Scalability and Design-Space Analysis of a 1T-1MTJ Memory Cell for STT-RAMs

Richard Dorrance, *Student Member, IEEE*, Fengbo Ren, *Student Member, IEEE*, Yuta Toriyama, Amr Amin Hafez, *Student Member, IEEE*, Chih-Kong Ken Yang, *Fellow, IEEE*, and Dejan Marković, *Member, IEEE*

***Abstract*—We present a design-space feasibility region, as a function of magnetic tunnel junction (MTJ) characteristics and target memory specifications, to explore the design margin of a one-transistor–one-magnetic-tunnel-junction (1T-1MTJ) memory cell for spin-transfer torque random access memories (STT-RAMs). Data from measured devices are used to model the statistical variation of an MTJ's critical switching current and resistance. The sensitivity of the design space to different design parameters is also analyzed for the scaling of both the MTJ and the underlying transistor technology. A design flow, using a sensitivity-based analysis and an MTJ switching model based on the Landau–Lifshitz–Gilbert equation, is proposed to optimize design margins for gigabit-scale memories. Design points for improved yield, density, and memory performance are extracted from MTJ-compatible complementary metal–oxide–semiconductor (CMOS) technologies for 90-, 65-, 45-, and 32-nm processes. Predictive technology models are used to explore the future scalability of STT-RAMs in upcoming 22- and 16-nm technology nodes. Our analysis shows that, to achieve Flash-like densities ($< 6F^2$) in advanced CMOS technologies, aggressive scaling of the critical switching current density will be required.**

***Index Terms*—Magnetic tunnel junction (MTJ), magnetoresistive random access memory (MRAM), process–voltage–temperature (PVT), spin-transfer torque (STT), spin-transfer torque random access memory (STT-RAM), variability.**

I. Introduction

MAGNETORESISTIVE RANDOM ACCESS MEMORIES (MRAMs) have attracted a significant amount of interest as a commercially viable universal memory technology. With the density of the dynamic random access memory (DRAM), the speed of the static random access memory (SRAM), and the nonvolatility of Flash, it is easy to see why [1]. MRAMs require zero standby power and boast a nearly unlimited programming endurance ($> 10^{15}$ cycles) [2]. Such a memory would eliminate the need for multiple application-specific memories, improving system performance and reliability while also lowering cost and power consumption from mobile devices to datacenters [3]; see Fig. 1.

Manuscript received September 16, 2011; revised December 6, 2011; accepted December 17, 2011. Date of publication January 26, 2012; date of current version March 23, 2012. This work was supported in part by the Defense Advanced Research Projects Agency Spin Torque Transfer–Random Access Memory (DARPA STT-RAM; HR0011-09-C-0114) Program. The review of this paper was arranged by Editor V. R. Rao.

The authors are with the Department of Electrical Engineering, University of California, Los Angeles, CA 90095 USA (e-mail: dorrance@ee.ucla.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/TED.2011.2182053

	SRAM	DRAM	Flash (NOR)	Flash (NAND)	MRAM	STT-RAM
Non-volatile	No	No	Yes	Yes	Yes	Yes
Cell Size [F^2]	50-120	6-10	10	5	16-40	6-20
Read Time [ns]	1-100	30	10	50	3-20	2-20
Write/Erase Time [ns]	1-100	15	1µs/1ms	1ms/0.1ms	3-20	2-20
Endurance	10^{16}	10^{16}	10^5	10^5	$>10^{15}$	$>10^{15}$
Write Power	Low	Low	Very High	Very High	High	Low
Other Power Consumption	Leakage	Refresh	None	None	None	None
High Voltage Required	No	3V	6-8V	16-20V	3V	<1.5V

Fig. 1. Performance of various memory technologies [3].

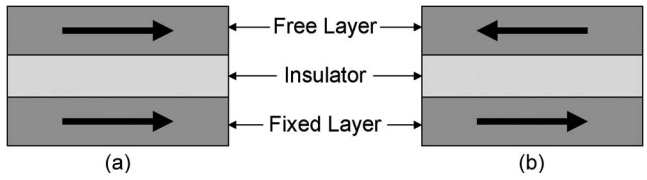


Fig. 2. MTJ ferromagnetic layers in (a) parallel and (b) antiparallel configurations.

The nonvolatile storage element of an MRAM is the magnetic tunnel junction (MTJ). Structurally, an MTJ is a pair of ferromagnets separated by a thin insulating layer. Data storage is achieved by exploiting the magnetic orientation of these ferromagnetic layers [4]. Only the following two magnetic states are stable: 1) the parallel combination [see Fig. 2(a)] and 2) the antiparallel combination [see Fig. 2(b)]. The parallel configuration leads to a low resistive state R_P, whereas the antiparallel configuration leads to a high resistive state $R_{\rm AP}$. Tunnel magnetoresistance (TMR), the ratio of the difference between R_P and $R_{\rm AP}$, is a metric for determining the efficiency of the spintronic operation of an MTJ [5]. TMR is defined as

$$TMR = \frac{R_{\rm AP} - R_P}{R_P}. \tag{1}$$

First-generation MRAMs used field-induced magnetic switching (FIMS) to toggle the MTJ between its parallel and antiparallel states [3]. FIMS works by organizing word and bit

lines into a crosspoint architecture [6]. When a synchronized pulse of current is applied to the desired word and bit lines, a strong magnetic field is created at the intersection of the two wires. This magnetic field then causes the MTJ to switch to the desired state. A small access transistor is also required to read the state of each MRAM cell [6]. Aside from suffering from a serious write disturbance problem (the half-select problem), the major drawback of conventional MRAM is the increase in writing current as technologies scale [3].

The discovery of spin-transfer torque (STT)-based switching has enabled MRAMs to scale below 90 nm. Rather than using an indirect current to generate a magnetic field, STT uses a spin-polarized current through the MTJ to accomplish device switching [7]. Toggling of the MTJ is roughly determined by the current density [8]. As the area of the MTJ device decreases, so does the writing current. Spin-transfer torque random access memories (STT-RAMs) have the added benefit of being architecturally much simpler than conventional MRAMs [9]. The simplest of STT-RAM architecture uses the one-transistor–one-magnetic-tunnel-junction (1T-1MTJ) structure.

Despite the importance of the 1T-1MTJ structure for the future success of STT-RAM, very little comprehensive analysis has been done on the subject. Analysis in the work of Raychowdhury *et al.* [10], [11] considers MTJs but not the underlying transistor technology. In fact, the design of the MTJ and the access transistor are intertwined. A given complementary metal–oxide–semiconductor (CMOS) technology constrains the design space of the MTJ due to the overhead and impact of the access transistor in each memory cell. This condition, in turn, affects the performance of the MTJ, which further impacts the design of the access transistor. Ono *et al.* [12] used a stochastic MTJ model, later verified with on-chip measurements, to optimize the design of a 32-Mb test chip in the presence of asymmetric access transistor behavior. Similarly, Chen *et al.* [13] discuss how a statistical model for the MTJ, which ignores the role of the access device, produces a suboptimal memory cell in both area and yield. Furthermore, the feasibility and yield of the memory depend on the design space and the variation of the MTJs [14].

In this paper, we present a comprehensive analysis of the design space of a 1T-1MTJ memory cell for STT-RAMs. We use a precessional-based switching model, modified to include thermally activated switching, to capture the dynamic nature of the MTJ. The effects of both CMOS and MTJ device variability across process–voltage–temperature (PVT), which is notably absent in prior works, are demonstrated with our analysis. These effects are used to characterize STT-RAMs that scale down to a 32-nm technology with measured device data and extrapolated down to 16 nm using predictive technology models (PTMs).

The remainder of this paper is organized as follows. Section II introduces a model for describing MTJ device variability and predicting MTJ behavior with device scaling. The design space of a 1T-1MTJ STT-RAM memory cell is defined in Section III. Subsequently, Section IV presents a sensitivity-based analysis of the design space for optimizing the yield of a memory array. Section V then explores the future scalability of 1T-1MTJ STT-RAMs, and Section VI concludes this paper.

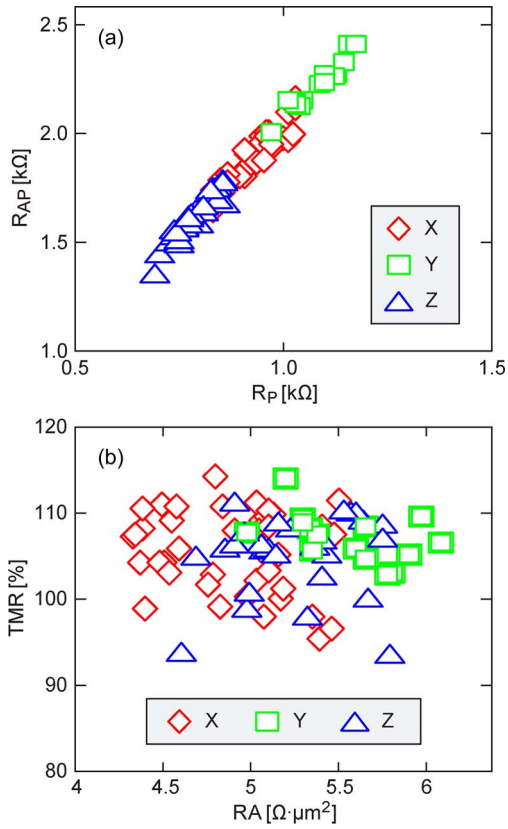


Fig. 3. Measured (a) R_{AP} versus R_P and (b) TMR versus RA for MTJ nanopillars measuring 150×45 nm^2 (X), 130×50 nm^2 (Y), and 170×45 nm^2 (Z) at room temperature (300 K).

II. Modeling MTJ Variability and Scaling

This section describes the MTJ model and characteristics that are used in the subsequent sections to explore the design space for several scaled CMOS technologies. We only consider the scaling of in-plane MTJ devices, because more extensive measurement data are available to us for these types of devices. Functionally, there is no difference between in-plane and perpendicular MTJs, and the analysis framework presented in Section IV does not assume either device. In addition, the critical switching current density of in-plane and perpendicular devices is comparable [15]. Furthermore, in-plane devices show excellent scalability well below 20 nm [16], whereas perpendicular devices are more difficult to fabricate and also suffer from high damping constants [17].

A. MTJ Device Variability

Although the statistical variation of CMOS is generally well understood, similar characteristics for MTJs have not been well documented. This paper uses a combination of fundamental equations and measured device characteristics to model the statistical behavior of MTJs.

1) Resistance: Variations in MTJ resistance and TMR are due to the small geometric differences between fabricated nanopillars. These typically arise from a combination of lithographic variations in the physical dimensions of the nanopillar, as well as minute fluctuations in the thicknesses of up to 20 different layers in state-of-the-art MTJ processes [18]. Fig. 3(a)

TABLE I
MEASURED DEVICE STATISTICS

	X	Y	Z
TMR [%]	105.7	107.3	105.3
σ_{TMR} [%]	4.7	2.7	4.6
RA [Ω·μm^2]	4.88	5.51	5.22
σ_{RA} [Ω·μm^2]	0.342	0.297	0.311

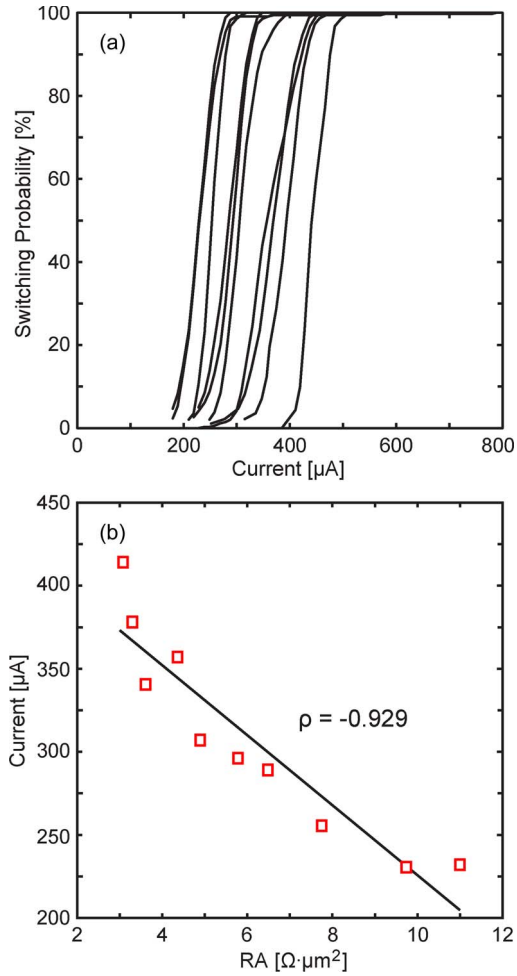


Fig. 4. Measurements of (a) probability of switching versus current and (b) RA versus current (at 50% switching probability) for MTJ nanopillars measuring 135 × 65 nm^2.

contains a plot of measured R_P versus $R_{\rm AP}$ for 105 MTJ nanopillars of varying size and target resistance-area (RA) products. The cumulative effects of random geometric variation on MTJ resistance can be condensed into random Gaussian variation on RA and TMR [19]. Fig. 3(b) and Table I show the calculated statistics for our MTJ nanopillars. Variation on TMR is on the order of 3%–5%, and variation in RA is on the order of 0.3 $\Omega\mu$m^2.

2) Switching Current: Variation in the MTJ critical switching current is the result of two different mechanisms. The first mechanism is thermal agitation, which leads to probabilistic switching in MTJ nanopillars at finite temperatures [20]. An example of this probabilistic switching behavior is shown in Fig. 4(a) for several MgO-based MTJ nanopillars. The second cause of switching current variation in MTJs is due to process-related geometric variation [21]. The effects of geometric variation are clearly evident in Fig. 4(a) as the varying offsets between the probability of switching curves for each MTJ. The general shape of the probability of switching curve for an MTJ has been shown, both theoretically and experimentally, to depend on the thermal stability Δ of the MTJ [22], [23].

To measure the effects that geometric variation has on the critical switching current, MTJ nanopillars were purposely fabricated with large geometric variation. Several critical layers in the MTJ were deposited as a wedge, with their thickness systematically varying by several nanometers from chip edge to chip edge. The resulting induced geometric variation is more than ten times greater than typical random process variation. A strong correlation ($\rho = -0.929$) was found to exist between the RA and the switching current of each device [see Fig. 4(b)]. This case allowed us to use fewer devices to measure the statistical variation of the critical switching current. Based on our measurements, the σ of the switching current due to geometric variation was estimated to be 7 μA or about 2% of the critical switching current. This result is in good agreement with measurements from the work of Driskill-Smith *et al.* (3% variation) [17], Huai *et al.* (3% variation) [20], and Pakala *et al.* (3.5% variation) [22].

B. Scaling of MTJ Current and Resistance

The resistance and switching current can be modeled using a precessional-based switching model, modified to include thermally activated switching [24]. The switching current of an MTJ in the precessional region, for a constant pulse of duration τ, is given by

$$I_C = I_{C0}\left[1 - \frac{\ln(\tau/\tau_0)}{\Delta}\right] \tag{2}$$

where τ_0 is the natural time constant, and I_{C0} is the critical switching current. This critical switching current [25] is given by

$$I_{C0} = \frac{\alpha 4\pi e}{\eta\hbar} M_S^2 V \tag{3}$$

where α is the Gilbert damping constant, η is the factor of spin polarization, $\hbar$ is the reduced Planck constant, e is the elemental charge of an electron, M_S is the magnetization saturation of the free layer, and V is the volume of the free layer.

For an MTJ with free-layer dimensions $l > w \gg d$ [26], [27], as shown in Fig. 5, the thermal stability of an MTJ is approximately

$$\Delta = \frac{E}{k_B T} = \frac{H_K M_S}{2k_B T} V \approx d\left(\frac{1}{w} - \frac{1}{l}\right)\frac{M_S^2}{k_B T} V \tag{4}$$

where k_B is the Boltzmann constant, T is the absolute temperature in Kelvin, H_K is the out-of-plane uniaxial anisotropy, and E is the energy of anisotropy [28], [29].

In this paper, dimensional scaling is performed to maintain a constant Δ to ensure the long-term nonvolatility of the STT-RAM. Therefore, dimensions l and w of the MTJ are scaled

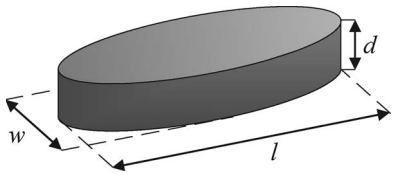


Fig. 5. MTJ free-layer dimensions.

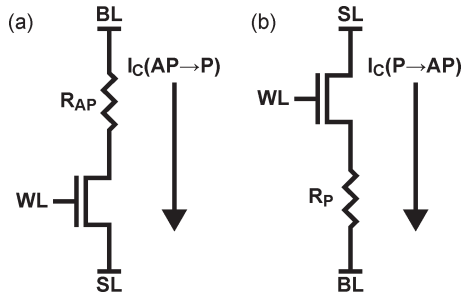


Fig. 6. 1T-1MTJ cell architecture showing MTJ switching current for (a) AP–P and (b) P–AP.

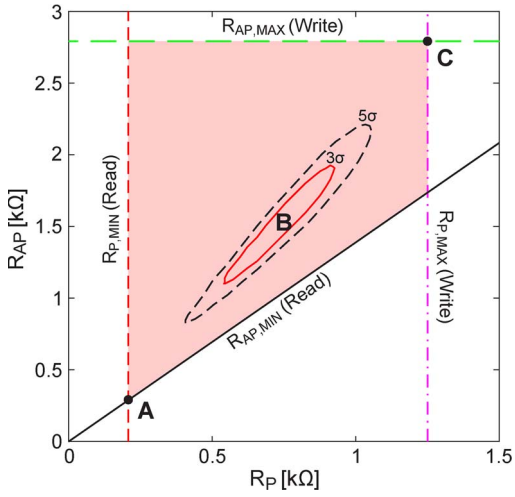


Fig. 7. Design space in a 65-nm process for $W_N = 2.0$ μm, $I_C(P \rightarrow AP) = 500$ μA, and $I_C(AP \rightarrow P) = 375$ μA, with an overlay of device X based on Table I.

by a factor λ to manipulate I_{C0} and $R_{P/\mathrm{AP}}$, and then, to keep Δ constant, d must scale by $\lambda^{-1/2}$. This case results in $I_{C0} \propto lwd \rightarrow \lambda^{3/2}$ and $R_{P/\mathrm{AP}} \propto l^{-1}w^{-1} \rightarrow \lambda^{-2}$.

III. Design Space

The analysis in this paper uses a conventional 1T-1MTJ cell architecture, as shown in Fig. 6. The minimum writing currents, to ensure a target write error rate (WER), for flipping the cell resistance are defined as $I_C(P \rightarrow AP)$ and $I_C(AP \rightarrow P)$. The design space of a single STT-RAM memory cell can be illustrated using an R_{AP} versus R_P plot, as shown in Fig. 7. The feasibility region is indicated by the shaded region. It contains all points (R_P, R_{AP}) in the design space so that a memory cell made with such an MTJ is functional. In the design space, the two lower bounds are set by the read margin of the cell, whereas the two upper bounds are set by the write margin of the cell.

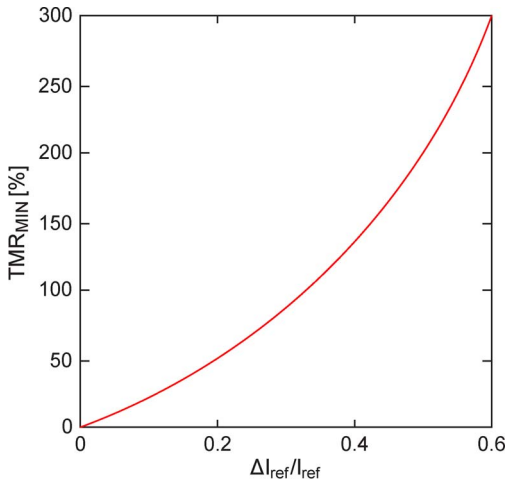


Fig. 8. Design-space lower bound TMR_{MIN} versus $\Delta I_{\mathrm{ref}}/I_{\mathrm{ref}}$ for a current-sensing read circuit with ideal reference resistance $2(R_P \parallel R_{\mathrm{AP}})$.

The lower bound $R_{P,\mathrm{MIN}}$ depends on the implementation of the sense amplifier and represents the minimum resistance required for reliable circuit operation. Parasitic resistances from the access transistor and column-mux, as well as the bit and source lines, all contribute to $R_{P,\mathrm{MIN}}$. In addition, $R_{\mathrm{AP,MIN}}$ is determined by TMR_{MIN} (Fig. 8), the minimum TMR required for the read amplifier to differentiate between R_P and R_{AP}. Regardless of the specifics of the implementation, all sense amplifiers are either a voltage- or a current-sensing topology. For a generic current-sensing read circuit, a read margin of ΔI_{ref} results in

$$(\mathrm{Current})TMR_{\mathrm{MIN}} = \frac{2\Delta I_{\mathrm{ref}}/I_{\mathrm{ref}}}{1 - \Delta I_{\mathrm{ref}}/I_{\mathrm{ref}}}. \tag{5}$$

For I_{ref} flowing through the reference resistance R_{ref}, $I_{\mathrm{ref}} + \Delta I_{\mathrm{ref},1}$ flows through R_P, and $I_{\mathrm{ref}} - \Delta I_{\mathrm{ref},2}$ flows through R_{AP}. When $\Delta I_{\mathrm{ref},1} = \Delta I_{\mathrm{ref},2} = \Delta I_{\mathrm{ref}}$, TMR_{MIN} is minimized. Under this condition, $R_{\mathrm{ref}} = 2(R_P \parallel R_{\mathrm{AP}})$, and we can express TMR_{MIN} as a function of the normalized fractional sensing current $(\Delta I_{\mathrm{ref}}/I_{\mathrm{ref}})$. In (5), ΔI_{ref} must be chosen so that the read amplifier correctly evaluates across all transistor PVT variations.

Similarly, TMR_{MIN} for a generic voltage-sensing topology is

$$(\mathrm{Voltage})\ TMR_{\mathrm{MIN}} = \frac{2\Delta V}{R_P I_{\mathrm{ref}}} = \frac{\Delta R}{R_P} \tag{6}$$

where ΔV, the voltage-reading margin, is the minimum difference in sensing voltage between the MTJ and the reference resistance, and $\Delta R = 2\Delta V/I_{\mathrm{ref}}$ is the minimum difference in resistance between R_P and R_{AP}. The difference between voltage- or current-sensing topologies is shown in Fig. 9. Voltage sensing is better suited for devices with larger RAs, where a small TMR can still translate into a large resistance difference. Alternatively, current-sensing topologies can better differentiate low-RA MTJs. Note that the lower bounds $R_{P,\mathrm{MIN}}$ and $R_{\mathrm{AP,MIN}}$, although critical to the readability of the cell, are almost completely independent of the MTJs used. The only requirement is that the sensing time and the current I_{ref}

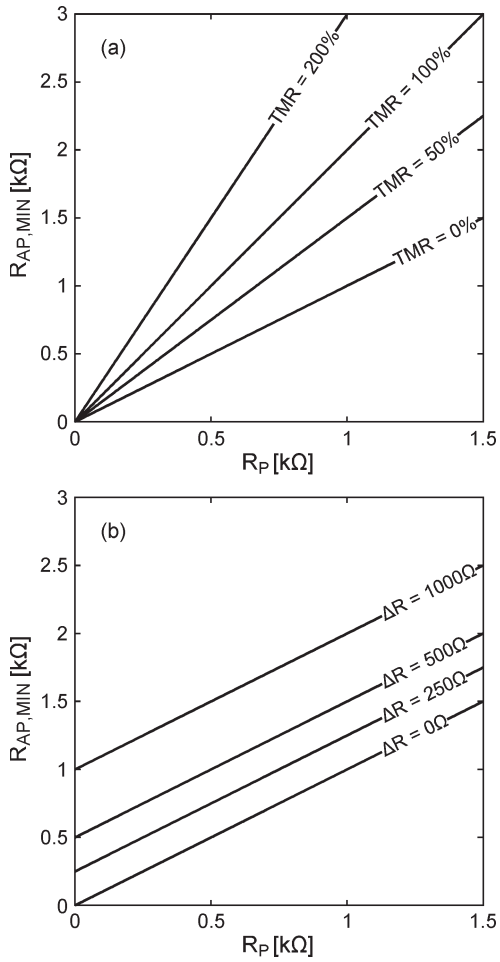


Fig. 9. Difference in $R_{\mathrm{AP,MIN}}$ between (a) current sensing and (b) voltage sensing.

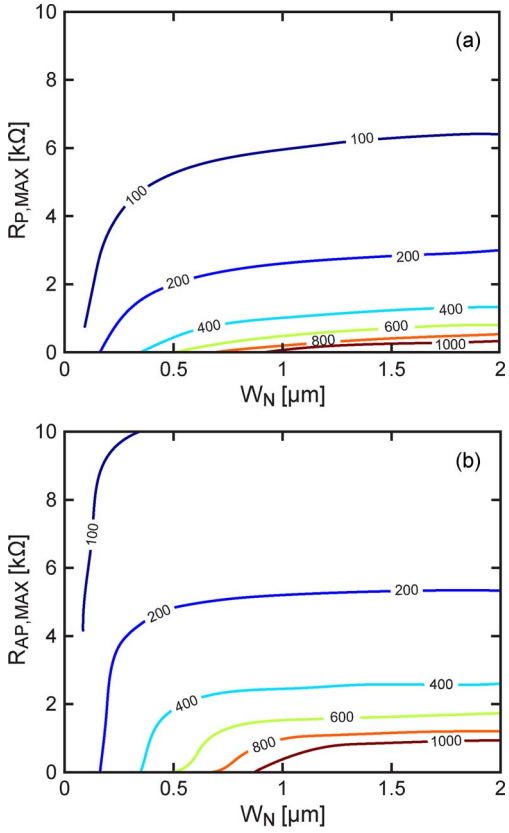


Fig. 10. (a) $R_{P,\mathrm{MAX}}$ and (b) $R_{\mathrm{AP,MAX}}$ at nominal V_{DD} for a 65-nm process (I_C contours are measured in microamperes).

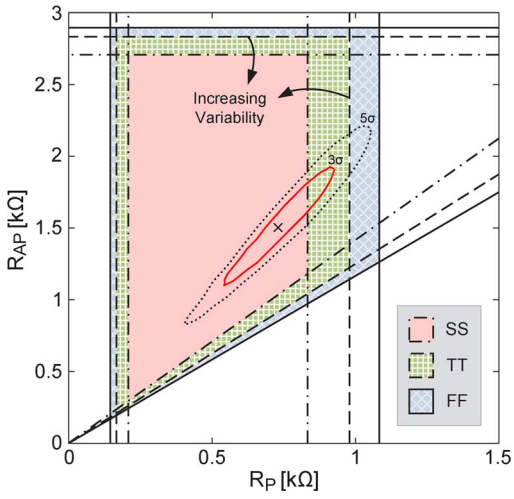


Fig. 11. Design space in a 65-nm process for $W_N = 750$ nm, $I_C(P \rightarrow AP) = 300\ \mu$A, and $I_C(AP \rightarrow P) = 300\ \mu$A, with an overlay of device X based on Table I, for Slow-Slow (SS), Typical-Typical (TT), and Fast-Fast (FF) corners.

are small enough to avoid disturbing the cell during the read operation.

The upper bounds $R_{P,\mathrm{MAX}}$ and $R_{\mathrm{AP,MAX}}$ are the maximum allowable resistances such that the access transistor, in a 1T-1MTJ configuration, can still provide the minimum critical writing currents $I_C(P \rightarrow AP)$ and $I_C(AP \rightarrow P)$. These upper bounds are subsequently very sensitive to the specific characteristics of the MTJ device used. As such, to ensure a sufficiently low WER, the effects of stochastic thermal fluctuations [30], self-induced heating [31], and backhopping [32] on the probability of switching should not be overlooked. Transistor-level simulations are used to determine the relationship between $R_{P/\mathrm{AP,MAX}}$, I_C, and cell size (transistor width W_N) for a technology. Fig. 10 shows an example of such a simulation in a 65-nm process. Using the conventional configuration in Fig. 6, W_N is swept along with R_{MAX}. The contours of the simulated current are shown.

Fig. 7 shows a specific MTJ cell and its associated statistical variation (the concentric ovals around point B) overlaid on the design space. The design-space margin (DSM) can be defined as the number of σ's of MTJ variation before crossing any of the previously defined bounds. Defining the DSM in terms of σ simplifies feasibility characterization to a single variable and thus allows yield to quickly be calculated. To a first order, 3σ–6σ of the design margin roughly correspond to reliably producing 1-kb, 32-kb, 4-Mb, and 1-Gb memory arrays.

Fig. 11 highlights the effects of CMOS variability on the design-space bounds. To more clearly illustrate the effects, a $35F^2$ cell in a 65-nm process is chosen. As expected, the more stringent constraints of the SS corner cause the design space to shrink. This shift is caused by an increase in the threshold voltage of the access transistor. Environmental variables such as

temperature also have a significant effect on the design space. A consumer-grade STT-RAM is expected to operate over a range of more than 100 °C, in which TMR can drop by more than 30% [33], degrading the DSM for readability. These sources of technological and environmental variability must also be considered in the design process.

IV. 1T-1MTJ Cell Optimization Using a Sensitivity Analysis

Many variables, at both the circuit and device levels, affect the design space. To optimize all variables for a target memory specification, we must determine how each variable impacts the design space. This section introduces a design-space sensitivity (DSS) as a metric of quantifying the behavior of the change in design space as a function of various design parameters (e.g., $V_{\rm DD}$, λ, J_C, RA, TMR, and W_N). We then present a sensitivity-based design flow that uses DSS to optimize the DSM of a 1T-1MTJ memory cell. A short design example using a 65-nm CMOS technology is provided.

A. DSS Analysis

First, consider the points A–C in Fig. 7. Points A and C correspond to the corner values of R_P and $R_{\rm AP}$ in the feasible design space. Point B represents the nominal MTJ at the center of the MTJ device distribution. For a positive design margin to exist, point B must fall somewhere between points A and C.

A "better" design space can be achieved from altering a design parameter when a larger distribution of the MTJs (the number of σ) falls within the feasible region. Note that the improved design space increases not only the area of the feasibility region but also the number of sigma enclosed by the feasible region. Recall that point A depends only slightly on the MTJ parameters. Therefore, the improvement (or deterioration) of the design space mostly depends on the change in DSM between points B and C as a function of a particular design variable.

Therefore, we define the DSS to the parameter X as

$$DSS(X) = \frac{\partial \left(\frac{R_C - R_B}{\sigma}\right)_{P/\rm AP}}{\partial X} \tag{7}$$

where R_B and R_C are taken as either R_P or $R_{\rm AP}$ at points B and C, thus defining the DSS along each dimension of the design space. $(R_C - R_B/\sigma)$ is the normalized distance between points B and C in the design space along the $R_{P/\rm AP}$ dimension. Intuitively, $DSS(X)$ describes the instantaneous rate of change in DSM to a particular design parameter X. The derivative loses positional information, and therefore, we used the DSS in conjunction with the original plot of the design space to determine the benefit of tuning the design parameter X. For both the R_P and $R_{\rm AP}$ dimensions, if $DSS(X) > 0$, then the DSM is improved by increasing X, and if $DSS(X) < 0$, then the DSM is improved by decreasing X. When the DSSs for the two dimensions conflict, the DSM in each dimension should then be taken into account.

The design flow is shown in Fig. 12.

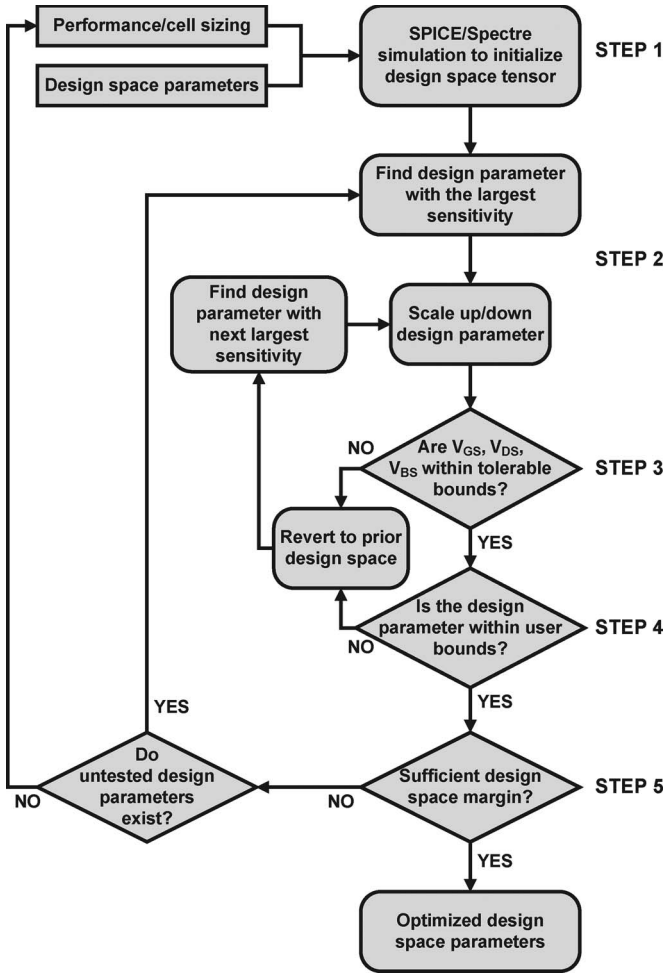


Fig. 12. Sensitivity-based design flow of a 1T-1MTJ memory cell for DSM optimization.

B. Design Flow for DSM Optimization

Using the aforementioned DSS analysis, the design flow for a 1T-1MTJ STT-RAM memory cell is described as follows.

- Step 1) Characterize the design space of the memory cell for a given cell size and some performance requirements using device-level simulations [e.g., Simulation Program With Integrated Circuit Emphasis (SPICE) or Spectre]. For N design-space variables, build an N-dimensional tensor of the design space.
- Step 2) Select the design-space parameter with the largest sensitivity, in terms of magnitude, to the limiting design bound and scale it by an incremental amount in the direction of DSS (up if positive and down if negative).
- Step 3) Check if the design parameter is within the specified bounds. If not, revert to the prior design space and repeat step 2 using the design parameter with the next highest sensitivity.
- Step 4) Check if the $V_{\rm GS}$, $V_{\rm DS}$, and $V_{\rm BS}$ voltage constraints are met. If not, revert to the prior design space and repeat step 2 using the design parameter with the next highest sensitivity.

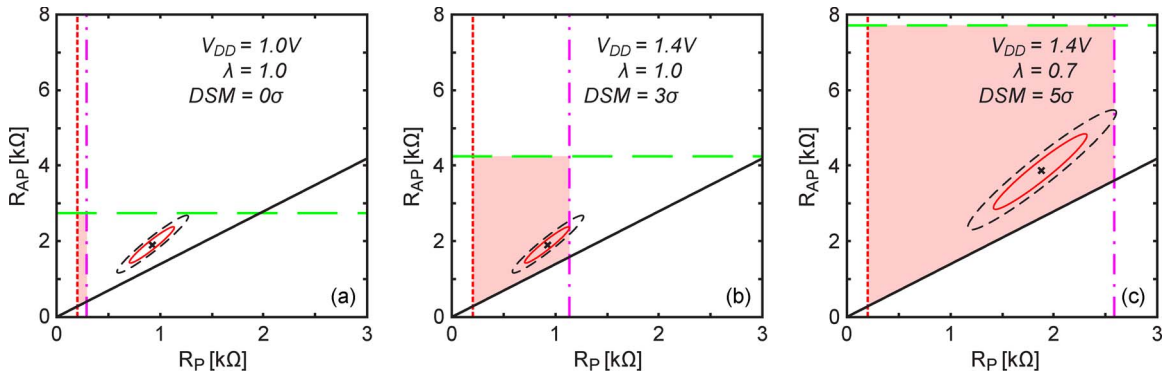


Fig. 13. Design space in a 65-nm process for a $30F^2$ cell ($W_N = 0.65$ μm) for device X based on Table I: $I_C(P \rightarrow AP) = 450$ μA, $I_C(AP \rightarrow P) = 300$ μA. The inner red oval represents 3σ of MTJ device variation. The dashed black oval corresponds to 5σ of MTJ variation.

Step 5) Check the DSM and repeat steps 2–4 if insufficient. If no additional design parameters exist, relax cell sizing or performance requirements and repeat steps 1–5.

C. Design Example

In this section, we use the sensitivity analysis to design a 4-Mb STT-MRAM with a cell size of $30F^2$ (comparable to eDRAM) in a 65-nm technology. Device X in Table I, with $I_C(P \rightarrow AP) = 450$ μA and $I_C(AP \rightarrow P) = 300$ μA, is the nominal MTJ and can be scaled by λ. In addition, approximately 5σ of the design margin is required for reasonable yield.

Fig. 13(a) shows the design space for a nominal $V_{\rm DD} = 1.0$ V and $\lambda = 1.0$. The inner red oval is the 3σ variation of the MTJ, whereas the dashed black oval represents the 5σ variation of the MTJ. Clearly, with nominal $V_{\rm DD}$ and λ, the memory is not functional. Fig. 14 shows that the design space is much more sensitive to $V_{\rm DD}$ than it is to λ. Therefore, we choose to scale $V_{\rm DD}$ to 1.4 V. Note that, at 1.4 V, much of the voltage is dropped across the MTJ, leaving the $V_{\rm GS}$ and $V_{\rm DS}$ of the access transistor below 1 V. Fig. 13(b) shows the new design space, with the 3σ bound at the edge of the design boundary.

Scaling $V_{\rm DD}$ alone proves insufficient to meet the 5σ design margin required, and therefore, we simultaneously scale λ. Fig. 14(b) shows that scaling λ results in conflicting DSS. The $R_{\rm AP}$ margin improves more by scaling λ up, whereas the R_P margin improves by scaling λ down. However, Fig. 13(b) indicates that the $R_{\rm AP}$ dimension has considerable margin and we can trade off some of that margin for improved margin in R_P. Therefore, we choose to scale λ down to 0.7. As shown in Fig. 13(c), the desired 5σ bound on MTJ variation is essentially enclosed within the design space.

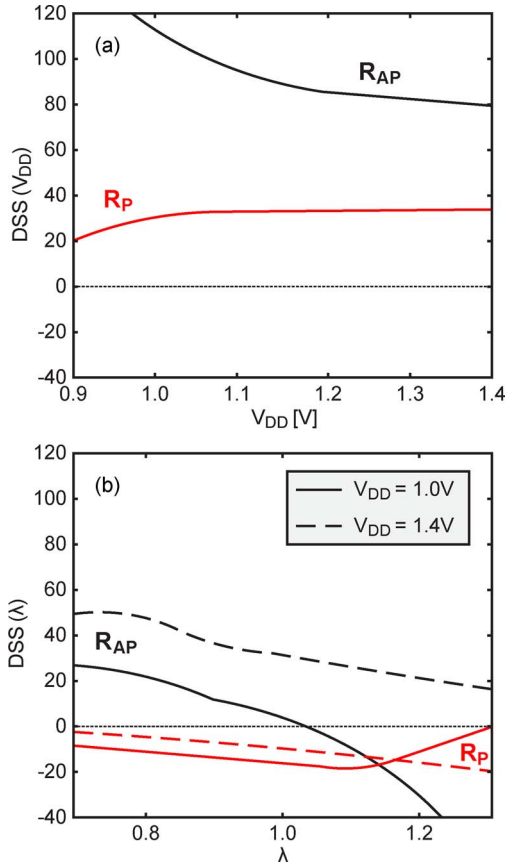


Fig. 14. DSS of parameters (a) $V_{\rm DD}$ and (b) λ in a 65-nm technology.

V. Future Scalability

Scalability is an important feature for the success of a memory technology. However, the continued scaling of SRAM, DRAM, and Flash memories has become increasingly more difficult. The growing severity of random dopant fluctuation (RDF) in 65- and 45-nm technology nodes has led to the phenomena of "reverse scaling," which is expected to become much more severe below 32 nm [34]. Although eight- and ten-transistor SRAM cells show better scalability than a six-transistor cell, their future beyond a 22-nm node is questionable at best [35]. Similarly, scaling DRAM and Flash technologies below 22 nm is also at risk. It has become increasingly difficult for DRAM to guarantee data retention while contending with exponentially increasing CMOS leakage and falling cell capacitances [36]. Deteriorating reliability and retention times, as well as decreased programming speeds, have already begun to show up in sub-45-nm Flash technologies [36].

Before STT, exponentially increasing critical current densities presented a major roadblock, preventing MRAM from scaling below 90 nm. However, with STT, the critical switching current density remains constant between successive

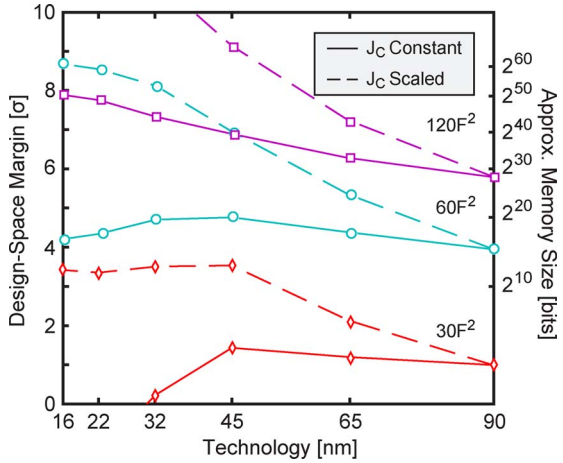


Fig. 15. Design margin versus technology node for constant $J_c(P \rightarrow AP) = 6 \times 10^6$ A/cm^2 and $J_c(P \rightarrow AP)$ scaling by 4.5% each technology generation.

TABLE II
$J_C(P \rightarrow AP)$ FOR AN RA OF 5 $\Omega \cdot \mu$m^2

Equivalent Cell Size	J_C vs. Technology Node (10^6 A/cm^2)					
	90nm	65nm	45nm	32nm	22nm‡	16nm‡
FLASH ($6F^2$)	2.72	2.62	2.65	2.20	2.04	1.86
eDRAM ($30F^2$)	4.60	4.58	4.57	4.31	4.23	4.12
SRAM ($120F^2$)	6.67	6.76	6.86	7.07	7.17	7.30

‡Predicted

generations of CMOS technologies. Fig. 15 shows how the scaling of only the CMOS technology (access transistor) impacts the design margin for different cell sizes when using the same MTJ with constant critical current density ($J_C = 6 \times 10^6$ A/cm^2 for 10-ns $P \rightarrow AP$ switching with an RA of 5 $\Omega \cdot \mu$m^2). SRAM-equivalent sizes ($120F^2$) scale quite well, with increasing DSM more than sufficient to construct gigabit memories ($> 6\sigma$) for the same performance. However, as we decrease the cell size, the DSM begins to degrade and practically disappears once we reach an eDRAM-equivalent cell size ($30F^2$). In addition, note that, below 45 nm, we see a sharp decrease in DSM for smaller cell sizes. The effects of RDF become more pronounced in future technology nodes, resulting in the stagnation of transistor current density scaling. However, if the MTJ current density scales with technology, also shown in Fig. 15, then this trend is reversed. By scaling J_C by as little as 4.5%, a constant design margin can be achieved between each technology node below 45 nm.

Alternatively, we can explore how critical switching current densities scale for constant DSM. Tables II–IV contain the critical switching current densities for Flash-, eDRAM-, and SRAM-equivalent cell sizes for RAs of 5, 10, and 15 $\Omega\mu$m^2, respectively. In each table, values correspond to 5σ of DSM for sub-10-ns switching times. Each table also corresponds to the low, mid, and high sides of reported MTJ RAs [37]. Note that, although larger RAs require smaller current densities (to meet voltage headroom constraints), they scale much better between successive technology nodes.

TABLE III
$J_C(P \rightarrow AP)$ FOR AN RA OF 10 $\Omega \cdot \mu$m^2

Equivalent Cell Size	J_C vs. Technology Node (10^6 A/cm^2)					
	90nm	65nm	45nm	32nm	22nm‡	16nm‡
FLASH ($6F^2$)	2.22	2.14	2.19	1.82	1.69	1.56
eDRAM ($30F^2$)	3.27	3.28	3.33	3.20	3.18	3.16
SRAM ($120F^2$)	4.27	4.33	4.44	4.59	4.68	4.78

‡Predicted

TABLE IV
$J_C(P \rightarrow AP)$ FOR AN RA OF 15 $\Omega \cdot \mu$m^2

Equivalent Cell Size	J_C vs. Technology Node (10^6 A/cm^2)					
	90nm	65nm	45nm	32nm	22nm‡	16nm‡
FLASH ($6F^2$)	1.83	1.81	1.82	1.67	1.62	1.56
eDRAM ($30F^2$)	2.56	2.59	2.63	2.58	2.59	2.61
SRAM ($120F^2$)	3.17	3.21	3.33	3.47	3.55	3.65

‡Predicted

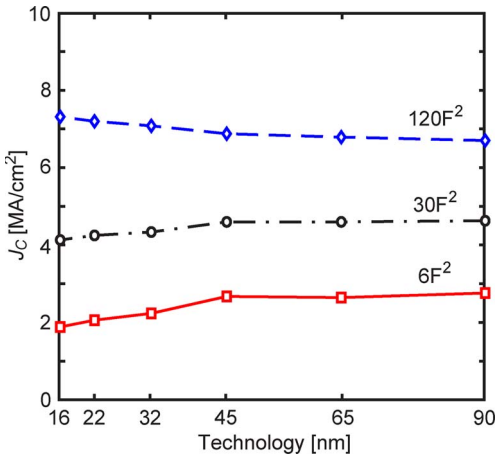


Fig. 16. Maximum critical switching current density for 5σ of DSM in Flash-, eDRAM-, and SRAM-equivalent cell sizes ($6F^2$, $30F^2$, and $120F^2$) for an RA of 5 $\Omega \cdot \mu$m^2.

Fig. 16 graphically represents Table II. Again, we can see that SRAM-like cell sizes scale very well with CMOS. The critical current density for eDRAM-like cell sizes remains constant until sub-45-nm technologies, where it begins to experience a very gradual decay. This decay becomes much more pronounced in Flash-like sizes due to the larger effect that RDF has on smaller transistors. Beyond a 16-nm technology node, the severity of RDF and the inability of metal pitches to scale will force STT-RAMs to use multilevel cells to maintain FLASH-like densities [15].

In Tables II–IV, a 45-nm technology requires current densities well below 3×10^6 A/cm^2 and less than 2×10^6 A/cm^2 in upcoming 22- and 16-nm technology nodes. State-of-the-art MTJs, with thermal stability factors large enough to support data retention for ten years or more, have current densities between $2-4 \times 10^6$ A/cm^2 [23]. These devices are well suited to replace SRAMs and eDRAMs. However, the aggressive scaling of MTJ switching currents is still required to achieve Flash-like densities in current and future technology nodes.

VI. Conclusion

In this paper, we have shown that the joint optimization of multiple design parameters is essential in the design of an STT-RAM memory array. We have derived the necessary framework to allow for such a systematic design procedure. In addition, the analytical methodology presented in this paper has been utilized to show that the mild scaling of MTJ J_C is required to enable Flash-like memory densities in upcoming CMOS technologies. Such densities, coupled with low write energies and the nonvolatility of STT-RAM, make STT-RAM a possible contender for next-generation memories.

Acknowledgment

The authors would like to thank P. Khalili, Z. Zeng, H. Park, and H. Chen of the University of California, Los Angeles (UCLA), as well as I. Krivorotov and G. Rowlands of the University of California, Irvine (UCI), for their contributions.

References

[1] B. Cockburn, "The emergence of high-density semiconductor-compatible spintronic memory," in *Proc. Int. Conf. MEMS, NANO Smart Syst.*, Jul. 2003, pp. 321–326.

[2] S. Tehrani, J. Slaughter, M. Deherrera, B. Engel, N. Rizzo, J. Salter, M. Durlam, R. Dave, J. Janesky, B. Butcher, K. Smith, and G. Grynkewich, "Magnetoresistive random access memory using magnetic tunnel junctions," *Proc. IEEE*, vol. 91, no. 5, pp. 703–714, May 2003.

[3] S. A. Wolf, J. Lu, M. R. Stan, E. Chen, and D. M. Treger, "The promise of nanomagnetics and spintronics for future logic and universal memory," *Proc. IEEE*, vol. 98, no. 12, pp. 2155–2168, Dec. 2010.

[4] M. E. Flatte, "Spintronics," *IEEE Trans. Electron Devices*, vol. 54, no. 5, pp. 907–920, May 2007.

[5] J. Z. Sun, "Spin angular momentum transfer in current-perpendicular nanomagnetic junctions," *IBM J. Res. Develop.*, vol. 50, no. 1, pp. 81–100, Jan. 2006.

[6] T. Sugibayashi, N. Sakimura, T. Honda, K. Nagahara, K. Tsuji, H. Numata, S. Miura, K. Shimura, Y. Kato, S. Saito, Y. Fukumoto, H. Honjo, T. Suzuki, K. Suemitsu, T. Mukai, K. Mori, R. Nebashi, S. Fukami, N. Ohshima, H. Hada, N. Ishiwata, N. Kasai, and S. Tahara, "A 16-Mb toggle MRAM with burst modes," *IEEE J. Solid-State Circuits*, vol. 42, no. 11, pp. 2378–2385, Nov. 2007.

[7] J. C. Slonczewski, "Current-driven excitation of magnetic multilayers," *J. Magn. Magn. Mater.*, vol. 159, no. 1/2, pp. L1–L7, Jun. 1996.

[8] E. Chen, D. Apalkov, Z. Diao, A. Driskill-Smith, D. Druist, D. Lottis, V. Nikitin, X. Tang, S. Watts, S. Wang, S. Wolf, A. Ghosh, J. Lu, S. Poon, M. Stan, W. Butler, S. Gupta, C. Mewes, T. Mewes, and P. Visscher, "Advances and future prospects of spin-transfer torque random access memory," *IEEE Trans. Magn.*, vol. 46, no. 6, pp. 1873–1878, Jun. 2010.

[9] M. Hosomi, H. Yamagishi, T. Yamamoto, K. Bessho, Y. Higo, K. Yamane, H. Yamada, M. Shoji, H. Hachino, C. Fukumoto, H. Nagao, and H. Kano, "A novel nonvolatile memory with spin torque transfer magnetization switching: Spin-RAM," in *IEDM Tech. Dig.*, Dec. 2005, pp. 459–462.

[10] A. Raychowdhury, D. Somasekhar, T. Karnik, and V. De, "Design space and scalability exploration of 1T-1STT MTJ memory arrays in the presence of variability and disturbances," in *IEDM Tech. Dig.*, Dec. 2009, pp. 1–4.

[11] A. Raychowdhury, D. Somasekhar, T. Karnik, and V. De, "Modeling and analysis of read (RD) disturb in 1T-1STT MTJ memory bits," in *Proc. Device Res Conf.*, Jun. 2010, pp. 43–44.

[12] K. Ono, T. Kawahara, R. Takemura, K. Miura, H. Yamamoto, M. Yamanouchi, J. Hayakawa, K. Ito, H. Takahashi, S. Ikeda, H. Hasegawa, H. Matsuoka, and H. Ohno, "A disturbance-free read scheme and a compact stochastic-spin-dynamics-based MTJ circuit model for gigabit-scale SPRAM," in *IEDM Tech. Dig.*, Dec. 2009, pp. 1–4.

[13] Y. Chen, X. Wang, H. Li, H. Xi, Y. Yan, and W. Zhu, "Design margin exploration of spin-transfer torque RAM (STT-RAM) in scaled technologies," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 18, no. 12, pp. 1724–1734, Dec. 2010.

[14] M. Motoyoshi, I. Yamamura, W. Ohtsuka, M. Shouji, H. Yamagishi, M. Nakamura, H. Yamada, K. Tai, T. Kikutani, T. Sagara, K. Moriyama, H. Mori, C. Fukamoto, M. Watanabe, R. Hachino, H. Kano, K. Bessho, H. Narisawa, M. Hosomi, and N. Okazaki, "A study for 0.18-μm high-density MRAM," in *Proc. Symp. VLSIT*, Jun. 2004, pp. 22–23.

[15] A. Driskill-Smith, D. Apalkov, V. Nikitin, X. Tang, S. Watts, D. Lottis, K. Moon, A. Khvalkovskiy, R. Kawakami, X. Luo, A. Ong, E. Chen, and M. Krounbi, "Latest advances and roadmap for in-plane and perpendicular STT-RAM," in *Proc. IEEE IMW*, May 2011, pp. 1–3.

[16] D. Apalkov, S. Watts, A. Driskill-Smith, E. Chen, Z. Diao, and V. Nikitin, "Comparison of scaling of in-plane and perpendicular spin transfer switching technologies by micromagnetic simulation," *IEEE Trans Magn.*, vol. 46, no. 6, pp. 2240–2243, Jun. 2010.

[17] A. Driskill-Smith, S. Watts, D. Apalkov, D. Druist, X. Tang, Z. Diao, X. Luo, A. Ong, V. Nikitin, and E. Chen, "Nonvolatile spin-transfer torque RAM (STT-RAM): An analysis of chip data, thermal stability and scalability," in *Proc. IEEE IMW*, May 2010, pp. 1–3.

[18] S. R. Min, H. N. Cho, K. W. Kim, Y. J. Cho, S.-H. Choa, and C. W. Chung, "Etch characteristics of magnetic tunnel junction stack with nanometer-sized patterns for magnetic random access memory," *Thin Solid Films*, vol. 516, no. 11, pp. 3507–3511, Nov. 2008.

[19] R. Beach, T. Min, C. Horng, Q. Chen, P. Sherman, S. Le, S. Young, K. Yang, H. Yu, X. Lu, W. Kula, T. Zhong, R. Xiao, A. Zhong, G. Liu, J. Kan, J. Yuan, J. Chen, R. Tong, J. Chien, T. Torng, D. Tang, P. Wang, M. Chen, S. Assefa, M. Qazi, J. DeBrosse, M. Gaidis, S. Kanakasabapathy, Y. Lu, J. Nowak, E. O'Sullivan, T. Maffitt, J. Sun, and W. Gallagher, "A statistical study of magnetic tunnel junctions for high-density spin torque transfer-MRAM (STT-MRAM)," in *IEDM Tech. Dig.*, Dec. 2008, pp. 1–4.

[20] Y. Huai, M. Pakala, Z. Diao, and Y. Ding, "Spin-transfer switching current distribution and reduction in magnetic tunneling junction-based structures," *IEEE Trans. Magn.*, vol. 41, no. 10, pp. 2621–2626, Oct. 2005.

[21] Y. Katoh, S. Saito, H. Honjo, R. Nebashi, N. Sakimura, T. Suzuki, S. Miura, and T. Sugibayashi, "Analysis of MTJ edge deformation influence on switching current distribution for next-generation high-speed MRAMs," *IEEE Trans. Magn.*, vol. 45, no. 10, pp. 3804–3807, Oct. 2009.

[22] M. Pakala, Y. Huai, T. Valet, Y. Ding, and Z. Diao, "Critical current distribution in spin-transfer-switched magnetic tunnel junctions," *J Appl. Phys.*, vol. 98, no. 5, pp. 056 107-1–056 107-3, Sep. 2005.

[23] A. Driskill-Smith, S. Watts, V. Nikitin, D. Apalkov, D. Druist, R. Kawakami, X. Tang, X. Luo, A. Ong, and E. Chen, "Nonvolatile spin-transfer torque RAM (STT-RAM): Data, analysis and design requirements for thermal stability," in *Proc. Symp. VLSIT*, Jun. 2010, pp. 51–52.

[24] T. Moriyama, T. J. Gudmundsen, P. Y. Huang, L. Liu, D. A. Muller, D. C. Ralph, and R. A. Buhrman, "Tunnel magnetoresistance and spin torque switching in MgO-based magnetic tunnel junctions with a Co/Ni multilayer electrode," *Appl. Phys. Lett.*, vol. 97, no. 7, pp. 072 513-1–072 513-3, Aug. 2010.

[25] J. Z. Sun, "Spin-current interaction with a monodomain magnetic body: A model study," *Phys. Rev. B*, vol. 62, no. 1, pp. 570–578, Jul. 2000.

[26] H. Chang and J. Burns, "Demagnetizing and stray fields of elliptical films," *J. Appl. Phys.*, vol. 37, no. 8, pp. 3240–3245, Jul. 1966.

[27] J. A. Osborn, "Demagnetizing factors of the general ellipsoid," *Phys. Rev.*, vol. 67, no. 11/12, pp. 351–357, Jun. 1945.

[28] V. Korenivski and R. Leuschner, "Thermally activated switching in nanoscale magnetic tunnel junctions," *IEEE Trans. Magn.*, vol. 46, no. 6, pp. 2101–2103, Jun. 2010.

[29] J. Sun and D. Ralph, "Magnetoresistance and spin-transfer torque in magnetic tunnel junctions," *J. Magn. Magn. Mater.*, vol. 320, no. 7, pp. 1227–1237, Apr. 2008.

[30] T. Devolder, J. Hayakawa, K. Ito, H. Takahashi, S. Ikeda, P. Crozat, N. Zerounian, J.-V. Kim, C. Chappert, and H. Ohno, "Single-shot time-resolved measurements of nanosecond-scale spin-transfer induced switching: Stochastic versus deterministic aspects," *Phys. Rev Lett.*, vol. 100, no. 5, p. 057 206, Feb. 2008.

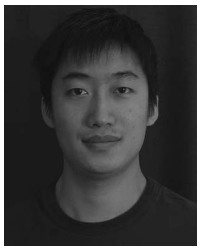
[31] Y. Zhang, Z. Zhang, Y. Liu, Z. Kang, B. Ma, and Q. Y. Jin, "Micromagnetic study of hotspot and thermal effects on spin-transfer switching in magnetic tunnel junctions," *J. Appl. Phys.*, vol. 101, no. 10, pp. 103 905-1–103 905-6, May 2007.

[32] J. Z. Sun, M. C. Gaidis, G. Hu, E. J. O'Sullivan, S. L. Brown, J. J. Nowak, P. L. Trouilloud, and D. C. Worledge, "High-bias backhopping in nanosecond time-domain spin-torque switches of MgO-based magnetic tunnel junctions," *J. Appl. Phys.*, vol. 105, no. 7, pp. 07D109-1–07D109-3, Apr. 2009.

[33] K. Lee and S. Kang, "Design consideration of magnetic tunnel junctions for reliable high-temperature operation of STT-MRAM," *IEEE Trans. Magn.*, vol. 46, no. 6, pp. 1537–1540, Jun. 2010.
[34] S. Kosonocky, A. Bhavnagarwala, and L. Chang, "Scalability options for future SRAM memories," in *Proc. ICSICT*, Oct. 2006, pp. 689–692.
[35] H. Yamauchi, "A discussion on SRAM circuit design trend in deeper nanometer-scale technologies," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 18, no. 5, pp. 763–774, May 2010.
[36] S. Hong, "Memory technology trend and future challenges," in *IEDM Tech. Dig.*, Dec. 2010, pp. 12.4.1–12.4.4.
[37] Z. M. Zeng, P. K. Amiri, G. Rowlands, H. Zhao, I. N. Krivorotov, J.-P. Wang, J. A. Katine, J. Langer, K. Galatsis, K. L. Wang, and H. W. Jiang, "Effect of resistance-area product on spin-transfer switching in MgO-based magnetic tunnel junction memory cells," *Appl. Phys. Lett.*, vol. 98, no. 7, pp. 072 512-1–072 512-3, Feb. 2011.

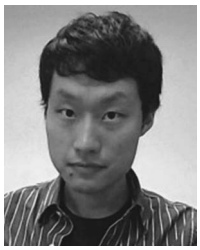
Richard Dorrance (S'09) received the B.S. degree in electrical engineering and computer science in 2009 from the University of California, Berkeley, and the M.S. degree in electrical engineering in 2011 from the University of California, Los Angeles, where he is currently working toward the Ph.D. degree in electrical engineering in the Department of Electrical Engineering.

His research interests include the modeling and integration of post-CMOS devices for VLSI circuit design.



Fengbo Ren (S'10) was born in Shenyang, China, on November 18, 1985. He received the B.Eng. degree in electrical engineering in 2008 from Zhejiang University, Hangzhou, China, and the M.S. degree in electrical engineering in 2010 from the University of California, Los Angeles, where he is currently working toward the Ph.D. degree in the Department of Electrical Engineering.

From August to December 2006, he was an Exchange Student with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Kowloon, Hong Kong, SAR. Since September 2008, he has been specializing in circuit and embedded systems. From September to December 2009, he was an Engineer Intern with the Digital ASIC Group, Qualcomm Inc., San Diego, CA. His research interests include circuit and system design with post-CMOS devices and design optimization.

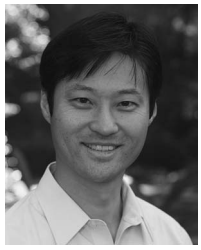


Yuta Toriyama received the B.S. degree in electrical engineering and computer science in 2009 from the University of California, Berkeley, and the M.S. degree in electrical engineering in 2011 from the University of California, Los Angeles, where he is currently working toward the Ph.D. degree in electrical engineering in the Department of Electrical Engineering.

His research interests include VLSI and digital architecture design.

Amr Amin Hafez (S'10) received the B.Sc. and M.Sc. degrees in electrical engineering from Ain-Shams University, Cairo, Egypt, in 2004 and 2008, respectively. He is currently working toward the Ph.D. degree in integrated circuits and systems in the Department of Electrical Engineering, University of California, Los Angeles.

His research interests include high-speed mixed signal and radiofrequency circuits and systems.



Chih-Kong Ken Yang (S'94–M'98–SM'07–F'11) was born in Taipei, Taiwan. He received the B.S., M.S., and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 1992, 1992, and 1998, respectively.

In 1999, he joined the University of California, Los Angeles, where he is currently a Professor with the Department of Electrical Engineering. His current research includes high-performance mixed-mode circuit design for VLSI systems such as clock generation, high-performance signaling, low-power digital design, and analog-to-digital conversion.

Dr. Yang received the 2003–2005 IBM Faculty Development Fellowship and the 2003 Northrup–Grumman Outstanding Teaching Award.

Dejan Marković (S'96–M'06) received the Dipl.Ing. degree in electrical engineering from the University of Belgrade, Serbia, in 1998 and the M.S. and Ph.D. degrees in electrical engineering from the University of California, Berkeley, in 2000 and 2006, respectively.

In 2006, he joined the faculty of the Department of Electrical Engineering, University of California, Los Angeles, as an Assistant Professor. His current research interests include digital integrated circuits and DSP architectures for parallel data processing in future radio and healthcare systems, design with post-CMOS devices, design optimization methods, and CAD flows.

Dr. Marković received the CalVIEW Fellow Award in 2001 and 2002 for his excellence in teaching and mentoring of industry engineers through the University of California, Berkeley, Distance Learning Program. He is a corecipient of the Best Paper Award at the IEEE International Symposium on Quality Electronic Design in 2004 and the recipient of the David J. Sakrison Memorial Prize from the UC Berkeley in 2007 in recognition of the impact of his Ph.D. work and the Faculty Early Career Development (CAREER) Award from the National Science Foundation (NSF) in 2009.