Area, Power, and Latency Considerations of STT-MRAM to Substitute for Main Memory

Youngbin Jin, Mustafa Shihab, and Myoungsoo Jung

Computer Architecture and Memory Systems Laboratory
Department of Electrical Engineering
The University of Texas at Dallas
Takeaways

- **Challenge:** Memory Scaling and Leakage and Refresh Power consumption issues are constraining DRAM’s long-term future as main memory.
- **Probable Solution:** Replacing DRAM with new non-volatile memories such as STT-MRAM.
- **Underlying Issues:**
  - Cell area of STT-MRAM is nearly four times that of DRAM, restraining it from achieving cell density required by main memory.
  - Write latency and power is too large and inhibit STT-MRAM’s implementation as main memory.
- **Our Contribution:**
  - Classified the critical device parameters, including thermal stability factor, that can make STT-MRAM better suited for main memory.
  - Introduced an optimized STT-MRAM that can offer DRAM-like density, lower power consumption, and shorter latency.
- **Results:** Our early-stage optimized STT-MRAM can offer shorter latency and lower power consumption than a baseline DRAM by 18.4% and 66.2%, respectively.
Overview

• Motivation
• MTJ, STT-MRAM Read and Write
• Critical Parameters
• Memory Level Optimizations
• Conclusion
For around 3 decades DRAM has been THE main memory of choice

However....

- Cell leakage current is making scaling down for larger memory a critical bottleneck
- Refresh power for large memory is becoming almost as high as the processing power
Probable Solutions

Short-Term: Low-leakage, energy-efficient DRAMs

Long-Term: Replacing DRAM with newer and more efficient memory technology

**STT-MRAM** is a popular *Heir Apparent*

**STT-MRAM is great:**
- Highly Scalable
- Zero Leakage
- High Endurance

**But, nothing comes for free…**
- Cell Area: 30F² STT-MRAM cell is 4x larger than DRAM (4 to 8 F²)
- Write power and latency is too high as well
What do we do?

It is not a dead-end, rather an opportunity for exploration.

We classify and tweak the Thermal Stability Factor and other critical parameters of STT-MRAM/MTJ to find Pareto optimal points.

Optimized STT-MRAM can offer:

i) Higher density
ii) Lower power consumption on writes
iii) Shorter write latency

All comparable to modern DRAM technology.
MTJ, STT-MRAM Read and Write

Magnetic Tunnel Junction (MTJ):
- Two ferromagnets separated by a thin insulator
- Stores data in terms of resistance level

Read Operation:
- Just sensing of the cell resistance
- Non-destructive read involves no write-back
- Only takes a couple of nanoseconds

Write Operation:
- Use polarized electrons to torque the magnetic state
- This process makes STTMRAM cell so large
  If the transistor is not large enough to drive enough current then write failure occurs
Critical Parameters

Thermal Stability Factor (Δ):
- The stability is intrinsically linked to the magnitude of the energy barrier between the two orientations.
- During a write process, energy barrier must be overcome.

The higher the thermal stability factor, the higher the current (energy) or longer time it requires to write.

\[ \Delta = \frac{E_b}{k_B T} = \frac{H_K M_s V}{2k_B T} \]

where, \( E_b \) = Energy Barrier, \( k_B \) = constant, \( T \) = Temperature, \( H_K \) = Anisotropy Field Term, \( M_s \) = Saturation Magnetization, and \( V \) = Volume of the MTJ = Area of the MTJ * Thickness of the MTJ = A. \( t_h \)
Critical Parameters

Critical Current (IC):
- Determines the amount of current access transistor has to drive

If the critical current can be reduced, access transistor size can be reduced. We can achieve smaller cell area for STT-MRAM

\[
I_{c0} = \left( \frac{4e k_B T}{h} \right) \cdot \frac{\alpha}{\eta} \cdot \Delta \cdot \left( 1 + \frac{4\pi M_{eff}}{2H_K} \right)
\]

Where, \( e \) = electron charge, \( h \) = Plank’s Constant, \( \alpha \) = damping constant, \( \eta \) = STT-MRAM efficiency parameter, \( 4\pi M_{eff} \) = Effective demagnetization field, and \( \Delta \) = Thermal Stability Factor
Critical Parameters

Retention Time:
- The expected time until a random bit-flip occurs in the cell
- Is determined by the thermal stability of the MTJ
High stability indicates the cell will have longer retention time, at the cost of higher write power and latency

\[ \tau = \tau_0 \cdot \exp \left( \frac{E_b}{k_BT} \right) \]

where, \( \tau_0 = \) operating frequency
Memory Level Optimization

DRAM Baseline:

Process technology: 45 nm

Cell Size:
$6F^2 \rightarrow 12150 \text{ nm}^2$ (for 45 nm technology)

Write power: 88 $\mu$W*

*taken from [15]
Memory Level Optimization

Area Optimization:

The diagram shows the area optimization of storage with varying retention times for different technologies (16nm, 22nm, 32nm, 45nm, and DRAM). The x-axis represents the retention time (1us, 1ms, 1s, 1min, 1h, 1d, 1mo, 1y, 10y), and the y-axis represents the area in nm². The graph illustrates how the area increases over time for each technology, with 16nm showing a steeper increase compared to 22nm and 32nm, which have a more moderate increase. The 45nm technology exhibits a slower increase, and DRAM shows the least increase in area. The 232ms and 27.29us are specific time points highlighted, indicating the retention times for these technologies.
Memory Level Optimization

Power Optimization:

- 60s
- 232ms
- 27.3us Retention Time
- 16nm
- 22nm
- 32nm
- DRAM 88uW
Memory Level Optimization

Write Latency Optimization:

- 60s Retention Time: 3.7ns latency, 88uW power
- 232ms Retention Time: 0.8ns latency, 100uW power
- 27.3us Retention Time: 88uW power
Summary

Our early-stage optimized STT-MRAM:
- Reduced power by 98.64%, 67.48%, and 32.48%
- Reduced normalized latency to 0.31, 0.17, and 0.074, compared to baseline DRAM

Potential Implementations:

<table>
<thead>
<tr>
<th>Process Technology</th>
<th>Thermal Stability Factor</th>
<th>Area (= 45nm DRAM)</th>
<th>Write Power (μW)</th>
<th>Write Latency (ns)</th>
<th>Retention Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>32 nm</td>
<td>27.26</td>
<td>12150 nm²</td>
<td>86.81</td>
<td>4.20</td>
<td>60 s</td>
</tr>
<tr>
<td>22 nm</td>
<td>19.26</td>
<td>12150 nm²</td>
<td>59.39</td>
<td>2.30</td>
<td>232 ms</td>
</tr>
<tr>
<td>16 nm</td>
<td>10.21</td>
<td>12150 nm²</td>
<td>28.59</td>
<td>1.00</td>
<td>27.3 μs</td>
</tr>
</tbody>
</table>
Related Prior Works

Zhenyu Sun et al. proposed both L1 cache and lower level cache designs using multi-retention level STT-MRAM cache, which can significantly reduce the total energy, while improving write for both level 2 and level 3 caches\(^{[16]}\).

Clinton W. Smullen et al. claimed that retention-relaxed STT-MRAM can replace SRAM in processor caches by reducing the high dynamic energy and slow write latencies\(^{[12]}\).

Emre Kultursay et al. explored the possibility of using STT-MRAM technology to completely replace DRAM in main memory by using partial write and row buffer write bypass\(^{[17]}\).
Thank you

Questions?
References

References