

The STeTSiMS STT-RAM Simulation and Modeling System

Clinton W. Smullen, IV ^{*}, Anurag Nigam [†], Sudhanva Gurumurthi ^{*}, Mircea R. Stan [†]

^{*} Department of Computer Science and

[†] Department of Electrical and Computer Engineering

University of Virginia

Email: cws3k@cs.virginia.edu, an2z@virginia.edu, gurumurthi@virginia.edu, mrs8n@virginia.edu

Abstract—There is growing interest in emerging non-volatile memory technologies such as Phase-Change Memory, Memristors, and Spin-Transfer Torque RAM (STT-RAM). STT-RAM, in particular, is experiencing rapid development that can be difficult for memory systems researchers to take advantage of. What is needed are techniques that enable designers to explore the potential of recent STT-RAM designs and adjust the performance without needing a detailed understanding of the physics. In this paper, we present the STeTSiMS STT-RAM Simulation and Modeling System to assist memory systems researchers.

After providing background on the operation of STT-RAM magnetic tunnel junctions (MTJs), we demonstrate how to fit three different published MTJ models to our model and normalize their characteristics with respect to common metrics. The high-speed switching behavior of the designs is evaluated using macromagnetic simulations. We have also added a first-order model for STT-RAM memory arrays to the CACTI memory modeling tool, which we then use to evaluate the performance, energy consumption, and area for: (i) a high-performance cache, (ii) a high-capacity cache, and (iii) a high-density memory.

I. INTRODUCTION

Spin-Transfer Torque RAM (STT-RAM) is an emerging non-volatile memory technology that stores data as the magnetic orientation of a magnetic tunnel junction (MTJ) and is being actively explored by industry [1]–[3]. STT-RAM has significantly better write endurance (at least 10^{12} cycles) than other non-volatile memory technologies such as Flash ($> 10^6$ cycles) and Phase-Change Memory (PCM) ($> 10^9$ cycles) [4]. Though it is not as dense as these other technologies, STT-RAM is capable of high performance operation and is CMOS compatible, which makes it suitable for use in a wide range of applications. In particular, the combination of high endurance and the lack of cell leakage makes it an ideal candidate for use within the die of a microprocessor for cache or memory.

Many papers are published that evaluate the designs for STT-RAM MTJs, but their results cannot be directly adapted to meet high-level design goals. In this paper, we present a methodology and tool-chain for evaluating and comparing MTJ designs. We first demonstrate how we extrapolate MTJ technology parameters from a technical evaluation to produce a complete model of the MTJ. We use the parameters to perform Monte-Carlo macromagnetic simulation of a MTJ to characterize its write latency and energy behavior. When combined with existing analytic models, this provides a complete characterization of the MTJ behavior. We demonstrate its use in evaluating STT-RAM memory designs by interfacing it to CACTI, a widely used, high-level, cache and memory array

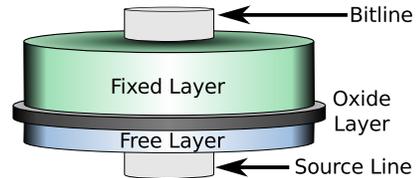


Fig. 1. Structure of Magnetic Tunnel Junction (MTJ)

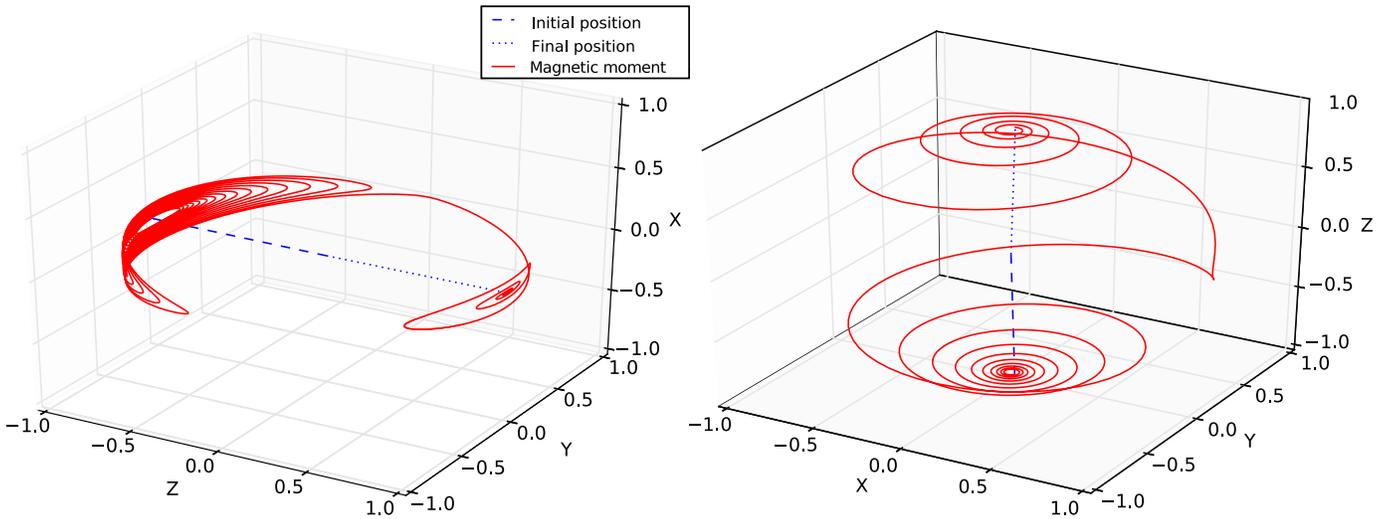
modeling tool developed by HP Labs [5], [6].

We demonstrate our methodology by applying it to three different types of MTJs: a traditional in-plane MTJ, a perpendicular MTJ, and a hybrid in-plane MTJ with partial perpendicular anisotropy (PPA). The technology parameters encompass the geometry and materials properties of the MTJ and are rarely given in complete detail in the literature. As such, we extrapolate the missing parameters using common figures-of-merit that are related to these parameters via analytic models and knowledge of the expected range for each parameter. After estimating the parameters for each design type, we show how to adjust the parameters to normalize the estimated behavior with respect to Δ , J_{c0} , or other figures-of-merit. This allows us to directly compare the three MTJ design types for use as: (i) a high-speed upper level cache, (ii) a high-capacity last-level cache, and (iii) a high-density memory. The use of normalization allows us to ask “what if” questions that would otherwise be impossible without extensive resources and technical expertise in making MTJs. However, it is to be expected that significant deviations from the fitted parameters can easily result in erroneous designs, as we discuss later in the paper.

The organization of the rest of the paper is as follows. Section II describes the operation of STT-RAM and the associated technology parameters and figures-of-merit used in describing MTJ behavior. Section III demonstrates our method for parameterizing published MTJs and normalizing their behavior to enable direct comparison and tuning for three use-cases. Section IV describes the models we have added to CACTI to enable the modeling of STT-RAM memories and comparison of the three MTJs for each use-case. Section V discusses related work on performing high-level modeling of MTJs, and Section VI gives our conclusions.

II. BACKGROUND

MTJs act as the storage element for STT-RAM memory arrays. They consist of at least two ferromagnetic layers with an oxide barrier (insulator layer) between them, as shown in Figure 1. One of the two magnetic layers is called the *hard* or



(a) In-plane MTJ
 (b) Perpendicular MTJ
 Fig. 2. Precession of the magnetic moment with spin-transfer torque from anti-parallel to parallel
 (In both graphs, the Z-axis represents the easy-axis, and the lower plane represents the plane of the MTJ)

pinned layer and has a fixed direction of magnetization, while the other is called the *soft* or *free layer* and has a variable magnetic orientation.

When the two layers are oriented in the same direction, the MTJ is in the *parallel (P)* state and exhibits a low resistance (R_P), and when the two layers are oriented in opposite directions, it is in the *antiparallel (AP)* state and exhibits a high resistance (R_{AP}). To change the state of the MTJ, a large, directional current is passed through the MTJ for a sufficient amount of time, the *write pulse width*, to force the free layer's orientation to change.

A. Magnetic Tunnel Junctions

All MTJs have a parallel and an anti-parallel state that corresponds to the two directions of the *easy-axis*, though the physical orientation depends on the type of MTJ. Figures 2a and 2b show the precession (change in orientation) of the magnetic moment under spin-transfer torque from the anti-parallel state to the parallel state for the two main types of MTJs. The magnetic moment has been normalized using the *saturation magnetization*, M_s , and the rate of precession is determined by the damping coefficient, α . The presence of *easy-axis anisotropy* attempts to keep the moment oriented in either the parallel ($Z=+1.0$) or anti-parallel ($Z=-1.0$) positions and is proportional to the *uniaxial anisotropy* factor, H_k , and M_s . Both H_k and M_s depend on the materials and design of the MTJ and are derived from empirical measurements.

1) In-plane MTJ

For the in-plane MTJ shown in Figure 2a, the plane of the MTJ lies in the Z-Y plane and current flows up or down in the direction of the X-axis. The flattened shape of the motion is caused by *easy-plane anisotropy*, which attempts to keep the magnetic moment within the plane of the MTJ. As the easy-axis lies within the easy-plane, the two types of anisotropy work in tandem to maintain the state of the MTJ, though the easy-plane anisotropy makes spin-transfer torque more difficult, as we will discuss shortly. Partial perpendicular anisotropy (PPA) partially offsets the easy-plane anisotropy,

which allows switching to occur more easily.

2) Perpendicular MTJ

Figure 2b shows a perpendicular MTJ, for which the plane of the MTJ lies in X-Y plane while current flows in the direction of the Z-axis. As the free layer is significantly wider and longer than it is thick, a *demagnetization* force attempts to pull the moment back into the plane of the MTJ. Since the easy-axis is perpendicular to the plane of the MTJ, the easy-axis anisotropy must be strong enough to overcome the demagnetization force ($H_k > 4\pi M_s$) to maintain the orientation of the magnetic moment, since they are in direct competition. However, the demagnetization force assists the spin-transfer torque in flipping the magnetic moment. Perpendicular MTJs typically have higher-density and faster switching, but their fabrication is more difficult to integrate with CMOS logic processes.

3) Spin-Transfer Torque

In 1996, Slonczewski showed how a spin-polarized current passing through the plane of a thin free layer could be used to change its state [7]. A fraction of the electrons flowing through a MTJ will become spin-polarized by a fixed magnetic layer and, with enough current, can overcome the anisotropy and demagnetization forces and flip the free layer's orientation. This works most straightforwardly with one fixed layer for each orientation (these are often called *spin filters*). This also works with a single fixed layer, though it requires significantly higher currents to switch to the anti-parallel state than the parallel state.

The spin-polarized current applies torque on the magnetic moment, pulling it in the corresponding direction. Since increasing the total current increases the amount of spin-polarized current, the MTJ will flip its orientation faster, resulting in a shorter path with fewer rotations around the Z-axis. However, without the current, the moments in Figures 2a and 2b would have remained near the initial, anti-parallel, position. The characteristics of this effect will be discussed further in Section II-B.

B. Figures of Merit

The magnetic parameters M_s , H_k , and α are determined by the materials used to make the MTJ. Also important are the geometry, defined by the free layer thickness (t_F) and planar area (A), the operating temperature, T , measured in kelvin, and the parameters for the oxide barrier that separates the ferromagnetic layers. The oxide barrier determines the resistance presented by the MTJ, which controls how much current can be passed through it and thus how fast the MTJ can switch. These parameters are not always independent, but we must first explain the key figures-of-merit for MTJs and how they can be used to determine unknown parameters. We will revisit the question of independence in Section III.

1) Tunneling Magnetoresistance (TMR)

The Tunneling Magnetoresistance (TMR) determines how distinguishable the two states are from one-another, and a high value allows read operations to be both faster and more reliable. The TMR is determined by the design of the oxide barrier and its interface with the ferromagnetic elements, and it is often analyzed using Equation (1), which puts it in terms of the high (anti-parallel) and low (parallel) resistance states. Values above 100% are preferred, but many MTJs have extremely low TMRs due to a very narrow separation in the energy levels of the two electron spin bands.

$$\text{TMR} = \frac{R_{ap} - R_p}{R_p} \quad (1)$$

Even when the TMR or resistance values are not explicitly stated by a paper, they can often be estimated from hysteresis plots of resistance versus the voltage, current, or an applied magnetic field that many publications include. Using the transport model created by Nigam et al. [8], we fit the parameters to produce the same resistance and TMR values and to match resistance-voltage or resistance-current plots, when they are given. Given a voltage, the transport model can calculate both the total current flowing as well as the amount of spin-current, given the orientation of the magnetic moment.

2) Thermal Stability (Δ)

For temperatures above absolute zero, the moment will never remain at exactly $\pm Z$ due to thermal noise that prevents the moment from reaching the minimum energy position. It can be modeled as a Langevin thermal field whose variance is determined by the *thermal stability*. The thermal stability, Δ , can be estimated by Equation (2), where k_B is Boltzmann's constant. Storage-class STT-RAM, which can retain data for at least ten years, requires Δ to be at least 40 [9]. However, $\Delta \geq 47$ is required to allow elevated temperatures of up to 350K, and it has been shown that $\Delta \geq 75$ is necessary to meet the requirements for a 1 Gb STT-RAM array in the absence of error-correction [10].

$$\Delta \approx \frac{A \cdot t_F \cdot H_k \cdot M_s}{2k_B \cdot T} \quad (2)$$

When Δ is provided along with the geometry of the free layer, it can be used to determine the range of possible values for H_k and M_s . Halving the free layer's volume would cut the write current in half, but at the cost of also halving Δ . This

would take a ten-year retention time and cut it to less than one second. Since most research has focused on producing storage-class STT-RAM, write-energy reduction using such techniques has not been explored in detail, though Smullen et al. did perform a high-level evaluation of reducing the area to lower the write latency and energy for use in on-die caches [11]

3) Write Current ($I_c(\tau)$)

The write current (I_c) for a given *write pulse width* (τ) is the (magnitude) threshold of current above which the free layer will reliably change its state in less than τ time. Analytically modeling $I_c(\tau)$ is difficult because it simultaneously depends on the magnetic parameters, the geometry, the oxide barrier, and it also behaves differently depending on τ itself. Since this makes using measured write currents to compare MTJs difficult, researchers instead use the *critical current density*, J_{c0} , as a key figure-of-merit.

Equation (3) gives a formulation of J_{c0} , where the constant e is the charge of an electron, \hbar is the reduced Planck's constant, and η is the spin-transfer efficiency. The constant X is calculated according to Equation (4) depending on whether the MTJ is in-plane, in-plane with PPA, or fully perpendicular. The PPA constant measures the fraction of easy-plane anisotropy that is negated by the partial perpendicularity. The spin-transfer efficiency is not actually a constant (it depends on the orientation of the magnetic moment), but it is treated as such when presenting J_{c0} . Using the known area of the MTJ, one can easily calculate the critical current with $I_{c0} = A \cdot J_{c0}$.

$$J_{c0} = \frac{2e}{\hbar} \cdot \frac{\alpha}{\eta} \cdot t_F \cdot M_s \cdot (H_k + 2\pi \cdot M_s \cdot X) \quad (3)$$

$$X = \begin{cases} 1 & , \text{In-plane} \\ 1 - \text{PPA} & , \text{In-plane PPA} \\ -2 & , \text{Perpendicular} \end{cases} \quad (4)$$

Given J_{c0} , Δ , and the MTJ geometry and type, one can solve Equations (2) and (3) to obtain formulae for H_k and M_s in terms of α . We assume $\eta = 1$ in the absence of specific details. Using the knowledge of the typical range for H_k and M_s for the given type of MTJ, this makes it possible to estimate values for the three magnetic parameters that are consistent with both J_{c0} and Δ . Table I shows the expected range of each parameter for the three MTJ types. In Section III, we will use these ranges to validate extrapolated parameters.

TABLE I
TYPICAL PARAMETER RANGES BY MTJ TYPE

	In-plane	In-plane PPA	Perpendicular
H_k	[200, 1000] Oe		[4, 21] kOe
M_s	[800, 2000] emu/cm ³		[200, 500] emu/cm ³
α	[0.005, 0.02]		
J_{c0}	[1, 6] MA/cm ²	[0.5, 2] MA/cm ²	[0.5, 2] MA/cm ²
Δ	[40, 70]		

$$I_c(\tau) = I_{c0} \cdot \left(1 - \frac{\ln(\tau/\tau_0)}{\Delta} \right), \quad \tau > 10 \text{ ns} \quad (5)$$

I_c has been shown to follow Table I, where $\tau_0 = 1$ ns, for write pulse widths in the *thermally activated switching* region above 10 ns [2], [3]. Though models have been proposed, it is difficult to analytically model I_c within the high-speed

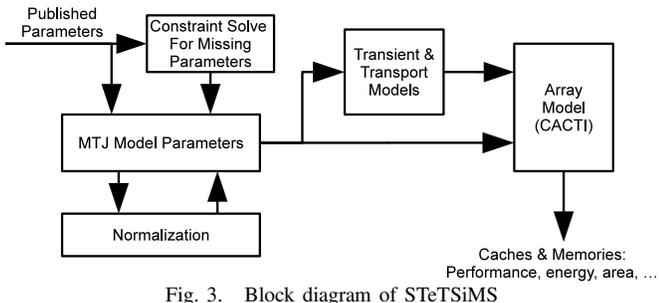


Fig. 3. Block diagram of STeTSiMS

precessional switching region below 3 ns and the *dynamic reversal* region in between [1], [3]. However, fast switching is necessary to enable the use of STT-RAM as a replacement for SRAM. To evaluate these shorter pulse widths, we combine the transport model from above with the transient behavior model, also by Nigam et al., to perform macromagnetic simulation of the free layer [8]. We perform Monte-Carlo simulation to estimate the maximum latency and write energy for a MTJ design given a specific write voltage. We will use this modeling along with Table I in Section III to evaluate the performance of both high-density and high-performance STT-RAM.

III. MODELING

Figure 3 shows all of the steps involved with our STeTSiMS methodology. The set of parameters expands as the user moves from stage to stage, though the normalization step may either be skipped or repeated multiple times, as is required. Published demonstrations of STT-RAM MTJs usually include only a subset of the parameters describes in Section II. For example, Yakushiji et al. focused on the innovative fabrication techniques used to create their perpendicular MTJs but did not provide dynamic performance characteristics other than the TMR [12]. Section III-A demonstrates how we use constraint solving to fill in these gaps to produce consistent models for three published MTJs. In Section III-B, we normalize these designs to have the same retention time and oxide barrier performance, which we later use to model STT-RAM structures in Section IV.

A. Fitting

For this work, we have modeled three single-barrier MTJs: (i) an in-plane MTJ by Diao et al. [13], (ii) another with partial perpendicular anisotropy (PPA) by Chen et al. [2], [14], and (iii) a perpendicular MTJ by Yukushiji et al. [12]. The published parameters for these designs are given in Table II, with question marks indicating information that is unknown. When fitting parameters, we will try to maximize the performance characteristics of the MTJ while respecting these constraints.

In-plane: As the in-plane design is missing H_k , t_F , and Δ , it is impossible to precisely extrapolate additional parameters. As such, we start by making the assumption that $\Delta = 60$, which allows us to estimate that $H_k \in [126, 535]$ Oe and $t_F \in [1.86, 0.44]$ nm. We prefer higher H_k to improve performance, and thus we choose the latter set, which corresponds to $\alpha = 0.02$. The MTJ is fully modeled after aligning the transport model to the given R_p , R_{ap} , and TMR.

TABLE II
PUBLISHED PROPERTIES FOR THE THREE MTJ TYPES

	In-plane [13]	In-plane PPA [2], [14]	Perpendicular [12]
H_k	?	?	21 kOe
M_s	1050 emu/cm ³	?	530 emu/cm ³
PPA	N/A	$\geq 80\%$	N/A
α	?	?	?
t_F	?	2.2 nm	1.2 nm
A	$\pi/4 \cdot 125 \times 205 \text{ nm}^2$	$\pi/4 \cdot 90 \times 180 \text{ nm}^2$	$\pi/4 \cdot 20^2 \text{ nm}^2$
J_{c0}	2 MA/cm ²	1 MA/cm ²	?
Δ	?	60 @ 300 K	?
R_p	2.5 k Ω	3.8 k Ω	?
R_{ap}	6 k Ω	7.2 k Ω	?
TMR	150%	100%	62%

(*? indicates an unknown parameter, while 'N/A' means it is not applicable)

In-plane PPA: None of the magnetic parameters are given for the PPA in-plane MTJ, though Chen et al. state that the PPA effect is at least 80% [14]. Since J_{c0} , Δ , and the geometry are given, one can solve Equations (2) and (3) to see that $M_s \in [1492, 673]$ emu/cm³ and $H_k \in [119, 264]$ Oe. As lower M_s values will improve switching performance, we use $H_k = 220$, $M_s = 808$, and $\alpha = 0.015$. As before, we align the transport model's parameters to produce the given resistance and TMR values.

Perpendicular: The given magnetic and geometry parameters can be directly used to calculate that $\Delta = 51$. Since neither α nor J_{c0} were given, we first use our range of values for α to see that $J_{c0} \in [1.4, 5.5]$ MA/cm². Since perpendicular MTJs are expected to require much less current to switch, we use the lowest value of $\alpha = 0.005$ to get $J_{c0} = 1.4$ MA/cm², though α is usually higher for perpendicular MTJs than for in-plane MTJs. We use the default values for the transport model and adjust them to match the TMR. This produces resistances of $R_p = 38$ k Ω and $R_{ap} = 61$ k Ω . High resistances are to be expected due to the extremely low area of this MTJ design.

Independence: As previously mentioned, these parameters are not always independent. For their MTJ design, Yakushiji et al. showed that M_s is affected by thickness while H_k is not [12]. As the relative strength of the different forces affect performance more than the actual values, it is difficult to predict the impact of changing these parameters. In the next Section, we will use planar dimension scaling to adjust these MTJ designs. Simply note that changes to the previously fitted parameters would very likely result in significantly different real-world MTJ behavior.

B. Normalization

Each of the three MTJs that we have parameterized differs significantly from the others. All of the MTJs have high resistance values with respect to J_{c0} , and are thus incapable of high-speed operation. The perpendicular MTJ also has significantly lower Δ than the others. To adjust for these disparities and to enable high-speed operation, we now *normalize* the designs to achieve the desired performance characteristics.

We first increase the planar dimensions of the perpendicular MTJ by 2 nm along each axis (to $\pi/4 \cdot 22 \times 22 \text{ nm}^2$), which

gives $\Delta = 61$. Since the perpendicular MTJ also has the fastest switching, we next adjust its oxide barrier to allow up to $10 \cdot J_{c0}$ at 1.1 V for the anti-parallel (low-resistance) state. This reduces R_p to $14\text{k}\Omega$ and R_{ap} to $23\text{k}\Omega$. Applying the same idea to the in-plane MTJ makes $R_p = 140\Omega$ and $R_{ap} = 360\Omega$, and, for the in-plane PPA MTJ, it gives $R_p = 570\Omega$ and $R_{ap} = 1140\Omega$. As the TMR remains as it was for all three MTJs, the voltage-current relationship is only nominally equivalent between the different designs. The three types have now been normalized with respect to retention time and current-carrying performance, though the actual switching performance will still differ, as we demonstrate next.

High-Performance: Using these variants, we perform Monte-Carlo simulation with 10,000 runs, each with a 10 ns warmup period to randomize the initial state. The simulation is run until the moment has completed two full rotations around the easy-axis as it approaches the target orientation. We have observed that, for high-speed switching, the energy is an approximately linear function of τ . For a high-performance write voltage of 1.1 V, the perpendicular MTJ reliably completes writes in $< 2.5\text{ ns}$, with room for error, and required less than 0.056 pJ/ns of energy. On average, the writes complete in $\approx 0.8\text{ ns}$, but leveraging this fact would require early write-termination circuitry, as proposed by Zhou et al. [15]. The in-plane MTJ takes less than 1 ns on average but requires upwards of 9 ns to be reliably finished, with $\approx 9\text{ pJ/ns}$. The in-plane PPA MTJ requires up to 8 ns to perform the operation ($\approx 2.5\text{ ns}$ average) and 1.9 pJ/ns .

These results demonstrate the fundamentally different behavior of each MTJ type. Though in-plane MTJs are capable of extremely fast switching, they require large amounts of energy to perform it. In-plane PPA MTJs will always require less time and energy to switch than an in-plane MTJ (when all other parameters are kept the same). The low H_k significantly raises the expected average latency, though it achieves a lower maximum latency and significantly lower write energies due to the reduced J_{c0} . The overall superiority of the perpendicular MTJ in every respect is dampened only by the challenge of integrating such MTJs with CMOS logic processes.

High-Density: For storage applications, ultra-fast write performance is much less critical than density. To facilitate this, we use Table I to determine I_c (20 ns). The in-plane MTJ requires $383\ \mu\text{A}$, the in-plane PPA MTJ requires $121\ \mu\text{A}$, and the perpendicular MTJ requires $5\ \mu\text{A}$. These values are all more than an order-of-magnitude less than their high-performance counterparts, and should thus permit significantly more dense memory arrays to be designed.

IV. ARRAY-LEVEL EVALUATION

We now look at performing a first-order evaluation of designing STT-RAM arrays that use the three normalized MTJ designs from Section III-B. We have incorporated the modeling of STT-RAM memory arrays into CACTI 6.5 [5], [6]. CACTI is a high-level tool created by HP Labs that is widely used to estimate the latency, area, and energy consumption of caches and memories. We utilize its 32 nm

technology node throughout this paper.

A. Array Modeling

CACTI models both traditional and non-uniform banked caches and memories using SRAM, embedded DRAM, or commodity DRAM. It uses a combination of analytic models along with parameters extracted from ITRS roadmaps to model the tag and data arrays of the desired cache or memory device [5], [16]. Each bank is capable of supporting parallel accesses and is comprised of one or more identical *subbanks*, which are themselves comprised of an array of identical *mats*. Given the total capacity, the number of banks, the associativity (for a cache), and the technology parameters, CACTI computes all legal permutations for dividing each bank into subbanks and mats.

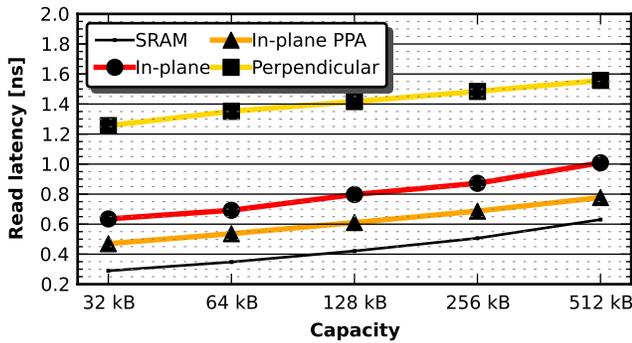
A mat has four identical subarrays which share pre-decoding logic, and each subarray is a basic array of memory cells combined with decoding logic, senseamps, multiplexers, and drivers. CACTI supports the addition of ECC bits within the subarrays as well as the addition of extra subarrays for redundancy. It selects the best candidate using a user-provided optimization function that establishes an ordering over all possible designs.

We have incorporated support for using STT-RAM primarily as part of the mat and subarray models. Though we provide our own technology parameters for the MTJ cell, we leverage the built-in ITRS high-performance N-channel transistor to model the MTJ access transistor. The access transistor is important for STT-RAM as it helps to prevent write disturbs and to eliminate wasteful energy consumption [1]. We now discuss the specific models we have added to CACTI to enable the modeling of STT-RAM read and write operations.

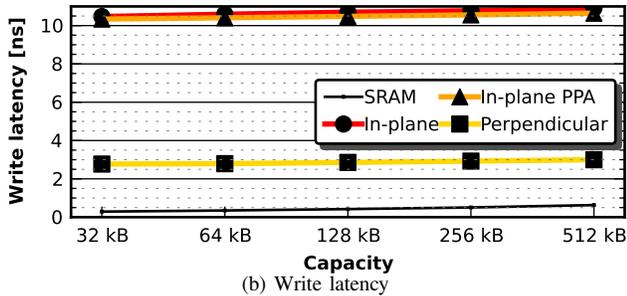
Read Operations: To read the state of the MTJ, a user-specified voltage is applied to the transport model (from Section II-B) to produce a current. CACTI does not presently have models for current-based senseamps, so it is necessary to adapt it to a voltage. We utilize two reference cells per bitline as part of the read circuitry proposed by Natarajan, et al. [17] Using SPICE, we found that an implementation of this circuit at 45 nm requires $\approx 50\text{ ps}$ to stabilize, which should also be a conservative estimate for the circuit's delay at 32 nm and smaller processes. We incorporate this time as part of the senseamp delay for STT-RAM devices, and we add two additional rows to each subarray to account for the addition of the reference cells.

Higher read voltages will reduce the read latency by swinging the bitlines more quickly, but it also increases the likelihood of causing read disturbs [2]. The normalized MTJ transport models permit significantly higher currents at both high and low voltages. Though a read voltage of $\approx 0.3\text{ V}$ gives good performance, our MTJs would conduct $> 2 \cdot J_{c0}$, making read disturbs very likely. To ensure correctness, we keep the read voltage at a low 0.1 V.

Write Operations: We use high-performance N-channel access transistors and disable bitline multiplexing to maximize the current carrying capability. This will negatively impact



(a) Read latency



(b) Write latency

Fig. 4. High-performance cache designs

the design of low-speed STT-RAM as extra bitline drivers will be added. We have assumed that the voltages used to estimate latency in Section III-B remain constant on the MTJ throughout the write operation and that they are identical for both orientations. CACTI does not perform transient modeling, so we use the maximum write current to size the access device.

B. High-Performance Cache

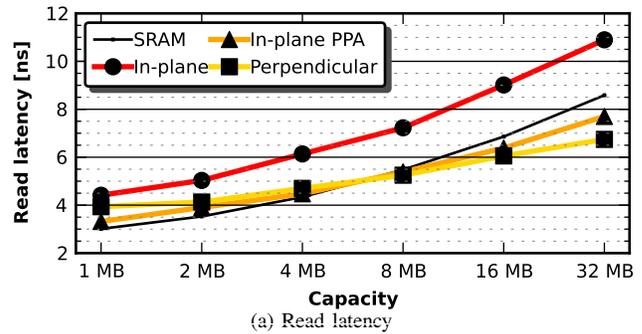
We start by using our high-performance MTJ models to build high-performance, eight-way set-associative caches ranging in size from 32 kB to 512 kB, for which Figures 4a and 4b show the read and write latencies. Each cache has a single bank with a single read-write port and a 64 b data interface with no error-correction. The caches use high-performance peripheral circuitry to maximize performance. In general, the write latency for a STT-RAM data array is equal to the read latency plus the write pulse width. This holds for both the in-plane and in-plane PPA MTJs but not for the perpendicular MTJ. This is caused by the extremely high resistance that it presents to the bitline, which requires strong drivers even though the required current is the lowest of the three. This can only be resolved by increasing the read voltage, which significantly raises the risk of read disturbs, or by renormalizing the MTJ to accept reduced write performance.

TABLE III

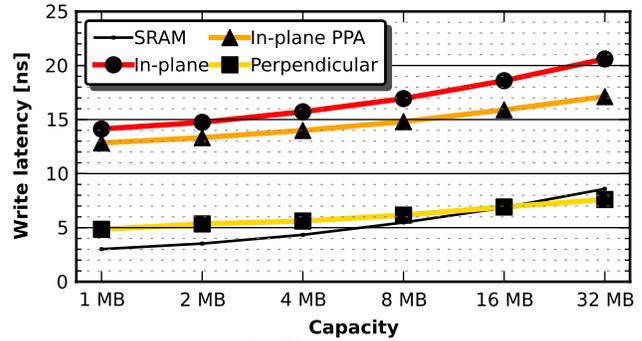
DETAILED INFORMATION FOR THE 32 kB HIGH-PERFORMANCE CACHE

	Read Energy	Write Energy	Area
SRAM	0.21 pJ/bit	0.13 pJ/bit	0.063 mm ²
In-plane	0.18 pJ/bit	62.0 pJ/bit	0.11 mm ²
In-plane PPA	0.14 pJ/bit	16.2 pJ/bit	0.043 mm ²
Perpendicular	0.90 pJ/bit	1.04 pJ/bit	0.053 mm ²

Table III shows more detailed information for the 32 kB cache designs. The energy-per-bit includes the cost of utilizing both the peripherals and tag array. Looking at the write energy, the in-plane MTJ is the highest by far, which we expect from it



(a) Read latency



(b) Write latency

Fig. 5. High-capacity cache designs

having a significantly higher J_{c0} than the other two MTJ types. In addition to reducing the performance, the large amount of peripheral circuitry necessary to support the high write current also results in the area being nearly double that of the SRAM design. The perpendicular MTJ has significantly high read energy than SRAM, though both of the in-plane MTJs use less. This is caused by the read latency penalty due to the extremely high resistance, which affects neither the write latency nor energy.

C. High-Capacity Cache

Figures 5a and 5b show the read and write latency for high-capacity, sixteen-way set-associative caches ranging in size from 1 MB to 32 MB, each with four banks. As in the previous evaluation, these were designed with the high-performance MTJ designs. Each has a single read-write port with a 576 b data interface that includes standard single-bit error correction. These caches use low power and leakage peripheral circuitry to maximize the density while minimizing power consumption.

The density improvements that STT-RAM arrays can achieve over SRAM allows the in-plane PPA and perpendicular MTJs to achieve significantly lower read latencies for capacities above 8 MB. Though its impact appears to have diminished, the fact that the 32 MB design has faster write performance than read indicates that penalty from high resistance continues to have a larger impact than the write pulse width does. Despite this, the 32 MB perpendicular design is still able to exceed both the read and write performance of the SRAM design by a sizable margin.

The continued poor performance for the in-plane MTJ and the stellar performance of the perpendicular MTJ can be seen more directly in Table IV. Compared to SRAM, the in-plane MTJ requires more energy to read and almost 300 \times the energy to write, all while occupying almost twice as much space.

The in-plane PPA MTJ still requires a great deal of energy to write, though its almost $4\times$ improvement in density could make it suitable to replace on-die caches that use embedded DRAM [18]. For the perpendicular MTJ, the almost $10\times$ reduction in area works to mitigate both the read latency penalty due to high resistance and the 2.5 ns write pulse width. The 16 and 32 MB designs end up being strictly better than their SRAM counterparts.

TABLE IV
DETAILED INFORMATION FOR THE 32 MB HIGH-CAPACITY CACHE

	Read Energy	Write Energy	Area
SRAM	3.69 pJ/bit	3.62 pJ/bit	65.2 mm ²
In-plane	4.81 pJ/bit	883.9 pJ/bit	115.9 mm ²
In-plane PPA	2.48 pJ/bit	60.0 pJ/bit	29.0 mm ²
Perpendicular	1.27 pJ/bit	1.40 pJ/bit	12.5 mm ²

The reduced performance requirements for these designs enable the in-plane PPA and perpendicular designs to use their density advantage to shorten the bitlines and wordlines. This provides better latency scaling than the in-plane MTJ or SRAM, which allows them to (eventually) catch up to the performance of the SRAM designs. However, further increases in capacity would require the use of lower leakage cells, which acts to reduce the demands placed on the peripherals, greatly extended the time before the in-plane PPA designs have faster write latencies.

D. High-Density Main Memory

STT-RAM is not currently dense enough to match either PCM or Flash memory. However, it is possible that it could be used to augment, such as Qureshi et al. did with PCM [19], or replace the commodity DRAM used as main memory. Figure 6a shows the read latency for a set of main-memory style memory chips that have eight banks, a 2 kb page size, and 64 b data interface. These use the low-speed, $\tau = 20$ ns MTJ designs, and all peripheral circuitry use the lowest power and leakage transistors. The write latency is reliably found by adding 20 ns to the read latency, and all three MTJ types track one another consistently. However, the perpendicular design shows increased read latency due to the high resistance penalty for the designs below 256 MB.

For these array sizes, the major limiting factor is simply geometry: the in-plane MTJ cells are at least 58% larger than the PPA cells, which are themselves 33x larger than the perpendicular MTJ cells. This enables the perpendicular MTJ designs to approach the density of DRAM, since they have significantly shorter bitlines and wordlines than the other two MTJ types. However, the high resistance penalty prevents their read latency from approaching that of DRAM. This shows that the sensitivity to high resistance persists even at much larger scales, despite having been subsumed for high-capacity caches.

E. Implications of the Results

At the cell level, perpendicular MTJs appear to have the most desirable characteristics for STT-RAM: high-density, high retention-time, and fast switching. However, the high resistance that is a consequence of the high-density requires the designer to choose between having fast writes and having fast reads. As expected, the in-plane PPA MTJs perform

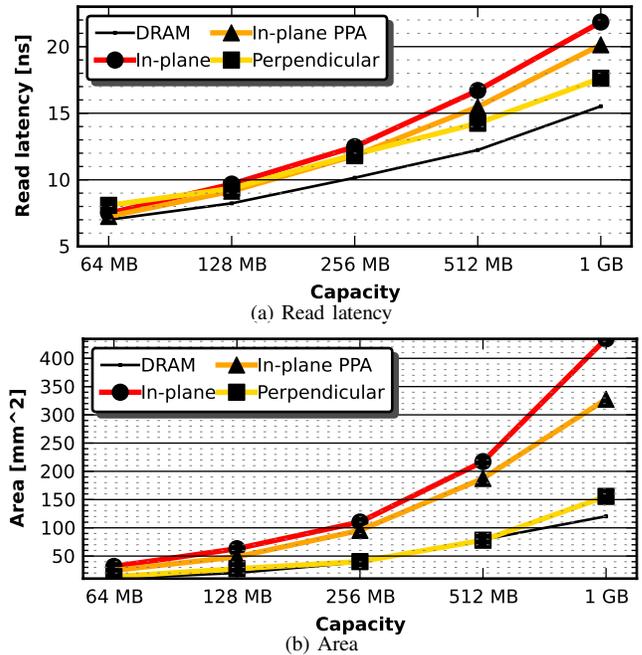


Fig. 6. High-density main memory chip designs

consistently better than their in-plane MTJ counterparts, but a chasm remains between PPA and fully perpendicular MTJs with respect to high-speed writes.

V. RELATED WORK

STT-RAM requires very high write energy when compared to SRAM and DRAM. As a result, circuit and architecture level studies that explore STT-RAM have focused on reducing the write energy of STT-RAM [11], [15], [20]–[22]. Other studies have focused on using magnetoresistive RAM (MRAM)-based memory technology to design disk caches and main memories [23], [24]. Guo et al. developed a fixed analytical STT-RAM model in CACTI to analyze the power savings of replacing CMOS components with STT-RAM in a modern processors [22]. More comprehensive tools and methodologies are necessary to provide flexibility for design experiments. This paper attempts to fill this void by providing a methodology for extrapolating a coherent set of MTJ parameters from a publication that meets the required characteristics and by presenting a first-order STT-RAM cache and memory model based on CACTI, a widely used high-level cache and memory modeling tool [5], [6].

Existing STT-RAM device models are based on both analytical and physics-based evaluations of magnetic tunnel junction (MTJ) properties. Lee et al. developed a HSPICE compatible model for MTJs with Gaussian curve fitting to simulate the DC current-voltage characteristics [25], Zhao et al. used a conductance tunnel current model to study the I-V characteristics of MTJs [26], and Nigam, et al., developed a model capturing both the steady state and transient properties of the MTJ, which we use here [8]. After fitting the parameters for a MTJ, a modified Simmons tunnel current is used to calculate the current, resistances, TMR, and the amount of spin-polarized current. This interfaces with the transient model that is used to solve the stochastic Landau-Lifshitz-Gilbert

(LLG) equation for the magnetic moment. The caveat remains that changing parameters in isolation can easily result in physically impossible designs. Further research is needed on analyzing the relationship with the various parameters, which would make our approach more robust and more powerful.

Device research has focused on resolving the issues of high write current with improved material properties [27]. Dual-barrier MTJs have been shown to significantly lower J_{c0} , though it is usually accompanied by a sizable reduction in the TMR [1], [28]. The use of early write termination would make it possible to leverage the fact that the average switching time is significantly (up to 8x) less than the maximum [15]. Similarly, the use of reduced retention-time MTJs could directly lower the necessary I_c without requiring circuit modifications [11]. With further extensions, our methodology may be extended to these types of approaches.

VI. CONCLUSION

Spin-Transfer Torque RAM (STT-RAM) is a promising non-volatile memory (NVM) technology with CMOS compatibility, high endurance, and low intrinsic leakage. We have demonstrated how to extrapolate a complete set of magnetic tunnel junction (MTJ) parameters from research publications. Though we only presented normalization with respect to Δ and J/J_{c0} , the same techniques apply to any other parameter, such as J_{c0} , I_{c0} , R_p , R_{ap} , etc. We have created a tool that allows directly performing Monte-Carlo simulation of MTJ designs to analyze their high-speed switching behavior. Using the precessional switching information and the fitted parameters, we demonstrate a STT-RAM memory modeling tool built on CACTI, which we use to analyze three different memory system examples. In each case, we found that the in-plane MTJ performed strictly worse than the in-plane partial perpendicular anisotropy (PPA) MTJ. The perpendicular MTJ appears to have the greatest potential for power and performance, but it was ultimately constrained by its extremely high resistance which limited its read performance. In this paper, we have presented the STeTSiMS STT-RAM Simulation and Modeling System which enables high-level evaluation of STT-RAM by researchers in memory devices and systems.

ACKNOWLEDGEMENT

This research has been supported by DARPA's STT-RAM program, NSF CAREER Award 0643925, and gifts from Google.

REFERENCES

- [1] Z. Diao, Z. Li, S. Wang, Y. Ding, A. Panchula, E. Chen, L.-C. Wang, and Y. Huai, "Spin-transfer torque switching in magnetic tunnel junctions and spin-transfer torque random access memory," *Journal of Physics: Condensed Matter*, vol. 19, p. 165209, 2007.
- [2] E. Chen, D. Apalkov, Z. Diao, A. Driskill-Smith, D. Druist, D. Lottis, V. Nikitin, X. Tang, S. Watts, S. Wang, S. A. Wolf, A. W. Ghosh, J. W. Lu, S. J. Poon, M. Stan, W. H. Butler, S. Gupta, C. Mewes, W. Mewes, and P. B. Visscher, "Advances and future prospects of spin-transfer torque random access memory," *IEEE Transactions on Magnetics*, vol. 46, no. 6, pp. 1873–1878, 2010.
- [3] A. Raychowdhury, D. Somasekhar, T. Karnik, and V. De, "Design space and scalability exploration of 1t-1st mtj memory arrays in the presence of variability and disturbances," in *IEEE International Electron Devices Meeting*, 2009, pp. 1–4.
- [4] W. Arden, P. Coge, M. Graef, R. Mahnkopf, H. Ishiuchi, T. Osada, J. Moon,

- J. Roh, C. H. Diaz, B. Lin, P. Apte, B. Doering, P. Gargini *et al.*, *International Technology Roadmap for Semiconductors*. <http://www.itrs.net/>: Semiconductor Industries Association, 2009.
- [5] S. Thoziyoor, N. Muralimanohar, and N. P. Jouppi, "CACTI 5.0," HP Laboratories, Tech. Rep. HPL-2007-167, 2007.
- [6] N. Muralimanohar, R. Balasubramanian, and N. P. Jouppi, "CACTI 6.0: A Tool to Model Large Caches," HP Laboratories, Tech. Rep. HPL-2009-85, 2009.
- [7] J. C. Slonczewski, "Current-driven excitation of magnetic multilayers," *Journal of Magnetism and Magnetic Materials*, vol. 159, no. 1–2, pp. L1–L7, 1996.
- [8] A. Nigam, K. Munira, A. Ghosh, S. Wolf, and M. R. Stan, "Self Consistent Parameterized Physical MTJ Compact Model for STT-RAM," in *Proceedings of the 2010 International Semiconductor Conference*, October 2010.
- [9] N. D. Rizzo, M. DeHerrera, J. Janesky, B. Engel, J. Slaughter, and S. Tehrani, "Thermally activated magnetization reversal in submicron magnetic tunnel junctions for magnetoresistive random access memory," *Applied Physics Letters*, vol. 80, no. 13, pp. 2335–2337, 2002. [Online]. Available: <http://link.aip.org/link/?APL/80/2335/1>
- [10] A. Driskill-Smith, S. Watts, V. Nikitin, D. Apalkov, D. Druist, R. Kawakami, X. Tang, X. Luo, A. Ong, and E. Chen, "Non-volatile spin-transfer torque RAM (STT-RAM): Data, analysis and design requirements for thermal stability," in *Symposium on VLSI Technology*, 2010, pp. 51–52.
- [11] C. W. Smullen, IV, V. Mohan, A. Nigam, S. Gurumurthi, and M. R. Stan, "Relaxing Non-Volatility for Fast and Energy-Efficient STT-RAM Caches," in *Proceedings of the 17th IEEE International Symposium on High Performance Computer Architecture*, February 2011, pp. 50–61.
- [12] K. Yakushiji, T. Saruya, H. Kubota, A. Fukushima, T. Nagahama, S. Yuasa, and K. Ando, "Ultrathin Co/Pt and Co/Pd superlattice films for MgO-based perpendicular magnetic tunnel junctions," *Applied Physics Letters*, vol. 97, no. 23, p. 232508, 2010.
- [13] Z. Diao, M. Pakala, A. Panchula, Y. Ding, D. Apalkov, L.-C. Wang, E. Chen, and Y. Huai, "Spin-transfer switching in MgO-based magnetic tunnel junctions (invited)," *Journal of Applied Physics*, vol. 99, no. 8, p. 08G510, 2006.
- [14] E. Chen, D. Lottis, A. Driskill-Smith, D. Druist, V. Nikitin, S. Watts, X. Tang, and D. Apalkov, "Non-volatile spin-transfer torque RAM (STT-RAM)," in *Device Research Conference*, June 2010, pp. 249–252.
- [15] P. Zhou, B. Zhao, J. Yang, and Y. Zhang, "Energy reduction for stt-ram using early write termination," in *IEEE/ACM International Conference on Computer-Aided Design*, November 2009, pp. 264–268.
- [16] W. Arden, P. Coge, M. Graef, H. Ishiuchi, T. Osada, J. Moon, J. Roh, H.-C. Sohn, W. Yang, M.-S. Liang, C. H. Diaz, C.-H. Lin, P. Apte, B. Doering, P. Gargini *et al.*, *International Technology Roadmap for Semiconductors*. <http://www.itrs.net/>: Semiconductor Industries Association, 2007.
- [17] S. Natarajan, S. Chung, L. Paris, and A. Keshavarzi, "Searching for the dream embedded memory," *IEEE Solid-State Circuits Magazine*, vol. 1, no. 3, pp. 34–44, 2009.
- [18] R. Matic and S. Schuster, "Logic-based eDRAM: Origins and rationale for use," *IBM Journal of Research and Development*, vol. 49, no. 1, January 2005.
- [19] M. K. Qureshi, V. Srinivasan, and J. A. Rivers, "Scalable High Performance Main Memory System Using Phase-Change Memory Technology," in *Proceedings of the 36th International Symposium on Computer Architecture*, 2009, pp. 24–33.
- [20] G. Sun, X. Dong, Y. Xie, J. Li, and Y. Chen, "A novel architecture of the 3D stacked MRAM L2 cache for CMPs," in *Proceedings of the 15th International Symposium on High Performance Computer Architecture*, February 2009, pp. 239–249.
- [21] M. Rasquinha, D. Choudhary, S. Chatterjee, S. Mukhopadhyay, and S. Yalamanchili, "An energy efficient cache design using spin torque transfer (stt) ram," in *Proceedings of the 16th ACM/IEEE International Symposium on Low Power Electronics and Design*, 2010, pp. 389–394.
- [22] X. Guo, E. Ipek, and T. Soyata, "Resistive computation: Avoiding the power wall with low-leakage, stt-mram based computing," in *Proceedings of the 37th annual International Symposium on Computer Architecture*, 2010, pp. 371–382.
- [23] C. W. S. IV, J. Coffman, and S. Gurumurthi, "Accelerating Enterprise Solid-State Disks with Non-Volatile Merge Caching," in *Proceedings of the First International Green Computing Conference*, August 2010, pp. 203–214.
- [24] R. Desikan, C. R. Lefurgy, S. W. Keckler, and D. Burger, "On-chip mram as a high-bandwidth, low-latency replacement for dram physical memories," University of Texas at Austin, Tech. Rep. TR-02-47, 2002.
- [25] S. Lee, S. Lee, H. Shin, and D. Kim, "Advanced HSPICE Macromodel for Magnetic Tunnel Junction," *Japanese Journal of Applied Physics*, vol. 44, no. 4B, pp. 2696–2700, 2005. [Online]. Available: <http://jap.jp/link?JJAP/44/2696/>
- [26] W. Zhao, E. Belhaire, Q. Mistral, C. Chapped, V. Javerliac, B. Dieny, and E. Nicolle, "Macro-model of Spin-Transfer Torque based Magnetic Tunnel Junction device for hybrid Magnetic-CMOS design," in *Proceedings of the 2006 IEEE International Behavioral Modeling and Simulation Workshop*, September 2006, pp. 40–43.
- [27] Y. Huai, "Spin-Transfer Torque MRAM (STT-MRAM): Challenges and Prospects," *Association of Asia Pacific Physical Societies Bulletin*, vol. 18, p. 33, 2008.
- [28] Z. Diao, A. Panchula, Y. Ding, M. Pakala, S. Wang, Z. Li, D. Apalkov, H. Nagai, A. Driskill-Smith, L.-C. Wang, E. Chen, and Y. Huai, "Spin transfer switching in dual MgO magnetic tunnel junctions," *Applied Physics Letters*, vol. 90, p. 132508, 2007.