Motivation

- **Trends**
  - Demand for storage continues to be robust
  - Storage performance gain has not kept pace with that of servers so proportionally more disks will be needed in the data center
  - The amount of power consumed in data centers is becoming an issue
  - New storage technologies: Flash, PCM, …

- **Questions**
  - Where will disk storage be in 2020?
  - Will new storage technologies help?
Agenda

- **Technology and power consumption**
  - Disks
  - Flash
  - Phase Change Memory

- **Illustrative power comparison -- 2020**
  - High data-rate and high transaction-rate scenarios
  - Power comparisons for technologies

HDDs

- Invented in the 1950s
- Mechanical device consisting of a rotating magnetic media disk and actuator arm with 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...
HDD strong suit is Areal Density

Future of HDD

Higher densities through

- perpendicular recording
  - Jul 2008
  - 610 Gb/in² → ~4 TB
- patterned media

Future of HDD

Higher densities through

- perpendicular recording
  - Jul 2008
  - 610 Gb/in² → ~4 TB
- patterned media
But, problems with Disk Drive Maximum Sustained Data Rate

Bandwidth [MB/s]

Date

And Disk Drive Latency

Latency [ms]

Date
So,

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

### Typical Storage Device Power

<table>
<thead>
<tr>
<th>Drive type</th>
<th>size</th>
<th>Watts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>stand-by</td>
</tr>
<tr>
<td>3.5&quot; 15K RPM FC/SAS</td>
<td>300 GB</td>
<td>2.0</td>
</tr>
<tr>
<td>3.5&quot; 10K RPM FC/SAS</td>
<td>300 GB</td>
<td>2.0</td>
</tr>
<tr>
<td>3.5&quot; 10K RPM FC</td>
<td>300 GB</td>
<td>9.2</td>
</tr>
<tr>
<td>3.5&quot; 7200 RPM SATA</td>
<td>500 GB</td>
<td>2.0</td>
</tr>
<tr>
<td>3.5&quot; 7200 RPM SATA NL</td>
<td>500 GB</td>
<td>7.4</td>
</tr>
<tr>
<td>3.5&quot; 7200 RPM SATA</td>
<td>500 GB</td>
<td>9.6</td>
</tr>
<tr>
<td>2.5&quot; 15K RPM SAS</td>
<td>73 GB</td>
<td></td>
</tr>
<tr>
<td>2.5&quot; 10K RPM SAS</td>
<td>73 GB</td>
<td>2.0</td>
</tr>
<tr>
<td>2.5&quot; Mobile 7200 RPM SATA</td>
<td>100 GB</td>
<td>0.3</td>
</tr>
<tr>
<td>2.5&quot; Mobile 5400 RPM SATA</td>
<td>100 GB</td>
<td>0.2</td>
</tr>
<tr>
<td>USB Flash Disk</td>
<td>32 GB</td>
<td>0.1</td>
</tr>
<tr>
<td>2.5&quot; laptop SSD</td>
<td>73 GB</td>
<td>2.4</td>
</tr>
<tr>
<td>3.5&quot; Enterprise SSD</td>
<td>155 GB</td>
<td>5.0</td>
</tr>
</tbody>
</table>
Illustrative Disk Drive Power Profile

- **Standby power**
  - Both spindle and actuator motors off
  - Base electronics powered
- **Idle**
  - Spindle motor running
  - Actuator motor off
  - Most electronics powered
- **Typical**
  - Spindle motor running
  - Actuator motor in use periodically
  - All electronics powered
- **Startup – spindle motor starting**
  - ~30 seconds for startup

Typical pwr

idle power: always being consumed regardless of whether operations are in progress or not

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 term
- **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

Generic SCM Array
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.
- Single Level vs Multi-level cell
- Scaling issues

Feeds and Speeds for typical NAND Flash

<table>
<thead>
<tr>
<th></th>
<th>NAND</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell Size</td>
<td>$4 F^2$ (2 $F^2$ virtual x 2-bit MLC)</td>
</tr>
<tr>
<td>Read Access Time</td>
<td>20-50 us</td>
</tr>
<tr>
<td>Read</td>
<td>15-25 MB/s</td>
</tr>
<tr>
<td>Write</td>
<td>5-8MB/sec</td>
</tr>
<tr>
<td>Erase</td>
<td>2ms</td>
</tr>
<tr>
<td>Start Up Time</td>
<td>50-100 us</td>
</tr>
<tr>
<td>Market Size (2007)</td>
<td>$14.2B</td>
</tr>
<tr>
<td>Applications</td>
<td>Multimedia</td>
</tr>
</tbody>
</table>
Representative NAND Flash Device

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

ONFI → Open NAND Flash Interface

Illustrative Flash Read Power Profile

- Read Access
- 2 KB Data transfer
Illustrative Flash Write Power Profile

Typical Storage Device Power

<table>
<thead>
<tr>
<th>Drive type</th>
<th>size</th>
<th>Watts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>stand-by</td>
</tr>
<tr>
<td>3.5” 15K RPM FC/SAS</td>
<td>300 GB</td>
<td>2.0</td>
</tr>
<tr>
<td>3.5” 10K RPM FC/SAS</td>
<td>300 GB</td>
<td>2.0</td>
</tr>
<tr>
<td>3.5” 10K RPM FC</td>
<td>300 GB</td>
<td>9.2</td>
</tr>
<tr>
<td>3.5” 7200 RPM SATA</td>
<td>500 GB</td>
<td>2.0</td>
</tr>
<tr>
<td>3.5” 7200 RPM SATA NL</td>
<td>500 GB</td>
<td>7.4</td>
</tr>
<tr>
<td>3.5” 7200 RPM SATA</td>
<td>500 GB</td>
<td>9.6</td>
</tr>
<tr>
<td>2.5” 15K RPM SAS</td>
<td>73 GB</td>
<td>2.0</td>
</tr>
<tr>
<td>2.5” 10K RPM SAS</td>
<td>73 GB</td>
<td>0.3</td>
</tr>
<tr>
<td>2.5” Mobile 7200 RPM SATA</td>
<td>100 GB</td>
<td>0.2</td>
</tr>
<tr>
<td>USB Flash Disk</td>
<td>32 GB</td>
<td>0.1</td>
</tr>
<tr>
<td>2.5” laptop SSD</td>
<td>73 GB</td>
<td>2.4</td>
</tr>
<tr>
<td>3.5” Enterprise SSD</td>
<td>155 GB</td>
<td>5.0</td>
</tr>
</tbody>
</table>

30% less than 15K 3.5” SAS
Illustrative Flash SSD Design

![Diagram of Flash SSD Design]

- CPU
- DRAM
- Flash Control
- Flash

Idle power: 5 W  Typical operating power: 8 W

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 term
- 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

![Diagram of Generic SCM Array]

bistable material
plus on-off switch

Generic SCM Array
Phase-Change RAM

PCM performance and power SWAG

- **Performance**
  - For devices optimized for storage applications:
    - Read access time of 1 us
    - Write access time of 3-5 us
    - Interface LPDDR or similar

- **Power**
  - Transfer power 15mW
  - Read power 20 mW
  - Write: 85 mW
    - Write-in-place
    - S0, no erase
Illustrative PCM SSD Design

Idle power: 1 W  Typical operating power: 3 W

Device comparison: energy and power

<table>
<thead>
<tr>
<th></th>
<th>disk</th>
<th>Flash SSD today</th>
<th>PCM SSD projected</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Today</td>
<td>2020</td>
<td></td>
</tr>
<tr>
<td>IOPS</td>
<td>3.5” 15K</td>
<td>1.8” 15K</td>
<td>3.5” 1.8”</td>
</tr>
<tr>
<td>BW</td>
<td>200</td>
<td>400</td>
<td>50,000</td>
</tr>
<tr>
<td>Idle power</td>
<td>100 MB/s</td>
<td>300 MB/s</td>
<td>250 MB/s</td>
</tr>
<tr>
<td>Typ. power</td>
<td>9 W</td>
<td>4 W</td>
<td>5 W</td>
</tr>
<tr>
<td>Energy per Rd op</td>
<td>16 W</td>
<td>6 W</td>
<td>8 W</td>
</tr>
<tr>
<td>10% utilization</td>
<td>485,000 uJ</td>
<td>105,000 uJ</td>
<td>1,060 uJ</td>
</tr>
<tr>
<td>50%</td>
<td>125,000 uJ</td>
<td>25,000 uJ</td>
<td>260 uJ</td>
</tr>
<tr>
<td>90%</td>
<td>85,000 uJ</td>
<td>16,000 uJ</td>
<td>171 uJ</td>
</tr>
</tbody>
</table>

- Energy is what you pay for
- In 2020 a disk will cost ~3000x more per operation than a PCM SSD
- IOPS, cost, power \rightarrow power, cost IOPS
- Flash SSDs may not exist in 2020
- Power for controllers, etc. viewed as second order effects
Extrapolate Storage Performance to 2020

- Start with ASCI Purple class machine
- Trend for storage performance growth: 70% CAGR
  - Driven by application requirements and investment
  - Could be impacted by changing in architecture
- Assume 50 yr disk trends continue and no game changing technology invention
- Semiconductor technologies stay on a roughly 40% CAGR
- PCM edges out FLASH for the solid state storage crown

- Bandwidth: 0.4 TB/s → 400 TB/s
- Transaction rate: 2 MIOPS → 2000 MIOPS

The Promise of Solid State Disk

- By 2020, Storage Class Memory should revolutionize data centers
The Promise of Solid State Disk

- By 2020, Storage Class Memory should revolutionize data centers

### Bandwidth Driven Storage System: 400 TB/s

**Floor Space**

- SCM

- 85 Square Feet

**Power**

- 42 KW

### Transaction Rate Driven Storage System: 2000 MOP/s

**Floor Space**

- DISKS

- 23,000 Square Feet

**Power**

- 22,000 KW
The Promise of Solid State Disk

- By 2020, Storage Class Memory should revolutionize data centers

Transaction Rate Driven Storage System: 2000 MOP/s

SCM

Floor Space

11 Square Feet

Power

KW

My Conclusions

- Idle power for disks drives up their energy costs significantly
- Energy per op for SSDs is much better than for HDDs
  - SSDs gain because of their very high transaction rates and high bandwidth
  - Combined with generally lower idle power
- But, even so, SSDs have a higher idle power that one would expect for solid state technology
  - Careful design of SSDs should alleviate much of this
- Current trends would seem to indicate that Enterprise disks may not be widely used in 2020 and their position will be taken by Solid State Storage
Questions?

Break Time

Further reading

- IBM Journal of Research and Development, special issue on Storage Technologies and Systems
  - Volume 52, Number 4/5 July/September 2008

- Other papers
Input from the device cost crystal ball

Price Trends: Magnetic disks and Solid State Disks

SSD Price = multiple of device cost
- 10x SLC @ -40% CAGR (4.5 → 1.8)
- 3x MLC @ -40% CAGR (2.3 → 1.1)
- 3x PCM @ -40% CAGR (1.0 → ?)

Input form the Subsystem Price Crystal Ball