Aggressive Leakage Current Reduction for Embedded MRAM Using Block-Level Power Gating

[1]Anh Tuan Do [2]Xuanyao Fong and [1]Fei Li
[1]Institute of Microelectronics, A*STAR, Singapore
[2]School of Electrical and Computer Engineering, National University of Singapore, Singapore

***Abstract*—This paper exploits circuit techniques to realize a power-gated MRAM with the instant-on characteristic. Each block in a large-capacity MRAM is gated by a separate power transistor, allowing it to be fully turned on after only 1.3 ns. As a result, the whole MRAM is always in sleep mode, except the selected block. This eliminates the need for access pattern prediction as well as the requirement that the MRAM must be in idle for a significant of time before it is put in sleep or deep sleep mode. Our simulation shows that even in the worst-case-scenario, 86% of leakage current is saved. In typical cases, there are 96.5% leakage and 69% total power reduction. Our proposed scheme's implementation is straight forward and incurs less than 0.5% area overhead, including power transistors and control circuits.**

***Index Terms*—MRAM, low-power, low-leakage, normally-off system**

I. INTRODUCTION

Spin-transfer torque magnetoresistive random access memory (STT-MRAM) is a promising memory technology that can satisfy many of strict requirements of low-power memories for future applications [1]. In STT-MRAM, data is represented as "0" and "1", depending on the magnetic state of the magnetic tunnel junction (MTJ). STT-MRAM gained the most attention from both research and the industries due to its excellent scalabilities and low-voltage operation [2], [3-6].

Most works to date have focused on reducing the large write energy in STT-MRAM. Nevertheless, leakage power in STT-MRAM can be the dominant component of power dissipation in large capacity STT-MRAM arrays. Thus, several works have proposed power-gating schemes to minimize its leakage power [3, 6]. The STT-MRAM usually consists of several sub-arrays, can be gated at different levels of design hierarchy. For example, an SRAM-like MRAM cell with word-level power-gating was proposed in [7]. However, implementing word-level power-gating incurs significant area overhead due to the complexity of the control circuitry. Alternatively, a sub-array can be turned off if its next access is predicted to be a long time in the future [8]. An analysis of the access pattern to the last-level-cache (LLC) shows that non-performance-critical sub-arrays can be turned off immediately after access without significantly impacting the overall system performance [8]. Moreover, an access history table may be used to eliminate any performance degradation by predicting the future access to a sub-array and putting it into sleep or deep-sleep mode accordingly [3, 6]. A common theme among these designs is that a particular sub-array is only turned off completely or partially if its next access is predicted to be significantly far into the future. Thus, an access predictor [8] or history table [3, 6] is required, which incurs area overhead. Furthermore, the accuracy of access prediction depends heavily on benchmarks and algorithms, resulting in inefficient leakage power suppression. On top of the performance degradation mentioned earlier, the sub-arrays are not aggressively power-gated because power dissipated to switch the power-gating device on/off may outweigh the leakage power saved.

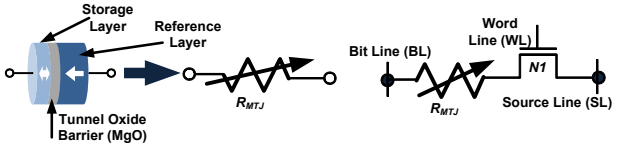
We propose an aggressive power gating scheme without any memory access prediction. In our design, only the sub-array being written/read is turned on during memory access while all other sub-arrays remain in the off state to save leakage power. Systemmatic simulation and transistor-level optimization will also be presented to minimize power and performance overhead of the design.

II. MTJ AND MRAM FUNDAMENTALS

This section reviews the fundamentals of STT MRAM [2, 9]. We will also discuss basic MRAM cell topologies and their read/write operation.

A. MTJ

The magnetic tunnel junction (MTJ – Fig. 1a) is a two terminal device which consists of three layers: two separate ferromagnetic layers (called the reference and the storage layers, respectively) that sandwich a tunnel oxide barrier. The magnetization of the reference layer is pinned such that it remains fixed during operation and can be used as a reference. In contrast, the magnetization of the storage layer is engineered such that it may be manipulated to point parallel or anti-parallel (P or AP state, respectively) to the magnetization of the reference layer.



(a) *(b)*

Fig. 1 (a) Simplified MTJ structure (b) 1T1MTJ MRAM cell topology

Due to the tunneling magnetoresistance effect, the magnetic

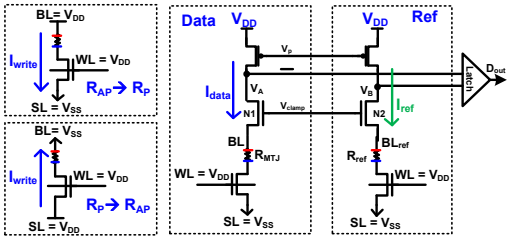
The next few paragraphs should contain the authors' current affiliations, including current address and e-mail.

configuration of the MTJ modulates the resistance (R_{MTJ}) between its terminals [10]. The ferromagnetic layers in the MTJ behave as spin-filters that polarize the spin of electrons constituting the tunneling current across the oxide barrier. As a result, R_{MTJ} is low when the magnetic moments of the storage and reference layers are in parallel (i.e. P state, $R_{MTJ} = R_P$). On the other hand, R_{MTJ} is high when the MTJ is in the anti-parallel state (i.e. AP state, $R_{MTJ} = R_{AP} > R_P$). The MTJ also exhibits the *spin-transfer torque* (STT) effect when a current is passed through it [11]. Thus, the cell can be programmed by using a sufficiently large current which generates a torque that overcomes the anisotropy energies in the storage layer to switch the MTJ between P and AP states.

B. MRAM-based embedded memory

Embedded MRAMs have similar interface and array organization as in SRAM [12]. It is usually partitioned into multiple sub-arrays. Each sub-array is a self-contained architecture including a local row decoder, memory cel array, and read/write circuitries. Numerous MRAM cell topologies have been proposed to address different optimization constraints [13-17] but the most popular designs are the 1T1M [16, 18] and 2T2M [3] due to their regular structure, compactness, and simple control circuitry.

The 1T1M bit-cell is formed by connecting an access transistor (N_1) to the MTJ as shown in Fig. 1(b). An NMOS is usually used because it has better driving capability as compared to the PMOS. The source terminal of N_1 is connected to the source line (SL), while the MTJ is connected to the bit line (BL). The gate of N_1 is controlled by a word line (WL). Every row of cells in the memory array has its own WL, which is connected to the memory cells in that row. Every column of cells in the memory array is also connected to its own SL and BL.



(a) (b)

Fig. 2 MRAM cell configuration during (a) write operation (b) read operation

The access transistor N_1 is normally OFF. The WL turns it ON when the memory bit-cell is accessed for read or write operation. The MTJ state is programmed during write operations by applying appropriate voltages to BL and SL and passing a current through the MTJ. To write a "0", current flows from BL to SL (i.e. R_{MTJ} switches from R_{AP} to R_P). Thus, BL and SL are set to V_{DD} and V_{SS} (Fig. 2a), respectively. Conversely, BL and SL are set to V_{SS} and V_{DD}, respectively, to write a "1" (i.e. R_{MTJ} switches from R_P to R_{AP}).

During the read operations, the MTJ state is determined by sensing the resistance of the bit-cell between BL and SL. A reference generation circuit is employed to assist the sense amplifier in differentiating between R_P and R_{AP} states of the MTJ, as shown in Fig. 2(b). However, due to PVT variations, a good reference and reliable sensing is difficult to achieve for a large memory array in nano-scale technologies. To overcome this, 2T2M was proposed [3]. The 2T2M cell consists of a pair of 1T1M cells to store both data bit and the complementary bit. Hence, differential read can beused which improves the sensing margin. The speed and reliability of read operations of the 2T2M cell are also more robust, albeit with additional area overhead. Writing to 2T2M is quite similar to that of 1T1M but an additional pair of BLB and SLB is required. In this work, we use 2T2M cell to realize fast sensing speed and eliminate the need for reference circuit.

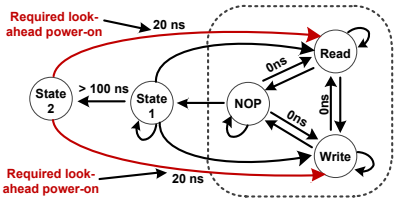


Fig. 3 MRAM power gating state machine using access predictor [12]

C. Power gating with access predictor

Power-gating is a popular approach to eliminate leakage power in MRAM. Due to its non-volatility, power-gating in MRAM is more straight forward when compared SRAM. In [3, 12], local power switch was also proposed, in combination with global power-gating.

Fig. 3 illustrates the state diagram used in in one of such schemes [3]. The controller keeps track of the current state of the memory. During active, it switches between Read, Write and Not-In-Operation (i.e. NOP). If it stays in NOP for a few cycles, the state jumps to "State 1" where the memory is put in sleep Mode. In this mode, leakage is reduced but not aggressively so that the memory can be woken up quickly. If it stays in "State 1" for a significant amount of time (e.g. 100 ns), the state will jump to "State 2" (i.e. Deep Sleep Mode). This deep sleep mode allows aggressive leakage reduction but requires look-ahead power-on whose delay is 20 ns. As we can see, this scheme requires both look-ahead power on and pattern recognition to decide when is the optimum time to jump to "State 1" and "State 2"

In this implementation, most peripheral circuits (such as sense amplifier, read/write drivers) are put in the global domain. Fast wake-up is only realized in the local array with the minimal peripheral. The mid-level and global circuits are put in sleep mode (i.e. state 1 and state 2) only when significantly long sleep duration is predicted. This limits the effectiveness of the power gating scheme. Furthermore, once it is in state 2, the memory takes a very long time to wake up. A more agresstive power gating scheme can be used to further reduce leakage current.

III. Instant-on Power Gating Technique

For the sake of clarity, our analysis assumes a 4Mb MRAM consists of 128 blocks, each contains 32 kb of MRAM cells. The 32 kb-MRAM block is self-contained design unit with row decoder, local read/write circuit and a 256-row x 128-column cell array as shown in Fig. 4. All power and speed performances are evaluated in 28nm CMOS process, 1 V supply, 200 MHz.

A. System overview and operating principles

Fig. 4 presents the proposed 4Mb cache organization. Each block has its own power-gating transistor P_i. There is no requirement to predict the future access pattern. Whenever a macro is not accessed, its power transistors are turned off to save leakage power. This is done regardless of the standby time.

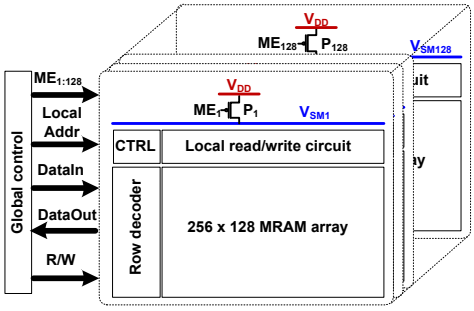


Fig. 4 MRAM architecture with one power gating transistor (ME) per macro

During a read or write, one particular block is selected. Correspondingly, a macro enable signal (i.e. ME_i) is triggered to turn on the power transistor P_i. Unlike the conventional design where the actual read/write operation is performed during the high clock phase, in our design, the clock signal is inverted in each sub-array so the actual read/write operation is performed during the low clock phase. *Thus, the block must wake up within half a clock cycle so that its supply voltage is fully on*. Compared to the un-gated design, our approach only incurs half cycle additional access latency while *throughput* and *maximum operating frequency* are not affected. Unselected blocks are put into idle mode immediately to save power. This instant-on power gating scheme allows transparent cache operation without cache access prediction.

B. Power transistor sizing strategy and area overhead

The power transistor must be sized properly to provide a fast wake-up time, supply enough write current to the write circuits and at the same time minimize both area overhead and leakage current. Note that in STT-MRAM, write operation is current-mode and thus a significant DC current must be maintained during each write cycle. Our MTJ model requires an average of 60 µA to complete its switching within 5 ns. Writing a word of 16 bits requires at least 1.2 mA DC current which includes both MTJ and other circuit's switching currents.

Fig. 5 analyzes both active and leakage current of various power transistor options (different sizing and V_{th} options). It is apparent that at the same on-current, a smaller LVT device is required. Another observation from Fig. 5 is that although the *leakage current ratio* between different device choices is large, their absolute values are similar and in the range of only a few nW. Table I summarizes the minimum transistor size requirement for each case to ensure 1.2 mA active current From this, we chose the LVT device with 100 nm channel length as the baseline power transistor design because it minimizes both area and dynamic switching power overhead.

Once, minimum transistor size is identified, we use larger transistor to meet wake-up time requirement (i.e. less than 2.5 ns). Fig. 6 shows the Monte-Carlo wake-up time simulation using two different power gating devices (i.e. L = 100 nm, W = 30 µm and 40µm). Post-layout extracted view of the 32 Kb MRAM block was used to estimate realistic parasitic resistances and capacitances. It can be seen that with W = 30µm, wake-up time requirement is met but with little margin. As a result, W = 40µm is a chosen. Its corresponding performances are 4 nW leakage, 1.25 ns mean wake-up time and 2.3 mA active current@V_{DS}=50mV. The estimated dynamic power consumed by switching this device is 10uW@200MHz, including ME_i signal buffers.

TABLE I: MINIMUM TRANSISTOR SIZE CONSTRAINT

	SVT			HVT			LVT		
$Min_{W/L}$	30µm / 100nm	46µm / 200nm	68µm / 300nm	32µm / 100nm	48µm / 200nm	70µm / 300nm	**23µm / 100nm**	36µm / 200nm	48µm / 300nm
Leakage (nW)	1.1	1.0	1.14	0.97	1.0	1.16	**2**	1.8	2.2

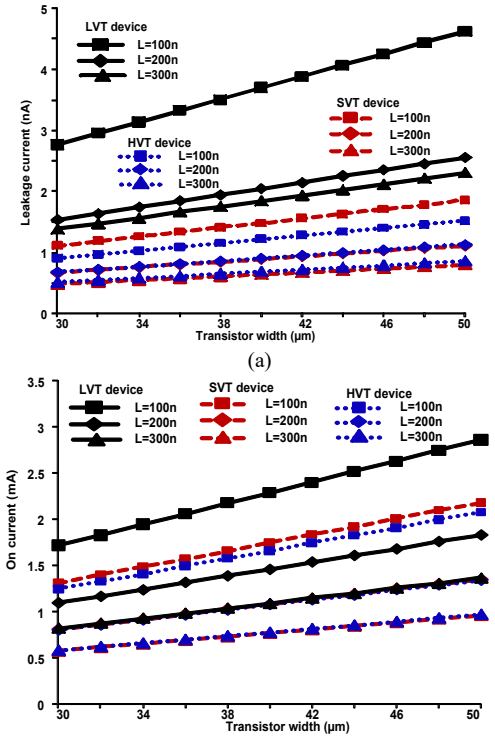


Fig. 5 (a) Active and (b) leakage currents of different transistor sizes and V_{th} options at 80 ºC, TT corner.

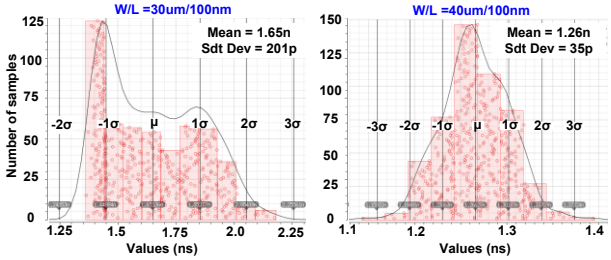


Fig. 6 Monte-Carlo simulatino showing the wake uptime using two different power transistor sizes (a) W = 30 µm and (b) W = 40 µm

IV. Energy Efficiency and Area Overhead

As mentioned, our design immediately puts a block into sleep mode whenever it is not accessed. In the worst-case scenario, this "sleep" period only lasts one cycle. Thus, additional power

may be consumed by just switching the *power gating device* while no leakage is saved as the idle time is too short. We need to evaluate the scheme to validate that on average, it is beneficial regardless of the "sleep" time of a particular macro

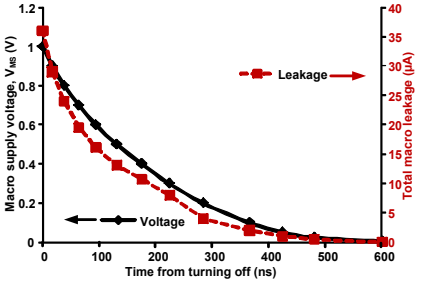


Fig. 7 Supply voltage to one MRAM block versus time after being turned off.

A. Energy efficiency evaluation

Fig. 7 simulates how voltage at the power rail of the MRAM block (V_{SMi} in Fig. 4) discharges versus time after it is turned off. During this period, the MRAM block continues to leak a certain amount of current and only settles to a resting voltage after about 120 cycles (i.e. 600 ns). If it is accessed again anytime during this period, the gating device must supply a large current to quickly bring it back to V_{DD}. This surge current does nothing but to compensate the leakage by the array. In addition, although the current sourced from the global V_{DD} network through the power gating device is nearly zero (i.e. a few nA), the effective leakage current in the array reduces gradually and must be accounted for. As a result, the average energy saved from that sub-array is insignificant if the "off" duration is short.

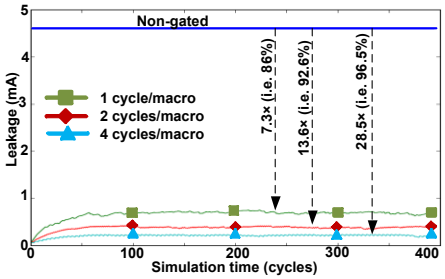


Fig. 8 Simulated total leakage of 4Mb MRAM using andom access pattern.

Figure 8 illustrates the benefit of our proposed scheme using extracted leakage data from Spectre simulation and offline random access modeling. The 4Mb MRAM is used for simulation. For each time step, a random block is chosen while leakage currents of other blocks are updated accordingly. At time step i^{th}, a block B_i is randomly selected and its power rail is reset to 1 (i.e.V_{Bi}=1). At the same time, power rail voltages and leakage currents of all other macros are updated and kept track for the next time step. In Figure 8, the total leakage current of non-gated design is represented by the dark navy line. At the bottom of the figures are three different current profiles corresponding to three cases where on average each sub-array is accessed for 1, 2 and 4 cycles everytime it is chosen, respectively. For the case of 1 cycle/macro, the simulator selects one sub-array, stay there for 1 cycle and move to a random one in the next cycle. Similarly, for the case of 2 cycles/macro (4 cycles/macro), when a sub-array is selected, it is access for 2 (4) cycles continuously.

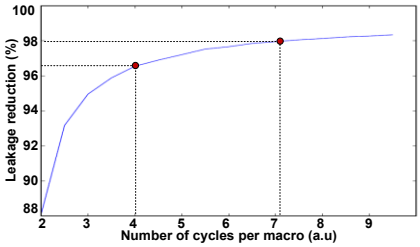


Fig. 9 Leakage current saving versus average access duration to each block

The simulation starts from the idle state. From that, its leakage power increases because more and more new blocks are selected while leakage currents from previous blocks are still comparable to un-gated leakage. However, this quickly saturates because macros that are accessed a long time ago does not contribute much to the present leakage current calculation. Our simulation (Fig. 8) shows that the total 4Mb leakage is only around 600 µA to 700µA for the case of 1cycle/macro. In this pessimistic scenario, leakage current saving is already 86%.

In more realistic scenarios, usually each block is accessed for a few cycles continuously. This greatly improves the leakage saving. Even with a modest value of 4 cycles per block, our proposed scheme saves more than 96% leakage current. Fig. 9 summarizes the total leakage savings versus different average access duration per block.

So far, our analysis assumes that the memory is 100% active and only concerns the leakage power of the sub-array. If both leakage power of the global peripheral circuits and dynamic power are included, the effective power reduction of the proposed scheme is 69%, as shown in Fig. 10. In this Figure, we also estimate the power reduction in cases where the memory is only active 50% and 10%, respectively. As the memory goes into the "normally-off" range (i.e. 50% active and 10% active cases), the leakage current of the peripheral circuit becomes significant and it is worth turning them off as well.

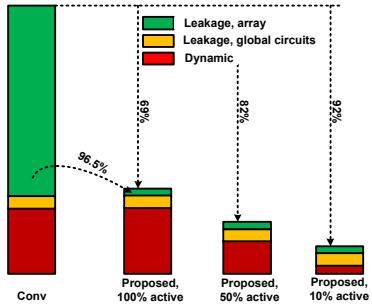


Fig. 10 Leakage current saving versus the percentage of active duration

B. Area overhead

Our original 256x128 memory block occupies 150µm x 120µm in 28 nm CMOS process. Adding one power gating transistor (40µm/100nm) incurs less than 0.5% area overhead, including global writing and enable signal buffers. In layout, 4 transistors, each sizes 10µm/100nm, are used to distribute power evenly around the supply rail of the block. Furthermore, power gating transistors of nearby blocks are abutted to ease the power supply network routing.

CIRCUIT IMPLEMENTATION

C. MRAM cell

We use 2T2MTJ cell to achieve both high access speed and reliability operation. The memory cell schematic is shown in Fig. 11, in conjunction with the array architecture, write driver and sense amplifier. We adjust the MTJ model parameter so that a low MTJ resistance (i.e. R_P) is 2.5 kΩ and a high MTJ resistance (i.e. R_{AP}) is 5 kΩ. The required current to switch from $R_P \rightarrow R_{AP}$ is 70 uA while that for $R_{AP} \rightarrow R_P$ is 50 uA. This gives an average switching current of 60 uA for our 2T2MTJ cell.

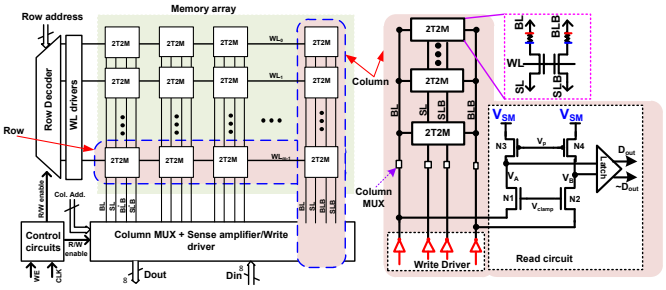


Fig. 11 MRAM macro circuit implementation

D. Read/write circuits

Schematic of the MRAM read/write circuits are also shown in Fig. 11. These circuits are adopted directly from our un-gated design. The most critical period during the write operation is when significant direct current is drawn to simultaneously switch the MRAM cells. During this period, V_{supply} actually drops from 1V to 0.96 V. Fortunately; all signals are digital and resilient against noise. Thus this marginal voltage drop does not affect the reliability of the write operation.

During read, within each selected cell, the SL and SLB are grounded while the WL is hold at V_{DD} to turn on the access transistor. At the same time the column MUX turns on the read circuit as shown in Fig. 11. Thanks to V_{clamp}, the BL and BLB voltages will be clamped to V_{clamp}-V_{th}, inducing cell currents to flow along the BL and BLB:

$$I_{BL} = \frac{V_{BL}}{R_{MTJL}} = \frac{V_{clamp}-V_{th}}{R_{MTJL}} \quad (1)$$

$$I_{BL} = \frac{V_{BL}}{R_{MTJR}} = \frac{V_{clamp}-V_{th}}{R_{MTJR}} \quad (2)$$

Where R_{MTJL} and R_{MTJR} are the corresponding resistances of the MTJs on the left and right, respectively. Since the MTJ pair always holds complementary values (i.e. R_P and R_{AP}), there is a differential current across the BL and BLB. This differential current multiplies with the equivalent input output resistance of the active load N3/4 to create differential inputs (i.e. VA/VB) to the latch sense amplifier.

Fig. 12 represents the alternative write and read simulation of our design. Fig. 12(a) shows how R_{MTJ} changes after each write operation while Fig. 12(b) details the read operation. Initially, both MTJs in the selected cells are initiated to 4 KΩ. Thus in the first two read operations, no differential signal is available at node V_A/V_B. As a result, the latch outputs are not consistent. After the first write (i.e WRITE 0), complementary data are written to the cells and the third read shows large differential signal between V_A and V_B before D_{out} = 0 is available. The fourth read is performed after a WRITE 1, resulting in a new D_{out}=1, as predicted. This simulation confirms that the read/write circuits function properly with the power gating device.

To further investigate the reliability of the sense amplifier, a Monte-Carlo simulation was performed. As shown in Fig. 13, our design copes well with process variations and offers good separation between V_A/V_B during a read operation. It can be seen that an input voltage gap of at least 400 mV is available to the latch, giving 100% correct output in a 5000-iteration Monte-Carlo simulation.

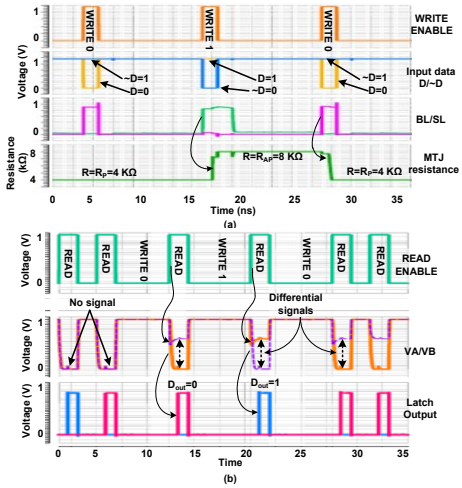


Fig. 12 (a) simulated write waveforms. (b) simulated read waveforms

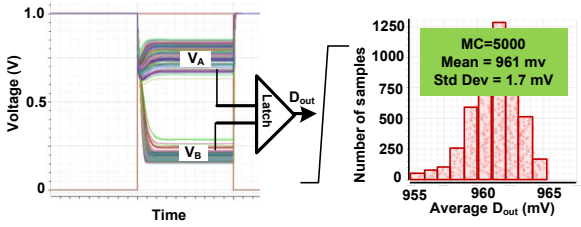


Fig. 13 Monte-Carlo simulation to evaluate the sensitivity of the latched amplifier

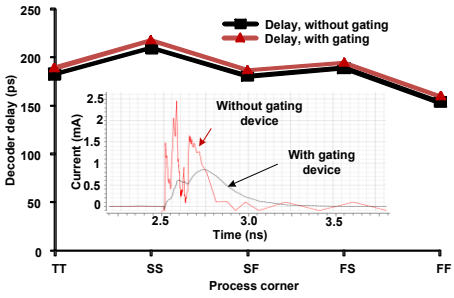


Fig. 14 Row decoder delay with and without the power gating device. The inset shows the instantaneous current profile sunk by the decoder upon the arrival of new address

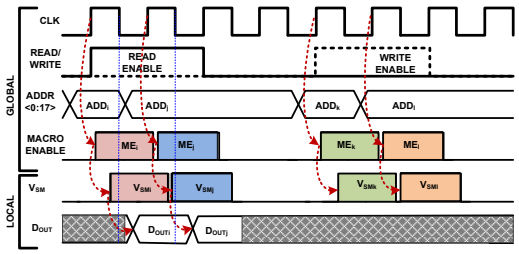


Fig. 15 Global timing diagram of the proposed MRAM.

E. Decoder and Wordline driver

A static 8-to-256 decoder is used for each sub-array. Fig. 14 shows the surge current and delay of the decoders at the rising

edge of the clock with and without the power gating device. There is only less than than decoder delay 4% across all five process corners. Therefore, it is safe to conclude that the power gating scheme has minimal impact on the speed of the design.

F. Global Control

Our global control is implemented using static gates. For simplicity, the global control circuitries are not power gated but has two levels and clock gated. The 128 macros are divided into 8 groups, each has 16 macros. The first level group decoder uses 3 MSB address bits $ADDR_{15:17}$. The second level decoder uses the next 4 address bits $ADDR_{11:14}$. Local address (i.e. $ADDR_{0:10}$) and input/output data are MUX-ed accordingly. Fig. 15 shows the timing diagram of the proposed design. At the rising edge of CLK, if the memory is active (i.e. either RE or WE is activated), the global decoder will turn on the corresponding macro enable signal (i.e. ME_i). As a result, local power network (i.e. V_{SMi}) is raised to V_{DD}. This process happens within half a cycle so thatV_{SM} reaches V_{DD} before the falling edge of CLK. During the low clock phase, the selected macro actually operates. In a read, D_{OUT} will be available after a short delay and MUX-ed out, also shown in Fig. 15. The use of two-level decoder reduces switching power of input/output data and local address bus. Their leakage power is slightly higher than the total leakage of one macro and thus contributes about 1.2% of total leakage of the un-gated design. After including the global control circuit, our effective leakage reduction is 95% in a typical scenario.

TABLE II: PERFORMANCE COMPARISION

	Proposed	Trans. VLSI Sys. 2011 [19]	ISSCC 2015 [6]
Technology	28 nm	N.A	65 nm
Cell type	2T2M	1T1M	2T2M
Capacity	4Mb	4MB	1Mb
Approach	Always power gate	No power gating	Power gating with access pattern prediction
Leakage per Mb (μW)	100	~3000	178 (w/o PG) / 82 (state 0) / 44.6 (state1)

V. CONCLUSION

This paper proposes a power gating scheme in combination with a half clock cycle phase shift to achieve instant-on MRAM characteristic. The use of half clock cycle shift allows the gated MRAM to operate at the same frequency as the non-gated conventional design. Our analysis shows that, with a 4 Mb memory capacity, the proposed scheme offers up to 96.5% leakage power reduction with minimum impact on its dynamic behavior. Therefore, it is the most suitable for high-speed low-power embedded memory applications.

REFERENCES

[1] R. Buhrman, "Spin Torque Writing for Next Generation MRAM - Challenges and Prospects," in *Device Research Conference, 2008*, 2008, pp. 223-224.

[2] X. Fong, Y. Kim, R. Venkatesan, S. H. Choday, A. Raghunathan, and K. Roy, "Spin-Transfer Torque Memories: Devices, Circuits, and Systems," *Proceedings of the IEEE,* vol. 104, pp. 1449-1488, 2016.

[3] H. Noguchi, K. Ikegami, S. Takaya, E. Arima, K. Kushida, A. Kawasumi, H. Hara, K. Abe, N. Shimomura, J. Ito, S. Fujita, T. Nakada, and H. Nakamura, "4Mb STT-MRAM-based cache with memory-access-aware power optimization and write-verify-write / read-modify-write scheme," in *2016 IEEE International Solid-State Circuits Conference (ISSCC)*, 2016, pp. 132-133.

[4] T. Na, J. P. Kim, S. H. Kang, and S. O. Jung, "Read Disturbance Reduction Technique for Offset-Canceling Dual-Stage Sensing Circuits in Deep Submicrometer STT-RAM," *IEEE Transactions on Circuits and Systems II: Express Briefs,* vol. 63, pp. 578-582, 2016.

[5] K. Jo and H. Yoon, "Variation-Tolerant Sensing Circuit for Ultra-Low-Voltage Operation of Spin-Torque Transfer Magnetic RAM," *IEEE Transactions on Circuits and Systems II: Express Briefs,* vol. PP, pp. 1-1, 2016.

[6] H. Noguchi, K. Ikegami, K. Kushida, K. Abe, S. Itai, S. Takaya, N. Shimomura, J. Ito, A. Kawasumi, H. Hara, and S. Fujita, "A 3.3ns-access-time 71.2uW/MHz 1Mb embedded STT-MRAM using physically eliminated read-disturb scheme and normally-off memory architecture," in *2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers*, 2015, pp. 1-3.

[7] J. P. Kim, T. Kim, W. Hao, H. M. Rao, K. Lee, X. Zhu, X. Li, W. Hsu, S. H. Kang, N. Matt, and N. Yu, "A 45nm 1Mb embedded STT-MRAM with design techniques to minimize read-disturbance," in *VLSI Circuits (VLSIC), 2011 Symposium on*, 2011, pp. 296-297.

[8] E. Arima, H. Noguchi, T. Nakada, S. Miwa, S. Takeda, S. Fujita, and H. Nakamura, "Immediate sleep: Reducing energy impact of peripheral circuits in STT-MRAM caches," in *Computer Design (ICCD), 2015 33rd IEEE International Conference on*, 2015, pp. 149-156.

[9] D. Apalkov, A. Khvalkovskiy, S. Watts, V. Nikitin, X. Tang, D. Lottis, K. Moon, X. Luo, E. Chen, A. Ong, A. Driskill-Smith, and M. Krounbi, "Spin-transfer torque magnetic random access memory (STT-MRAM)," *J. Emerg. Technol. Comput. Syst.,* vol. 9, pp. 1-35, 2013.

[10] T. Kishi, H. Yoda, T. Kai, T. Nagase, E. Kitagawa, M. Yoshikawa, K. Nishiyama, T. Daibou, M. Nagamine, M. Amano, S. Takahashi, M. Nakayama, N. Shimomura, H. Aikawa, S. Ikegawa, S. Yuasa, K. Yakushiji, H. Kubota, A. Fukushima, M. Oogane, T. Miyazaki, and K. Ando, "Lower-current and fast switching of a perpendicular TMR for high speed and high density spin-transfer-torque MRAM," in *2008 IEEE International Electron Devices Meeting*, 2008, pp. 1-4.

[11] J. C. Slonczewski, "Current-driven excitation of magnetic multilayers," *Journal of Magnetism and Magnetic Materials,* vol. 159, pp. L1-L7, 1996/06/01 1996.

[12] N. Sakimura, R. Nebashi, H. Honjo, S. Shinsaku, K. Yuko, and S. Tadahiko, "A 500-MHz MRAM macro for high-performance SoCs," in *Solid-State Circuits Conference, 2008. A-SSCC '08. IEEE Asian*, 2008, pp. 261-264.

[13] K. Huang, R. Zhao, N. Ning, and Y. Lian, "A Low Power Localized 2T1R STT-MRAM Array With Pipelined Quad-Phase Saving Scheme for Zero Sleep Power Systems," *IEEE Transactions on Circuits and Systems I: Regular Papers,* vol. 61, pp. 2614-2623, 2014.

[14] A. Vatankhahghadim, W. Song, and A. Sheikholeslami, "A Variation-Tolerant MRAM-Backed-SRAM Cell for a Nonvolatile Dynamically Reconfigurable FPGA," *IEEE Transactions on Circuits and Systems II: Express Briefs,* vol. 62, pp. 573-577, 2015.

[15] S. Huda and A. Sheikholeslami, "A Novel STT-MRAM Cell With Disturbance-Free Read Operation," *IEEE Transactions on Circuits and Systems I: Regular Papers,* vol. 60, pp. 1534-1547, 2013.

[16] H. Koike, S. Miura, H. Honjo, T. Watanabe, H. Sato, S. Sato, T. Nasuno, Y. Noguchi, M. Yasuhira, T. Tanigawa, M. Muraguchi, M. Niwa, K. Ito, S. Ikeda, H. Ohno, and T. Endoh, "1T1MTJ STT-MRAM Cell Array Design with an Adaptive Reference Voltage Generator for Improving Device Variation Tolerance," in *2015 IEEE International Memory Workshop (IMW)*, 2015, pp. 1-4.

[17] J. W. Ryu and K. W. Kwon, "A Reliable 2T2MTJ Nonvolatile Static Gain Cell STT-MRAM With Self-Referencing Sensing Circuits for Embedded Memory Application," *IEEE Transactions on Magnetics,* vol. 52, pp. 1-10, 2016.

[18] T. Na, J. Kim, J. P. Kim, S. H. Kang, and S. O. Jung, "Reference-Scheme Study and Novel Reference Scheme for Deep Submicrometer STT-RAM," *IEEE Transactions on Circuits and Systems I: Regular Papers,* vol. 61, pp. 3376-3385, 2014.

[19] Z. Sun, X. Bi, H. Li, W. F. Wong, and X. Zhu, "STT-RAM Cache Hierarchy With Multiretention MTJ Designs," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems,* vol. 22, pp. 1281-1293, 2014.