Future Memories and Today’s Opportunities

Tom Coughlin, Coughlin Associates
Jim Handy, Objective Analysis
Cataclysmic Changes Coming Soon

• Scaling Limits
  – We can’t make transistors any smaller

• New Storage Hierarchies
  – You think SSDs were disruptive? Just wait!

• More Layers Will Be Added
  – It’s all about touch rates and response time

• Processors Must Adapt
SCALING LIMITS
Scaling Limits Create Opportunities for New Memories

![Graph showing relative cost vs. process geometry from 500nm to 2nm, with lines for Flash and New Tech technologies.]
No Shortage of Options
### Today’s Memories Are Limited

<table>
<thead>
<tr>
<th>Feature</th>
<th>SRAM</th>
<th>DRAM</th>
<th>ROM</th>
<th>EEPROM</th>
<th>NOR</th>
<th>NAND</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonvolatile</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Erasable</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Programmable</td>
<td>Yes</td>
<td>Yes</td>
<td>Factory</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Smallest Write</td>
<td>Byte</td>
<td>Byte</td>
<td>N/A</td>
<td>Byte</td>
<td>Byte</td>
<td>Page</td>
</tr>
<tr>
<td>Smallest Read</td>
<td>Byte</td>
<td>Page</td>
<td>Byte</td>
<td>Byte</td>
<td>Byte</td>
<td>Page</td>
</tr>
<tr>
<td>Read Speed</td>
<td>V Fast</td>
<td>Fast</td>
<td>Fast</td>
<td>Fast</td>
<td>Fast</td>
<td>Slow</td>
</tr>
<tr>
<td>Write Speed</td>
<td>V Fast</td>
<td>Fast</td>
<td>N/A</td>
<td>Slow</td>
<td>Slow</td>
<td>Slow</td>
</tr>
<tr>
<td>Sleep Power</td>
<td>V Low</td>
<td>High</td>
<td>Zero</td>
<td>Zero</td>
<td>Zero</td>
<td>Zero</td>
</tr>
<tr>
<td>Price/GB</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
<td>High</td>
<td>Med</td>
<td>V Low</td>
</tr>
<tr>
<td>Applications</td>
<td>Small Fast</td>
<td>Main Memory</td>
<td>Stable Code Volume</td>
<td>Serial #, Trim</td>
<td>Code</td>
<td>Data</td>
</tr>
</tbody>
</table>

© 2016 Coughlin Associates & Objective Analysis
## Emerging Memories Perform Better

<table>
<thead>
<tr>
<th></th>
<th>MRAM</th>
<th>ReRAM</th>
<th>FRAM</th>
<th>PCM</th>
<th>XPoint</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Nonvolatile</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Erasable</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Programmable</strong></td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Smallest Write</strong></td>
<td>Byte</td>
<td>Byte</td>
<td>Byte</td>
<td>Byte</td>
<td>Byte</td>
</tr>
<tr>
<td><strong>Smallest Read</strong></td>
<td>Byte</td>
<td>Byte</td>
<td>Byte</td>
<td>Byte</td>
<td>Byte</td>
</tr>
<tr>
<td><strong>Read Speed</strong></td>
<td>Fast</td>
<td>Fast</td>
<td>Fast</td>
<td>Fast</td>
<td>Fast</td>
</tr>
<tr>
<td><strong>Write Speed</strong></td>
<td>Fast</td>
<td>Fast</td>
<td>Fast</td>
<td>Fast</td>
<td>Fast</td>
</tr>
<tr>
<td><strong>Active Power</strong></td>
<td>Low</td>
<td>Med</td>
<td>Low</td>
<td>High</td>
<td>High?</td>
</tr>
<tr>
<td><strong>Sleep Power</strong></td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td><strong>Price/GB</strong></td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>High?</td>
</tr>
<tr>
<td><strong>Applications</strong></td>
<td>Niche</td>
<td>TBD</td>
<td>Low Power</td>
<td>Obsolete</td>
<td>Main Memory</td>
</tr>
</tbody>
</table>

© 2016 Coughlin Associates & Objective Analysis
NEW STORAGE HIERARCHIES
Memory/Storage Hierarchy

From: Objective Analysis: Solid State Drives in the Enterprise
NAND Flash SSDs Today Make Sense

From: Objective Analysis: Solid State Drives in the Enterprise
Cost Brought Flash Into Computing

Average Price per Gigabyte

From: Objective analysis: Hybrid Drives: How, Why, & When?

© 2016 Coughlin Associates & Objective Analysis
3D XPoint Will Do The Same In 2017

Price per Gigabyte

Bandwidth (MB/s)

From: Objective Analysis: Solid State Drives in the Enterprise
Intel’s & Micron’s 3D XPoint Intro

**3D XPoint™ Technology:** An Innovative, High-Density Design

**Cross Point Structure**
Perpendicular wires connect submicroscopic columns. An individual memory cell can be addressed by selecting its top and bottom wire.

**Stackable**
These thin layers of memory can be stacked to further boost density.

**Non-Volatile**
3D XPoint™ Technology is non-volatile—which means your data doesn’t go away when your power goes away—making it a great choice for storage.

**Selector**
Whereas DRAM requires a transistor at each memory cell—making it big and expensive—the amount of voltage sent to each 3D XPoint™ Technology selector enables its memory cell to be written to or read without requiring a transistor.

**High Endurance**
Unlike other storage memory technologies, 3D XPoint™ Technology is not significantly impacted by the number of write cycles it can endure, making it more durable.

**Memory Cell**
Each memory cell can store a single bit of data.

**Transforming the Memory Hierarchy**
For the first time, there is a fast, inexpensive and non-volatile memory technology that can serve as system memory and storage.

**~8x to 10x Greater Density than DRAM**
3D XPoint™ Technology’s simple, stackable, transistor-less design packs more memory into less space, which is critical to reducing cost.

© 2016 Coughlin Associates & Objective Analysis
3D XPoint Slashes Latency

Source: Storage Technology Group, Intel

SSD NAND technology offers ~100X reduction in latency versus HDD

NVMe™ eliminates 20 μs of latency today

3D XPoint™ technology reduces NVM latency offering ~10x reduction in latency vs NAND SSD

© 2016 Coughlin Associates & Objective Analysis
WHY MORE STORAGE LAYERS
Mapping This To Storage Technologies


© 2016 Coughlin Associates & Objective Analysis
PROCESSORS MUST ADAPT
Context Switches Become The Issue

Doug Voigt, HP, 2015 FMS

Min, Max Latencies For Example Technologies

Context Switch

Non-Uniform Memory Access

Latency (Log)

2 uS

200 nS

HDD

SATA SSD

NVMe Flash

Persistent Memory

© 2016 Coughlin Associates & Objective Analysis
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
Summary

• Scaling limits open doors to new memories
  – New architectures will create other opportunities

• NAND is **NOW**
  – 3D XPoint is coming soon

• Performance drives need for new layers

• New layers will drive new processor architectures
  – Application programs will also adapt
Thank You!

Tom Coughlin, Coughlin Associates
Jim Handy, Objective Analysis