BLOCK-LEVEL POWER GATING MRAM WITH 96.5% LEAKAGE REDUCTION

ANH TUAN DO

Institute of Microelectronics,Agency for Science, Technology and Research (A*STAR), Singapore,
Email: doat@ime.a-star.edu.sg

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. The whole MRAM memory can always be put 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. 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

Thanks to its non-volatility and many of strict requirements of low-power memories [1], STT-MRAM is a promising memory technology for ubiquitous future battery-powered applications. In STT-MRAM, data are represented as "0" and "1" as the magnetic state of the MTJ. To date, STT-MRAM gained the most attention from both research and the industries due to its excellent scalabilities[2]. Several STT-MRAMimplementations reported in the literaturedemonstrated desirable low-power and high-density characteristics[3-6]. Most works to date have focused on reducing the large dynamic energy in STT-MRAM due to large write current. Since the non-volatility of STT-MRAM is attractive for embedded applications, there is a need to eliminate leakage power in STT-MRAM, which can be the dominant component of power dissipation in large capacity STT-MRAM arrays. Several works have proposed power-gating schemes to minimize leakage power in STT-MRAM[3, 6].For a practical application, on-chip memory 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 complexityof the control circuitry. Alternatively, a sub-array may be turned off if its next predicted access is far 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, anaccess 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 modeaccordingly[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. Next, 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 a STT-MRAM design for extreme leakage power saving using an aggressive power gating scheme without any memory access prediction. In our proposed design, only the sub-array being accessed is turned on during memory access while all other sub-arrays remain in the off state. It can be shown that our scheme always saves leakage power. In Section III, we discuss our proposed instant-on power-gating scheme. Energy efficiency and area overhead of our proposed STT-MRAM design will be discussed in Section IV. Section V concludes the paper.

II. MTJ AND MRAM FUNDAMENTALS

A. MTJ

MTJ [2, 9]is a two-terminal magnetoresistive device and is used as the basic storage element in MRAM. The electrodes of the MTJ are connected to separate ferromagnetic layers (called the reference and the storage layers, respectively) that sandwich a tunnel oxide barrier (Fig. 1(a)). Due to the tunneling magnetoresistance effect, the magnetic configuration of the MTJ modulates the resistance (R_{MTJ}) between its terminals[10]. The ferromagnetic layers in the MTJ behave as filters that only allow electrons with specific spin polarization to pass through. When the magnetic moments of both layers are in parallel, the resulting MTJ resistance R_{MTJ} is low (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 thespin-transfer torque (STT) effect when a current is passed through it[11]. A sufficiently large current flow through the MTJ generates a torque that overcomes the anisotropy energies in the storage layer to switch the MTJ into the AP or P state, depending on the direction of the current. Thus, the MTJ state can be programmed by just passing current through the MTJ.

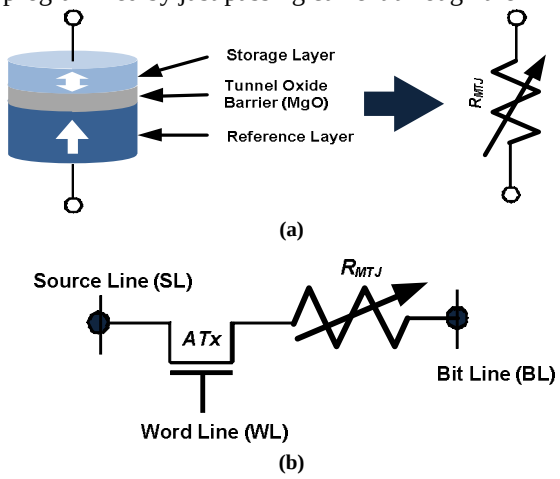


Fig. 1 (a) Simplified MTJ structure (b) 1T1MTJ MRAM cell topology

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 about 16 kb to 64 kb. The sub-array is the building block to construct larger memory size. Each sub-array is a self-contained architecture including alocalrowdecoder, memory array, and read/write circuitries. Every memory array is formed by organizing the STT-MRAM memory cells into rows and columns.Numerous cell topologies have been proposed to address different optimization constraintsbut the most popular designs are the 1T1M[1] and 2T2M[3] due to their regular structure, compactness, and simple control circuitry. The 1T1M bit-cell is formed by connecting an NMOS (ATx) to the MTJ as shown in Fig. 1(b). An NMOS ATx is usually used because it hasbetter driving capability as compared to the PMOS ATx.The gate of ATx is connected towordline(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.

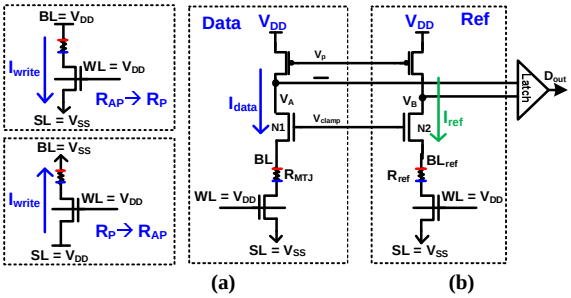


Fig. 2 MRAM cell configuration during (a) write operation (b) read operation

The access transistor is normally OFF.The WL turns it ON only during the read operation.To program the MTJ, the SL and BL are driven to appropriate voltages and passing a current through the MTJ. As a convention, to write a "0" the SL driven to "0" while the BL is driven to "VDD". (i.e. R_{MTJ}switches from $R_{AP}$to R_P). Conversely, BL and SL are set to $V_{SS}$and V_{DD}, respectively, to write "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). This circuitry is usually shared in the entire array. Also, abit-line clamp voltage is used to limit the current flowing through the memory cells so that they are not accidentally written intoduring read operation.For a correct sensing, the reference signal (i.e. I_{ref}) must lie between the low and high value of the actual data signal (i.e. $I_{data,AP}$ and $I_{data,P}$). However, due to PVT variations, agoodreferenceand reliable sensing is crucial to achieve for a large memory array in nano-scale technologies.

C. Power gating with access predictor

As mentioned earlier, power-gating is a popular approach to eliminate leakage power in MRAM. Unlike SRAM, the MRAM cell is non-volatile and the MRAM array can be completely powered down without data loss. Thus, 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.

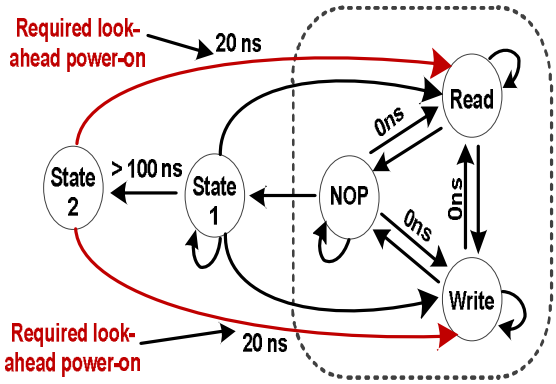


Fig. 3 MRAM power gating state machine using access predictor [12]

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"

Furthermore, 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 theminimal peripheral. The mid-leveland 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.

III. INSTANT-ON POWER GATING TECHNIQUE

In this section, we discuss how to efficiently power-gate the memory array to minimize both its access latency and leakage power. For the sake of clarity, our analysis assumes a 4Mb MRAM consists of 128blocks, each contains 32 kbof MRAM cells. The 32 kb-MRAM blockis self-contained design unit with row decoder, local read/write circuit and a 256-rowx 128-column cell array as shown in Fig. 4.

A. System overview and operating principles

Fig. 4 presents the proposed 4Mb cache organization. Each block has its own power-gating transistor. In our proposed design, 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. During a read or write, one particular block is selected by the sub-array address decoder using MSB of the address. 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 during the low clock phase. 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.

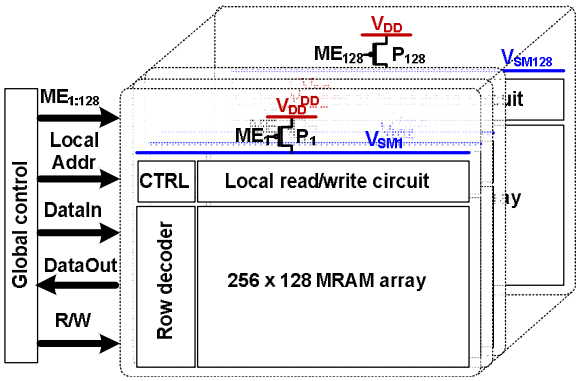


Fig. 4 MRAM architecture with one power gating transistor (ME) per macro

This instant-on power gating scheme allows transparent cache operation without cache access prediction. Its behavior is similar to word-level power-gating in. However, our approach incurs much less area overhead with seamless integration.

IV. ENERGY EFFICIENCY AND AREA OVERHEAD

A. Energy efficiency evaluation

Fig. 5 shows the simulated voltage at the power rail of one MRAM block versus time after it is turned off (V_{SMi} in Fig. 4). Since there is a significant parasitic capacitance at this node, it takes thetime to discharge to ground. 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.

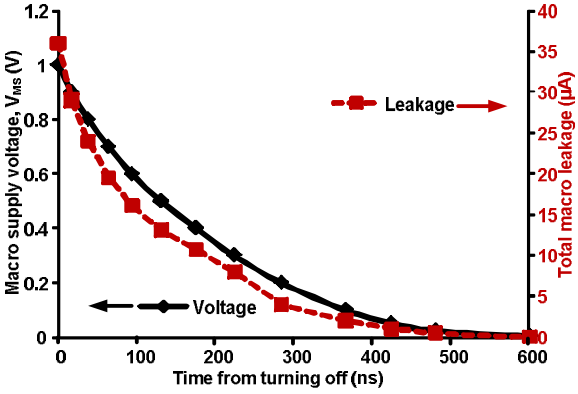


Fig. 5 The supply voltage to one MRAM block versus time after being turned off.

However, intuitively speakingif the number of macros is large enough, in order for a few macros to be repeatedly turned off then accessed again in a

short while, some others must be in idle mode for a long time. The worst case scenario happens when on average the operating system access each macro in only one cycle, jump to the next and so on. Even in this situation, statistically,each macro will be put in idle mode for about 127 cycles before it is activated again. As a result, our scheme is still beneficial, regardless of the memory access pattern. Figure 6 illustrates the benefit of our proposed scheme using extracted leakage data from Spectre simulation and offline random access modeling. The 4Mb MRAM above 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. The discharge speeds and instantaneous leakage currents are extracted from Fig. 5. Due to the large number of macros simulated in the array, we observed no macros were accessed in consecutive cycles. 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.

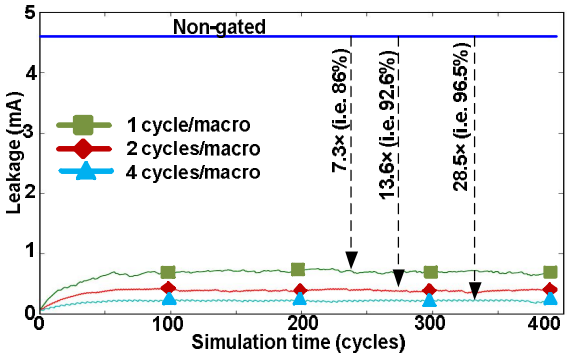


Fig. 6 Simulated total leakage of 4Mb MRAM using leakage profile in Fig. 5 and random access pattern.

The simulation starts from theidle state so its initial total leakage is close tozero. From that, it 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. 6) shows thatthe 4Mb leakage approaches and oscillates around 600 µA to 700µAfor the case of 1cycle/macro. In this pessimistic scenario, leakage current saving is already 86%.

B. Area overhead

Our original 256x128 memory block occupies 150µm x 120µm in 28 nm CMOS process. Thus 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.

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 at least86% leakage power reduction with minimum impact on its dynamic behavior. Therefore, it is the most suitable for high-speed low-power embedded memory applications.

ACKNOWLEDGEMENTS

The work was supported by the Science and Engineering Research Council of A*STAR (Agency for Science, Technology and Research), Singapore, under the grant number: 1423400027.

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.

★★★