Integration of STT-MRAM model into CACTI simulator

S. Arcaro, S. Di Carlo, M. Indaco, D. Pala, P. Prinetto, Elena I. Vatajelu
Dip. di Automatica e Informatica
Politecnico di Torino
Turin, Italy
{firstname.lastname}@polito.it

Abstract—In the last decade, academies and private companies have actively explored emerging memory technologies. STT-MRAM in particular is experiencing a rapid development but it is facing several challenges in terms of performance and reliability. Several techniques at cell level have been proposed to mitigate such issues but currently few tools and methodologies exist to support designers in evaluating the impact that specific micro-level design choices can determine on the STT-MRAM macro design. In this paper we present a system-level tool based on CACTI simulator to assist memory system designers. We use our tool to generate high-performance and low-power cache memories comparing performance, energy consumption, and area with traditional SRAM.

Keywords—STT-MRAM, CACTI, Emerging Memories

I. INTRODUCTION

The focus of emerging memories is placed on non-volatile technologies which should meet the high demands of tomorrow applications. That includes non-volatility, high performance and high density similar to SRAMs and DRAMs respectively, good endurance features, small devices sizes, good integration, low power profile, resistance to radiation effects, and ability to scale below 20nm.

One of the most promising candidate as embedded memory is the spin-transfer torque magnetic random access memory (STT-RAM) [1] offering faster read and write access time (nanoseconds) and better CMOS integration compared to other proposed technologies such as Phase-Change RAM (PCRAM) [2], Resistive RAM (RRAM) [3] and Ferromagnetic RAM (FeRAM) [4]. The key building block of STT-MRAM cell is the magnetic tunneling junction (MTJ) that is integrated with CMOS circuitry using 3-D technology [5]. The smallest STT-MRAM cell design is a 1T1MTJ (one transistor, one magnetic tunneling junction) device. Logical data is stored by applying the spin polarized current through the MTJ element to switch the memory states.

Anyway, with scaling, STT-MRAM cell is facing a set of challenges that strongly influence performances and reliability, severely affecting the yield of the memory array. Such issues are mainly related to a) process variations of MOS and MTJ devices involving the variation of geometry size, threshold voltage, and magnetic materials [5], [6] b) the high write cost due to high switching current required to flip the MTJ state [7], and c) the thermal fluctuations in the MTJ switching [8].

To tackle such issues, efficient design paradigm at cell level from circuit and/or architecture perspective to improve the cell robustness and integration density have been proposed. However, achieved results for STT-MRAM cell design may be not directly adapted to meet high-level design requirements.

It is of utmost importance to quantify and to assess the performance degradation in terms of write/read latency, power consumption, and area that can potentially affect the behavior of the whole memory array when specific requirements-driven designs at cell level are targeted.

For this reason, more comprehensive tools and methodologies are necessary to provide flexibility for design experiments. In this context, Smullen et al. present a methodology and tool-chain for evaluating and comparing MTJs design [15]. In [11] authors propose a fixed analytical STT-MRAM model in CACTI, to analyze the power reduction in modern microprocessors when SRAM is replaced with STT-MRAM. CACTI is a widely used high-level cache and memory modeling tool [9] [10].

In this paper we present a system-level tool based on CACTI simulator to estimate area, energy consumption and write/read latency of STT-MRAM based cache memories. The tool supports a parameterizable interface where a wide set of physical parameters of STT-MRAM technology can be specified. The implemented extensions enable our tool to be integrated with system-level emulation tools such as QEMU, as well. In order to prove the correctness of our tool, we generate STT-MRAM based cache memories with different sizes comparing the resulting performances with SRAM technology. The proposed tool, thus, can support the design of cache or main memories by evaluating the impact that specific micro-level design choices can determine on the STT-MRAM macro design. The tool is made available and it can be freely downloadable from the website of our research group: http://www.testgroup.polito.it/.

The paper is organized as follows: Section II describes operation principles of STT-MRAM technology and shortly CACTI tool. In Section III modeling and parameterization of STT-MRAM technology that we implemented in CACTI is discussed while in Section IV a comparison of three MTJ configurations for each use-case is given. Section V concludes the paper.
II. BACKGROUND

In this section, an overview about STT-MRAM technology in terms of operation principles and electrical model is given. Finally, the main features of CACTI tool are described.

A. Basic Principles

STT-MRAM technology is built up upon the magnetic tunneling junction (MTJ) device which aims at persistently store logic data. Commonly, an MTJ device is composed of two ferromagnetic layers (FLs) interleaved with one oxide barrier layer. FLs are characterized by their magnetic orientation: one has a fixed magnetic orientation (fixed layer) and the other has a freely rotating magnetic orientation (free layer). By applying a sufficiently dense current pulse through the MTJ device, the free layer magnetic direction can be dynamically switched.

B. Electrical Model

When the FLs exhibit the same magnetic orientation, the MTJ has a low electrical resistance, whereas MTJ experiences high electrical resistance in presence of antiparallel configuration. Typically, the low electrical resistance \((R_{MTJ} = R_L)\) is associated with logic state ‘0’ and the high electrical resistance \((R_{MTJ} = R_H)\) is associated with the logic state ‘1’, as depicted in Fig. 1.

![MTJ configurations](image)

According to the relative magnetic orientations of the two layers, the electrical resistance of the MTJ is different. The tunneling magnetoresistance (TMR) is defined as the relative resistance change between the two magnetized states. TMR is a figure of merit of MTJ design and it is often analyzed by resorting to Equation (1):

\[
TMR = \frac{R_H - R_L}{R_L}
\]

(1)

An higher TRM value is commonly preferred since it means that a more robust read operation can be performed. Values above 100% are typically preferred.

Despite of the wide set of STT-MRAM cell designs, the most popular is the 1T-1MTJ whose structure is composed of one NMOS transistor and one MTJ device connected in series. Due to wide set of technological information that are available in literature, we target in-plane 1T-1MTJ cell in this paper whose equivalent electric circuit is provided in Fig. 2. Bit Line (BL), Source Line (SL), and Word Line (WL) aim at operate cell access.

![STT-MRAM electrical model](image)

C. Writing Operation

Many device-related parameters (e.g., MTJ area, material property) determine the write current amplitude that is required to change the free layer magnetic direction. Moreover, it behaves differently according to the current pulse width. Generally, if a longer current pulse is applied, a lower current density is required to switch the MTJ state. Based on the trade-off between write current amplitude and write pulse width, three distinct switching modes were identified [12]: thermal activation (TH), processional switching (PR), and dynamic reversal (DY) (Fig. 3). The equations are prompted as follows:

\[
J_{c,TH}(\tau) = J_{co}(1 - \frac{1}{2})\ln\left(\frac{\tau}{\tau_0}\right)
\quad (\tau > 20\text{ns})
\]

(2)

\[
J_{c,PR}(\tau) = J_{co} + \frac{C}{\tau_0}
\quad (\tau < 3\text{ns})
\]

(3)

\[
J_{c,DY}(\tau) = \frac{J_{c,TH}(\tau) + J_{c,PR}(\tau)e^{-k(\tau-C)}}{1 + e^{-k(\tau-C)}}
\quad (3\text{ns} < \tau < 20\text{ns})
\]

(4)

where \(J_{co}\) is the critical switching current density (i.e., the current density in presence of zero temperature), \(\tau_0\) is inverse of attempt frequency (typically equals to 1ns). \(C, C, k, \text{ and } \gamma\) are fitting constants. The thermal stability \(\Delta\) is a key factor of the MTJ. It depends on thickness or area of free layer and on magnetic properties of MTJ materials.

![Dependence of switching current density on write pulse width](image)
Looking at Figure 3, it is evident that when operating in processional switching zone small differences in write pulse width determine wide variation in current density. On the other hand, in the thermal activation area the required switching current increases very slowly even though the current pulse width is dramatically increased.

**D. Reading Operation**

When a read operation is performed a small bias voltage is applied on the control lines, resulting in a current (IR). This current is, then, compared against a reference value (IREF) to discriminate the stored logic state. When IR is higher than the IREF it means that the cell stores a logic value ‘0’, whereas if IR is lower than IREF the cell stores a logic value ‘1’.

It is worth noticing here that both reading currents used to discriminate the logic state have the same order of magnitude. For this reason, a Sense Amplifier is commonly used to compare IR and IREF to determine the actual logic state of the cell.

Different circuitual schemas can be implemented to generate the reference current. In [13] a pinned MTJ device is designed to have an electrical resistance equals to the average value of \( R_L \) and \( R_P \). Another approach to generate the reference current requires to adopt two MTJ cells. One of the reference cells is in the parallel (low resistance) state while the other is in the anti-parallel (high resistance) state. In this case, the resulting reference resistance is computed as the average between the low and high resistance values [14].

**E. Data Retention**

One of the most important parameter characterizing storage class memory devices is the amount of time the information is reliably stored into a cell. The data retention time of an STT-MRAM bit-cell depends on thermal stability of the MTJ. It is usually evaluated by Equation (5):

\[
R_t = \tau_0 e^\Delta
\]

The dependence of the retention time from \( \Delta \) is exponential: the higher thermal stability, the longer retention time. Nevertheless, designing MTJ to increase the thermal stability corresponds to higher write energy.

**F. CACTI**

CACTI is a widely used open-source high-level cache and memory modeling tool [13] [14] supported by HP Labs. CACTI has analytical models for all the basic building blocks of a memory: decoder, sense-amplifier, crossbar, on-chip wires, DRAM/SRAM cell and latch. CACTI models both traditional and non-uniform banked caches and memories using SRAM, and DRAM of which it can compute delay, power, and area. For a user-specified set of input parameters (e.g., energy/delay, memory size), the tool performs an exhaustive design space exploration across different array sizes and on-chip interconnections to identify, if existing, an optimal configuration that meets the input constraints.

**III. MODELING**

Our research work aims at extending CACTI to support in-plane STT-MRAM technology. By modeling bit-line, read circuitry, delay, area and energy consumption, additional parameters are combined with existing analytical models and seamlessly integrated with CACTI. The first release supports the simulation of set-associative cache memories.

**A. Array Modeling**

By integrating analytical models along with parameters extracted from ITRS roadmaps [17], CACTI supports modeling of array of targeted cache or memory devices. Memory is divided into an array of banks. Each bank is composed of one or more subbanks which are comprised of identical mats. A Mat has 4 subarray which share pre-decoding logic and each subarray contain a set of wordlines and bitlines to access the basic memory cells. To support STT-MRAM technology, we mainly focus on mat and subarray.

**B. MTJ Model**

The 1T-1MTJ cell is modeled by considering a NMOS access transistor connected in series with a MTJ device. MTJ is then modeled as a resistance whose values depends on the relative magnetization of the free layer. We provide a fully parameterized MTJ model to give the capability to explore a wide set of designs. Table I shows the model input parameters.

**Table 1:** MTJ parameters integrated into CACTI

<table>
<thead>
<tr>
<th>MTJ Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>StrType</td>
<td>Type of MTJ. This version supports only in-plane</td>
</tr>
<tr>
<td>Jc0</td>
<td>Critical current at zero temperature</td>
</tr>
<tr>
<td>( \Delta )</td>
<td>Thermal Stability</td>
</tr>
<tr>
<td>MTJArea</td>
<td>Area of MTJ</td>
</tr>
<tr>
<td>Rp</td>
<td>MTJ resistance in parallel magnetization</td>
</tr>
<tr>
<td>Rap</td>
<td>MTJ resistance in anti-parallel magnetization</td>
</tr>
<tr>
<td>Vbitline</td>
<td>Write voltage</td>
</tr>
<tr>
<td>Raccess</td>
<td>Equivalent resistance of the access transistor</td>
</tr>
</tbody>
</table>

The \( \text{Delta} \) parameter is used to compute the resulting retention time by resorting to Eq. (5). The aforementioned MTJ parameters are integrated in CACTI to model STT-MRAM cell and to figure out read and write latency as described further on.

**C. Read Latency Model**

A read operation involves several phases. A specified voltage is applied to a bitline and the resulting current passing through MTJ is compared to a reference value. In order to estimate read latency we model both the bitline and the sense amplifier (SA). In STT-MRAM memories, the sensing operation is performed by means of current-based SA. Nevertheless, CACTI currently has only models for voltage-based SA. Therefore, we adapt the current-based sensing operation of the MTJ to the existing voltage-based SA. The circuitual schema involves two reference cells and three PMOS transistor to implement the current-to-voltage converter. Interested readers can refer to [16], for further details. This circuit is modeled using SPICE at 45nm and it requires about 50ps for stabilization. It is included into CACTI as additional delay to the existing SA. The additional area and energy due to MTJ reference cells are also accounted.
D. Write Latency Model

The difference between read and write latency is quite relevant in STT-MRAM memories. Performing a write operation is typically slower. Moreover, the required write voltage is between 1 and 2 volts whereas a smaller bias voltage (0.1V ~ 0.3V) is needed for reading.

There exist a strong dependence between the write voltage and the expected write latency. Such a relationship is modeled by Eq. (2), Eq. (3), and Eq. (4) that provide an accurate MTJ write time estimation. The voltage used to estimate latency in the analytical model is supposed to be constant during the write operation and identical for both free layer orientations. Moreover, since CACTI does not provide a mechanism to input a distribution of desired logic values to be written, we only consider the switching case from parallel to anti-parallel magnetization of the free layer that is the worst case in terms of latency.

But this contribution is not sufficient to estimate the overall latency as each STT-MRAM is connected to an access transistor (see Figure 2) to mitigate write disturbs and to reduce the energy consumption. Therefore, without losing accuracy, the computation of the overall write latency for a STT-MRAM data array is equal to the read latency added to the MTJ write time.

E. Area Estimation Model

The area of STT-MRAM cell strongly depends on the design of the access transistor. Let us consider that a cell is composed of an access transistor and a MTJ stacked in a 3D structure. The resulting area is mainly dominated by the element that requires the larger planar surface that is generally the access transistor. Determining the proper size of the access transistor is one of the most critical aspects of the cell design. Due to technological constraints, a small size improves reading latency whereas a large size enhance write performances. The analytical model integrated in CACTI for cell area estimation is given in the Equation (6).

\[ A_{cell} = 3\left(\frac{W}{L} + 1\right)F^2 \]  

where \( F \) is the minimum feature size and \( W \) and \( L \) are the width and length, respectively. The equivalent resistance of the access transistor influences the length. There is an inverse proportionality between them: a high resistance corresponds to a small cell area and high storage density, instead a low resistance increases considerably memory area.

The computation of the total area of the memory is not dependent only from the size of cells. Interconnections considerably impact on resulting memory size, as well. For this reason, according to user requirements, CACTI attempts to optimize on-chip memory interconnections to meet latency or energy constraints.

F. Energy Estimation Model

For sake of completeness, we consider write and read energy model individually. Read energy per operation is evaluated by computing the Equation (7):

\[ E_{read} = C_{tot}V_{read}^2 \]

where \( C_{tot} \) depends on the total capacitance of the bitline, on the all wire contributions and on the access transistor. \( V_{read} \) is the read voltage. A lower read voltage reduces the probability of read disturb while a high value privileges read latency.

The computation of write energy can be divided in two main contributions (see Equation (7)). The former is related to the energy consumption due to the current flowing through MTJ device while the latter is similarly computed by exploiting the model in Eq. (6):

\[ E_{write} = \frac{V_{write}^2}{R_{MTJ}R_{ACC}}t_{write}C_{tot}V_{write} \]

where \( V_{write} \) is the write voltage, \( R_{MTJ} \) is the equivalent MTJ resistance, \( R_{ACC} \) is the equivalent NMOS resistance and \( t_{write} \) is the MTJ switching time. It is worth noticing here, that the computation of write energy is performed accounting for the worst case: the MTJ switches from parallel to anti-parallel state.

IV. EXPERIMENTAL RESULTS

In the previous section, we described modeling and integration of in-plane STT-MRAM technology into CACTI tool. In order to prove the correctness of our tool we generate high-performance and low-power cache memories for different MTJ configurations compared to SRAM technology. Considered MTJ input parameters are listed in Table 2. MTJ configurations differ in terms of parallel and anti-parallel resistance, the write voltage, and the equivalent resistance of the access transistor.

<table>
<thead>
<tr>
<th>Table 2: MTJ configurations</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A )</td>
</tr>
<tr>
<td>StrType</td>
</tr>
<tr>
<td>Jc0 [mA/cm²]</td>
</tr>
<tr>
<td>( \Delta )</td>
</tr>
<tr>
<td>MTJArea [cm²]</td>
</tr>
<tr>
<td>Rp [kΩ]</td>
</tr>
<tr>
<td>Rap [kΩ]</td>
</tr>
<tr>
<td>Vbitline [V]</td>
</tr>
<tr>
<td>Raccess [kΩ]</td>
</tr>
</tbody>
</table>

A. High-Performance Cache Memories

For this study we generate high-performance, eight-way set-associative cache memories with no error correction mechanism which range in size from 32 kB to 512 kB. Each cache has 64 b IN/OUT data interface with a single read-write port.Transistors are modeled by resorting to high performance cells (itrs-hp) for both the data and tag array and peripheral circuit. The usage of itrs-hp maximizes performances at expense of power consumption.

Figure 4 (h) compares the read latency of the three different MTJ configurations with respect to SRAM. The fastest read latency is achieved by SRAM. Among all the MTJ configurations, the configuration A show the best timing.
This is due to its small cell area given by the high resistance of the access transistor. Indeed, the area of the 1T1MTJ cell is proportional to the width of access transistor that is inversely proportional to the resistance. Typically, area and timing are strictly correlated and so smaller memories achieve faster performances. The discrepancy between SRAM and STT-MRAM is more evident when write latency is targeted (see Figure 4 (i)). The configurations A and B have a similar write time of around 10 ns that is quite worse than SRAM. By reducing the parallel and anti-parallel resistances the current density flowing through MTJ device is higher, so the configuration C approaches SRAM write latency.

The plot of energy consumption due to write and read operations is shown in Figure 4 (f) and in Figure 4 (g), respectively. About read energy, the configuration A shows the best results also outperforming SRAM. This is more evident for a 512 kB memory size. In addition to input parameters described in Table II, configuration A takes advantage also from read voltage set to 0.1 V in this paper. On the other hand, write energy drained by configuration A is one order of magnitude worse than SRAM. Write energy consumption mainly depends on both current density and pulse width needed to flip magnetic orientation of MTJ device. For this reason, the configuration
featuring best write latency is expected to consume more write energy.

Area consumption (see Figure 4 (f)) is dominated by the width of access transistor that is determined by its resistance. Configurations B and C have a similar trend that is almost 2x worse than configuration A.

B. Low-Power Cache Memories

We consider low-power high-capacity set-associative cache memories ranging in size from 1 MB to 32 MB. They feature sixteen-way with four banks. The input interface is 576 bits wide that include standard single-bit error correction. We use low standby power cells (itrs-lstp) for designing array and interconnections. The usage of this kind of cells allow reducing the leakage power of peripheral circuitry.

In this test, we consider just configuration A and B since configuration C is not relevant. Indeed, configuration C is designed to have a MTJ write time around 2 ns. This results in a considerably area, read and energy penalty for high-capacity memories.

Figure 4 (c) shows the read latency for low-power cache memories. The observed trend is quite similar to the one previously described in Figure (h). A remarkable difference is that for very large arrays, configuration A and SRAM are quite close in terms of read latency. Delay penalty is mostly due to the interconnections and not to the cell itself. For this reason, on equal high-capacity memory sizes, STT-MRAMs exhibits a smaller density than SRAM. Nevertheless, the performances of configuration B and C do not show a similar trend. The motivation is that CACTI performs several optimizations, according to user constraints, that can change the internal partition of the array. This can impact on length of bitlines and wordlines and on the size of the interconnection circuit resulting in a potential memory performance delay. User can force CACTI to adopt a fixed partition to avoid this issue.

The density improvements that STT-MRAM arrays can attain over SRAM arrays allow in-plane STT-MRAM to be a valid technology solution to design low-power cache memories compared to SRAM when read intensive applications are targeted (see Figure 4 (a), and Figure 4 (b)).

On the other hand, the in-plane MTJ still requires a great deal of energy to write, about 3x more than SRAM counterparts, as plotted in Figure 4 (e).

V. CONCLUSIONS

STT-MRAM is one of the most promising emerging memory technologies. It features low latency and high endurance with respect to existing memories. In this paper we presented an in-plane STT-MRAM memory modeling tool by integrating into CACTI a fully parameterized MTJ model. Bitline and read circuitry models have been implemented to give the capability to explore a wide set of designs by emulating memory arrays Targeting three different MTJ configurations, we used the tool to generate high-performance and low-power cache memories analyzing their relative performances compared to SRAM. Experimental results showed that up to now in-plane STT-MRAMs are not able to compete against SRAM even though acceptable results in terms of energy and area consumption have been achieved in presence of read operations.

ACKNOWLEDGMENT

This research has been partly supported by the 7th Framework Program of the European Union through the CLERECO Project, under Grant Agreement 611404.

REFERENCES


