# Tunneling Negative Differential Resistance-Assisted STT-RAM for Efficient Read and Write Operation

Shaodi Wang, *Student Member, IEEE,* Andrew Pan, *Member, IEEE,* Chi On Chui, *Senior Member, IEEE,* and Puneet Gupta, *Senior Member, IEEE*

***Abstract*—The adoption of spin-transfer torque random access memory (STT-RAM) into non-volatile memory systems faces three major obstacles: high write energy, low sensing margin, and high read disturbance. Many designs have been suggested to resolve each of these challenges separately and at the cost of significant overhead. We propose a single low-overhead solution to all these problems without changing the underlying memory architecture by using negative differential resistance (NDR) devices like tunnel diodes (TDs) or tunnel field-effect transistors (TFETs) to assist the STT-RAM write and read process. We show through simulations that the proposed designs can dramatically improve the write and read energy efficiency and sensing margins while minimizing the read disturbance, even after accounting for process variations. Our results open a design path for energy-efficient and reliable STT-RAM technologies.**

***Index Terms*—MRAM, negative differential resistance, tunneling diode, tunneling FET, write termination, read margin, read disturbance, reliability.**

## I. Introduction

RANDOM access memory (RAM) using spin-transfer torque (STT) magnetic tunnel junctions (MTJs) is a promising data storage technology [1–3] due to its non-volatility, zero leakage power, high endurance, immunity to single-event soft errors, and high thermal budget [4] while potentially matching the speed and area of dynamic RAM (DRAM) [5]. It is therefore a possible replacement for current memory technologies, including static RAM (SRAM) cache [6, 7] and DRAM main memories [8], as well as a candidate for new memory architectures utilizing its non-volatility, e.g., fast persistent memory systems for instant recovery from the off-state [9].

However, STT-RAM faces several key challenges: 1) high write currents (up to $100\,\mu A$ at the 45nm node [5]), 2) low sensing margins [10], which force trade-offs between read error rate with read time and energy, and 3) susceptibility to read disturbance [3, 11, 12], *i.e.*, MTJ false switching during sensing, which unfortunately increases with sensing margin. These limitations are intrinsically due to the low tunnel magnetoresistance (TMR) of STT-MTJs, which is defined by the ratio of high and low resistances $R_H$ and $R_L$ $TMR = (R_H - R_L)/R_L$ and typically ranges from 50% to 200%.

For both in-plane and out-of-plane STT-MTJs, write current density does not decrease with scaling [13]. Moreover, small STT-MTJs face reliability problems from existing commercial etching techniques [14]. As a result, the use of STT-RAM is currently limited by high write energy density requirements. To make matters worse, write time may be further extended by process and temperature variations [10, 15]. Similarly, read errors and read disturbance are major reliability concerns in STT-RAM [10–12, 16], wherein sensing current can falsely switch the MTJ and create soft errors.

In this paper, we offer an unified low-overhead solution to simultaneously resolve all of these challenges by utilizing the peak-to-valley ratio (PVR) of negative differential resistance (NDR) devices like tunnel field-effect transistors (TFETs) [17] and tunnel diodes (TDs) [18] within the read and write circuitry. We propose for the first time the use of NDR devices to perform early magnetic RAM (MRAM) write termination and we present new NDR-based read circuit designs to improve read margin and reduce read disturbance. Our proposals limit redundant write current and allow the sensing current ratio of the high and low MTJ states to be amplified up to the NDR PVR, which can be much larger. Our results show that the proposed designs greatly reduce write energy and read disturbance and increase the sensing margin and simplifying sensing circuitry, enabling truly low power STT-RAM technology. The basic issues and modeling of STT-MTJs and NDR devices are introduced in Section II. We present the interaction of NDR devices with MTJs and our novel read and write designs for STT-RAM using NDR in Section III. We perform a Monte Carlo analysis of the impact of process variations on the proposed designs in Section IV. We summarize our conclusions in Section V.

## II. MTJ and NDR Background and Modeling

### A. MTJ Modeling and Memory Operation Concerns

The STT-MTJ is a resistive memory device whose resistance is determined by the magnetization directions of two ferromagnetic layers. The direction of one layer (referred to as the reference layer) is fixed while the other one (the free layer) can be switched by driving a write current through the device. A low resistance $R_L$ is present when the magnetization directions of the two layers are parallel (the $P$ state), while high resistance $R_H$ is realized when the two directions are anti-parallel (the $AP$ state). For memory applications, we will denote these two resistance conditions as the "1" and "0" states, respectively. Through this work, the electrical characteristics of STT-MTJs (with assumed diameter of 50 nm) are modeled using a Landau-Lifshitz-Gilbert (LLG) differential equation solver [15]. The switching characteristics of MTJs leads to several technological challenges for MRAMs, which are briefly reviewed below.

The manuscript was submitted for review August 2, 2016.

The authors are with the Department of Electrical Engineering, University of California at Los Angeles, Los Angeles, CA, 90095 USA (e-mail: shaodiwang@g.ucla.edu).

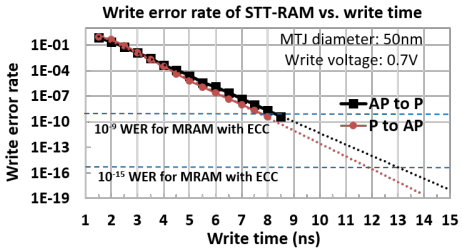



Fig. 1: MRAM write error rate as a function of write time assuming 0.7 V write voltage extracted from 10 billion Monte Carlo circuit simulations using the methodology of Section IV. The circuit includes MTJ, access transistor, and 256-1T1M bit-line capacitance load.

*1) Wasted Power During Write Cycles:* MTJ switching is a stochastic process whose probability increases with the duration of the write current pulse. The write error rate (WER) as a function of write time is shown in Fig. 1. Generally, switching from 1 to 0 ($P$ to $AP$) requires a larger write current than 0 to 1 [19]. In our simulation, the switching efficiency ratio for 1-to-0 and 0-to-1 is 0.75 to 1. However, this asymmetry is offset by the different resistances of the two states; hence when the same voltage is used for both switching directions, similar write times (time to achieve a required WER) are observed in Fig. 1. For robust function of memories without error-correcting code (ECC), the WER should be below $10^{-18}$, which needs 15 ns write time. In this paper we will consider designs with ECC, which still require a WER of $10^{-9}$ or better, necessitating at least 9 ns long write pulse [20]. While long write times are necessary to maintain accuracy, over 90% of switching events (i.e., WER $< 0.1$) are completed within the first 3 ns indicating to a dramatic waste of energy. In particular, write-1 consumes 1.4X the energy of write-0 because the MTJ stays longer in the low resistance 1 state, leading to higher leakage.

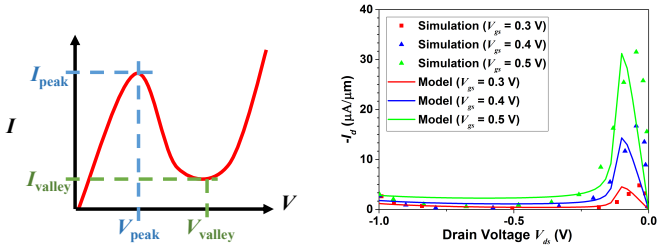
*2) Read Margin and Read Disturbance Limits:* The low TMR of STT-MTJs limits MRAM read margins and read disturbance, causing reliability problems. In particular it is difficult to simultaneously improve both read margin and read disturbance rate in conventional designs. This is because higher margins require higher read voltages, which increase the read current and can lead to unwanted MTJ switching even if the current is smaller than the critical switching current. For example, to obtain a read margin of 150 mV in a conventional design with 100 fF bit-line load, a WER of over $10^{-8}$ is needed, exceeding the ECC error tolerance (see Fig. 10). Non-uniformity in device characteristics due to process variations can worsen this problem.

### B. NDR Device Characteristics

*1) Tunneling-Based NDR Devices:* To address the challenges facing STT-based MRAM, we will introduce NDR devices into the read and write circuitry. As illustrated in Fig. 2a, NDR devices have the property that within a certain bias range (between $V_{peak}$ and $V_{valley}$), the absolute current decreases with increased absolute voltage. The ratio of the maximum and minimum currents ($I_{peak}$ and $I_{valley}$, respectively) within this range is known as the PVR. A variety of two- and three-terminal devices utilizing quantum tunneling such as Esaki diodes, resonant tunneling diodes (RTDs), and reverse-biased TFETs can be used to generate this effect. While TDs are relatively mature devices developed specifically to implement NDR for various applications, TFETs have the advantage that the drain current NDR can be tuned by gate voltage. In Table I we summarize the experimental characteristics of some representative TDs and TFETs. RTDs with good performance have already been demonstrated on Si/SiGe [18], which is already a CMOS-compatible platform [21]. Many of other best-performing NDR devices thus far are based on III-V materials. Non-commercialized technologies like heteroepitaxy [22] and nano-transfer [23] can integrate III-V MOSFET and FinFET with Si CMOS at the expense of cost increase. The integration of III-V on silicon is already a high priority in industry (for instance, by using III-V MOSFETs or TFETs to supplement or replace silicon transistors in logic), and commercial advances in that direction will also ease integration of NDR devices.

TABLE I: Experimental characteristics of selected NDR tunneling devices from literature. Peak current is expressed in terms of per unit width for TFETs and per unit area for TDs.

| Device | Material | Substrate | Peak Current | PVR |
|---|---|---|---|---|
| TD [18] | Si/SiGe | Si | $50\,\mu A/\mu m^2$ | 6 |
| TD [24] | InGaAs | Si | $2.5\,\mu A/\mu m^2$ | 56 |
| TD [25] | InGaAs/InAlAs | InP | $2\,\mu A/\mu m^2$ | 144 |
| TFET [26] | InAs/AlSb/GaSb | GaSb | $\leq 230\,\mu A/\mu m$ | $\leq 5.5$ |
| TFET [27] | InGaAs/InAs | InP | $\leq 4\,\mu A/\mu m$ | $\leq 6.2$ |




(a) (b)

Fig. 2: a) Schematic $I - V$ for typical NDR device. b) $I_d - V_d$ of analytical TFET model and simulated device data of [17]. For the TFET model, parameters are $A_{TFET}$ = 1.3E-8 A/um, $B_{TFET}$ = 4E6 eV/cm, $E_g$ = 0.74 eV, $\lambda$ = 6E-7 cm, $A$ = -0.02, $B$ = 0.0456, $C$ = 0.04, $n$ = 0.3, and $D = 0.0025$.

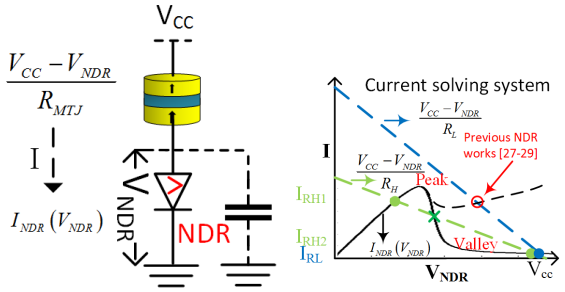
*2) Modeling of NDR devices:* In this work we explore MRAM designs using either TDs or TFETs to implement NDR. To describe diode characteristics, we adapt the compact model and model parameters in [28], which were chosen to fit the InAs/AlSb/GaSb TD characteristics presented therein. For n-type TFETs, while compact models have been developed to describe the positive drain-source voltage device operation, most simple models neglect the NDR characteristics under negative drain-source voltage. In this work, we model TFET NDR by fitting device data to the equation

$$I_{drain} = A_{TFET}(V_{gs}, V_{ds}) f_{NDR}(V_{ds}) + I_{diode} \quad (1)$$

where $A_{TFET}(V_{gs}, V_{ds})$ is an existing gate- and drain-voltage TFET $I - V$ analytical model [29], $f_{NDR}(V_{ds})$ is a function describing two-terminal TD current [28], and $I_{diode}$ is a standard expression for intrinsic diode current. By adjusting the model coefficients we can match both the ordinary and NDR characteristics of simulated and experimental TFETs. In this paper, we discuss results for TFETs using the simulated device characteristics presented in [17], which is compared with our analytical model in Fig. 2b.

## III. NDR-ASSISTED MRAM WRITE AND READ

### A. Behavior of Series-Connected MTJ and NDR Devices



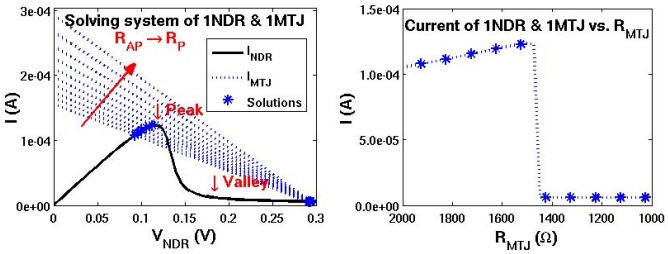

(a) (b)

Fig. 3: (a) Series connection of MTJ and NDR device (TFET or TD). Note that each NDR device is shared by multiple bit-lines in the proposed design. (b) NDR (solid line) and MTJ current (dashed lines) as a function of voltage drop on NDR ($V_{NDR}$). Green dashed line shows the requirement for previous NDR-MRAM designs [30–32].

The unique characteristics of NDR devices enable them to enhance MRAM performance when integrated into the read and write circuitry. The reasons and conditions for these improvements can be understood by examining the behavior of series-connected MTJ and NDR devices, as shown in Fig. 3a. For simplicity the MTJ is treated as a resistor whose current is $(V_{cc} - V_{NDR})/R_{MTJ}$, where $R_{MTJ} = R_H$ (0 state) or $R_L$ (1 state) depending on its state. In Fig. 3b, we plot the current through the NDR device (solid black line) and the MTJ (dotted lines) as functions of the voltage drop across the NDR device $V_{NDR}$. The intersections of the NDR and MTJ curves represent the possible steady-state solutions of the circuit: there are three solutions in the 0 state but only one solution in the 1 state. The target operating conditions are that 1) $|-R| < R_L$, where $-R$ is the effective NDR between $V_{peak}$ and $V_{valley}$, and 2) the NDR device $I_{peak}$ is greater than the current through $R_H$ but below that of $R_L$ at $V_{Peak}$. Under these conditions, the circuit current in the 1 state is limited by the (minimal) NDR valley current.

The idea that negative resistance can differentiate between MRAM states has previously been used in a few proposals for read voltage margin improvement or write energy saving [30–32]. However, those proposals design the NDR-MTJ circuit for only one solution in both the 0 and 1 states (illustrated by the dashed NDR valley curve in Fig. 3b). This requires that the NDR curve only intersect the $R_H$ load line once, forcing both MTJ states to have comparatively high current and making the design vulnerable to NDR or MTJ device variations. Such nominal designs can improve the voltage margin but not necessarily the current margin between the MTJ states, whereas our concept amplifies both the voltage and current differences since the 1 state is forced into the much lower current NDR valley region. Therefore, such proposals do not offer all the operational advantages of our concept, as further discussed below.

In our design, the system should reside in the leftmost solution (below the NDR peak voltage) when the MTJ is in the 0 state. This requires that state to be stable; fortunately it can be easily shown using Lyapunov's second method that the leftmost and rightmost states (solid green dots in Fig. 3b) are asymptotically stable in the sense of being convergent over time, whereas the middle state (the green cross in the figure) is unstable because fluctuations around this point will drive the system towards the stable points. For the low resistance state, only one stable solution exists in the current valley as illustrated by the blue dot in Fig. 3b.




(a) (b)

Fig. 4: (a) $I_{NDR}$ versus $V_{NDR}$ (solid lines) and $I_{MTJ}$ versus $(V_{CC} - V_{NDR})$ (dashed lines) for different $R_{MTJ}$. (b) Current of the series connection of MTJ and NDR vs. $R_{MTJ}$.

To illustrate the behavior under switching, we show the NDR and MTJ currents for different MTJ resistance under constant $V_{CC}$ in Fig. 4a. If the MTJ is initially in the 0 state, the current is close to $I_{peak}$. Upon switching to the 1 state, the MTJ resistance will decrease and the stable solutions approach $I_{peak}$, beyond which the NDR current suddenly drops to the valley region. The current vs. resistance of this process is also plotted in Fig. 4b. This demonstrates that, given proper choice of peak current, $|-R|$, and $V_{CC}$, the NDR device can sense the different MTJ states and switching therein and adjust the current through the circuit accordingly. For sufficiently large PVR, the resulting difference in the series-connected current can be much greater than the ratio of the MTJ resistances. This means that the operating margins are no longer limited by the TMR (typically 0.5-2X) but by the NDR PVR (which can be 5-100X, as shown in Table I).

### B. STT-RAM Write Energy Reduction

Having established the scenarios in which NDR devices can clamp the current of a serially connected MTJ, we propose using this effect to perform early write current cutoff when STT-MTJs switch to or stay in the 1 state. During writing from the 0 to 1 state, this configuration can cause write termination by automatically cutting off the write current once the 1 state is attained and the NDR device ($V_{NDR}$) enters its valley region. Similarly, if a write-to-1 operation is performed on an MTJ already in the 1 state, the NDR forces a very low current during the whole write cycle, saving energy.

This can be accomplished by integrating NDR devices within the memory write circuitry in the manner shown on the right side of Fig. 5. An NDR device (TFET or TD)is added to a regular STT-RAM circuit for write assistance. Because only one MTJ cell is activated at a time during a write operation, the required overhead is minimal since, much like the sense amplifier, a single NDR device can be shared by multiple MTJ bit-lines. For a write-0 operation, *Write1*, *Read*, and *Pre-charge* are set to GND (0 voltage), and *Write0* is set to $V_{cc}$ to activate a write path without NDR (the same as conventional write). For a write-1 operation however, which dissipates the most power, only *Write1* is set to $V_{cc}$ and hence an NDR

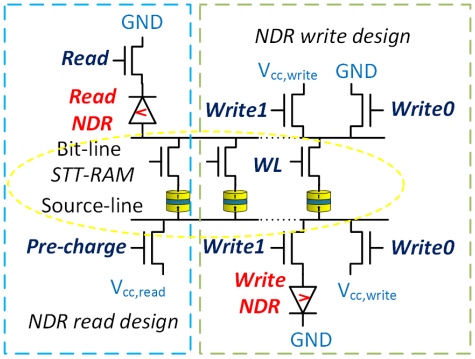



Fig. 5: The proposed NDR read and write circuitry designs. Yellow dotted line denotes the STT-RAM array.

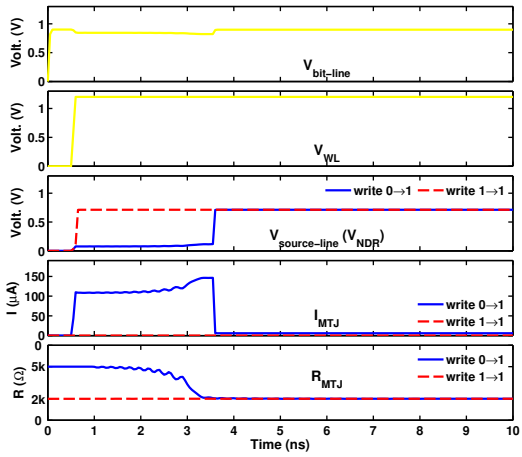



Fig. 6: SPICE simulated waveforms for a write-1 termination in the memory design of Fig. 5. During write operation, a bit-line is first selected and charged to $V_{CC}$, then a MTJ bit is selected by $WL$, and write current is high or low depending on the MTJ initial state.

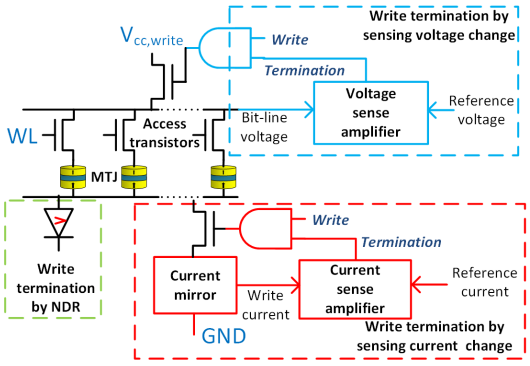
device is connected in the write path and acts to reduce the write energy by curtailing write current after switching.

The advantage of the proposed NDR-based design can be simply understood by comparing its energy dissipation $E_{NDR}$ with that of conventional MRAM designs ($E_{conv}$):

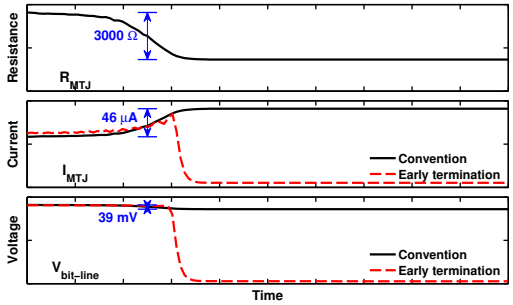
$$E_{conv} = V_{CC}^2 C_{BL} + V_{CC}(t_{AP} I_{AP} + t_P I_P) \quad (2)$$

$$E_{NDR} = (V_{CC} + V_{Peak})^2 C_{BL} + (V_{CC} + V_{Peak})(t_{AP} I_{Peak} + t_P I_{Valley}) \quad (3)$$

where $C_{BL}$ is the bitline capacitance, $I_{AP,P}$ are the currents when the MTJ is in the $AP$ or $P$ state, respectively, and $t_{AP,P}$ are the corresponding time intervals when the MTJ is in either state. In the proposed design, the applied voltage is increased slightly by the peak voltage of the NDR device (which is typically of order 0.1 V) and the write current in the $AP$ state approaches the NDR peak current $I_{Peak}$, but the write current in $P$ state is drastically reduced to the NDR valley current $I_{Valley}$. Comparing $E_{conv}$ and $E_{NDR}$, we see that most of the energy savings comes from the reduced consumption in the $P$ state and hinges on the ratio of $I_P$ and $I_{Valley}$, which is bounded by the PVR of the NDR device; higher PVR leads to more efficient operation. As the discussion in Section III-A shows, one prerequisite for current termination operation in our design is that $V_{NDR}$ at the beginning of the write process is lower than $V_{peak}$, so that current through NDR converges




(a)




(b)

Fig. 7: (a) Three early write termination designs using bit-line voltage change sensing [33], current change sensing [34], and the proposed NDR. (b) Simulated waveforms of MTJ resistance (AP: 5000 Ω, P: 2000 Ω, bit-line voltage, and write current as functions of time. The black line is for the conventional write, and the read dash line is for the early write terminations.

to the stable high current solution. This can be guaranteed by pre-discharging the source-line voltage to zero.

To quantify the effect of the proposed design, we simulate the waveforms of bit-line voltage ($V_{bit-line}$), bit selection (WL), voltage drop on NDR ($V_{source-line}$), write current ($I_{MTJ}$), and MTJ resistance ($R_{MTJ}$) under early write termination in Fig. 6. If the MTJ starts in the 0 state, the NDR device voltage and current increase to near $V_{peak}$ and $I_{peak}$, respectively, after the WL goes high and stays there until the MTJ switches, whereupon the NDR voltage approaches $V_{CC}$, turning off the write current. If the MTJ is initially in the 1 state, the NDR directly goes to $V_{CC}$ after the WL goes high and cuts off the write current.

Our proposed concept has significant advantages over previous attempts at write termination which used additional sense and control circuitry as illustrated in Fig. 7a. In [33] (see blue dashed box in Fig. 7a), sensing circuitry is added to sense the voltage change on the bit-line and terminate the write. However, such voltage changes are small (see Fig. 7b) in general because 1) the MTJ resistance change is small, and 2) the MTJ-bit selection transistor resistance changes inversely with that of the MTJ (*e.g.*, a MTJ resistance decrease leads to a voltage increase on the transistor and its equivalent transistor resistance), partially canceling bit-line voltage change. The resulting low sensing margin leads to long sensing times and large sensing energy and is susceptible to process variations. Another write termination design using current change sensing is proposed in [34] (red dashed box in Fig. 7a); this requires a current mirror in the write path to copy the current change to sensing circuit, which increases write $V_{CC}$ and adds redundant

write energy. Moreover, both voltage and current sensing designs require sense amplifiers and reference voltage/current generation (usually created by writing a reference MTJ in parallel), which add large energy overhead. One design uses different write pulses for the asymmetric write directions to save energy [35], though this method cannot safely write MTJs at the variation corners. Yet another proposal also introduces NDR into the MRAM write circuitry to avoid a current increase after MTJ switching to 1 [30]; however, as discussed in Section III-A, that method cannot fully terminate write current because both MTJ states will still have relatively large current, in contrast to our design where the 1 state current is truly minimized in the NDR valley region.

### C. STT-RAM read assistance using NDR

The sensing margin in STT-RAM is fundamentally bounded by the MTJ TMR ratio and is further reduced in practice by process variations; this has led to many proposals to maximize sensing margin within these limitations [36–40]. However, we propose a new read design that boosts the current difference ratio beyond the TMR ratio up to the PVR of the NDR device, substantially increasing sensing margins and reducing the sensitivity to process variations.

*1) Read margin improvement:* Similar to our write assistance design we propose a new read design, shown on the left of Fig. 5, in which multiple bit-lines share a single NDR device, minimizing overhead. During a read cycle, a normal pre-charge operation first charges the bit-line and source-line so that *Pre-charge* and *Read* are set to $V_{cc}$; the charge thus stored is then discharged through the MTJ and NDR series connection, where the latter amplifies the current difference between the 1 and 0 MTJ states. In this paper, read margin is defined as the voltage difference on the bit-line during discharging of the two MTJ states, which is sensed by a differential sense amplifier [41]. For conventional designs not using NDR, the source-line is used rather than the bit-line. NDR read requires the initial voltage of the NDR to be below $V_{peak}$ before the pre-charge, which is guaranteed by discharging any remaining charge on the bit-line after each read or write operation. As in the write case, the NDR device should be selected such that $I_{peak}$ is larger than the MTJ current in the 0 state but smaller than that of the 1 state. Therefore, as shown in Fig. 8, when the 0 state is sensed, the NDR device always stays in its low resistance region below the peak and the bit-line cannot be charged up. When the 1 state is sensed, the transient current through the MTJ in the pre-charge stage will exceed $I_{peak}$, pushing the NDR device into its high-resistance valley region so that the discharging (sensing) current is cut off.

$$V_{M,conv} \approx \left(1 - \frac{I_{AP}}{I_P}\right) V_{BL} \tag{4}$$

$$V_{M,NDR} \approx \left(1 - \frac{I_{Valley}}{I_{Peak}}\right) V_{SL} = \left(1 - \frac{1}{PVR}\right) V_{SL} \tag{5}$$

where $V_{SL}$ is the source-line voltage. Whereas the read margin in ordinary designs is limited by the discharging current ratio of the 1 and 0 states (typically around 0.7-0.9), in the NDR design $V_{M,NDR}$ is limited by $1 - 1/PVR$ which approaches 1 for well-designed NDR devices, dramatically increasing the read margin.

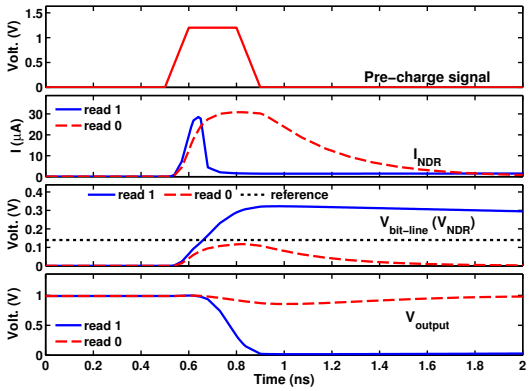
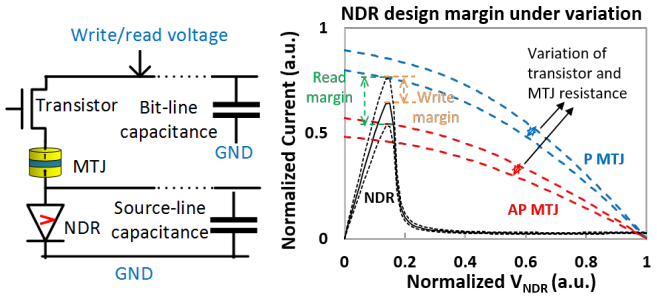



Fig. 8: SPICE simulation of read operations using NDR-assisted design in Fig. 5. The discharging current ($I_{NDR}$) difference for sensing MTJ states is significantly increased by the high PVR of NDR. A large and constant voltage margin is achieved on the bit-line ($V_{bit-line}$), which is sensed by a constant reference voltage leading to a stable sense amplifier output ($V_{output}$).

Our design has several key advantages over previous proposals. For instance, [36–38] proposed local-reference and self-reference read designs to improve immunity to process variations, but could not improve the margin beyond the MTJ TMR limits. Others have proposed amplifying the read margin using NDR [30–32], but because of the limited read current difference in those designs (see Fig. 3b), but the achievable margins are smaller and read disturbance is not minimized in contrast to our proposal. Refs. [42, 43] propose introducing a TD into each MTJ cell to amplify the read margin, but such a requirement is not applicable to current-switched technologies like STT-RAM and greatly increases area overhead.

*2) Read disturbance minimization:* Read disturbance, another primary reliability concern, is caused by false switching of a MTJ via sensing currents through the cell. Any transient current pulse has some possibility of switching a MTJ, but the probability of such switching increases dramatically with the amplitude and duration of the current pulse. To tame this, conventional designs have to reduce read current as well as sensing margin and bit-line size, resulting in larger sensing circuits (large gate sizes and more sensing transistors), longer sensing time, and higher sensing energy. In the proposed NDR-assisted design, the PVR simultaneously reduces read disturbance and improves read margin while introducing little overhead. As Fig. 5 shows, an *AP* state MTJ cannot be falsely switched due to the sensing current direction and read disturbance can occur only during read-1 operation when the sensing current accidentally switches a STT-MTJ from the 1 to 0 state. However, the probability of such disturbances in our design is almost zero since the switching probability scales exponentially with current and duration. We can dramatically lower the sensing current through a 1-state MTJ to the near-0 NDR valley current, which is significantly lower than previous works [30–32]. The read disturbance rate is simulated for different bit-line size in Section IV.




(a) (b)

Fig. 9: (a) Simulated circuit implementation in CUDA. (b) NDR peak current margin variation analysis.

## IV. Evaluation of NDR-assisted designs with process variations

We next study the robustness of our proposed design under process variations using comprehensive Monte Carlo circuit simulations. To study variability effects accurately, massive simulations (over 10 billion runs) are needed to detect WER and read disturbance rates down to $10^{-9}$, which are impractical for conventional SPICE methods (indeed, we estimate such a calculation would take decades); instead, we implement our device physical model and a small circuit simulator using CUDA and run them on a Tesla M207 GPU, enabling billions of simulations within a few hours. The simulated circuit is shown in Fig. 9a. Process variations for transistors and MTJs are included in the manner of [15].

We first assess variation-imposed limits on the NDR design margin, as illustrated in Fig. 9b. Here the blue and red dashed lines show the corner cases for the MTJ high and low resistance states versus $V_{NDR}$. For read applications, the NDR $I_{peak}$ must lie between the high and low currents (the top and bottom black dotted lines). When writing 0-to-1, $I_{peak}$ must lie within the top half of the current margin (between the top and middle dotted lines) to guarantee that NDR cuts off once the MTJ reaches an intermediate resistance state (between $R_H$ and $R_L$) where it can converge to the 1 state without further assistance from a large switching current. Fluctuations in devices due to process variations like random dopant fluctuations or line edge roughness [44–46] can change the threshold voltage of TFETs and affect $I_{Peak}$. We analyze the design tolerance for such process variations through Monte Carlo simulation.

In total we perform over 100 billion Monte-Carlo simulations on the NDR-assisted write process for 0 to 1 switching. In addition to ordinary write errors caused by thermal fluctuations, we find a special error may also occur in the NDR-assisted write process for individual MTJs or transistors with very low resistance due to process variations (*e.g.*, when the intersection of the initial high resistance state approaches the NDR peak current too closely in Fig. 9b). In such cases, the NDR may turn off the write current when fluctuations drive the MTJ current curve past $I_{peak}$ but before the MTJ can switch to its low resistance state. To avoid such write errors, a higher NDR peak current is required, increasing write energy. The write energy and WER as functions of NDR peak current are shown in Fig. 10a. As peak current rises, the WER decreases but write energy increases. Fortunately, for the standard ECC requirement of WER $< 10^{-9}$, significant energy reductions ($> 50\%$ lower than conventional designs) are realized over a wide range of nominal $I_{peak}$ for both TFET ($161 - 146$ =15 µA or 20 mV threshold voltage shift) and TD ($153 - 126$ =27 µA) designs; this is the effective design tolerance of NDR devices for write circuits. We summarize the write performance results in Table II where, for fairness, we compare NDR-based and conventional circuits with the same write latency and WER. Comparing the energy usage of write-1 operations (0-to-1 and 1-to-1), we see 76% and 52% energy savings for TFET- and TD-based designs, respectively. Looking at the average write energy (assuming equal usage of the four switching directions), we still see major reductions of 52% and 36% for TFET- and TD-assisted writes, respectively.

TABLE II: Write energy, read margin, and read energy of NDR-assisted designs as extracted from Fig. 10. $C_{BL} = 25$ fF; nominal TFET $V_th = 0.25$ V. Since NDR does not affect write-0 operations (*0→0* and *1→0*), conventional designs are used for these cases. Effective PVR is the ratio of circuit current in the 1 and 0 states for chosen $V_{CC}$ and differs for write and read due to different bias.

| | | Conventional | TFET | TD |
|---|---|---|---|---|
| Write Energy (fJ) | ***0→1*** | 1040 | 248 | 498 |
| | ***0→0*** | 699 | 699 | 699 |
| | ***1→1*** | 1269 | 61 | 419 |
| | ***1→0*** | 838 | 838 | 838 |
| | Average | 964 | 462 | 613 |
| Write latency (ns) | | 9 | 9 | 9 |
| Read voltage (mV) | | 700 | 210 | 250 |
| Read margin (mV) | | 139 | 164 | 174 |
| Read energy | | 42 | 5.5 | 3.9 |
| NDR design variation tolerance | Peak current shift ($\mu A$), Fig. 10 | Write | 15 | 27 |
| | | Read | 8 | 11 |
| | Threshold voltage shift (mV) | Write | 20 | N/A |
| | | Read | 105 | N/A |
| Effective PVR | | Write | 23.6 | 2.64 |
| | | Read | 8.38 | 5.86 |

The dependence of the read margin (current margin on the source-line) on $I_{peak}$ and read voltage is shown in Fig. 10b. Again we observe good design tolerance for NDR device variations from the range of $I_{peak}$ for which read margin is large and nearly constant. At low $V_{read}$ (0.21 V for TFET and 0.25 V for TD), a read margin of over 150 mV can be maintained for the TFET design over an 8 µA variation range in peak current (equivalent to 105 mV threshold voltage shift), and for the TD design over an 11 µA variation in $I_{peak}$. The trend continues at higher read voltages as well. The read performance is summarized and compared with a representative conventional read design in Table II; we note that the latter requires a much larger $V_{read} = 0.7$ V to achieve comparable read margin, consuming much more energy.

Finally, we examine read disturbance rates under NDR and conventional designs for different source-line and bit-line loads. Longer bit-lines have larger loads, leading to more charging/discharging current during read. In the read operation, current flows from source-line to bit-line, which may falsely switch the MTJ from 1 to 0. Fig. 10c shows simulated read disturbance rates as functions of bit-line/source-line size (load). Compared with conventional designs with similar read margin, NDR-assisted designs enable vastly improved disturbance rates (over 10 million times lower). Moreover, the read disturbance rates of conventional designs cannot satisfy the ECC requirement ($< 10^{-9}$) for bit-line loads larger than 100 fF. The dramatic read disturbance reductions we observe

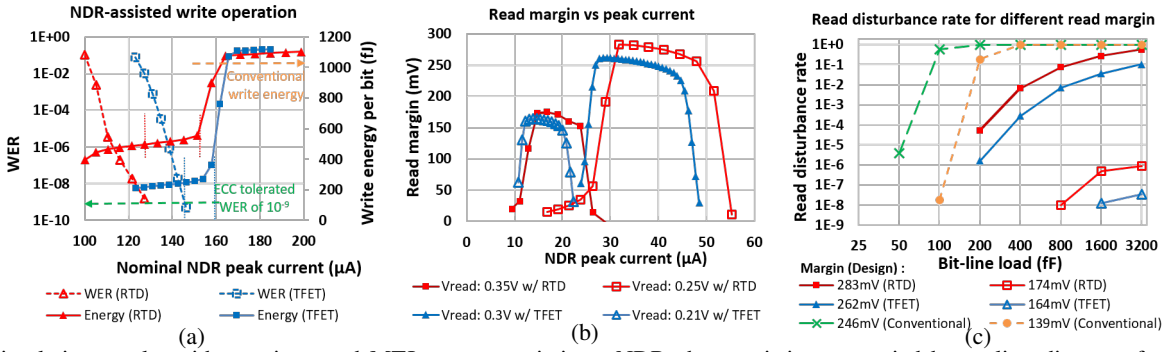



Fig. 10: Simulation results with transistor and MTJ process variations. NDR characteristics are varied by scaling diameter for TDs (0.4-0.55 μm) and threshold voltage for TFETs (assuming device width=1.95 μm in write and 0.195 μm in read circuitry).(a) WER and write energy versus nominal NDR peak current. Write energy includes bit-line pre-charge, access transistor, and MTJ, but excludes row/column decoders. (b) Read margin versus NDR nominal peak current. $C_{BL} = C_{SL}$ = 25 fF in (a) and (b). (c) Read disturbance rate as function of bit-line/source-line load and read margin. High/low margin designs are obtained using different $V_{read}$ (0.35 V/0.25 V for TD, 0.3 V/0.21 V for TFET, and 1.8 V/0.7 V for conventional designs). Read disturbance rates below $10^{-10}$ not detectable within sample size.

for the NDR-assisted design are due to its low $V_{read}$ and minimized discharge current (limited by $I_{valley}$).

In the proposed applications, we can use a large NDR device (e.g., gate width over 1 $\mu m$ for NDR write) to provide sufficient peak current and minimize process variation. This does not limit memory density since every NDR device can be shared by multiple bit-lines containing thousands of cells. Using results from a device variation analysis on 14nm TFETs [44], we estimate using the variation scaling rule [47] that the $6\sigma$ of threshold voltage shift is about 9 $mV$ if a $200nm$ x $1000nm$ TFET-NDR, well within the simulated tolerance level of our design.

The simulation parameters for NDR assisted write and read is shown in Table III. The PVR is the effective PVR determined by both NDR device and voltage bias.

TABLE III: Simulation parameters at 300K.

| **STT-RAM** | Diameter | $t_{MgO}$ | $t_{tfl}$ | $R_P$ | $R_{AP}$ | BL |
|---|---|---|---|---|---|---|
| | 50nm | 1.18nm | 1.1nm | 2k Ω | 5k Ω | 256 bits |
| **NDR** | Write | | High-margin read | | Low-margin read | |
| | Width | PVR | Width | PVR | Width | PVR |
| TFET | 1.95 $\mu m$ | 23 | 0.39 $\mu m$ | 17 | 0.2 $\mu m$ | 8 |
| RTD | 0.44 $\mu m$ | 2.7 | 0.23 $\mu m$ | 10.5 | 0.16 $\mu m$ | 5.5 |

## V. Conclusion

In this paper, we propose a novel STT-RAM write and read design with the assistance of NDR devices such as TFETs or TDs. In a write-to-1 operation, NDR can detect the MTJ state and cut off current flow after switching to avoid wasting energy. In the read operation, NDR can amplify the sensing current and voltage margin by detecting current through 1 and 0 state MTJs, reducing read voltage and read energy. Additionally, read disturbance current can also be reduced by NDR devices to low valley current values, preventing false switching. We show these advantages are robust against transistor and MTJ process variations within a reasonable NDR design margin. Our simulations show write energy reductions of 36% and 52% using prototypical TDs and TFETs, while read disturbance rate reduction over $10^7$X is shown for both TDs and TFETs with similar read margin.

## References


[1] Said Tehrani, JM Slaughter, E Chen, M Durlam, J Shi, and M DeHerren. "Progress and outlook for MRAM technology". *IEEE Transactions on Magnetics*, vol. 35, no. 5, pp. 2814–2819, 1999.

[2] C Heide. "Spin currents in magnetic films". *Phys. Rev. Lett.* Vol. 87, no. 19, p. 197201, 2001.

[3] Wang Kang, Liuyang Zhang, Jacques-Olivier Klein, Youguang Zhang, Dafiné Ravelosona, and Weisheng Zhao. "Reconfigurable codesign of STT-MRAM under process variations in deeply scaled technology". *IEEE Transactions on Electron Devices*, vol. 62, no. 6, pp. 1769–1777, 2015.

[4] B Fang, X Zhang, BS Zhang, ZM Zeng, and JW Cai. "Tunnel magnetoresistance in thermally robust Mo/CoFeB/MgO tunnel junction with perpendicular magnetic anisotropy". *AIP Advances*, vol. 5, no. 6, p. 067116, 2015.

[5] *ITRS*. http://www.itrs.net/about.html. 2008,2011.

[6] Zhenyu Sun, Xiuyuan Bi, Hai Helen Li, Weng-Fai Wong, Zhong-Liang Ong, Xiaochun Zhu, and Wenqing Wu. "Multi retention level STT-RAM cache designs with a dynamic refresh scheme". *Proc. MICRO*. ACM. 2011, pp. 329–338.

[7] Adwait Jog, Asit K Mishra, Cong Xu, Yuan Xie, Vijaykrishnan Narayanan, Ravishankar Iyer, and Chita R Das. "Cache revive: architecting volatile STT-RAM caches for enhanced performance in CMPs". *Proc. DAC*. ACM. 2012, pp. 243–252.

[8] Emre Kultursay, Mahmut Kandemir, Anand Sivasubramaniam, and Onur Mutlu. "Evaluating STT-RAM as an energy-efficient main memory alternative". *ISPASS*. IEEE. 2013, pp. 256–267.

[9] Steven Pelley, Peter M Chen, and Thomas F Wenisch. "Memory persistency". *Proc. ISCA*. IEEE. 2014, pp. 265–276.

[10] Jing Li, Charles Augustine, Sayeef Salahuddin, and Kaushik Roy. "Modeling of failure probability and statistical design of spin-torque transfer magnetic random access memory (STT MRAM) array for yield enhancement". *Proc. Design Automation Conference*. IEEE. 2008, pp. 278–283.

[11] Wang Kang, Liuyang Zhang, Weisheng Zhao, Jacques-Olivier Klein, Youguang Zhang, Dafiné Ravelosona, and Claude Chappert. "Yield and reliability improvement techniques for emerging nonvolatile STT-MRAM". *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 5, no. 1, pp. 28–39, 2015.

[12] Wang Kang, Zheng Li, Jacques-Olivier Klein, Yuanqing Chen, Youguang Zhang, Dafiné Ravelosona, Claude Chappert, and Weisheng Zhao. "Variation-tolerant and disturbance-free sensing circuit for deep nanometer STT-MRAM". *IEEE Transactions on Nanotechnology*, vol. 13, no. 6, pp. 1088–1092, 2014.

[13] Ki Chul Chun, Hui Zhao, Jonathan D Harms, Tae-Hyoung Kim, Jian-Ping Wang, and Chul Han Kim. "A scaling roadmap and performance evaluation of in-plane and perpendicular MTJ based STT-MRAMs for high-density cache memory". *IEEE Journal of Solid-State Circuits*, vol. 48, no. 2, pp. 598–610, 2013.

[14] Jong-Yoon Park, Se-Koo Kang, Min-Hwan Jeon, Myung S Jhon, and Geun-Young Yeom. "Etching of CoFeB Using CO/ NH3 in an Inductively Coupled Plasma Etching System". *Journal of The Electrochemical Society*, vol. 158, no. 1, H1–H4, 2011.

[15] Shaodi Wang, Hochul Lee, Farbod Ebrahimi, P. Khalili Amiri, Kang L. Wang, and Puneet Gupta. "Comparative Evaluation of Spin-Transfer-Torque and Magnetoelectric Random Access Memory". *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 6, no. 2, pp. 134–145, 2016.

[16] E Chen, D Apalkov, Z Diao, A Driskill-Smith, D Druist, D Lottis, V Nikitin, X Tang, S Watts, S Wang, et al. "Advances and future prospects of spin-transfer torque random access memory". *IEEE Transactions on Magnetics*, vol. 46, no. 6, pp. 1873–1878, 2010.

[17] Rui Li, Yeqing Lu, Guangle Zhou, Qingmin Liu, Soo Doo Chae, T. Vasen, Wan Sik Hwang, Qin Zhang, P. Fay, T. Kosel, M. Wistey, Huili Xing, and A. Seabaugh. "AlGaSb/InAs Tunnel Field-Effect Transistor With On-Current of 78 $\mu A/\mu m$ at 0.5 V". *Electron Device Letters, IEEE*, vol. 33, no. 3, pp. 363–365, Mar. 2012.

[18] K Eberl, R Duschl, OG Schmidt, U Denker, and R Haug. "Si-based resonant inter- and intraband tunneling diodes". *Journal of crystal Growth*, vol. 227, pp. 770–776, 2001.

[19] M Hosomi, H Yamagishi, T Yamamoto, K Bessho, Y Higo, K Yamane, H Yamada, M Shoji, H Hachino, C Fukumoto, et al. "A novel nonvolatile memory with

spin torque transfer magnetization switching: Spin-RAM". *Proc. International Electron Devices Meeting*. IEEE. 2005, pp. 459–462.
[20] Dmytro Apalkov, Alexey Khvalkovskiy, Steven Watts, Vladimir Nikitin, Xueti Tang, Daniel Lottis, Kiseok Moon, Xiao Luo, Eugene Chen, Adrian Ong, et al. "Spin-transfer torque magnetic random access memory (STT-MRAM)". *ACM Journal on Emerging Technologies in Computing Systems*, vol. 9, no. 2, p. 13, 2013.
[21] S. Anthony. *Beyond Silicon: IBM Unveils World's First 7 nm chip—With a Silicon-Germanium Channel and EUV Lithography, IBM Crosses the 10 nm Barrier. Ars Technica*. http://arstechnica.com/gadgets/2015/07/ibm-unveils-industrys-first-7nm-chip-moving-beyond-silicon. Accessed: 2016-10-04.
[22] Xin-Yu Bao, Cesare Soci, Darija Susac, Jon Bratvold, David PR Aplin, Wei Wei, Ching-Yang Chen, Shadi A Dayeh, Karen L Kavanagh, and Deli Wang. "Heteroepitaxial growth of vertical GaAs nanowires on Si (111) substrates by metal- organic chemical vapor deposition". *Nano letters*, vol. 8, no. 11, pp. 3755–3760, 2008.
[23] Greg Leung, Shaodi Wang, Andrew Pan, Puneet Gupta, and Chi On Chui. "Evaluation Framework for Nanotransfer Printing-Based Feature-Level Heterogeneous Integration in VLSI Circuits". *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 24, no. 5, pp. 1858–1870, 2015.
[24] SL Rommel, D Pawlik, P Thomas, M Barth, K Johnson, SK Kurinec, A Seabaugh, Z Cheng, JZ Li, J-S Park, et al. "Record PVCR GaAs-based tunnel diodes fabricated on Si substrates using aspect ratio trapping". *Proc. International Electron Devices Meeting*. IEEE. 2008, pp. 1–4.
[25] H.H. Tsai, Y.K. Su, H.H. Lin, R.-L. Wang, and T.L. Lee. "P-N double quantum well resonant interband tunneling diode with peak-to-valley current ratio of 144 at room temperature". *Electron Device Letters, IEEE*, vol. 15, no. 9, pp. 357–359, Sept. 1994.
[26] Yuping Zeng, Chien-I Kuo, Rehan Kapadia, Ching-Yi Hsu, Ali Javey, and Chenming Hu. "Two-dimensional to three-dimensional tunneling in InAs/AlSb/GaSb quantum well heterojunctions". *Journal of Applied Physics*, vol. 114, no. 2, 024502, 2013.
[27] Xin Zhao, A. Vardi, and J.A. Del Alamo. "InGaAs/InAs heterojunction vertical nanowire tunnel fets fabricated by a top-down approach". *Proc. International Electron Devices Meeting*. IEEE. Dec. 2014, pp. 25.5.1–25.5.4.
[28] J.N. Schulman, H.J. De Los Santos, and D.H. Chow. "Physics-based RTD current-voltage equation". *Electron Device Letters, IEEE*, vol. 17, no. 5, pp. 220–222, May 1996.
[29] A. Pan and Chi On Chui. "A Quasi-Analytical Model for Double-Gate Tunneling Field-Effect Transistors". *Electron Device Letters, IEEE*, vol. 33, no. 10, pp. 1468–1470, Oct. 2012.
[30] David Halupka, Safeen Huda, Wanjuan Song, Ali Sheikholeslami, Koji Tsunoda, Chikako Yoshida, and Masaki Aoki. "Negative-resistance read and write schemes for STT-MRAM in 0.13$\mu$m CMOS". *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2010 IEEE International*. IEEE. 2010, pp. 256–257.
[31] Yohei Umeki, Koji Yanagida, Shusuke Yoshimoto, Shintaro Izumi, Masahiko Yoshimoto, Hiroshi Kawaguchi, Koji Tsunoda, and Toshihiro Sugii. "STT-MRAM Operating at 0.38 V Using Negative-Resistance Sense Amplifier". *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 97, no. 12, pp. 2411–2417, 2014.
[32] Yohei Umeki, Koji Yanagida, Shusuke Yoshimoto, Shintaro Izumi, Masahiko Yoshimoto, Hiroshi Kawaguchi, Koji Tsunoda, and Toshihiro Sugii. "A negative-resistance sense amplifier for low-voltage operating STT-MRAM". *Proc. ASP-DAC*. IEEE. 2015, pp. 8–9.
[33] Ping Zhou, Bo Zhao, Jun Yang, and Youtao Zhang. "Energy reduction for STT-RAM using early write termination". *Proc. ICCAD*. IEEE. 2009, pp. 264–268.
[34] R. Bishnoi, F. Oboril, M. Ebrahimi, and M. B. Tahoori. "Self-Timed Read and Write Operations in STT-MRAM". *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 24, no. 5, pp. 1783–1793, May 2016.
[35] Rajendra Bishnoi, Mojtaba Ebrahimi, Fabian Oboril, and Mehdi B Tahoori. "Asynchronous asymmetrical write termination (AAWT) for a low power STT-MRAM". *Proceedings of the conference on Design, Automation & Test in Europe*. European Design and Automation Association. 2014, p. 180.
[36] UK Klostermann, M Angerbauer, U Griming, F Kreupl, M Ruhrig, F Dahmani, M Kund, and G Muller. "A perpendicular spin torque switching based MRAM for the 28 nm technology node". *Proc. International Electron Devices Meeting*. IEEE. 2007, pp. 187–190.
[37] Yiran Chen, Xiaobin Wang, Wenzhong Zhu, Wei Xu, Tong Zhang, et al. "A nondestructive self-reference scheme for spin-transfer torque random access memory (STT-RAM)". *Proc. DATE*. IEEE. 2010, pp. 148–153.
[38] Enes Eken, Yaojun Zhang, Wujie Wen, Rajiv Joshi, Hai Li, and Yiran Chen. "A New Field-assisted Access Scheme of STT-RAM with Self-reference Capability". *Proc. Design Automation Conference*. ACM. 2014, pp. 1–6.
[39] Seyedhamidreza Motaman, Swaroop Ghosh, and Jaydeep P Kulkarni. "A novel slope detection technique for robust STTRAM sensing". *Proc. ISLPED*. IEEE. 2015, pp. 7–12.
[40] H. Lee, C. Grèzes, S. Wang, F. Ebrahimi, P. Gupta, P. K. Amiri, and K. L. Wang. "Source Line Sensing in Magneto-Electric Random-Access Memory to Reduce Read Disturbance and Improve Sensing Margin". *IEEE Magnetics Letters*, vol. 7, pp. 1–5, 2016.
[41] Robert J Proebsting. *Differential sense amplifier circuit*. US Patent 6,154,064. Nov. 2000.
[42] Tetsuya Uemura, Satoshi Honma, Takao Marukame, and Masafumi Yamamoto. "Large enhancement of tunneling magnetoresistance ratio in magnetic tunnel junction connected in series with tunnel diode". *Japanese journal of applied physics*, vol. 43, no. 1A, p. L44, 2003.
[43] Tetsuya Uemura and Masafumi Yamamoto. "Proposal of four-valued MRAM based on MTJ/RTD structure". *Proc. International Symposium on Multiple-Valued Logic*. IEEE. 2003, pp. 273–278.
[44] Nattapol Damrongplasit, Sung Hwan Kim, and Tsu-Jae King Liu. "Study of random dopant fluctuation induced variability in the raised-ge-source TFET". *IEEE Electron Device Letters*, vol. 34, no. 2, pp. 184–186, 2013.
[45] Greg Leung and Chi On Chui. "Stochastic variability in silicon double-gate lateral tunnel field-effect transistors". *IEEE Transactions on Electron Devices*, vol. 60, no. 1, pp. 84–91, 2013.
[46] Shaodi Wang, Greg Leung, Andrew Pan, Chi On Chui, and Puneet Gupta. "Evaluation of digital circuit-level variability in inversion-mode and junctionless FinFET technologies". *IEEE Transactions on Electron Devices*, vol. 60, no. 7, pp. 2186–2193, 2013.
[47] David Burnett and Shih-Wei Sun. "Statistical threshold-voltage variation and its impact on supply-voltage scaling". *Microelectronic Manufacturing'95*. International Society for Optics and Photonics. 1995, pp. 83–90.

**Shaodi Wang** Shaodi (S'12) is currently fifth-year Ph.D. student in the NanoCAD lab at Department of Electrical Engineering, UCLA advised by Prof. Puneet Gupta. His research interests include emerging memory and device technology circuit- and system-level design, evaluation and optimization, and modeling for manufacturing.

Shaodi received his M.S degree in electrical engineering from UCLA, and his B.S. degree from Peking University Electronics Engineering and Computer Science department, China in the division of Microelectronic.

**Andrew Pan** Andrew Pan (S'12-M'15) is currently in a postdoctoral researcher at the Department of Electrical Engineering, University of California, Los Angeles. His research interests include electronic device modeling, transport phenomena, and solid state physics.

**Chi On Chui** Chi On Chui (S'00–M'04–SM'08) received his B.S. degree in physics and his M.S. and Ph.D. degrees in electrical engineering from the University of California, Los Angeles, where he is currently a lecturer and project scientist. His research interests include semiconductor device modeling, transport phenomena, and solid state physics.

**Puneet Gupta** Puneet Gupta (M'07-SM'16) (http://nanocad.ee.ucla.edu) is currently a faculty member of the Electri-cal Engineering Department at UCLA. He received the B.Tech degree in Electrical Engineering from Indian Insti-tute of Technology, Delhi in 2000 and Ph.D. in 2007 from University of California, San Diego. He co-founded Blaze DFM Inc. (acquired by Tela Inc.) in 2004 and served as its product architect till 2007.

He has authored over 150 papers, 17 U.S. patents, a book and a book chapter. He is a recipient of NSF CAREER award, ACM/SIGDA Outstanding New Faculty Award, SRC Inventor Recognition Award and IBM Faculty Award. He currently leads the IMPACT+ center (http://impact.ee.ucla.edu) which focuses on future semi-conductor technologies. Dr. Gupta's research has focused on building high-value bridges across application-architecture-implementation-fabrication interfaces for lowered cost and power, increased yield and improved predictability of integrated circuits and systems.