm and the PCM was filled perfectly without void.

[Im:2008 IEDM]

**IEEE J. Solid-State Circuits**

*Bedeschi:2009*
Outlook of PCM

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

? what happens when the # of atoms becomes countable?

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

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

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

February 4, 2008
[http://www.technologyreview.com/Infotech/20148/]

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

For more information (on PCRAM)


PCRAM
### In comparison…

<table>
<thead>
<tr>
<th>Knowledge level</th>
<th>Flash</th>
<th>SONOS Flash</th>
<th>Nanocystal Flash</th>
<th>FeRAM</th>
<th>FeFET</th>
</tr>
</thead>
<tbody>
<tr>
<td>Smallest demonstrated cell</td>
<td>4F² (2F² per bit)</td>
<td>4F² (1F² per bit)</td>
<td>16F² (@90nm)</td>
<td>15F² (@130nm)</td>
<td>—</td>
</tr>
<tr>
<td>Prospects for...</td>
<td>poor</td>
<td>maybe</td>
<td>unclear</td>
<td>poor</td>
<td>unclear</td>
</tr>
<tr>
<td>...scalability</td>
<td>yes</td>
<td>yes</td>
<td>unclear (enough stored charge?)</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>...fast readout</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>...fast writing</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>...low switching Power</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>...high endurance</td>
<td>NO</td>
<td>poor (1e7 cycles)</td>
<td>NO</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>...non-volatility</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>poor (30 days)</td>
</tr>
<tr>
<td>...MLC operation</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>difficult</td>
<td>difficult</td>
</tr>
</tbody>
</table>

### Knowledge level

<table>
<thead>
<tr>
<th>MRAM</th>
<th>Racetrack</th>
<th>PCRAM</th>
<th>RRAM</th>
<th>solid electrolyte</th>
<th>organic memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>product</td>
<td>basic research</td>
<td>advanced development</td>
<td>Early development</td>
<td>development</td>
<td>basic research</td>
</tr>
<tr>
<td>25F² @180nm</td>
<td>—</td>
<td>5.8F² (diode) 12F² (BJT) @90nm</td>
<td>—</td>
<td>8F² @90nm (4F² per bit)</td>
<td>—</td>
</tr>
<tr>
<td>Prospects for...</td>
<td>poor (high currents)</td>
<td>unknown (too early to know, good potential)</td>
<td>promising (rapid progress to date)</td>
<td>unknown</td>
<td>promising (filament-based, but new materials)</td>
</tr>
<tr>
<td>...scalability</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>...fast readout</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>...fast writing</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>sometimes</td>
<td>yes</td>
</tr>
<tr>
<td>...low switching Power</td>
<td>NO</td>
<td>uncertain</td>
<td>poor</td>
<td>sometimes</td>
<td>yes</td>
</tr>
<tr>
<td>...high endurance</td>
<td>yes</td>
<td>should</td>
<td>yes</td>
<td>poor</td>
<td>unknown</td>
</tr>
<tr>
<td>...non-volatility</td>
<td>yes</td>
<td>unknown</td>
<td>yes</td>
<td>sometimes</td>
<td>sometimes</td>
</tr>
<tr>
<td>...MLC operation</td>
<td>NO</td>
<td>yes (3-D)</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>
Outline

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

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

---

**Cost structure of silicon-based technology**

`Cost` determined by

- cost per wafer
- # of dies/wafer
- memory area per die [sq. μm]
- **memory density** [bits per 4F²]
- **patterning density** [sq. μm per 4F²]
Need a 10x boost in density **BEYOND** Flash!

Paths to ultra-high density memory

- Starting from standard $4F^2$...
- $2F$... $2F$...
- Add $N$ 1-D sub-lithographic "fins" ($N^2$ with 2-D) demonstrated (at IEDM 2005)
- Store $M$ bits/cell with $2^M$ multiple levels demonstrated (at IEDM 2007)
- 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
Multi-level phase-change memory

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?)
3-D stacking

- 3-D anti-fuse
  (Matrix semiconductor)

- 3-D Flash (Toshiba) [Tanaka:2007]

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...

<table>
<thead>
<tr>
<th>Density</th>
<th>Product</th>
</tr>
</thead>
<tbody>
<tr>
<td>43 Gb/cm^2</td>
<td>32GB</td>
</tr>
<tr>
<td>86 Gb/cm^2</td>
<td>64GB</td>
</tr>
<tr>
<td>344 Gb/cm^2</td>
<td>256GB</td>
</tr>
<tr>
<td>1376 Gb/cm^2</td>
<td>~1 TB</td>
</tr>
</tbody>
</table>

* 2006 ITRS Roadmap
For more information (on ultra-high density)


- ITRS roadmap, www.itrs.net

How does SCM compare to existing technologies?

Cost:
- NOR FLASH
- NAND FLASH
- DRAM
- SRAM
- HDD

Performance:
- STORAGE CLASS MEMORY
If you could have SCM, why would you need anything else?

Technology conclusions

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

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

- **If you build it, they will come**
  - With its combination of low-cost and high-performance, SCM could impact much more than just the server-room...
Thank you!
Outline

- Introduction
- NAND Flash: the current Storage Class Memory (SCM)
- Wear leveling
- What if? → Comparison of extrapolated disk and SCM storage systems
- SCM in the memory stack
- Applications
If you could have SCM, why would you need anything else?

$1 / GB

$10k / GB

$100k / GB

$1k / GB

$100 / GB

$10 / GB

$10 / GB

$0.10 / GB

$0.01 / GB


SCM

NAND

DRAM

Desktop HDD

Enterprise HDD

SCM

34nm 4-Layer 1-bit
22nm 4-Layer 2-bit
18nm 4-Layer 3-bit
14nm 4-Layer 3-bit
14nm 4-Layer 4-bit

Chart courtesy of Dr. Chung Lam, IBM Research.

Updated version of plot from IBM Journal R&D article.

Memory/Storage Stack Latency Problem

Century

Year

Month

Day

Hour

Second

Time in ns

Get data from TAPE (40s)

Access DISK (5ms)

Access FLASH (20 us)

Access PCM (100 – 1000 ns)

Get data from DRAM or PCM (60ns)

Get data from L2 cache (10ns)

CPU operations (1ns)
SCM in a large System

1980
- CPU
- RAM
- DISK
- TAPE

2008
- CPU
- RAM
- FLASH
- SSD
- DISK
- TAPE

2013
- CPU
- RAM
- SCM
- DISK
- TAPE

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
NAND Flash

Representative NAND Flash Device

- Chip size: 12mm x 20mm
- Power ≈ 100mW
- Interface: byte wide
- Page
  - 2112 Bytes
  - Moving to 4224 Bytes
- Block = 64 - 128 Pages
Representative NAND Flash Behavior

- Read copies Page into BUF and streams data to host
  - Read 20us access,
  - 20 MB/s transfer rate – sustained
  - Moving to 40 MB/s
- Write streams data from host into BUF
  - 6 MB/s transfer rate sustained
  - 20 MB/s burst → 40 MB/s
- Program copies BUF into 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

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
Flash Drive Channel

Host Intf. → Flash Chip → Channel

Channel

Flash 1

Flash 2

Flash n

2 KB Data transfer
Program page
Erase block

SCM: Generic Storage Design

CPU

MEMORY

DRAM

SCM Control

SCM

SCM

SCM

SCM

SCM

SCM

SCM

SCM

SCM

SCM
Classes for Flash SSDs

- **Sustained Read Bandwidth**
- **Sustained Write Bandwidth**
- **Maximum Random Read IOPs**
- **Maximum Random Write IOPs**

**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 connected hosts

SAN: Storage Area Network
SVC: San Volume Controller
Wear Leveling

Challenges with SCM

- **Asymmetric performance**
  - Flash: writes much slower than reads
  - Not as pronounced in other technologies

- **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) → $10^5$ writes/cell
    - Multi level cell (MLC) → $10^4$ writes/cell
  - Coping strategy → Wear leveling, etc.
Life-time of SCM devices

In a device that wear out on writes,

\[ T_{\text{life}} = \text{Endurance} \times \text{Fill-Time} = E \times T_{\text{fill}} \]

\[ T_{\text{fill}} = \frac{C}{B} \quad \text{(Fill-Time)} \]

= time to write all C elements, given write-rate of B

\[ T_{\text{fill}} \sim 1 \text{ sec for DRAM, } \sim 10,000 \text{ seconds for disks} \]

Consider an SLC flash chip with \( C = 8 \text{ M blocks}, E = 10^5 \) and \( B = 600 \text{ b/s} \) (blocks per second where a block = 2 KB)

Without any wear-leveling and looping on one block, \( C = 1 \) (not 8 M blocks) and

\[ T_{\text{life}} = \frac{E}{B} = \frac{10^5}{600} \text{ b/s} = 170 \text{ seconds} \]

(Perfect) Wear-leveling improves \( T_{\text{life}} \) (for 1 block) by the capacity \( C \)

\[ T_{\text{life}} = E \times T_{\text{fill}} = E \times \frac{C}{B} = 10^5 \times 8 \text{ Mblocks/600 b/s} = 1.36 \times 10^9 \text{ seconds} \]

From ~3 minutes to more than 42 years!

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
  - Overprovisioning
  - Compress, de-duplicate & write elimination…
  - \( E^* = E \times q \times f(\text{error correction}) \times g(\text{overprovisioning}) \times h(\text{compress})… \)
  - With S included, \( T_{\text{life}(\text{System})} = T_{\text{fill}} \times E^* \)
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.

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
SCM Extrapolation

SCM module ‘Specs’ in 2020

- SCM modules are (small?) block oriented storage devices

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Capacity</strong></td>
<td>1 TB</td>
</tr>
<tr>
<td><strong>Read or Write</strong></td>
<td>&lt;1 us</td>
</tr>
<tr>
<td><strong>Access Time</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Data Rate</strong></td>
<td>&gt;1GB/s</td>
</tr>
<tr>
<td><strong>Sustained transaction rate</strong></td>
<td>200,000 IOPS</td>
</tr>
<tr>
<td>(-1us + 4K / 1GB/s = 5us)</td>
<td></td>
</tr>
<tr>
<td><strong>Sustained bandwidth</strong></td>
<td>800MB/s</td>
</tr>
<tr>
<td>(-4KB/5us = &gt;800MB/s)</td>
<td></td>
</tr>
</tbody>
</table>

* Cntrl
  * SCM
  * SCM
  * SCM

A D A D A D A D ...
Basic 2020 Storage Package

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

2020 SCM Storage System Package

- 16 modules per FLP
- 200 FLPs per 2U drawer
- 21 2U Drawers per rack
2020 Comparison

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

<table>
<thead>
<tr>
<th>TODAY</th>
<th>2020</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.4 TB/s</td>
<td>0.4 PB/s</td>
</tr>
<tr>
<td>2 MIOP/s</td>
<td>2 GIOP/s</td>
</tr>
<tr>
<td>(10,000 disks)</td>
<td></td>
</tr>
</tbody>
</table>

Disk Assumptions for 2020

- Long term disk drive technology trend
  - Areal density growth has flattened to ~40% CAGR
  - Bandwidth improvement is ~10% CAGR
  - Access time improvement is ~5% CAGR
- Enterprise disk: 1.8” diameter
- Sustained bandwidth of 300 MB/s
- 400 IOP/s
- 4 Watts
- 256 drives packaged in a standard 4U (7 inch high) drawer.
- Ten such drawers packaged in a standard 19-inch rack.
Results of Extrapolation

<table>
<thead>
<tr>
<th></th>
<th>Compute centric</th>
<th>Data centric</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Devices</strong></td>
<td>1.3 M Disks</td>
<td>5 M Disks</td>
</tr>
<tr>
<td><strong>space</strong></td>
<td>4500 sq.ft.</td>
<td>16,500 sq.ft.</td>
</tr>
<tr>
<td><strong>power</strong></td>
<td>6,000 kW</td>
<td>22,000 kW</td>
</tr>
<tr>
<td><strong>SCM</strong></td>
<td>406 K modules</td>
<td>8 K modules</td>
</tr>
<tr>
<td><strong>Disk</strong></td>
<td>85 sq. ft.</td>
<td>12 sq. ft.</td>
</tr>
<tr>
<td><strong>SCM</strong></td>
<td>41 kW</td>
<td>1 kW</td>
</tr>
</tbody>
</table>

Disk ≡ SCM

SCM in the Memory Stack
CPU & Memory System (Node) in 2008

Logical Address > Address Translation > Physical Address

CPU & Memory System (Node) in 2008

Logical Address > Address Translation > Physical Address

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 alternatives

- CPU
- VM
- PMC

SCM Controller
Wear Leveler

- DRAM
- DIMM

SCM Card

CPU & Memory System alternatives

- CPU
- VM
- PMC

SCM Card

SCM Controller

- X86, PowerPC, SPARC...

- DRAM
- DIMM
## Uses of SCM in overall memory/storage stack

<table>
<thead>
<tr>
<th>Access Mode</th>
<th>Use Mode</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>synchronous</td>
<td>Cache (e.g. Level 4)</td>
<td>Wear level too high?</td>
</tr>
<tr>
<td>Memory-like</td>
<td>Main memory - version (a)</td>
<td>Separate WL/SCM controller</td>
</tr>
<tr>
<td></td>
<td>Main memory - version (b)</td>
<td>Integrated WL/SCM/RAM controller</td>
</tr>
<tr>
<td></td>
<td>Main memory - version (c)</td>
<td>SCM Wear level managed by software &amp; VM manager (dangerous)</td>
</tr>
<tr>
<td>Cache-line?</td>
<td>Via legacy I/O busses</td>
<td>Easy, but wastes SCM performance</td>
</tr>
<tr>
<td></td>
<td>Via new interfaces</td>
<td>Good for memory mapping use model</td>
</tr>
<tr>
<td></td>
<td>Paging Device</td>
<td>Very promising use</td>
</tr>
<tr>
<td></td>
<td>I/O Cache and/or meta-date</td>
<td>Act as NVRAM, good use</td>
</tr>
<tr>
<td></td>
<td>storage for a disk controller</td>
<td></td>
</tr>
</tbody>
</table>
Shift in Systems and Applications

- **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

Main Memory:

- Active data on disk
- Inactive data on tape
- SANs in heavy use

Storage:

- Compute centric
- Focus on hiding disk latency

Applications:

- Active data on SCM
- Inactive data on disk/tape
- DAS ??

- Data centric comes to fore
- Focus on efficient memory use and exploiting persistence
- Fast, persistent metadata

Summary

- **Storage Class Memory** is a new class of data storage/memory technology → many technologies are competing to be the ‘best’ SCM

- Flash, which has may SCM characteristics, is available now and PCM is in the wings.

- **SCM blurs the distinction** between memory and storage

- SCM will impact on the design of computer systems and applications

- How will you use SCM?
Questions

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 allocation
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

Implications on Traditional Commercial Databases

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

- **Long term, deep Impact: Random access replaces paging**
  - DB performance depends heavily on good guesses what to page in
  - Random access eliminates column/row access tradeoffs
  - Reduces energy consumption (big effect)

- **Existing trend is to replace ‘update in place’ with ‘appends’**
  - that’s good – helps with write endurance issue

- **Reduce variability of data mining response times**
  - from hours and days (today) to seconds (SCM)
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
Price/MB for DRAM-NAND FLASH- SCM - HDD

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