Spin-Hall Effect MRAM Based Cache Memory: A Feasibility Study

Jongyeon Kim, Bill Tuohy, Cong Ma, Won Ho Choi, Ibrahim Ahmed, David Lilja, and Chris H. Kim

University of Minnesota
Dept. of ECE
Overview

1. Basic Concepts of SHE-MRAM

2. SHE-MRAM Device Design
   - Material and Device Parameters set-up

3. SHE-MRAM Circuit Design
   - SPICE Modeling and Macro Design

4. Cache-level Technology Benchmarking
   - Feasibility of SHE-MRAM as L2 Cache

5. Conclusion
STT-MRAM: Pros & Cons

Pros: 1. Zero static power with nonvolatility, 2. Compact cell size
3. Shorter latency for large caches (e.g. L3, L4) from reduced global interconnect delay owing to compact bit-cell size

Cons: 1. High write energy, 2. Limited TMR 3. Read/write conflict with scaling

Spin-Hall Effect MRAM (SHE-MRAM)

- Low $I_c/\Delta$ with efficient spin generation (i.e. $I_{\text{spin}}/I_{\text{charge}} > 100\%$)
- Longer device lifetime owing to the decoupled read and write paths
- A comprehensive study showing the feasibility of SHE-MRAM for large on-die cache memory not reported yet

Thermal Stability Criterion

- $\Delta = 64$ is set by considering a fixed bit-cell failure rate (0.01%) under a 10 year data retention time.
- The maximum read current ($I_{\text{read}}/I_{\text{write}} = 15\%$) is also determined based on the same read disturbance failure rate.

\[ F_{\text{bi-cell}} = 1 - \exp\left[-m \frac{t}{\tau_0} \exp\left\{ \frac{E}{k_B T} (1 - \frac{I_{\text{cell}}}{I_{\text{write}}}) \right\} \right] \]

- Failure mode: 10yr data retention \([1]\)
- $m$: total memory size
- $t$: 10yrs, $I_{\text{cell}} = 0$
- Failure mode: Read disturbance
- $m$: number of bits per read
- $t = t_{\text{read}}/t_{\text{cycle}} \times 10\text{yrs}$, $I_{\text{cell}} = I_{\text{read}}$

Parameter Set-up: STT-MRAM vs. SHE-MRAM

<table>
<thead>
<tr>
<th>Parameters</th>
<th>STT-MRAM</th>
<th>SHE-MRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTJ type</td>
<td>Interface perpendicular</td>
<td>In-plane</td>
</tr>
<tr>
<td>Thermal stability factor</td>
<td>65</td>
<td>65</td>
</tr>
<tr>
<td>Free layer material</td>
<td>CoFeB</td>
<td>CoFeB</td>
</tr>
<tr>
<td>Free layer dimensions, $W_f \times L_f \times t_f$</td>
<td>40nm$\times$40nm$\times$1.34nm</td>
<td>22nm$\times$77nm$\times$2.7nm</td>
</tr>
<tr>
<td>Saturation magnetization, $M_s$</td>
<td>$1.077 \times 10^3$A/m</td>
<td>$1.077 \times 10^2$A/m</td>
</tr>
<tr>
<td>Damping factor, $\alpha$</td>
<td>0.018</td>
<td>0.006</td>
</tr>
<tr>
<td>Polarization factor, $P$</td>
<td>0.63</td>
<td>0.63</td>
</tr>
<tr>
<td>Critical thickness, $t_c$</td>
<td>1.5nm</td>
<td>-</td>
</tr>
<tr>
<td>TMR</td>
<td>130%</td>
<td>130%</td>
</tr>
<tr>
<td>RA ($\Omega \cdot \mu m^2$)</td>
<td>5</td>
<td>5.5</td>
</tr>
<tr>
<td>SHM dimensions, $W_{SHM} \times L_{SHM} \times t_{SHM}$</td>
<td>-</td>
<td>77nm$\times$44nm$\times$2.2nm</td>
</tr>
<tr>
<td>SHM spin diffusion length, $\lambda_{ch}$</td>
<td>-</td>
<td>1.5nm</td>
</tr>
<tr>
<td>SHM resistivity, $\rho_{SH}$</td>
<td>-</td>
<td>200$\mu \Omega \cdot cm^2$ (for W)</td>
</tr>
<tr>
<td>Spin Hall angle, $\theta_{SH}$</td>
<td>-</td>
<td>0.3 (for W)</td>
</tr>
</tbody>
</table>

* $\Delta$ and material parameters are extracted based on 85°C.
* RA of SHE-MRAM includes $R_{SHM}$. Thickness dependency of $\alpha$ is also considered.

• Parameters determined based on $\Delta$ requirement for the 22nm node
• SHE-MRAM: In-plane MTJ + Tungsten (W) SHM
• Maximum spin generation takes place when $t_{SHM}=2.2$nm
SHE-MRAM SPICE Model Framework

**LLG:**
\[
\frac{1+\alpha^2}{\gamma} \frac{dM}{dt} = -M \times H_{\text{Keff}} - \alpha \cdot M \times (M \times H_{\text{Keff}}) + \frac{\hbar I_s}{2eWL_i F M_s} \cdot M \times (M \times M_p)
\]

**SHE:**
\[
I_s = \frac{A_{\text{MTJ}}}{A_{\text{SHM}}} \theta_{\text{SH}} (1 - \text{sech} \left( \frac{t_{\text{SHM}}}{\lambda_{\text{sf}}} \right)) I_{\text{ch}}
\]

- SPICE-compatible SHE-MTJ model was implemented by incorporating the spin current from SHM into LLG equation
Memory Macro for Circuit Simulation

- 22nm CMOS Predictive Technology Model (PTM) [1]
- Features bi-directional write current drivers, dual-voltage WL drivers, reference circuit using $I_{\text{Ref}} = (I_{\text{AP}} + I_{\text{P}})/2$
- Read current in the AP to P direction to minimize read disturbance

22nm FinFET Based Bit-Cell Layout

* 22nm FinFET design rule: [1]
  \[ W_{\text{fin}} = 8\text{nm}, L_{\text{fin}} = 24\text{nm}, H_{\text{fin}} = 34\text{nm}, P_{\text{fin}} = 60\text{nm}, \]
  \[ W_m = 33\text{nm}, W_c = 22\text{nm}, W_{g2c} = 22\text{nm} \]

(a) STT-MRAM: \( N_{\text{fin}} = 4 \) for access TR ↑
(b) SHE-MRAM: \( N_{\text{fin}} = 2 \) for read/write TRs →

- Both SHE-MRAM and STT-MRAM are roughly 3x denser
- Two fins for read/write transistors makes the cell area of SHE-MRAM comparable to that of a standard 1T1R STT-MRAM cell

256Kbit Sub-array Performance

Results indicate that SHE-MRAM will always outperform STT-MRAM regardless of the cache size.

- For large caches such as L3 and L4, SHE-MRAM will have a shorter access latency than SRAM due to the denser cache size.
- Can SHE-MRAM outperform SRAM for smaller L2 caches (1Mbit)?
Read Performance Boosting with High $\Delta$

- **SHE-MRAM** shows small write overhead with increase in $\Delta$
- Higher $\Delta$ allows larger read current resulting in shorter read delay
SHE-MRAM L2 Cache (1Mb) Performance

* 1Mbit, 8-way associativity, private bank, CACTI simulator

<table>
<thead>
<tr>
<th>Metrics</th>
<th>SRAM</th>
<th>STT-MRAM</th>
<th>SHE-MRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thermal stability (@85°C)</td>
<td>-</td>
<td>65</td>
<td>85</td>
</tr>
<tr>
<td>Bit-cell failure rate (%)</td>
<td>-</td>
<td>$10^{-2}$</td>
<td>$10^{-11}$</td>
</tr>
<tr>
<td>Read latency (ns)</td>
<td>0.42</td>
<td>0.71</td>
<td>0.43</td>
</tr>
<tr>
<td>Read energy (nJ)</td>
<td>0.07</td>
<td>0.22</td>
<td>0.44</td>
</tr>
<tr>
<td>Write latency (ns)</td>
<td>0.42</td>
<td>6.77</td>
<td>1.95</td>
</tr>
<tr>
<td>Write energy (nJ)</td>
<td>0.10</td>
<td>0.41</td>
<td>0.21</td>
</tr>
<tr>
<td>Leakage power (mW)</td>
<td>39.5</td>
<td>4.96</td>
<td>4.96</td>
</tr>
<tr>
<td>Area (mm²)</td>
<td>0.55</td>
<td>0.16</td>
<td>0.16</td>
</tr>
</tbody>
</table>

L2 cache performance summary

- Read latency comparable to that of SRAM but with a lower leakage power and denser area
- Higher TMR and efficient sensing circuits are necessary to reduce the high read energy incurred by the current-forcing read
Conclusion

- We explore the trade-off points across different levels of design abstraction (i.e. device, circuit, and architecture) to evaluate the feasibility of SHE-MRAM for large on-die cache memory.
- SHE-MRAM has a 4.7x shorter write time and 1.3x shorter read delay as compared to a standard STT-MRAM having the same cell size.
- Read/Write latencies of the denser SHE-MRAM can be comparable to those of SRAM for L2 caches but larger read/write energy is needed.

Acknowledgement

- This work was supported in part by C-SPIN, one of six centers of STARnet, a Semiconductor Research Corporation program, sponsored by MARCO and DARPA.