The Computer IS the Memory

Tom Coughlin, Coughlin Associates
Jim Handy, Objective Analysis
Outline

• The Shape of Things to Come
• Getting Memory & Processing Closer
• Spin-Based Memory and Processors
• Conclusions
The Shape of Things to Come
Memory/Storage Hierarchy

NVM/Storage trade-offs

- Storage costs ($/TB)
- Performance
  - IOPS
  - Data rates
Touch Rate

Source: Steve Hetzler, IBM
IOPS Required

79% between 10K -1M IOPS

Source: Coughlin Associates

How Many IOPS Do You Really Need?

© 2015 Coughlin Associates & Objective Analysis
Minimum Latency Requirement

36% at 10ms latency

Source: Coughlin Associates

How Many IOPS Do You Really Need?
Persistent Memory Implications

• Retains data during a power loss
  – Instant recovery of state before power down
• Lower latencies than disk
• Lower power than DRAM
• Allows persistent states for Remote Direct Memory Access (RDMA)
• Supports “logic-in-memory architecture”
  – Could lead to new distributed computer architectures
A Timeline For Change

Flash as Memory
Processor in DRAM
New Memory
Processor in New Memory

© 2015 Coughlin Associates & Objective Analysis
Flash as Memory

• Already exists on PCIe bus
  – Fusion-io pioneered “Extended Memory”
• Now migrating to memory bus
  – NVDIMM-N: DRAM backed by flash
  – NVDIMM-F: Flash on a DIMM
• Presents new problems to existing compute model
  – Memory is persistent
  – Flash is weird (wear, block erase, page write, etc.)
  – SNIA standards evolving to cover this
Flash Already More Economical than DRAM

Source: Objective Analysis

How PC NAND Will Undermine DRAM

© 2015 Coughlin Associates & Objective Analysis
Getting Memory & Processing Closer
Processor in Memory

• Pioneering efforts already exist:
  – Micron Technology Automata Processor
  – Venray Technology TOMI
  – Berkeley iRAM (1996)
  – Notre Dame PIM (1996)

Source: Micron Technology
## Trouble Mixing Logic & DRAM

<table>
<thead>
<tr>
<th>Logic Process</th>
<th>DRAM Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>• High speed</td>
<td>• Slow speed</td>
</tr>
<tr>
<td>• Variable capacitance</td>
<td>• High capacitance</td>
</tr>
<tr>
<td>• Multiple metal layers</td>
<td>• Few metal layers</td>
</tr>
<tr>
<td>• Weaker cost focus</td>
<td>• Cost-obsessed</td>
</tr>
<tr>
<td>• Muddies up a DRAM process</td>
<td>• Undermines a logic process</td>
</tr>
</tbody>
</table>

© 2015 Coughlin Associates & Objective Analysis
Alternative Memories to the Rescue

• Byte read/write
  – Read/write speeds roughly balanced

• No “Erase Before Write”

• Logic-friendly fabrication process
  – Supports easier integration of logic and memory

• Persistent, inexpensive, & fast
Yes, But *Which* New Memory Will Win?

• Many good alternatives:
  – STT MRAM
  – ReRAM
  – Memristor
  – Racetrack
  – PCM
  – FRAM

• Hard to know which will cross over DRAM & NAND pricing
ReRAM or RRAM Cross Point Array

- Likely NAND replacement
- Multi-film diodes eliminate need for larger CMOS selector

Source: Flash Memory Summit 2013
Phase Change Memory (PCRAM)

- Bit addressable, high-density arrays.
- The phase change causes resistance change:
  - Crystalline: Low resistance
  - Amorphous: Higher resistance.
Magnetic RAM (MRAM)

• 3 Major types
  – Toggle Mode (field driven)
  – Spin Torque Transfer (STT)
  – Magnetothermal MRAM (Heat Assisted)
Memory/Storage Performance/Density Roadmap

2014 Emerging Non-Volatile Memory and Storage Technologies and Manufacturing Report, Coughlin Associates
How Alternatives Will Win

![Graph showing relative cost per bit for different process nodes. The graph compares Flash and New Tech. Flash consistently has a lower relative cost per bit compared to New Tech across all process nodes (180nm, 65nm, 23nm, 8nm). The graph is sourced from Objective Analysis.]
The Impact of Moore’s Law

• “Doubling transistors per chip every year or two”
  – Put differently: “Making two chips into one every year or two”

• Using the 18-month rule of thumb
  – 2015 single-socket server with 96GB of RAM & an SSD
    • 200-300 chips
  – 2030 – 15 years later
    • About 1/5\textsuperscript{th} to 1/3\textsuperscript{rd} of one chip

• Today’s 10,000-server data center could fit in 10 server slots!
  – Computing structures will adapt to this evolution

• Will Moore’s Law stop? Stay tuned!
Summary

• First step: Persistent Memory
• Second Step: Processor in Memory
• Third Step: Processor in Persistent Memory
• Finally: Data Center in a Rack

*Architectures and code will have to adapt!*
Spin-Based Memory and Processors

© 2015 Coughlin Associates & Objective Analysis
Spintronics

Spintronics (Spin + Charge)

• Semiconductor (Charge)
• Magnetic materials (Spin)

Applications

- HDD (Hard Disc Drive)
- MRAM (Magnetic Random Access Memory)
- Spin-FET (Spin-Field Effect Transistor)

GMR
Large TMR + Low R
Large CPP-GMR

High spin injection efficiency into semiconductor


From Tohoku University in Japan

© 2015 Coughlin Associates & Objective Analysis
STT MRAM

- STT MRAM uses the electron spin to create a memory element
- STT MRAM market could exceed $2 B by 2019

<table>
<thead>
<tr>
<th></th>
<th>STT MRAM</th>
<th>DRAM</th>
<th>SRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Read Time (ns)</td>
<td>3-5</td>
<td>30</td>
<td>1-100</td>
</tr>
<tr>
<td>Write Time (ns)</td>
<td>3-15</td>
<td>50</td>
<td>1-100</td>
</tr>
<tr>
<td># Rewrites</td>
<td>&gt;$10^{15}$</td>
<td>$10^{16}$</td>
<td>$10^{16}$</td>
</tr>
<tr>
<td>Input Voltage</td>
<td>1.5</td>
<td>2</td>
<td>None</td>
</tr>
</tbody>
</table>
Everspin 64 Mbit STT MRAM Chip Used for Caching

Over 40 M MRAM Chips shipped

© 2015 Coughlin Associates & Objective Analysis
$/GB for Memory Technologies

(includes data from Jim Handy, Objective Analysis)

2014 Emerging Non-Volatile Memory and Storage Technologies and Manufacturing Report, Coughlin Associates
What if Spin Transfer Replaced Charge Transfer (Current)?

• Spin transfer would not generate the heat that electrical currents generate
• Build powerful devices in small spaces
• Spin could be used for both processing and memory/storage
Conclusions

• The storage/memory landscape has more options than ever
• Non-volatile memories will replace volatile memories
• NVM and processors want to come together
• Spin-based electronics are one example of a technology that could put memory and processing in one device
Thanks