A 22nm 20Mb Embedded MRAM with 5Gbps Read and 1Gbps Programming

Presenter: Nick Hendrickson

Introduction

• The data presented here is for Numem’s first generation TSMC 22nm MRAM
• The primary goal of this architecture is for high performance, high reliability embedded MRAM IP
• The testchip along with our custom test platform provides a low cost, scalable means to test large amounts of memory in parallel while still allowing discreet analog access to every cell on the chip.
Chip Overview

- The testchip makes use of 16 high-performance, high-density MRAM instances
  - TSMC 22nm
  - 0.0456um² bitcell
- Numem’s standard testchip interface enables very high throughput, flexibility and control at modest pincounts
  - 32 pin digital interface
  - 3 analog pins
  - Remainder are various power/ground
  - 100 pin total
Instance Overview

- Daisy chainable block; 2 deep for demonstration purposes
  - Center mounted Sense Amplifiers to share between two MRAM arrays without additional metal expense
  - 512WL x 640BL array
  - 40b Data Word
  - Easily adaptable to 80b Data Word for higher throughput and area efficiency

- Each instance interfaces by means of a standard SP SRAM interface
  - Simplifies integration into existing systems
  - Provides sufficient bandwidth to not bottleneck operation
  - Incurs no meaningful latency penalty at system level
Read Architecture

• Reference generation is the most critical element of an MRAM read architecture. There reference must track:
  - PVT
  - Bitline resistance/position
  - Sourceline resistance/position
  - Wordline voltage / access device resistance
  - Without compensation of these terms, a state separation of just 2.4 sigma is left from a starting separation of 8 sigma!

• Numem’s patented reference generation methodology has demonstrated near perfect compensation for all of these variations, retaining nearly all of the theoretical read window

• A forced current sense approach is able to settle within 6ns
  - Allowing for address/data propagation and sense resolve time, the total access time is 8ns, or 5Gbps
Programming Performance

- All 40 bits are programmed in parallel
- Verify is implemented to eliminate program soft errors
- Single pulse SER of <100PPM is achieved @ 32ns
  - Just 10% overdrive voltage required during programming
  - Increasing program time to 64ns only reduces the VPRG by 6%
  - Lower programming time and increased programming voltage generally improves overall power usage

<table>
<thead>
<tr>
<th>VPRG</th>
<th>32ns</th>
<th>64ns</th>
<th>96ns</th>
<th>128ns</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.14x</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1.12x</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1.10x</td>
<td>45</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1.08x</td>
<td>113</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1.06x</td>
<td>289</td>
<td>6</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1.04x</td>
<td>735</td>
<td>29</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>1.02x</td>
<td>2097</td>
<td>174</td>
<td>36</td>
<td>11</td>
</tr>
<tr>
<td>1.00x</td>
<td>6682</td>
<td>1193</td>
<td>408</td>
<td>184</td>
</tr>
</tbody>
</table>

* VPRG scaled to nominal process maximum voltage
Aging Results

• The bitline voltage data shown quantifies the total resistance of the cell including RMTJ, RACC, RBL, RSL, etc.
  ▪ Chip level features are mostly compensated for in the test itself
  ▪ While actual sense window distributions are better compensated, this provides a detailed analog view of cell aging

• Across cycling, virtually no drift in resistance for either \( R_p \) or \( R_{ap} \) is observed
  ▪ These values have been measured at time 0, 1e6, and 1e7 with no noticeable change in median or sigma values
  ▪ This is important in order to ensure reference placement / read windows are constant as the device ages
Yield

- Local yield on healthy die are very good, with the process moving into production fabs
- Column and Wordline repairs are optional, but generally recommended to optimize yield
  - On this testchip, over 90% of die show no need for either repair mechanism anywhere on the die
- Bit level repairs are reasonable and being actively monitored as the process continues to mature
  - This design implemented with 16 bit repairs per Mb; nearly all otherwise healthy die yield within that repair
  - Spatial mapping shows failing bits to be randomly distributed throughout the memory
  - Bit defectivity includes: opens, shorts, out-of-distribution resistance
  - All of these defect types are repaired using Word repair mechanism
Summary

• This test chip demonstrates a current generation MRAM technology, production ready, delivering on some of the most aggressive performance predictions

• A simple SP SRAM interface is ideal to harness the bandwidth and latencies available for embedded implementations

• High performance and low power for both reads and writes positions this MRAM as both a superior NVM replacement as well as a real SRAM competitor for low power, low speed sockets

• Yields are already sufficient for volume manufacturing and continue to improve