Basic Principles, Challenges and Opportunities of STT-MRAM for Embedded Memory Applications

Luc Thomas

TDK- Headway Technologies,  
463 S. Milpitas Boulevard, Milpitas CA 95035, USA
Guenole Jan, Son Le, Santiago Serrano-Guisan
Yuan-Jen Lee, Huanlong Liu, Jian Zhu,
Jodi Iwata-Harms, Ru-Ying Tong, Sahil Patel,
Vignesh Sundar, Dongna Shen, Yi Yang, Renren He,
Jesmin Haq, Jeffrey Teng, Vinh Lam, Paul Liu,
Magnetic Random Access Memories

-more than 20 years ago: Field-MRAM

1st research program: IBM / Motorola (1995)

From S. Parkin and K. Roche IBM
60 years ago: TDK first foray in MRAM technology

→ TDK’s 18x24 bit Magnetic Core Memory

→ MRAM was the predominant computer memory from the 50’s to the 70’s
Outline

- Basic principles of STT-MRAM
- STT-MRAM integration
- STT-MRAM in emerging memory landscape
Magnetic Tunnel Junction (MTJ) device

- Two ferromagnetic electrodes separated by a thin MgO tunnel barrier

- Tunnel Magnetoresistance (TMR): device resistance depends on the relative orientation of the magnetization of the two magnetic electrodes

From S. Parkin and K. Roche IBM

Yuasa et al. (AIST) Nature Materials 2004
Magnetic Tunnel Junction (MTJ) device

- Two ferromagnetic electrodes separated by a thin MgO tunnel barrier

- Tunnel Magnetoresistance (TMR): device resistance depends on the relative orientation of the magnetization of the two magnetic electrodes

Reproduced from website of MultiDimension Technology Co., Ltd.

Yuasa et al. (AIST) Nature Materials 2004
Perpendicular Magnetic Anisotropy (PMA) MTJ

- PMA is needed for data retention scaling and writing efficiency.

- PMA is based on interfacial anisotropy between MgO and CoFeB (Ikeda et al., Nature Mat. 2011, Worledge et al., APL 2012).

- Free layer sandwiched between MgO interfaces for the free layer for enhanced anisotropy.

- Dual reference layer for reducing dipolar fields and enhanced stability.

Ikeda et al., IEDM2014
High data retention in PMA-MTJs

- Developed a MTJ stack of high PMA and thermal stability to satisfy solder reflow requirement of 260°C for 90 seconds (2016 VLSI TSMC/TDK)

- Method of projecting error rate from chip level data in ppm regime

1ppm 10 years retention at 225°C
Resistance vs magnetic field hysteresis loops

Two well-defined resistance states depending on orientation of magnetic electrodes
Reading with Tunnel Magnetoresistance

- Read operation by probing the resistance of the device at low voltage bias

- True Binary device: no resistance drift of the 2 resistance state even after repeated cycling at maximum drive current


Writing with Spin-Transfer Torque

Transfer of spin-angular momentum from polarized conduction electrons to electrodes magnetization

Phenomenon discovered in 1996 by two theoreticians: John Slonczewski (IBM) Luc Berger (Carnegie Mellon)

Read: Tunnel Magnetoresistance

Write: Spin Transfer Torque

Reproduced from Quantumwise.com
Trade-offs of STT writing

- Switching Current scales with area (constant current density)
  - smaller device -> smaller current requirement

- Current inversely proportional to pulse width
  - faster -> higher current requirement
Trade-offs of STT writing (cont’d)

→ Write current scales with energy barrier for data retention

Energy barrier: \( E_B \sim K_u V \)

Write current: \( I_{c0} = \frac{4e}{\hbar} \left( \frac{\alpha}{P} \right) E_B \)

STT efficiency: \( \frac{E_B}{I_{c0}} \sim 1-2 \) in \( k_B T/\mu A \)

→ Writing is probabilistic

- STT vanishes for parallel alignment of PL and FL
- Switching time inversely proportional to angle between PL and FL
- Thermal fluctuations provide initial ‘kick’
Outline

→ Basic principles of STT-MRAM

→ STT-MRAM integration

→ STT-MRAM in emerging memory landscape
Integration of 8 Mb test chips at TDK - Headway

- 8Mbits (16x512k) 1T-1MTJ
- IBM’s 90nm CMOS technology
- 50F² cell size
- Sense Amplifiers for reading
- Redundancy and 2bit ECC
- FEOL in IBM foundry
- BEOL in TDK-Headway’s fab
STT MRAM process integration

- MRAM only add three additional layers (MTJ and electrodes) to standard CMOS BEOL: 3 to 4 mask adder
- MTJ stack is about 20 nm thick, can be easily integrated into CMOS backend process
Defect rate of 8 Mb chip

- Distribution of device current in the P state

**Quantile plot**

Distribution of device current in P state

**Log scale**

Distribution of device current in P state

\[ \Rightarrow \text{less than 0.4 ppm defect rate} \]
400C annealing after MTJ patterning

- 400C BEOL process can add up to several hours, depending on how many metal layers on top of MTJ
- Elemental movements and morphology changes can degrade anisotropy, exchange coupling, and defect level
  - selection of materials, diffusion barrier and interface/growth quality
  - Thorough engineering needed for electrodes, film stack, process, encapsulation

- 2.5 hours @400°C after MTJ etching
- Diameter ~ 30 nm (electrical)
- DRR = 175%
- RA of 8.5 Ω-μm²
- $H_C = 3300$ Oe with no offset
Error free writing in chip level (TDK VLSI 2014 & 2016)

- Error free writing on 8 Mb chips without ECC
  - Down to 6 ns write pulse
  - While keep data retention to 142°C for 10 years

![Graphs showing error count as a function of BL Voltage and temperature]

- 1ppm @ 142°C for 10 years
Temperature dependence (TDK VLSI2014)

Fast operation down to 4.5 ns demonstrated over wide temperature range

<table>
<thead>
<tr>
<th>Temperature</th>
<th>-25°C</th>
<th>0°C</th>
<th>25°C</th>
<th>55°C</th>
<th>85°C</th>
<th>125°C</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No ECC</strong></td>
<td><img src="image1" alt="Graph" /></td>
<td><img src="image2" alt="Graph" /></td>
<td><img src="image3" alt="Graph" /></td>
<td><img src="image4" alt="Graph" /></td>
<td><img src="image5" alt="Graph" /></td>
<td><img src="image6" alt="Graph" /></td>
</tr>
<tr>
<td></td>
<td>4.5</td>
<td>10</td>
<td>4.5</td>
<td>10</td>
<td>4.5</td>
<td>10</td>
</tr>
<tr>
<td><strong>2 bit ECC</strong></td>
<td><img src="image7" alt="Graph" /></td>
<td><img src="image8" alt="Graph" /></td>
<td><img src="image9" alt="Graph" /></td>
<td><img src="image10" alt="Graph" /></td>
<td><img src="image11" alt="Graph" /></td>
<td><img src="image12" alt="Graph" /></td>
</tr>
<tr>
<td></td>
<td>4.5</td>
<td>10</td>
<td>4.5</td>
<td>10</td>
<td>4.5</td>
<td>10</td>
</tr>
</tbody>
</table>
Outline

- Basic principles of STT-MRAM
- STT-MRAM integration
- STT-MRAM in emerging memory landscape
## STT-MRAM vs other memory technologies

### Stand Alone Memory Position in 2014: commercial products performances

<table>
<thead>
<tr>
<th></th>
<th>Emerging Memory</th>
<th>Established Memory</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>STT MRAM</td>
<td>PCM</td>
</tr>
<tr>
<td><strong>Non-Volatile</strong></td>
<td>YES</td>
<td>YES</td>
</tr>
<tr>
<td><strong>Endurance (Nb cycles)</strong></td>
<td>High ($10^{12}$)</td>
<td>Medium ($10^{10}$)</td>
</tr>
<tr>
<td><strong>2014 latest technological node produced (nm)</strong></td>
<td>90 nm</td>
<td>45 nm</td>
</tr>
<tr>
<td><strong>Cell size (cell size in $\mu$m)$^2$</strong></td>
<td>Medium (6-12)</td>
<td>Medium (6-12)</td>
</tr>
<tr>
<td><strong>Write speed (ns)</strong></td>
<td>High (10 ns)</td>
<td>Medium (75 ns)</td>
</tr>
<tr>
<td><strong>Power consumption</strong></td>
<td>Medium/low</td>
<td>Medium</td>
</tr>
<tr>
<td><strong>2014 price ($/Gb$)</strong></td>
<td>High ($100 - 50/Gb$)</td>
<td>Medium (few $$/Gb)</td>
</tr>
<tr>
<td><strong>Suppliers</strong></td>
<td>Everspin</td>
<td>Micron, Samsung</td>
</tr>
</tbody>
</table>

Emerging memory has distinctive technical features (endurance, speed, non-volatility), but price and scalability are obstacles to competing with dominant DRAM, NAND or SRAM memory.
STT-MRAM requirements

- Critical requirements depend on application

<table>
<thead>
<tr>
<th>Priority</th>
<th>Embedded STT-MRAM</th>
<th>Standalone STT-MRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Higher</td>
<td>Unified NVM</td>
<td>NV Working RAM</td>
</tr>
<tr>
<td></td>
<td>Retention</td>
<td>NV DRAM</td>
</tr>
<tr>
<td></td>
<td>Cell Size/Chip Size</td>
<td>NV SRAM</td>
</tr>
<tr>
<td></td>
<td>Density</td>
<td>Endurance</td>
</tr>
<tr>
<td></td>
<td>Process Integration</td>
<td>Performance/Speed</td>
</tr>
<tr>
<td></td>
<td>Endurance</td>
<td>Density</td>
</tr>
<tr>
<td></td>
<td>Performance/Speed</td>
<td>Endurance</td>
</tr>
<tr>
<td></td>
<td>Cell Size/Chip Size</td>
<td>Performance/Speed</td>
</tr>
<tr>
<td></td>
<td>Retention</td>
<td>Recovery</td>
</tr>
</tbody>
</table>

from S.H Kang, Qualcomm (Proc. VLSI 2014)
STT-MRAM Challenge

Cost is directly related to density & cell/chip size

Current available scales with transistor size

- Standalone DRAM: GB chips, cell size $\sim 4F^2$
  
  $F$ smallest feature at technology node (28, 20, 14/16nm, …)
  
  MTJ < 20 nm
  
  Write current < 20 $\mu$A
  
  TMR $\sim$ 300%


- Embedded Flash / DRAM: cell size $\sim 40-50F^2$

  MTJ $\sim$ 40-100 nm

  Write current > 100 $\mu$A

  TMR > 100%

L Thomas et al., MSST 2017 - Santa Clara, May 17th, 2017
Embeded STT-MRAM is cheaper and better!

- Lower cost
  - Similar or Smaller bit cell size
  - Very few added mask layers
  - Does not interfere with CMOS transistor performances (as a add-on in the backend metal layers)

<table>
<thead>
<tr>
<th></th>
<th>eMRAM</th>
<th>eFlash</th>
<th>eDRAM</th>
<th>SRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cell size (F^2)</td>
<td>30-50</td>
<td>30-50</td>
<td>30-90</td>
<td>100-300</td>
</tr>
<tr>
<td>Added mask layers</td>
<td>2-3</td>
<td>10-12</td>
<td>4-6</td>
<td>0</td>
</tr>
</tbody>
</table>

- Almost “universal memory”
  - Combines non-volatility, high speed, and infinite endurance
  - Can replace eFlash, eDRAM, and last-level cache (LLC) SRAM
  - Efficient system architectures, without moving data between code storage, and working memory, and data storage

- Higher energy efficiency (longer battery life)
  - Mobile and IoT applications have low duty cycles and need fast wake-up and low standby power
6-Transistor SRAM scaling challenge

- 22nm to 10 nm node:
  - Expected area scaling: 4.8X
  - Actual scaling: ~ 2X
- 400F^2 at 10nm vs 52F^2 at 40nm
- Complex design limits scaling

Dramatic increase of the area occupied by memory vs logic in performance SoC and CPU’s
Opportunity for eMRAM as Last Level Cache

- Compact design 1T-1MTJ
- 8 Mb written without error with 1.5 ns write pulse
Summary

- STT-MRAM combine low write current, data retention and write speed, and is compatible with BEOL processes.

- Working chips have been demonstrated

- MTJ device can be tailored to specific applications that require data retention or speed,

- Great opportunity for embedded applications from eFlash to SRAM replacement (both Samsung and TSMC have announced production)

- Many challenges remain: writing efficiency, read margin (TMR), process control (tight pitch, uniformity), …
1970: Magnetic memories lose the war to Silicon

2017: year of the comeback for MRAM?

Circa 1970 – Intel corporation - Computer history museum