Reducing Switching Latency and Energy in STT-MRAM Caches With Field-Assisted Writing

Ravi Patel, *Student Member, IEEE*, Xiaochen Guo, *Student Member, IEEE*, Qing Guo, *Student Member, IEEE*, Engin Ipek, *Member, IEEE*, and Eby G. Friedman, *Fellow, IEEE*

***Abstract*— A field-assisted spin-torque transfer magnetoresistive RAM (STT-MRAM) cache is presented for the use in high-performance energy-efficient microprocessors. Adding field assistance reduces the switching latency by a factor of 4. An array model is developed to evaluate the switching energy for different field currents and array sizes. Several STT-MRAM-based cells demonstrate a 55% energy reduction as compared with an SRAM cache subsystem. As compared with STT-MRAM caches with subbank buffering and differential writes, a field-assisted STT-MRAM cache improves the system performance by 28%, with a 6.7% increase in energy.**

***Index Terms*— Cache, magnetic tunnel junction (MTJ), magnetoresistive RAM (MRAM), spin-torque transfer (STT), STT-MRAM.**

I. Introduction

PERFORMANCE scaling of modern computing systems is largely constrained by conventional memory technologies. Six-transistor SRAM, which has long been the workhorse of high-performance caches, is projected to be replaced by 8T, 10T, and 12T variants to tolerate retention errors, variability, and read disturbance [1]. As a result, SRAM density has not increased commensurately with CMOS scaling.

Emerging resistive memories, which rely on resistance (rather than charge) to carry information, have the potential to scale to much smaller geometries than charge-based memories (e.g., SRAM). The smaller cell area, near-zero leakage power, and enhanced scalability make resistive memories viable alternatives to SRAM and DRAM in the next-generation memory systems. Among other resistive memories, spin-torque transfer magnetoresistive RAM (STT-MRAM) exhibits low access latency (<200 ps in 90 nm) [2], densities comparable with DRAM (8 F^2) [3], and practically unlimited endurance [4]. STT-MRAM is close to becoming a CMOS-compatible universal memory technology. The 64-Mb STT-MRAM products have already entered the marketplace [5]. Despite these advantages, STT-MRAM generally suffers long write latency and high write energy, which constrain the use of STT-MRAM to low activity caches (e.g., last-level cache).

The storage element in an STT-MRAM cell is a magnetic tunnel junction (MTJ), which is the primary factor limiting the speed of STT-MRAM due to the relatively long switching latency. In addition, the write energy of STT-MRAM is orders of magnitude higher than SRAM. A constant, large-amplitude current must be applied to the STT-MRAM during the entire switching period, which dissipates large static power.

To address these issues, an MRAM field-assisted mechanism is proposed for incorporation into STT-MRAM caches. The physical topology utilizes an assistive field current to destabilize the MTJ during switching, which reduces the switching latency of STT-MRAM by an order of magnitude, from 6.45 to 0.62 ns. The additional energy consumed by the field current can be amortized by applying the field over a row of STT-MRAM cells [along with the wordline (WL)], which leads to an 82% reduction in energy per cell. Evaluation of a microprocessor cache system demonstrates a 55% average energy reduction and a 5% speedup compared with a standard SRAM cache subsystem. Different from previous work [6] that trades off STT-MRAM retention time for improved write speed and energy, the approach described in this paper does not require modification of the MTJ structure nor is the data retention time compromised.

The rest of this paper is organized as follows. Background on STT-MRAM and cell topologies is provided in Section II. The field-assisted writing mechanism is described in Section III. Models of an STT-MRAM cell and array are presented, respectively, in Sections IV and V. Several STT-MRAM cell variants (with and without the applied field) are compared with SRAM within a microprocessor cache system in Section VI. Finally, the conclusions are drawn in Section VII.

II. MTJ Background

A. MTJ Structure and Operation

An MTJ is a two-terminal resistive element that operates on the principle of spin-dependent conduction through magnetic

Manuscript received February 17, 2014; revised October 23, 2014 and January 27, 2015; accepted January 29, 2015. Date of publication March 3, 2015; date of current version December 24, 2015. This work was supported in part by the Binational Science Foundation under Grant 2012139, in part by the National Science Foundation under Grant CCF-1329374 and Grant CCF-1054179, in part by the New York State Office of Science and Technology, and in part by the Research Grant through the IBM, Qualcomm, Cisco Systems, and Samsung.

The authors are with the Department of Electrical Engineering and Computer Engineering, University of Rochester, Rochester, NY 14627 USA (e-mail: rapatel0@gmail.com; xiaochen.guo@rochester.edu; qguo@cs.rochester.edu; ipek@cs.rochester.edu; friedman@ece.rochester.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TVLSI.2015.2401577

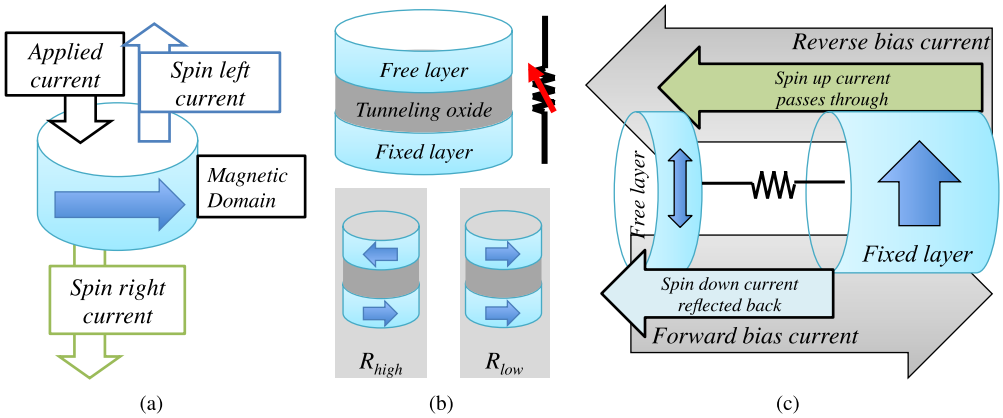


Fig. 1. Demonstration of (a) domain-dependent polarization effect, (b) MTJ stack, and (c) STT effect.

domains [4], [7], [8]. When applying a current to a magnetic domain, two spin currents (with opposite polarization) are generated across the device due to spin-dependent tunneling and reflection, as shown in Fig. 1(a). Electrons passing through the domain exhibit a net spin polarization aligned with the magnetic domain, whereas electrons reflecting off the domain have a net spin antiparallel to the domain.

An MTJ is a stack of two magnetic layers separated by a tunneling oxide, as shown in Fig. 1(b). One layer has a fixed magnetization direction, and the other (free layer) can flip between two opposite polarities, one parallel to the fixed layer and the other antiparallel to it. When domains in the two layers are aligned (in parallel), electrons passing through both layers are unimpeded; the MTJ exhibits a low resistance (R_{Low}). When domains in the two layers are antiparallel, however, an electron obtains a net polarity in one layer, and enters a layer with the opposite polarity. The electron reflects off the second domain. This effect increases the MTJ resistance (R_{High}).

Conventional MRAM circuits, such as Stoner–Wohlfarth MRAM [9], [10] and toggle MRAM [11], use two large orthogonal currents to generate magnetic fields within the free layer. These fields must be sufficiently strong to induce a torque on the magnetization, which eventually induces a reversal in the polarity of the free layer. STT-MRAMs, however, utilize spin-dependent currents to alter the polarity of the free layer, as shown in Fig. 1(c). With reverse bias, current passes through the fixed layer and attains a large net magnetic polarity. Electrons in the STT current transfer angular momentum to the electrons in the free layer, thereby inducing a net torque on the free layer polarity. When the magnitude of the STT current exceeds a threshold current, the generated torque switches the free layer to a parallel alignment with the fixed layer. The switching mechanism is similar to the forward bias case except that the free layer is subjected to a reflected spin current with a polarity antiparallel to the fixed layer. The free layer will, therefore, switch into an antiparallel alignment.

An MTJ can be created with either an in-plane or an out-of-plane structure. Out-of-plane devices, also known as perpendicular MTJs, organize the stack to ensure that both the pinned layer and the free layer are vertically aligned. Unlike in-plane devices, which rely on the geometric shape to provide a stable axis for the free layer, perpendicular devices rely on some combination of crystallographic orientation and interface characteristics of the magnetic thin film for stability. The mechanisms for tunneling and switching are the same for both device configurations.

B. MTJ Switching Dynamics

Spin polarization of electrons incident on a free layer induces a torque on the magnetic polarity. This torque, shown in Fig. 2(b), is immediately countered by a natural damping torque, which stabilizes the magnetic polarity along the long axis of the domain. When the current-induced torque is sufficiently large to overcome the damping torque, the domain polarity aligns with the short axis. The damping torque switches sides and assists the current-induced torque, which switches the polarity of the domain.

Note that this switching process is inherently stochastic. Since the current-induced torque is parallel or antiparallel to the resting polarity of the device, the effective torque on the polarity is zero (the cross product of two parallel or antiparallel torques is zero). If the polarity deviates slightly from a resting position, the cross product becomes nonzero. This deviation is due to thermal fluctuations within the MTJ device. The probability of STT switching is, therefore, based on the magnitude of the current, bias duration, and ambient temperature [12].

C. Field-Assisted Switching

Stochastic switching requires that random thermal fluctuations are sufficiently large to allow STT current-induced switching. A perpendicular magnetic field during the switching process directly addresses this issue. Field-assisted switching requires application of an orthogonally oriented magnetic field in addition to the STT current to reduce the switching latency. The magnetic field torque destabilizes the MTJ polarity toward the short axis, as shown in Fig. 2(c). As a result, the spin-transfer torque exhibits a larger effective magnitude. This method ensures that the process is less reliant on random thermal fluctuations for switching to occur.

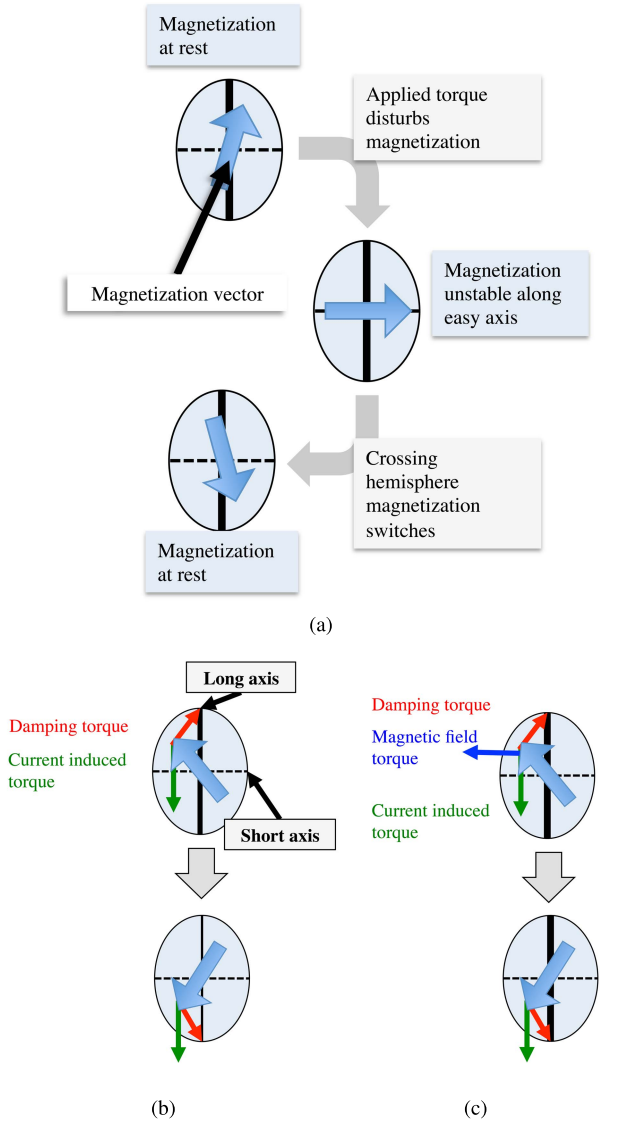


Fig. 2. Overview of (a) general switching process for an MTJ free layer with (b) standard STT switching and (c) field-assisted STT switching.

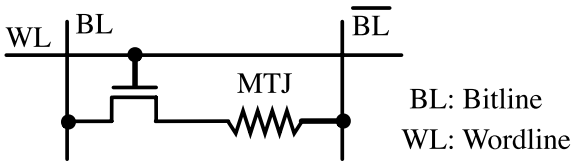


Fig. 3. 1T-1MTJ STT-MRAM cell.

D. STT-MRAM Cell Structure

STT-MRAM is CMOS compatible. A typical one transistor, one MTJ (1T-1MTJ) STT-MRAM cell is shown in Fig. 3. The MTJ serves as a storage element and the resistance represents a single data bit. The access transistor, in series with the MTJ, behaves as a gating element. To read a cell, the WL is asserted and the resistance of the MTJ is sensed. To write a cell, the WL is turned ON and the cell is driven by a write current. The direction of the write current determines the logic state of the bit written into the cell.

III. Field-Assisted STT-MRAM

Since the introduction of the STT effect into MTJ switching [4], MRAMs have primarily used STT for writing. Field-assisted excitation of the magnetic free layer, however, can complement the STT effect. Stoner–Wohlfarth and toggle MRAMs use two perpendicular currents with a single selected MTJ at the intersection to produce a magnetic field that acts on the free layer of an MTJ [Fig. 4(a)]. This approach suffers from two key issues: 1) the use of two currents to switch a single bit consumes a large amount of energy and 2) the MTJs in adjacent columns and rows are half-selected by the high fields caused by the write currents, constraining the design space to avoid erroneous writes [13].

The STT effect overcomes these problems using a single current that passes through the MTJ. This technique enables a row of MTJs (along the WL) to be written in parallel, as shown in Fig. 4(b). The direction of the applied current translates into the final state of the MTJs, i.e., a forward bias sets the device to 0, and a reverse current sets the device to 1. The switching current is much lower than in toggle-mode MRAMs, which alleviates the half-select problem. The write latency, however, remains significantly longer than the read latency, and the switching energy is also significantly greater than SRAM. Supplying a sufficiently large write current requires a large access transistor, which reduces the density of the circuit.

The approach proposed herein combines an STT-based current with a field-generating current. The field current produces a magnetic field that destabilizes the MTJs across a row. Each MTJ is biased with an STT current that controls the switching direction of the MTJs in each column. Use of a field current in this manner has two beneficial effects: 1) the alignment of the field with respect to the MTJ can destabilize the device, which reduces both the write latency and the energy, and 2) the field current is shared across the row, ensuring that the energy consumption of the field current is amortized across all of the cells within a row.

A. Related Work

External magnetic fields are used in toggle and Stoner–Wohlfarth MRAM as the primary switching mechanism. This paper shows that the superposition of an external magnetic field with local STT currents reduces both the switching latency and the energy while removing the issue of half-select disturbance in on-chip, write intensive memories. The use of both a magnetic field and an STT current for switching was demonstrated physically in [14] but considered discrete off-chip memories as a replacement for Stoner–Wohlfarth and toggle MRAM switching. The approach in [14] used a nascent STT device and an older CMOS technology. The small size is limited to DRAM-replacement applications with dense cell layouts. In the proposed method, the magnitude of the applied current and size of the memory reduce the switching latency of the MTJ device.

Andre *et al.* [15] presented a similar structure that utilizes a field current to set the MTJ device to an initial reset state

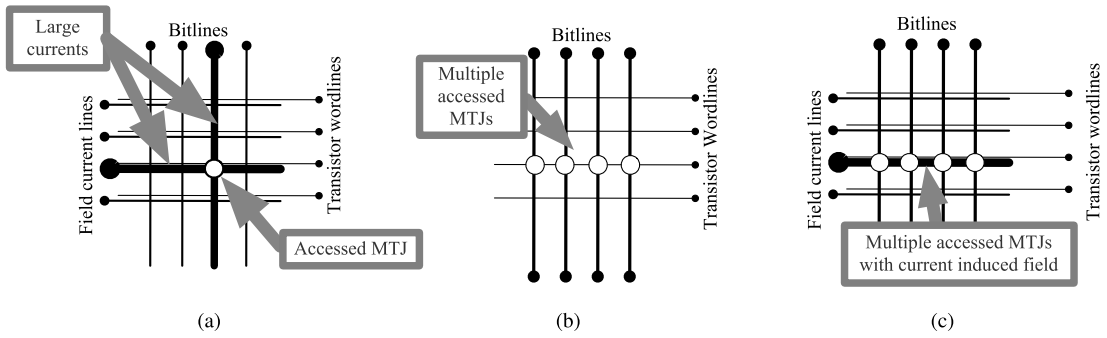


Fig. 4. Current biasing scheme for (a) Stoner–Wohlfarth and toggle MRAM, (b) STT-MRAM, and (c) proposed field-assisted STT-MRAM.

TABLE I
LLG SIMULATION PARAMETERS

γ	$1.76\text{x}10^{11}\ \frac{\text{rad}}{\text{s}-\text{T}}$
α	0.01
Temperature	350 K
Time step	0.25 ps
Initial angle (θ_0)	5 °

TABLE II
MTJ PARAMETERS

Saturation Magnetization (M_s)	$8\ \text{x}\ 10^5 A/\text{m}$
Long axis	70 nm
Short axis	20 nm
Thickness	2.9 nm
R_{on}	5 kΩ
TMR	150%
I_{crit}	39.4 μA

(either R_{ON} or R_{OFF}) prior to writing the device. This method enables the unidirectional cells and diodes to select the individual memory cells [15], which provide cell density advantages appropriate for DRAM-replacement memory applications. A reset process, however, requires the MTJ devices to undergo two switching events for every write, one to switch to a reset state (either R_{ON} or R_{OFF}), and a second switching event to write the correct state for the remaining bits. This process doubles the write latency of an MRAM array. The approach presented in this paper requires CMOS transistors for bipolar switching and utilizes magnetic fields to enhance the dynamic behavior of the switching process to reduce the energy of a write, while sharing the field current to amortize the energy across multiple columns. The device is not reset to a stable state but rather an additional torque is applied dynamically to enhance the switching process, reducing the overall write latency and enabling use in latency critical applications.

Ding [16], Wang *et al.* [17], and Cao *et al.* [18] describe individual cell structures used for field-assisted MRAM switching. Each of these publications describe structures and topologies for individual field-assisted MRAM cells. The key difference between these publications and the work presented in this paper is the notion of sharing the field current across multiple cells across the WL. System level sharing of the field current results in a significant reduction in energy.

IV. MODEL OF A FIELD-ASSISTED STT-MRAM CELL

An individual in-plane MTJ is modeled here using the classical Landau–Lifshitz–Gilbert (LLG) macrospin model with thermal agitation based on a Langvin random field using the M^3 simulator [19], with parameters listed in Table I. While the proposed field-assisted mechanism is applicable to both in-plane and perpendicular devices, only an in-plane device is considered here because of the relative maturity of the technology. The MTJ free-layer parameters are selected to ensure that the thermal stability factor (Δ) provides a 10-year retention of the device state ($\Delta = 40$). The MTJ parameters for the resistance and tunneling magnetoresistance ratio (TMR) (from ITRS 2011 [20]) are listed in Table II. The critical switching current of the MTJ is dependent on the geometric and material properties of the free layer, permitting the current to be determined from the free-layer geometry. The resultant critical current is in agreement with the switching current targeted by the ITRS [20]. Read simulations assume a worst case variation of 30% for both R_{ON} and R_{OFF}. Data for statistical variation of the thermal barrier are unavailable. Cache entries, however, exhibit lifetimes on the order of seconds. Higher thermal barriers may lengthen the switching time of an MTJ. The relatively high thermal barrier assumed in this paper is conservative.

The predictive technology model is used to characterize the cell access transistor [21]. A low threshold transistor is used for the selection device and is modeled with a 20% reduction in threshold voltage. The WL is bootstrapped to $V_{\text{DD}} + V_{\text{th}}$. The cell transistor width provides a switching current 1.5 times greater than the critical switching current. This width is selected to ensure that the device operates in precessional mode [12] while allowing the access transistor to be small.

Durlam *et al.* [10] present a toggle MRAM cell and memory. Measurements of the field observed by the free layer are demonstrated at a distance of 0.3 μm for a 0.6-μm process. Linear scaling of this dimension to a 22-nm process is assumed for the field line spacing to evaluate the field-assisted cell, as shown in Fig. 5. Simple linear scaling of this dimension is not sufficient, as the MTJ dimensions are proportionally larger than in Stoner–Wohlfarth and toggle MRAM. To compensate, the MTJ dimensions are scaled linearly and the thickness of the MTJ stack is assumed to occupy an additional 10 nm. This thickness is typical of many demonstrated STT-MTJ stacks [22], [23].

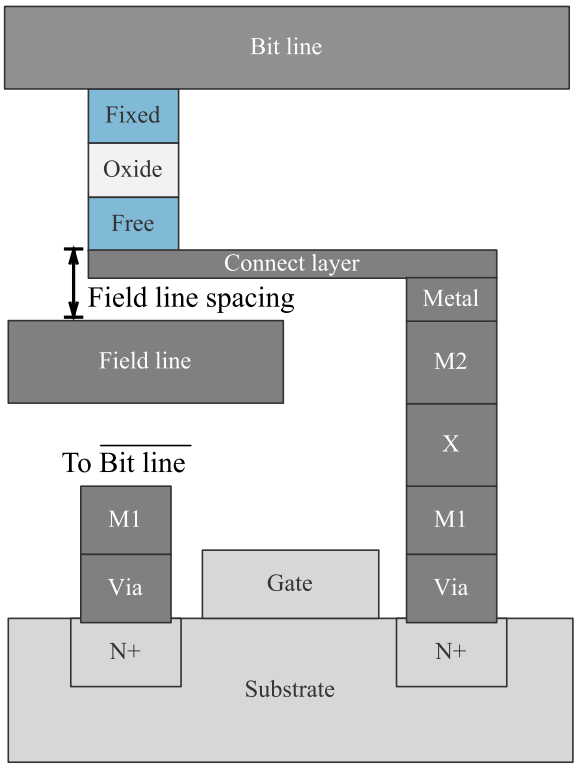


Fig. 5. Profile view of field-assisted STT-MRAM cell.

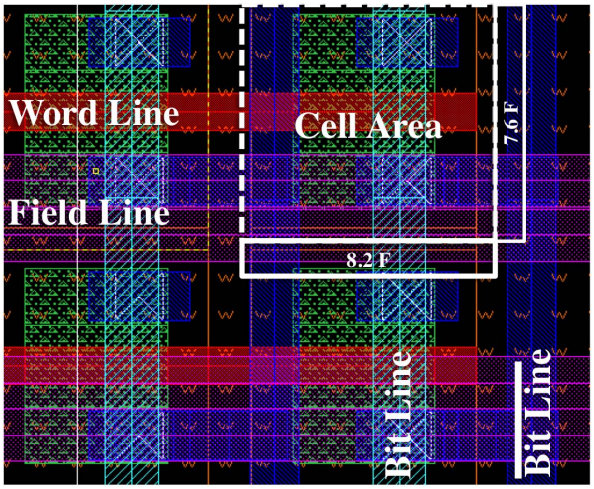


Fig. 6. Physical layout of field-assisted STT-MRAM cell.

The cell layout is based on 45-nm FreePDK design rules and scaled to 22 nm, as shown in Fig. 6. A spectrum of cell sizes is evaluated for performance. The base cell area is 55.5 F^2. In a prior work, the area of a conventional 1T-1MTJ cell is shown to be 49.9 F^2 with the same technology rules, indicating that the area overhead of the metal line supporting the additional field current is small [24], [25]. This cell has a relatively large cell density as compared with commodity STT-MRAM (6 F^2 [20]), since the layout rules originate from a logic oriented process. A standalone memory process with tighter design rules would provide greater density.

Three distinct physical configurations of a 1T-1MTJ memory cell are compared in Table III. The field-assisted STT-MRAM cell (Field-Assisted) is compared with a minimum-sized 1T-1MTJ cell capable of supplying the same nominal switching current (Minimum). The additional metal line devoted to the field current impedes contact sharing and consumes additional area as compared with the minimum cell. The third memory cell (Isometric) has the same total area as the field-assisted cell. Due to extra area consumed by the bit lines above the silicon substrate, the field-assisted cell can use a slightly larger transistor than the minimum cell without affecting cell density, resulting in a slightly larger STT switching current.

TABLE III
STT-MRAM CELL PARAMETERS

STT-MRAM Cell Type	Isometric	Minimum	Field-Assisted
Technology	22 nm		
Supply (V_{DD})	0.8 V		
Nominal switching current	59.1 μA		
STT switching current	75 μA	59.1 μA	66.2 μA
Field line spacing	N/A	N/A	21 nm
Cell length	119 nm	119 nm	167 nm
Cell width	252 nm	175 nm	180 nm
Cell Area	62.1 F^2	43 F^2	62.1 F^2

TABLE IV
MEMORY ARRAY PARAMETERS

R_{flcell} (Ω)	0.7
C_{flcell} (aF)	28.8

The magnetic field through a current loop can be estimated by the Bio–Savart's law [26]

$$B = \frac{\mu_0 I_{\text{field}}}{2\pi d}. \tag{1}$$

The current through the MTJ induces a spin torque on the free layer, generating a magnetic field that adds linearly to the magnetic field generated by the field current. The magnetic field produced by the STT is assumed to be negligible for two reasons. First, the STT current is almost two orders of magnitude smaller than the field current, making the field generated by the STT current relatively small. Second, the field current is applied to the MTJ before the STT current is applied, ensuring that the free-layer magnetization is in an unstable state prior to application of the STT current. As a result, the magnetic field of the STT current does not affect the destabilization process.

The deterministic switching latency with increasing field current in the absence of thermal noise is shown in Fig. 7. The latency decreases monotonically with increasing field current, indicating that the maximum available current improves circuit speed. Thermal noise is, however, an important concern, as discussed in Section V-A.

V. MODEL OF STT-MRAM ARRAY

Optimizing the energy consumed by an MRAM array with a field-assisted write produces a tradeoff between the size of the array and the current bias when minimizing the switching time of an MTJ. The parasitic impedances of the array, extracted from the cell layout, are listed in Table IV [21].

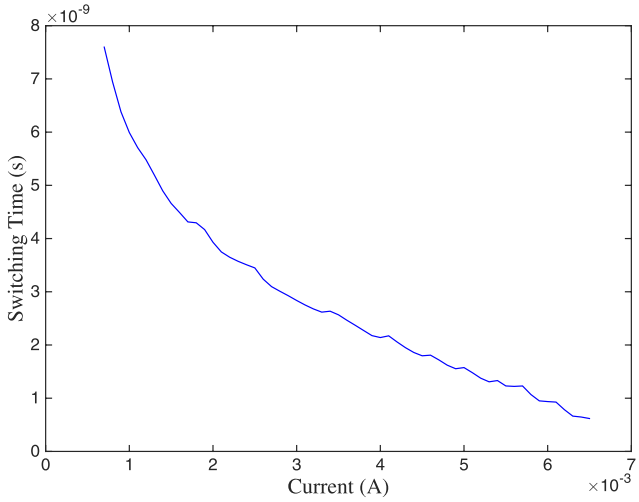


Fig. 7. Switching latency of a field-assisted MRAM cell. STT switching current is 59.1 μA.

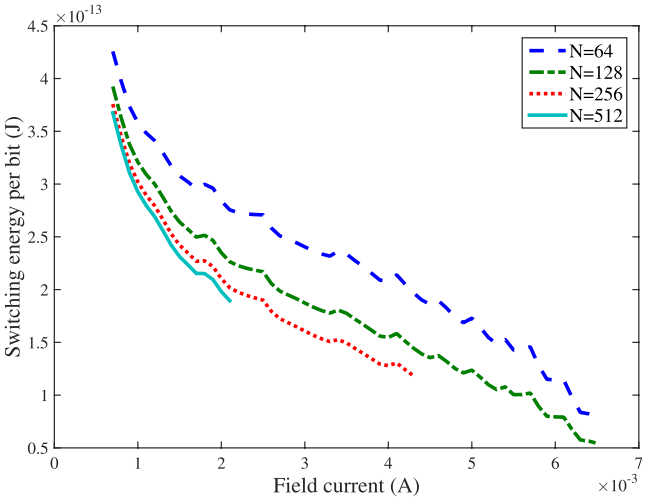


Fig. 8. Switching energy of a field-assisted MRAM cell.

The array is biased using a field current that traverses the entire row. As the size of the row increases, the energy associated with the field current is amortized across the entire row. The energy associated with the field current is the sum of the dynamic energy to charge the line and the static current to generate the magnetic field. Expression (2) quantifies this dependence, where R_{flcell} and C_{flcell} describe, respectively, the per cell parasitic resistance and capacitance, N describes the number of cells in a row, R_{driver} is the resistance of the drive transistor that supplies the field current, V_{DD} is the supply voltage, $t_{\text{switching}}$ is the MTJ switching latency, and I_{field} is the generated field current of the line. The first term in (2) describes the dynamic energy required to charge a field line, while the second term quantifies the static energy consumed by the field current. The dynamic component of the energy is, therefore, a function of array width and the dc voltage on the bit line during a write

$$E_{\text{field}} = \frac{1}{2} C_{\text{flcell}} N \left(\frac{N R_{\text{flcell}}}{N R_{\text{flcell}} + R_{\text{driver}}} V_{\text{DD}} \right)^2 + V_{\text{DD}} I_{\text{field}} t_{\text{switching}}. \tag{2}$$

The energy of the static current is a function of the field current, supply voltage, and switching time of the MTJ. The static component is independent of array size as the supply voltage is constant and the voltage drop is across the peripheral write drivers and the array. The array field current is also constrained by the resistance of the field line

$$I_{\text{field}} R_{\text{flCell}} N \leq V_{\text{DD}}. \tag{3}$$

The energy to switch a single MTJ (E_{switch}) is

$$E_{\text{switch}} = I_{\text{STT}} V_{\text{DD}} t_{\text{switching}} \tag{4}$$

where I_{STT} is the spin-torque switching current. E_{switch} is, therefore, only dependent on the switching time of the MTJ. The total energy per bit is

$$E_{\text{total}} = E_{\text{switch}} + \frac{E_{\text{field}}}{N}. \tag{5}$$

The switching energy is shown in Fig. 8. For comparison, the minimum energy to switch an MTJ, as described by (4), for a nonfield-assisted STT-MRAM cell, is 0.3 pJ/bit. The minimum switching energy of the field-assisted cell is 0.054 pJ/bit with a corresponding switching latency of 618 ps. Due to the bit line resistance, longer rows support a maximum field current at a specific supply voltage. A sufficiently high field cannot be generated to reduce the switching latency of the MTJ, ensuring that the energy consumption is higher than with a shorter row. An optimum row length, therefore, exists that minimizes the overall switching energy of an array during a write. For the configuration shown in Fig. 8, the optimum row length is 128 cells.

As shown by the figure, increasing the number of cells in a row produces a linear increase in energy consumed per bit. However, as the row length increases, the maximum current becomes bounded. For latency critical as well as energy critical circuits, the field currents should be maximized for a given row length, and a larger current should be used rather than a longer row, except for small row lengths.

Large field currents, however, have classically been associated with the half-select problem as magnetic fields may interact with MTJs in adjacent rows. The high thermal stability assumed in this paper, however, prevents errors in the cache. Under an applied field, the expression, $\Delta(H_{\text{app}}) = \Delta(1 - H_{\text{app}}/H_k)^2$, governs the thermal stability of an MTJ [27]. For a 10-ns applied pulse, the immediately adjacent rows exhibit a bit-error rate (BER) of $\sim 10^{-15}$. This BER is small as compared with the lifetime of data within a high-performance cache, and is sufficient for practical operation.

A. Effects of Stochastic Switching

As noted previously, STT switching is a stochastic operation [5]. While deterministic information is sufficient to determine a suitable design point, practical design methods require that the stochastic nature of the switching process be considered.

The energy and latency of each of the physical memory cells are listed in Table V. Each cell type is evaluated for a row length of 128 bits with a 6.5-mA field current applied to

TABLE V
ENERGY AND LATENCY OF STT-MRAM CELLS

Cell Type	Latency (ns)	σ (ns)	90% (ns)	Energy (fJ/bit)
Field-Assisted	0.47	0.481	0.996	93.4
Field-Assisted ($\Delta = 30$)	0.18	0.18	0.38	35.4
Minimum	4.96	1.62	6.65	316.9
Isometric	3.06	0.94	4.10	246.0

the MRAM device. The field-assisted cell exhibits a significant reduction in energy and latency as compared with the minimum and isometric STT cells. As the field is applied, the switching latency decreases; the standard deviation, however, falls disproportionately. A minimum-sized STT cell exhibits a switching latency of 4.96 ns with a 30% standard deviation. While the field-assisted cell exhibits a reduced mean write latency of 0.47 ns, the standard deviation of switching is 102% of the mean. Intuitively, with increasing applied field, the effect of the damping torque diminishes and the system becomes more unpredictable during switching. This effect causes greater variability in the switching latency. To compensate for this variability and to enhance circuit speed, a 90% write success rate is targeted. Write back circuits are used to ensure proper operation, as described in Section VI.

For comparative purposes, a field-assisted cell with reduced nonvolatility is also presented. Unlike the baseline cell, this cell assumes a reduced thermal barrier for the MTJ, which lowers the retention time of the MTJ to one day. This combination produces the shortest latency and the lowest energy configuration. The reduced thermal barrier also exhibits no additional variability as compared with the baseline field-assisted cell. In subsequent analyses, however, the baseline cell is designed to ensure that a typical industrial 10 year retention time is maintained [6].

VI. CACHE EVALUATION

The development of L1 and L2 caches with a field-assisted STT-MRAM is evaluated in this section. SRAM caches and caches using conventional STT-MRAM (without the field-assisted switching mechanism) are treated as a baseline for comparative purposes.

Naive replacement of SRAM arrays and sensing circuitry with STT-MRAM arrays degrades the performance in write critical caches due to the long switching latency, producing an unfair comparison. The baseline STT-MRAM (Minimum and Isometric) caches, therefore, incorporate two state-of-the-art architectural techniques to improve the system performance while tolerating write latency. The caches are typically divided into multiple subbanks to increase the parallel throughput of data accesses and to amortize the cost of the peripheral logic circuitry. Subbank buffering [28] adds an SRAM write buffer in front of each cache subbank [Fig. 9(a)], which locally buffers on-going writes. When data is stored within a subbank buffer, the H-Tree data bus, which is shared across all of the subarrays, serve the next cache access while the long latency STT-MRAM write is local within the subbank. Decoupling the access circuitry and interface bus from the long latency

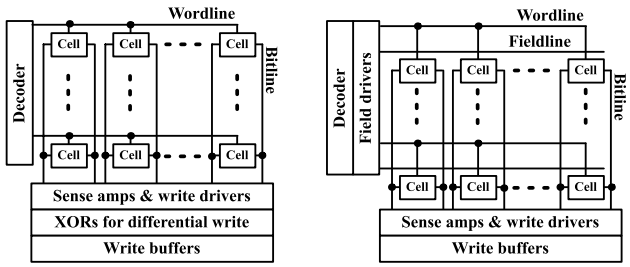


Fig. 9. Array organization for (a) baseline STT-MRAM and (b) field-assisted STT-MRAM.

TABLE VI
CACHE AND MEMORY PARAMETERS

L1 Caches	
iL1/dL1 size	32kB / 32kB
iL1/dL1 block size	64B / 64B
iL1/dL1 round-trip latency	2 / 2 cycles (uncontended)
iL1/dL1 ports	1 / 1
iL1/dL1 banks	1 / 1
iL1/dL1 MSHR entries	8 / 8
iL1/dL1 associativity	2-way / 2-way
Coherence protocol	MESI
Consistency model	Release consistency
Shared L2 Cache and Main Memory	
Shared L2 cache	4MB, 64B block, 8-way
L2 MSHR entries	64
L2 round-trip latency	20 cycles (uncontended)
Write buffer	64 entries
DRAM subsystem	DDR3-1600 SDRAM
Memory controllers	4

write significantly improves cache throughput. In addition, differential writes [29] is a technique commonly used to reduce write energy. Before a write, the stored data are read and compared with the to-be-written data. Only those STT-MRAM cells with different binary states actually switch.

Field-assisted STT-MRAM caches [Fig. 9(b)] also employ subbank buffering, but do not incorporate differential writes since all of the STT-MRAM cells in a row are affected by the field. To guarantee a successful STT-MRAM switching process, a checker read is issued after every write. Upon a write failure, a retry write is issued.

A. Simulation Setup

The cycle accurate SESC simulator [30] has been modified to model a chip multithreaded processor with eight cores and four threads per core operating at 4 GHz. The three configurations for the memory subsystem are listed in Table VI. A field-assisted STT-MRAM cache, an isometric cell cache, and a baseline SRAM configuration are explored in this paper. Isometric and field-assisted cells are used in the L1 cache to evaluate the performance impact of a field-assisted cache. In the L2 cache, where write latency is not a critical parameter, minimum-sized cells are used. Both of these configurations are normalized to an SRAM baseline configuration. CACTI [31] and NVSim [32] are used to estimate the cache energy and access latencies. The cache capacities are the same for both the STT-MRAM and the SRAM caches. The cache latencies

TABLE VII
STT-MRAM CACHE PARAMETERS (CYCLE: 250 ps)

	Baseline STT	Field-Assisted STT
iL1/dL1 latency	1 cycle	1 cycle
L1s write occupancy	17 cycles	4 cycles
L2 latency	6 cycles	7 cycles
L2 write occupancy	28 cycles	4 cycles

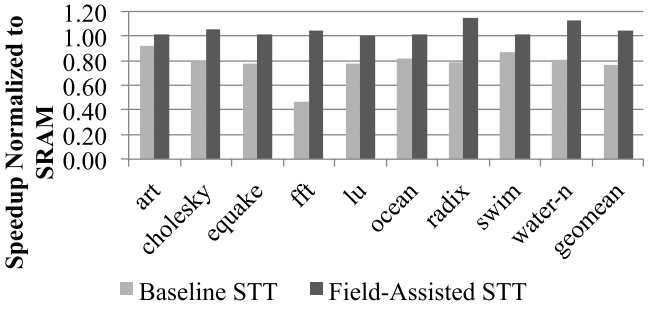


Fig. 10. System performance of STT-MRAM caches normalized to baseline SRAM caches for each cell type.

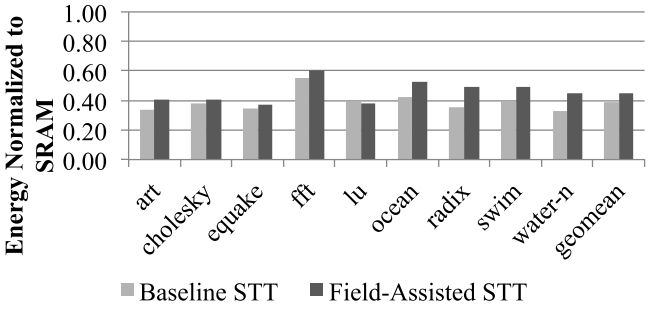


Fig. 11. Energy of STT-MRAM caches normalized to baseline SRAM caches for baseline and field-assisted cell types.

for these configurations are summarized in Table VII. For the baseline STT-MRAM cache configuration, the isometric cells are used for the L1 caches to minimize the MTJ switching latency, and minimum-sized STT-MRAM cells are used for L2 to decrease the cache area and read latency. The field-assisted STT-MRAM cache configuration uses the field-assisted cells for all of the caches within the hierarchy.

A wide range of parallel workloads have been simulated for each configuration. The benchmark suite includes nine software applications, among which three programs are from SPEC OMP2001 [33] and six programs are from SPLASH2 [34]. All workloads are executed in 32 threads on an eight core processor.

B. System Performance, Energy, and Area

The system performance and cache energy are shown in Figs. 10 and 11. All of the comparisons are normalized to the performance of the SRAM caches with the same capacity.

The field-assisted STT-MRAM caches exhibit a slight performance increase as compared with the SRAM caches (Fig. 10) since the STT-MRAM caches occupy less area while maintaining the same capacity, hence benefiting from a shorter wire delay. The baseline STT-MRAM caches exhibit an

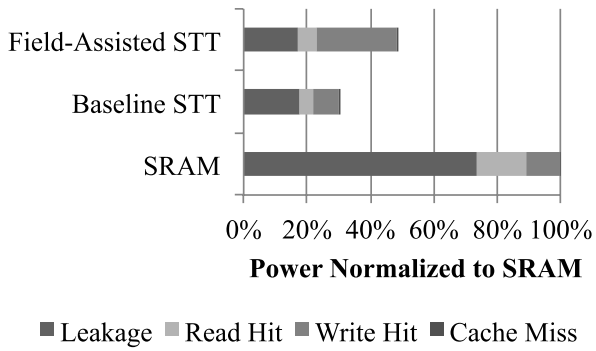


Fig. 12. Power dissipation of STT-MRAM and SRAM caches.

overall decrease in performance as compared with the baseline SRAM caches due to the long write latency. Despite subbank buffering, the reads can be blocked by writes when subbank conflicts occur.

For these applications, STT-MRAM-based caches require less energy (Fig. 11). The field-assisted STT-MRAM caches consume slightly higher energy as compared with the baseline STT-MRAM caches due to two reasons: 1) the field current consumes additional energy and 2) differential writes are applied to the baseline STT-MRAM but not to the field-assisted STT-MRAM. In the application LU, however, the field-assisted STT-MRAM caches consume less energy. This behavior occurs because LU uses a greater number of bit flips during writes. As a result, differential writes have less of an effect on the write energy as compared with other applications using isometric or minimum STT-MRAM cells.

The power dissipated by the benchmarks circuits is shown in Fig. 12 for STT-MRAM and SRAM caches. For all of the STT-MRAM caches, the leakage power is less than SRAM. The power dissipated by the read operations is also less due to the smaller array area and shorter wires. For the baseline STT-MRAM caches, the power dissipated during the write operations is comparable with the power dissipated during the SRAM writes because the MTJs consume greater switching power but the access time is smaller than the SRAM caches. The field-assisted STT-MRAM caches require higher write power due to the additional field currents applied during each write. The field-assisted STT-MRAM caches, however, provide faster write and shorter execution time; hence, the effect of the field currents on the total energy is amortized across the row.

Both caches are compared with the standard SRAM cache for multiple applications. While the baseline STT-MRAM cache exhibits a reduction in total energy of 61.4%, the performance drops by 23.1% as compared with SRAM. The field-assisted STT-MRAM cache exhibits a 54.7% reduction in energy as compared with SRAM, 6.7% more energy than the baseline STT-MRAM cache. Despite this small increase in energy, the field-assisted cache completes execution 4.8% faster than SRAM, a 28% performance improvement as compared with the baseline STT-MRAM cache.

The area and area efficiency (AE) of the field-assisted caches, isometric caches, and SRAM baseline caches are listed in Table VIII. The AE of a cache describes the area of the memory cells as compared with the total area of the cache,

TABLE VIII
AREA OF STT-MRAM CACHES

	Capacity	Cell Type	Cache Portion	Cell area (μm^2)	Subarray area (μm^2)	Total area (μm^2)	Total area efficiency
L1 Data/Instruction Cache	32KB	Field Assisted	Data	7893.9	84.4	40135.1	21.3%
			Tag	621.3	12.1		
L1 Data/Instruction Cache	32KB	Isometric	Data	7879.2	98.2	39714.6	21.2%
			Tag	615.5	7.0		
L2 Shared	4MB	Minimum	Data	87149.5	620.7	3152555.2	24.2%
			Tag	8170.2	18.4		
L1 Data/Instruction Cache	32KB	SRAM	Data	18524.1	1212.0	28819.0	57.5%
			Tag	1302.5	339.2		
L2 Shared	4MB	SRAM	Data	2371086.8	1212.0	5803978.0	44.4%
			Tag	129669.0	2121.1		

expressed as

$$\text{AE} = \frac{\text{Cell area}}{\text{Cache area}} \times 100. \quad (6)$$

The field-assisted and isometric cells exhibit similar area and AE. Notably, the AE drops significantly for MRAM memories. The SRAM baseline cache achieves the same capacity at a smaller area than either MRAM variant, since the overhead of the peripheral circuitry increases for MRAM. While the field-assisted cells require drive transistors for each row, the field-assisted cache exhibits less area for each subarray than the isometric cell cache due to the greater number of subbank buffers needed to manage the write latency. This area difference, however, is small as compared with the total area of the L1 cache. The area of both STT cells requires marginally greater area than an SRAM for a L1 cache. For the larger L2 cache, the density advantages of STT-MRAM are sufficient to reduce area. For the overall cache subsystem, both STT-MRAM configurations use 45.3% less area than an SRAM configuration due primarily to the significant area reduction of the L2 cache.

VII. CONCLUSION

A field-assisted approach is applied to MRAM cells to reduce the switching latency of an STT-MTJ. An array model of the switching latency and energy consumption for different field currents and array sizes is described. It is shown that the per bit switching latency is reduced by a factor of 4. If nonvolatility constraints are relaxed, the overall switching latency is reduced by a factor greater than 10.

Several field-assisted STT-MRAM cells are compared with minimum-sized and isometric area-based STT-MRAM cells. Each of these cells is evaluated for a variety of applications and compared with standard L1 and L2 SRAM cache. The field-assisted STT-MRAM cache demonstrates a 25% performance improvement as compared with a nonfield-assisted cache STT-MRAM cache and a 5% improvement as compared with an SRAM cache while reducing overall energy consumption by an average of 55% as compared with an SRAM cache. The overall cache subsystem exhibits a 42.5% reduction in total area as compared with an SRAM variant, however, a 33% increase in the area of the L1 cache is observed due to the additional peripheral circuitry required to interface with STT-MRAM. The reduction in both switching energy and latency support embedded high-performance STT-MRAM-based cache subsystems, enabling the use of STT-MRAM in upper level caches within high-performance microprocessors.

REFERENCES

[1] L. Chang *et al.*, "An 8T-SRAM for variability tolerance and low-voltage operation in high-performance caches," *IEEE J. Solid-State Circuits*, vol. 43, no. 4, pp. 956–963, Apr. 2008.

[2] W. Zhao, E. Belhaire, C. Chappert, and P. Mazoyer, "Spin transfer torque (STT)-MRAM–based runtime reconfiguration FPGA circuit," *ACM Trans. Embedded Comput. Syst.*, vol. 9, no. 2, pp. 1–16, Oct. 2009, Art. ID 14.

[3] S. Chung *et al.*, "Fully integrated 54 nm STT-RAM with the smallest bit cell dimension for high density memory application," in *Proc. IEEE Int. Electron Devices Meeting*, Dec. 2010, pp. 12.7.1–12.7.4.

[4] M. Hosomi *et al.*, "A novel nonvolatile memory with spin torque transfer magnetization switching: Spin-RAM," in *IEEE Int. Electron Devices Meeting, Tech. Dig.*, Dec. 2005, pp. 459–462.

[5] J. M. Slaughter *et al.*, "High density ST-MRAM technology (Invited)," in *Proc. IEEE Int. Electron Devices Meeting*, Dec. 2012, pp. 29.3.1–29.3.4.

[6] C. W. Smullen, V. Mohan, A. Nigam, S. Gurumurthi, and M. R. Stan, "Relaxing non-volatility for fast and energy-efficient STT-RAM caches," in *Proc. IEEE 17th Int. Symp. High Perform. Comput. Archit.*, Feb. 2011, pp. 50–61.

[7] T. Kishi *et al.*, "Lower-current and fast switching of a perpendicular TMR for high speed and high density spin-transfer-torque MRAM," in *Proc. IEEE Int. Electron Devices Meeting*, Dec. 2008, pp. 1–4.

[8] T. Kawahara *et al.*, "2 Mb SPRAM (SPin-transfer torque RAM) with bit-by-bit bi-directional current write and parallelizing-direction current read," *IEEE J. Solid-State Circuits*, vol. 43, no. 1, pp. 109–120, Jan. 2008.

[9] S. S. P. Parkin *et al.*, "Exchange-biased magnetic tunnel junctions and application to nonvolatile magnetic random access memory (invited)," *J. Appl. Phys.*, vol. 85, no. 8, pp. 5828–5833, Apr. 1999.

[10] M. Durlam *et al.*, "A low power 1 Mbit MRAM based on 1T1MTJ bit cell integrated with copper interconnects," in *IEEE Symp. VLSI Circuits Dig. Tech. Papers*, Jun. 2002, pp. 158–161.

[11] B. N. Engel *et. al*, "A 4-Mb toggle MRAM based on a novel bit and switching method," *IEEE Trans. Magn.*, vol. 41, no. 1, pp. 132–136, Jan. 2005.

[12] Z. Diao *et al.*, "Spin-transfer torque switching in magnetic tunnel junctions and spin-transfer torque random access memory," *J. Phys., Condens. Matter*, vol. 19, no. 16, p. 165209, Apr. 2007.

[13] R. P. Robertazzi, D. C. Worledge, and J. Nowak, "Investigations of half and full select disturb rates in a toggle magnetic random access memory," *Appl. Phys. Lett.*, vol. 92, no. 19, pp. 192510-1–192510-3, May 2008.

[14] W. C. Jeong, J. H. Park, J. H. Oh, G. T. Jeong, H. S. Jeong, and K. Kim, "Highly scalable MRAM using field assisted current induced switching," in *IEEE Symp. VLSI Technol. Dig. Tech. Papers*, Jun. 2005, pp. 184–185.

[15] T. Andre, S. Tehrani, J. Slaughter, and N. Rizzo, "Structures and methods for a field-reset spin-torque MRAM," U.S. Patent 8 228 715, Jul. 24, 2012.

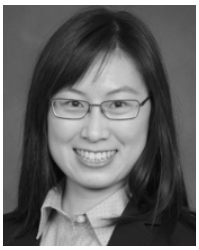
[16] Y. Ding, "Method and system for using a pulsed field to assist spin transfer induced switching of magnetic memory elements," U.S. Patent 7 502 249, Mar. 10, 2009.

[17] X. Wang *et al.*, "Magnetic field assisted STRAM cells," U.S. Patent 8 400 825, Mar. 19, 2013.

[18] X. Cao, H. Xi, W. Zhu, R. Lamberton, and K. Gao, "Field assisted switching of a magnetic memory element," U.S. Patent 8 422 277, Apr. 16, 2013.
[19] C. Mewes and T. Mewes. *M^3 Micromagnetic Simulator*. [Online]. Available: http://www.bama.ua.edu/~tmewes/Mcube/Mcube.shtml, accessed Nov. 2013.
[20] The ITRS Technology Working Groups. *International Technology Roadmap for Semiconductors (ITRS)*. [Online]. Available: http://public.itrs.net
[21] W. Zhao and Y. Cao, "New generation of predictive technology model for sub-45 nm early design exploration," *IEEE Trans. Electron Devices*, vol. 53, no. 11, pp. 2816–2823, Nov. 2006.
[22] H. Zhao *et. al*, "Low writing energy and sub nanosecond spin torque transfer switching of in-plane magnetic tunnel junction for spin torque transfer random access memory," *J. Appl. Phys.*, vol. 109, no. 7, p. 07C720, Apr. 2011.
[23] S. Ikeda *et. al*, "A perpendicular-anisotropy CoFeB–MgO magnetic tunnel junction," *Nature Mater.*, vol. 9, no. 9, pp. 721–724, Jul. 2010.
[24] R. Patel, E. Ipek, and E. Friedman, "STT-MRAM memory cells with enhanced on/off ratio," in *Proc. IEEE Int. Syst.-Chip Conf.*, Sep. 2012, pp. 148–152.
[25] R. Patel, E. Ipek, and E. G. Friedman, "2T–1R STT-MRAM memory cells for enhanced on/off current ratio," *Microelectron. J.*, vol. 45, no. 2, pp. 133–143, Feb. 2014.
[26] F. T. Ulaby, E. Michielssen, and U. Ravaioli, *Fundamentals of Applied Electromagnetics*. Englewood Cliffs, NJ, USA: Prentice-Hall, 2010.
[27] D. D. Tang and Y.-J. Lee, *Magnetic Memory: Fundamentals and Technology*. Cambridge, U.K.: Cambridge Univ. Press, 2010.
[28] X. Guo, E. Ipek, and T. Soyata, "Resistive computation: Avoiding the power wall with low-leakage, STT-MRAM based computing," in *Proc. IEEE/ACM Int. Symp. Comput. Archit.*, Jun. 2010, pp. 371–382.
[29] B. C. Lee, E. Ipek, O. Mutlu, and D. Burger, "Architecting phase change memory as a scalable dram alternative," in *Proc. IEEE/ACM Int. Symp. Comput. Archit.*, Jun. 2009, pp. 2–13.
[30] J. Renau *et al.* (Jan. 2005). *SESC: SuperESCalar Simulator*. Available: http://sesc.sourceforge.net
[31] N. Muralimanohar, R. Balasubramonian, and N. Jouppi, "Optimizing NUCA organizations and wiring alternatives for large caches with CACTI 6.0," in *Proc. 40th Annu. IEEE/ACM Int. Symp. Microarchitecture*, Dec. 2007, pp. 3–14.
[32] X. Dong, C. Xu, Y. Xie, and N. P. Jouppi, "NVSim: A circuit-level performance, energy, and area model for emerging nonvolatile memory," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 31, no. 7, pp. 994–1007, Jul. 2012.
[33] *SPEC OMP2001 Benchmark Suite*. [Online]. Available: http://www.spec.org/omp2001/, accessed Nov. 2013.
[34] S. C. Woo, M. Ohara, E. Torrie, J. P. Singh, and A. Gupta, "The SPLASH-2 programs: Characterization and methodological considerations," in *Proc. IEEE/ACM Int. Symp. Comput. Archit.*, Jun. 1995, pp. 24–36.

Ravi Patel (S'09) received the B.Sc. and M.Sc. degrees in electrical and computer engineering from the University of Rochester, Rochester, NY, USA, in 2008 and 2010, respectively, where he is currently pursuing the Ph.D. degree.

He was a Research Intern with Freescale Semiconductor Inc., Tempe, AZ, USA, in 2010 and 2012, and imec, Leuven, Belgium, in 2014, where he was investigating power network design sub-10-nm integrated circuits. His current research interests include memristors, STT-MRAM, and high performance memories.



Xiaochen Guo (S'09) received the master's degree in electrical and computer engineering from the University of Rochester, Rochester, NY, USA, and the bachelor's degree in computer science and engineering from Beihang University, Beijing, China. She is currently pursuing the Ph.D. degree with the University of Rochester.

Dr. Guo was a twice recipient of the IBM Ph.D. Fellowship.

Qing Guo (S'15) received the B.E. degree in automation from Xi'an Jiaotong University, Xi'an, China, in 2007, and the M.S. degree in computer science from the University of Rochester, Rochester, NY, USA, in 2012, where he is currently pursuing the Ph.D. degree in computer science.

His current research interests include broad area of computer architecture, with an emphasis on energy efficient computing and architectures exploiting resistive memory technologies.

Engin Ipek (M'09) received the B.S., M.S., and Ph.D. degrees from Cornell University, Ithaca, NY, USA, in 2003, 2007, and 2008, respectively, all in electrical and computer engineering.

He was a Researcher with the Computer Architecture Group with Microsoft Research, Redmond, WA, USA, from 2007 to 2009. He is currently an Assistant Professor of Electrical and Computer Engineering and Computer Science with the University of Rochester, Rochester, NY, USA, where he leads the Computer Systems Architecture Laboratory. His current research interests include energy-efficient architectures, high performance memory systems, and the application of emerging memory technologies to computer systems.

Dr. Ipek's research has been recognized by the IEEE Computer Society TCCA Young Computer Architect Award in 2014, two IEEE Micro Top Picks Awards, an Invited Communications of the ACM Research Highlights Article, the ASPLOS 2010 Best Paper Award, and the NSF CAREER Award.

Eby G. Friedman (F'00) received the B.S. degree from Lafayette College, Easton, PA, USA, in 1979, and the M.S. and Ph.D. degrees from the University of California at Irvine, Irvine, CA, USA, in 1981 and 1989, respectively, all in electrical engineering.

He was with Hughes Aircraft Company, from 1979 to 1991, where he became the Manager of the Signal Processing Design and Test Department and was responsible for the design and test of high performance digital and analog ICs. He has been with the Department of Electrical and Computer Engineering, University of Rochester, Rochester, NY, USA, since 1991, where he is currently a Distinguished Professor, and the Director of the High Performance VLSI/IC Design and Analysis Laboratory. He is also a Visiting Professor with the Technion—Israel Institute of Technology, Haifa, Israel. He has authored over 400 papers and book chapters, holds 12 patents, and has authored and edited about 16 books in high-speed and low-power CMOS design techniques, 3-D design methodologies, high-speed interconnect, and the theory and application of synchronous clock and power distribution networks. His current research interests include high performance synchronous digital and mixed-signal microelectronic design and analysis with application to high-speed portable processors and low-power wireless communications.

Dr. Friedman is a Senior Fulbright Fellow. He was a recipient of the IEEE Circuits and Systems 2013 Charles A. Desoer Technical Achievement Award, the University of Rochester Graduate Teaching Award, and the College of Engineering Teaching Excellence Award. He was the Editor-in-Chief and Chair of the Steering Committee of the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATION SYSTEMS, the Regional Editor of the *Journal of Circuits, Systems and Computers*, a member of the Editorial Board of the IEEE PROCEEDINGS, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II: ANALOG AND DIGITAL SIGNAL PROCESSING, the IEEE JOURNAL ON EMERGING AND SELECTED TOPICS IN CIRCUITS AND SYSTEMS, and the *Journal of Signal Processing Systems*, a member of the Circuits and Systems Society Board of Governors, and the Program and Technical Chair of several IEEE conferences. He is also the Editor-in-Chief of the *Microelectronics Journal*, a member of the Editorial Boards of the *Analog Integrated Circuits and Signal Processing*, the *Journal of Low Power Electronics*, and the *Journal of Low Power Electronics and Applications*, and a member of the Technical Program Committee of numerous conferences.