Chapter 4
Embedded STT-MRAM: Device and Design

Seung H. Kang and Seong-Ook Jung

Abstract Spin-transfer-torque magnetoresistive random access memory (STT-MRAM) is made of a combination of semiconductor integrated circuits (IC) and a dense array of nanometer-scale magnetic tunnel junctions (MTJ). This emerging memory is of growing technological interest due to its potential to bring disruptive device innovation to the world of electronics. STT-MRAM is capable of providing high speed, unlimited endurance, and nonvolatility simultaneously, which is often recognized as a unique advantage over conventional and other emerging memories. While the technology is at an early stage and evolving in multiple platforms, STT-MRAM is particularly compelling as an embedded memory for system-on-chip (SOC). STT-MRAM can be integrated into SOC without altering baseline logic platforms both in process and in design. This chapter overviews key device and circuit subjects from the perspective of co-designing logic and MTJ.

4.1 Introduction

Generic device scaling no longer secures the evolution of IC, causing the silicon-based technology to face unprecedented challenges in materials, devices, and processes. These challenges translate to compromises in power dissipation, performance, and cost for a wide range of IC products. While the end of physical scaling is not imminent, its value is being heavily eroded by the growing technological and economic concerns at the nanoscale. Some of promising innovations that can mitigate or overcome such problems may be found in spintronic IC. In the past few years, the spintronics community has achieved significant discoveries and breakthroughs [1]. Most recognized is the emergence of STT-MRAM [2–6]. Key discoveries and advances have triggered industry-wide R&D efforts in pursuit of an alternative memory in lieu of conventional memories that are not only facing acute

S.H. Kang (✉)
Qualcomm Technologies Inc., 5775 Morehouse Dr., San Diego, CA 92121, USA
e-mail: kang@qti.qualcomm.com

S.-O. Jung
Yonsei University, Seoul, South Korea
e-mail: sjung@yonsei.ac.kr

R.O. Topaloglu (ed.), *More than Moore Technologies for Next Generation Computer Design*, DOI 10.1007/978-1-4939-2163-8_4

tradeoffs in performance and power, but also nearing fundamental scaling limits. In parallel, various forms of MTJ-based novel logic devices and circuits have been demonstrated [7–9], opening a possible path for spintronic IC to expand beyond memory applications. Furthermore, a novel computing architecture concept, known as normally-off computer, was proposed as a way to reduce the energy consumption of modern microprocessors [10–12]. Still at an early stage in its endeavor, the global spintronics community continues to propel a plethora of innovations in materials, devices, circuits, and architectures.

STT-MRAM is particularly compelling as an embedded memory for SOC. In contrast to standalone commodity memories, each type of SOC requires a different combination of memory attributes such as speed, energy consumption, and reliability including cyclic endurance and data retention. STT-MRAM can be offered in a variety of macros whose designs are customized for application-specific SOC. In general, density requirements are found over a wide range (a few kbits to 256 Mbits). Yet, even in small densities, it can realize significant values in system performance, energy consumption, security, and cost, when device and circuit attributes are tailored at a system-architecture level. Furthermore, the memory element MTJ can be integrated in a fully logic-compatible way without altering or adversely impacting baseline logic platforms by adding two or three mask layers into a back-end-of-line (BEOL) flow [3].

Driving STT-MRAM beyond discrete devices and arrays toward SOC necessitates extensive learning cycles in device, circuit, yield, and reliability engineering. In order to produce variability- and fault-tolerant STT-MRAM, a systematic design methodology is required to assure robust functionality of STT-MRAM over a wide range of process-voltage-temperature (PVT) windows. This chapter overviews key device and circuit subjects from the perspective of co-designing logic and MTJ to enable STT-MRAM as a scalable custom embedded memory to serve advanced SOC.

4.2 Device Physics

4.2.1 Magnetic Tunnel Junction (MTJ)

A MTJ is a building block as a storage element for STT-MRAM. A MTJ consists of metallic ferromagnetic films separated by an oxide tunnel barrier, typically an ultra-thin magnesium oxide (MgO). The conductance of a ferromagnetic metal-insulator-ferromagnetic metal (FM1-I-FM2) structure is governed by tunnel magnetoresistance, a quantum mechanical phenomenon that results from spin-dependent tunneling [13]. When conduction electrons are emitted from one ferromagnetic metal electrode FM2, schematically illustrated in Fig. 4.1, they are spin-polarized to the magnetization direction of FM2 and tunnel through the thin tunnel barrier with their spin states conserved. The electron density of states in the opposite

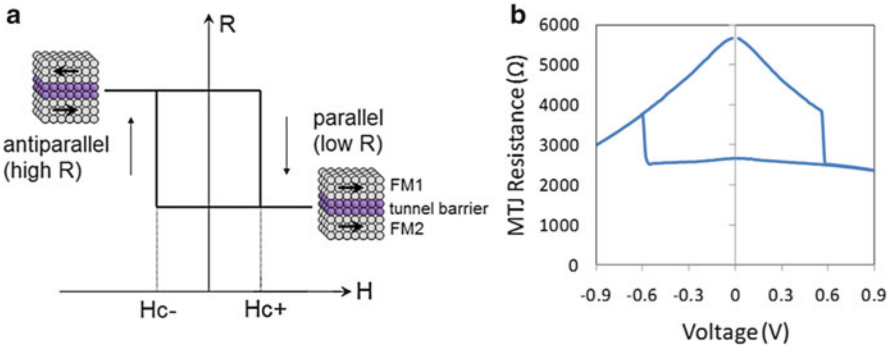


Fig. 4.1 (**a**) Conceptual illustration of a MTJ hysteresis curve. A binary resistance state is obtained through configuring the magnetization of FM1 (free layer) with respect to that of FM2 (pinned layer) either parallel or antiparallel at the switching field H_{c+} or H_{c-}, respectively; (**b**) a MTJ hysteresis curve driven by spin-transfer-torque (STT) switching

ferromagnetic metal electrode FM1 that these tunneling electrons encounter is dependent on the magnetization direction of FM1. Consequently, the electrical resistance (R) of FM1-I-FM2 structure is determined by relative orientations of the magnetizations, which is described by [14]

$$R = \frac{R_{\perp}}{1 + \frac{TMR}{2}\cos\theta} \tag{4.1}$$

where θ is the angle between the two configurations, $R_{\perp}$ is the resistance measured in the perpendicular magnetic configuration ($\theta = \pi/2$). R becomes minimum (R_p) for the parallel magnetization configuration ($\theta = 0$) and maximum (R_{ap}) for the anti-parallel configuration ($\theta = \pi$). Accordingly, a MTJ serves as a variable resistor that can be configured to have binary states (0 and 1) defined by two discrete resistance values (R_p and R_{ap}, respectively). The tunnel magnetoresistance ratio (TMR) is then defined as:

$$TMR = \frac{R_{ap} - R_p}{R_p} \times 100\,\% \tag{4.2}$$

TMR is one of critical device parameters for the design of STT-MRAM for error-free and high-speed read operations since the signal margin for sensing an array of MTJ is governed by TMR.

Figure 4.2 illustrates typical MTJ film stacks that essentially consist of metallic films separated by a tunnel barrier, most commonly MgO on the order of 1 nm in thickness. Depending on the orientation of the magnetization with respect to the film plane, two representative cases are shown here: (a) in-plane MTJ (i-MTJ); and (b) perpendicular MTJ (p-MTJ). The free layer is a soft ferromagnetic metal

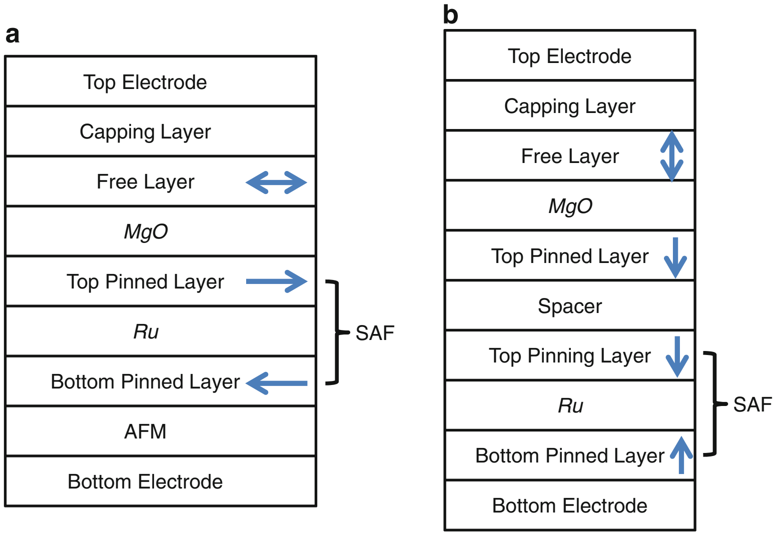


Fig. 4.2 Schematic illustration of typical MTJ film stacks: (**a**) in-plane MTJ (i-MTJ); (**b**) perpendicular MTJ (p-MTJ)

(e.g. CoFeB) whose magnetization can be switched by STT. The reference layer is commonly a synthetic structure to provide a reference magnetization fixed in one direction (e.g. top pinned layer in Fig. 4.2a) relative to the free layer magnetization. For i-MTJ, the magnetization of the bottom pinned layer is fixed by an antiferromagnet (AFM) pinning layer (e.g. PtMn) via the exchange bias effect. The top pinned layer is then antiferromagnetically coupled to the bottom pinned layer via interlayer exchange coupling with a non-magnetic spacer (e.g. Ru). This type of reference layer is called a synthetic antiferromagnet (SAF). In comparison, for p-MTJ, the out-of-plane magnetization of the bottom pinned layer can be developed inherently during film formation, hence the AFM pinning layer is not necessary (Fig. 4.2b). Ordinarily, the reference layer stack of p-MTJ is still structured in a SAF configuration. To promote high TMR, however, p-MTJ often necessitates an additional pinned layer (commonly in CoFeB) underneath the tunnel barrier which is exchange-coupled with the SAF.

The metallic films of MTJ are deposited by physical vapor deposition (PVD). The MgO barrier can be grown by PVD or a combination of PVD and oxidation. MTJ device properties are tailored through a selection of desired materials and a precise control of microstructure, film thickness, and cross-sectional feature size. Key MTJ parameters, critical to optimizing performance, energy consumption and reliability, include: TMR, resistance–area product (RA), energy barrier (E_B) and switching current density (J_c).

Experimental TMR reached ∼600 % at room temperature in a CoFeB/MgO/Co FeB junction [15]. From a microstructure perspective, the most critical factor in achieving high TMR is promoting strong MgO (001) texture. For practical device applications, high TMR needs to be achieved in conjunction with relatively low RA, preferably, $<10\ \Omega\ \mathrm{cm}^2$. TMR of 253 % at RA $= 5.9\ \Omega\ \mathrm{cm}^2$ has been demonstrated by inserting CoFe as a crystallization template to induce preferred grain growth in MgO and then to promote crystallization of CoFeB through annealing [16]. In-situ annealing of the MgO barrier has been known to promote the (001) texture further, resulting in high TMR (>170 %) even for MTJ films with an ultralow RA ($\sim 1\ \Omega\ \mathrm{cm}^2$) [17].

At a static mode, MTJ maintains its resistance state without power (i.e. non-volatile) as long as the magnetic anisotropy of its free layer is greater than the thermal excitation energy described by $k_B T$ where k_B is the Boltzmann constant and T is temperature. For i-MTJ which is typically patterned into an elliptically shaped cell, the free layer magnetic moment can have only two energetically favorable states aligned with the long-axis (called easy-axis) of the MTJ, thereby allowing either R_p or R_{ap}. For p-MTJ, the two states are determined by out-of-plane moments. Hence p-MTJ does not require a particular shape and is typically patterned in a circular shape.

Under the simplified assumption of a single-domain free layer, the energy barrier (E_B) between the two energetically favorable states is often given by

$$E_B = \frac{M_s H_k V}{2}\left(1 - \frac{H_{ext}}{H_k}\right)^2 \tag{4.3}$$

where M_s is the saturation magnetization of the free layer, V is the free layer volume, H_k is the effective uniaxial anisotropy field, and H_{ext} is the external field present along the easy-axis (which vanishes in the absence of any stray field). For MTJ to be non-volatile, E_B must be larger than the thermal excitation energy over a range of operating and storage temperatures. For example, for a single MTJ to retain its state for 10 years, E_B must be $40\,k_B T$ (1 eV) or greater. A recent report demonstrated that E_B can be $100\,k_B T$ or greater, which is remarkable for p-MTJ on the order of 30 nm in diameter [6].

4.2.2 Spin-Transfer-Torque (STT) Switching

A traditional way of programming MTJ is to apply a magnetic field to switch the free layer magnetization. A drawback of this method is the requirement of large current to induce sufficient magnetic field. It is also well understood that this method does not provide good scalability because decreasing the MTJ size entails larger switching fields, hence, even more current.

A breakthrough in physics of MTJ switching was accomplished in 1996 by the theoretical formulation that the free layer magnetization could be modulated by the direct transfer of spin angular momentum from spin-polarized electrons [18, 19]. This phenomenon, called spin-transfer-torque (STT) magnetization reversal, delivered a new means to control the free layer magnetization by directly applying electric current through MTJ without a need of magnetic field. The magnitude of STT scales with the current density (J). This is particularly beneficial for device scalability since the critical switching current (I_c) should scale proportionally to the size of MTJ. A breakthrough demonstration of STT-MRAM at an array level was first reported in 2005, including TMR of 160 % and switching speed as fast as 1 ns [2].

For i-MTJ, the intrinsic critical switching current (I_{c0}) is given by

$$I_{c0} = \frac{2e}{\hbar}\,\frac{\alpha}{\eta} M_S V \left(H_{k||} + \frac{H_d - H_{k\perp}}{2} \right) \tag{4.4}$$

where α is the damping constant, η is the spin polarization constant, $H_{k||}$ is the uniaxial anisotropy field in the film plane, H_d is the effective perpendicular demagnetization field that corresponds to the field required to saturate the free layer moment perpendicular to the film plane, and $H_{k\perp}$ is the anisotropy field perpendicular to the plane. The H_d term, given by $4\pi M_s$, represents an additional energy term that needs to be overcome during STT switching because the shape anisotropy induces an oscillatory motion of magnetization confined in the direction perpendicular to the film plane, resulting in an elliptical precession. Undesirably, H_d only greatly increases I_{c0} without contributing to E_B. A technological challenge in building STT-MRAM is to reduce I_{c0} while maintaining sufficient E_B. Hence, an effective way of reducing I_{c0} without degrading E_B is to introduce perpendicular anisotropy $H_{k\perp}$ to cancel a substantial portion of H_d.

Considered as an essential figure of merit, the STT switching efficiency is described by the ratio of E_B and I_{c0}. For i-MTJ, it is typically on the order of 0.5–1 $k_BT/\mu A$. This allows good scalability for i-MTJ as small as approximately 40 nm (short axis). However, the success of STT-MRAM is largely dependent on whether MTJ can be scaled to deep nanoscale nodes (30 nm and below) in conjunction with low switching energy and high stability. Unless the STT efficiency is raised significantly, i-MTJ may not provide sufficient E_B for nonvolatility, which limits physical scaling of i-MTJ for future nodes.

This scalability challenge can be overcome by adopting p-MTJ which provides much greater anisotropy even in small features. E_B of p-MTJ is determined by crystalline or interface perpendicular magnetic anisotropy (PMA), not by the shape anisotropy of i-MTJ. Various PMA materials have been investigated, which include L1_0-ordered FePt or FePd alloys, Co-based superlattices such as Co/Pt and Co/Ni laminates, rare-earth/transition metal alloys, etc. To build useful MTJ devices, however, these materials must be engineered for an optimal combination of materials properties like M_s and H_k and device properties like TMR and J_c. A prior report addressed that the anisotropy resulting from the CoFeB-MgO interface can induce large $H_{k\perp}$ [4]. When CoFeB is sufficiently thin (typically ~1.5 nm or thinner), such interface PMA can overcome the demagnetization field, i.e., $H_{k\perp} > H_d$. The film can then become magnetized fully perpendicular to the plane. With further tuning of the stack, such interface PMA can be achieved for even thicker CoFeB. Recently, p-MTJ devices utilizing interfacial PMA of CoFeB have successfully been engineered for fully functional 8 Mbit STT-MRAM [6].

Referring to Eq. (4.4), I_{c0} and E_B pertaining to p-MTJ with interfacial PMA are described by

$$I_{c0} = \frac{e}{\hbar}\,\frac{\alpha}{\eta}\,M_S V H_{k\perp}^{eff} \tag{4.5}$$

$$E_B = \frac{M_S V H_{k\perp}^{eff}}{2} \tag{4.6}$$

where $H_{k\perp}^{eff}$ is the effective perpendicular anisotropy field. In contrast to i-MTJ described by Eqs. (4.3) and (4.4), note that I_{c0} is directly proportional to E_B. The absence of the H_d term means that STT switching is far more efficient. This leads to substantially higher STT efficiency (E_B/I_{c0}). Recently, $E_B/I_{c0} \sim 5\,k_BT/\mu A$ has been reported from an array of ~30 nm p-MTJ [6], suggesting that the STT efficiency of p-MTJ could be an order of magnitude greater than that of in-plane MTJ. This is a significant breakthrough demonstrating the scalability of p-MTJ based on interfacial PMA of CoFeB, which is a preferable material to achieve high TMR as well.

4.3 Device Engineering

4.3.1 Bitcell and Array

STT-MRAM is a hybrid IC built on a combination of semiconductor logic and MTJ. Its bitcell which represents 1 bit is commonly architected in 1 transistor plus 1 MTJ (1T-1J). As shown in Fig. 4.3, a MTJ is connected in series to an n-type metal oxide semiconductor transistor (NMOS). This transistor is called an access transistor since it controls read and write access to the connected MTJ as a digital switch.

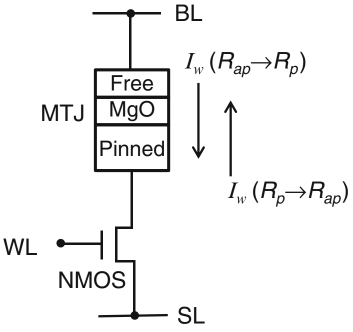


Fig. 4.3 Schematic representation of a 1T-1J bitcell

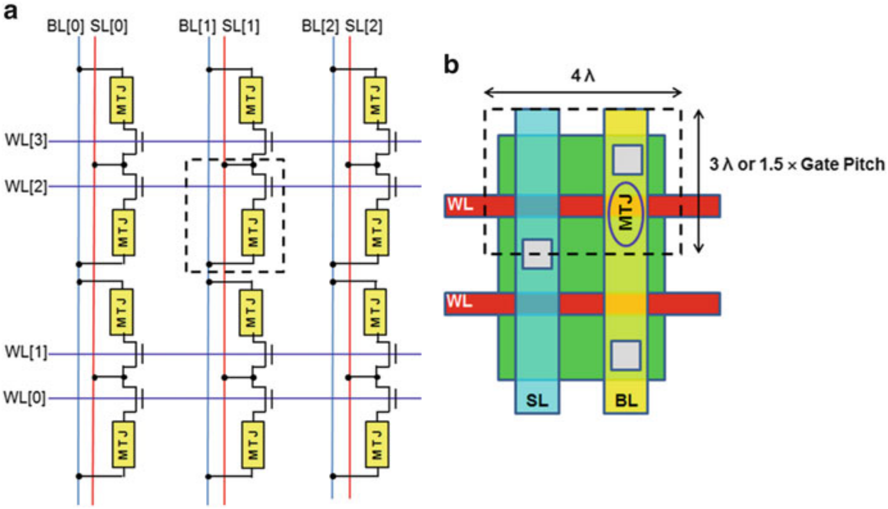


Fig. 4.4 STT-MRAM array architecture with SL parallel to BL: (**a**) 4 × 3 array; (**b**) 1T-1J bitcell layout

Figure 4.4 is a schematic representation of a typical STT-MRAM array that consists of 1T-1J bitcells. To read the information stored in a cell, the word line (WL) of the selected cell is turned on and a small read current is applied by a sensing circuit (Sect. 4.4) to either the selected bit line (BL) or the source line (SL) with the other end of the cell grounded (*GND*). A sense amplifier determines the cell state by sensing the difference between the cell resistance and the reference resistance predefined from a reference MTJ array. In comparison, the write operation requires bidirectional currents because the direction of write current determines which resistance state (R_p or R_{ap}) is programmed to MTJ. With the bitcell architecture shown in Fig. 4.4, for $R_{ap} \rightarrow R_p$, a write voltage is applied to BL ($V_{BL} = V_{DD}$) with WL turned on ($V_{WL} = V_{DD}$) and SL grounded ($V_{SL} = 0$ V, *GND*), and vice versa for $R_p \rightarrow R_{ap}$. For successful write operation, the write current (I_w) supplied to the MTJ in each bitcell must be larger than the MTJ critical switching current (I_c).

Figure 4.4b shows an example layout of 1T-1J bitcell with an array architecture illustrated in Fig. 4.4a. Provided that the minimum metal half-pitch is λ, two metal lines BL and SL running in parallel limit the minimum bitcell width to 4 λ. Then the metal plate connected to the source and the drain of the access transistor may limit the bitcell height to 3 λ or 1.5 times of the gate pitch. Assuming the metal pitch is larger than the gate pitch, the bitcell size can be as small as 12 λ^2.

The array architecture shown in Fig. 4.4 is simple to design and operate. One shortcoming of this structure is that every BL is coupled with its own SL, thereby causing a larger array footprint. A more compact array can be realized by placing SL orthogonal to BL, as shown in Fig. 4.5. SL is then parallel to WL and shared between two neighboring rows of WL. With this architecture, the bitcell size can be

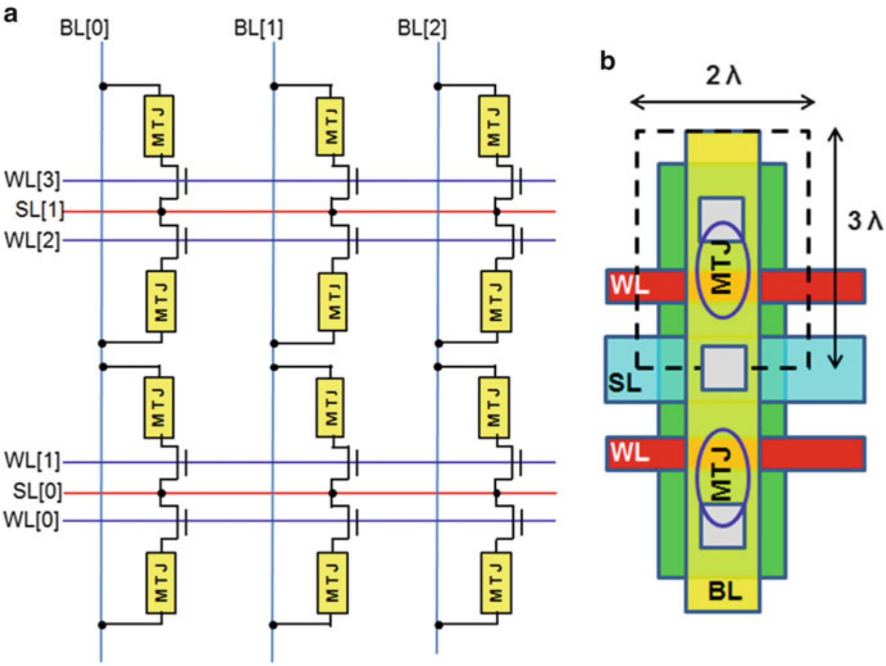


Fig. 4.5 STT-MRAM array architecture with SL orthogonal to BL: (**a**) 4 × 3 array; (**b**) 1T-1J bitcell layout

as small as 6 λ^2 (Fig. 4.5b), a half of the size in Fig. 4.4. However, this architecture results in more complex write operation for $R_p \rightarrow R_{ap}$. When SL is raised to a write voltage, the selected BL is grounded. Simultaneously, all the unselected BL associated with the selected WL must be raised to the same level of the write voltage to avoid unintentional current flows to the unselected MTJ. Consequently, this architecture consumes more power during write operation. Furthermore, this may even necessitate two separate write pulses to complete a full write cycle, since in chip-level operation each full cycle carries multiple bits (typically, 32, 64, or 128 bits) of R_p and R_{ap} concurrently. Accordingly, this architecture is not desirable for low power and high speed applications.

Table 4.1 describes an example of the attributes of a bitcell embedded for a 45-nm low-power logic platform [3]. The bitcell size is ~50 F^2, where F is 45 nm (minimum feature size of this node). When the half-metal pitch λ is used, the size is ~20 λ^2 since λ is 70 nm. This is significantly larger than that of an ideal layout of the same array architecture in Fig. 4.4b, which is attributed to the constraints of logic design rules of this particular logic technology.

Table 4.1 Key attributes of an embedded STT-MRAM bitcell demonstrated for a 45-nm low-power logic platform [3]

Feature size: F/λ	45/70 nm
V_{DD} (core/IO)	1.1/1.8 V
Cell architecture	1T-1J (reversely connected)
Bitcell size	0.1026 μm^2
Access NMOS (length/width)	40/270 nm
MTJ size	40 nm (short axis)
MTJ aspect ratio	2.5–3
TMR/RA	110 %/9 Ω μm^2
BEOL	Cu/low-k Seven metal layers

F is the minimum logic feature size, and λ is the minimum metal half-pitch

4.3.2 Writability

MTJ switching is a current-induced phenomenon, and the switching operation requires a bidirectional control of current. For 1T-1J, the currents supplied to MTJ are not symmetrical with respect to the polarity of current, owing to the phenomenon known as the source degeneration effect. This occurs when a resistive load is placed at the source side of a transistor. As a consequence, despite the same operating voltage (V_{DD}) applied to BL or SL, the transistor output currents are asymmetrical. This is illustrated in Fig. 4.6, where such asymmetry is simulated at a full circuit level. This causes a significant disadvantage which reduces the write margin of 1T-1J. Furthermore, the STT effect on a typical MTJ is also asymmetrical, which is described by I_c asymmetry (β), defined as $\left|\frac{I_c^{P\to AP}}{I_c^{AP\to P}}\right|$. Typical MTJ devices exhibit β of 1.5 or larger, presumably, due to smaller STT effect for $R_p \to R_{ap}$ (electrons flowing from the free layer to the reference layer). When these two effects are coupled in a conventional 1T-1J bitcell, it is much more difficult to switch the cell from R_p to R_{ap}, often results in increase in transistor size or operation voltage. Several approaches have been suggested to mitigate these problems: (1) I_c asymmetry reduction using dual spin polarizers [20]; (2) a “top-pinned” MTJ film stack [21]; and (3) a modified 1T-1J with a reversely connected MTJ [3].

In most cases of STT-MRAM targeted for fast switching, a primary challenge is to design for the capability of supplying sufficiently large driving current for MTJ switching. A simple alternative to 1T-1J is 2T-1J, for which one MTJ is coupled with two access transistors in parallel. The drive current can become significantly larger. Despite the fact that an additional transistor makes the effective transistor size twice as large as that of 1T-1J, the bitcell size increases only by ~33 % to 16 λ^2, as shown in Fig. 4.7. This is realized through an optimized 2T-1J layout by sharing the source line between neighboring bitcells and therefore eliminating the spacing between the active regions of neighboring bitcells. Compared with the 1T-1J bitcell (Fig. 4.4b) whose bitcell height is often 1.5 times of the gate pitch, the height of the 2T-1J bitcell is increased to 2 times of the gate pitch, thereby increasing the bitcell size by ~33 %.

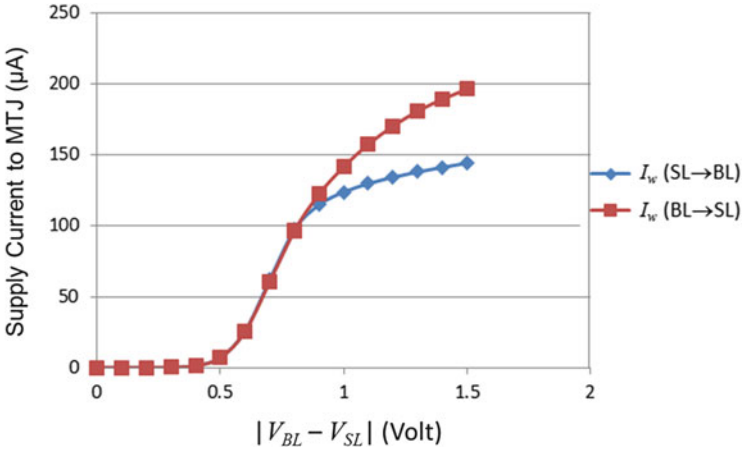
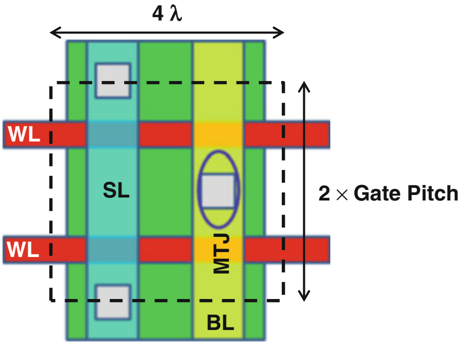


Fig. 4.6 Write supply current (I_w) simulated from a full STT-MRAM chip. Due to the source degeneration effect, I_w is asymmetrical with respect to the polarity of the current

Fig. 4.7 A 2T-1J bitcell layout. While the effective transistor width is doubled versus 1T-1J of Fig. 4.4b, the bitcell size is only ∼33 % larger



In STT-MRAM, I_c has a strong dependence on write pulse width, as illustrated in Fig. 4.8. Fast MTJ switching, often referred to as precessional switching (∼10 ns or below), requires substantially larger I_c than relatively slow switching. This leads to challenges in designing high-performance bitcells. Unless the MTJ size is substantially small, it is often difficult to realize sub-10 ns switching without enlarging the bitcell size. This is a primary reason why continuing innovations in MTJ materials engineering are still desired to reduce J_c. Recent advances have realized reliable switching in 1T-1J below 4 ns with write error rate lower than 10^{-6} [6].

Practically, it is necessary to tailor MTJ and bitcell attributes for varying write speed requirements depending on different STT-MRAM product applications. For example, for embedded Level 2 or Level 3 CPU cache memory, the MTJ switching speed is preferred to be on the order of a few nanoseconds, although this could often be relaxed significantly through various design optimization techniques. In contrast, for traditional embedded nonvolatile memory applications, the switching speed on the order of a microsecond is still compelling (a few orders of magnitudes faster than embedded Flash). An advantage of STT-MRAM is such that MTJ can be tuned for custom bitcells which can serve widely varying ranges of product applications.

4.4 Circuit Design

4.4.1 Write Circuit

The write operation of STT-MRAM is to switch the state of MTJ by supplying current higher than I_c. The polarity of the current determines the switched state, either 0 (R_p) or 1 (R_{ap}). As shown in Fig. 4.9 [22], a write driver is connected to BL and SL, respectively, which acts as a current source or a sink depending on the current polarity. Each write driver is realized by a tri-state inverter. The magnitude of the write supply current (I_w) is determined by the size of the write driver. To write 0, the current flows from the free layer to the pinned layer of MTJ, so that the write driver of BL operates as a current source and that of SL as a current sink. Accordingly, the D value of the write driver (Fig. 4.9c) is high for BL and low for SL. As shown in Fig. 4.8, I_c is a function of write pulse width. I_c becomes higher as the pulse width is shorter. Thus a write enable signal (WET, WEB) should be controlled precisely to prevent write failure (occurring when $I_w < I_c$). On the other hand, a wear-out reliability risk may arise when I_w is too high. Therefore, designing a write driver must consider two factors: precise control of write pulse width and optimal sizing of the driver.

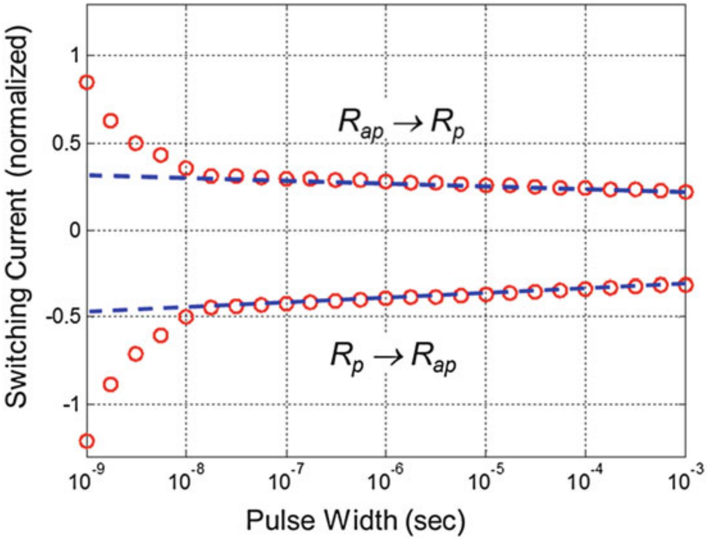


Fig. 4.8 An example characteristic of switching current (I_c) as a function of write pulse (I_w) width

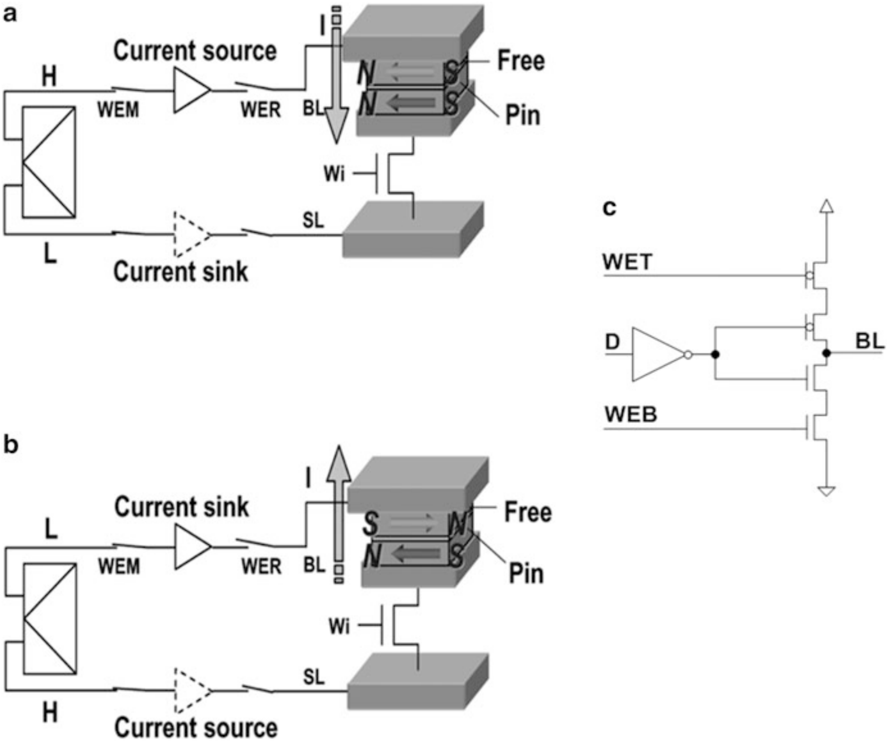


Fig. 4.9 Illustration of STT-MRAM write operation [22]: (**a**) 1 → 0; (**b**) 0 → 1; (**c**) a schematic of a tri-state write driver

Both I_w and I_c are dependent on process variation and can be modeled by Gaussian distributions. For a single cell, the write access pass yield (WAPY), expressed in sigma (standard deviation), is obtained by combining the distributions of I_w and I_c

$$WAPY_{Cell} = \frac{\mu_{I_w} - \mu_{I_c}}{\sqrt{\sigma_{I_w}^2 + \sigma_{I_c}^2}} \tag{4.7}$$

where μ_{I_W} and μ_{I_C} are the mean of I_w and I_c, respectively, and σ_w and σ_c are the standard deviation of I_w and I_c,respectively.

4.4.2 Read Circuit

4.4.2.1 Conventional Sensing Circuit

The read operation of STT-MRAM determines the resistance state of each cell with respect to the predefined state of a reference MTJ array. The operation relies on a sensing circuit and a sense amplifier which converts an output voltage of the sensing circuit to a digital signal. Figure 4.10 shows a conventional sensing circuit designed for MRAM [23]. The circuit is comprised of a data branch and two reference branches. Each branch includes a clamp NMOS (NC_D or NC_R) and a load PMOS (PL_D or PL_R). The sensing current (I_s) is controlled by the gate voltage of clamp NMOS (V_{G_clamp}). The clamp NMOS generates different currents according to the MTJ state 0 or 1. The source voltage of clamp NMOS is fixed in a saturation region. The saturation current of clamp NMOS is high at 0 and low at 1. The two clamp NMOS of the reference branches (NC_R) are designed to generate a saturation current at a medium level between 0 and 1 of the data branch, as shown in Fig. 4.11. The saturation current of NC_R is conveyed to PL_D through a current mirror circuit. Thus,

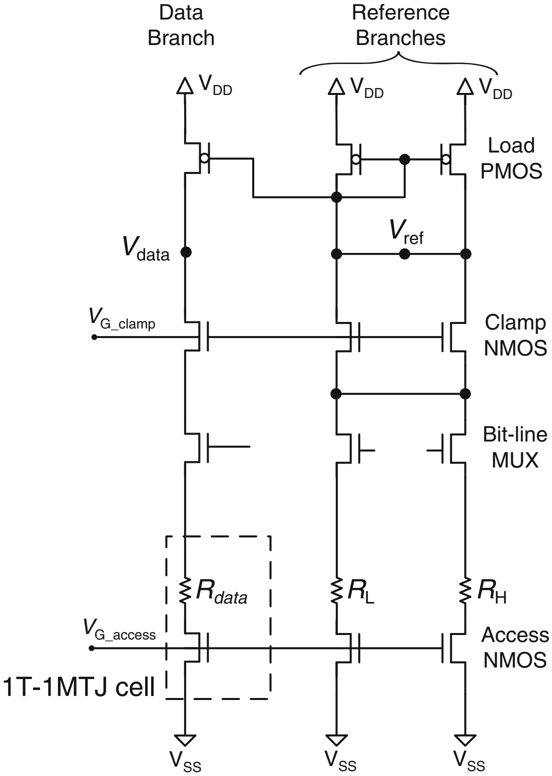


Fig. 4.10 Schematic illustration of a conventional sensing circuit for MRAM [23]

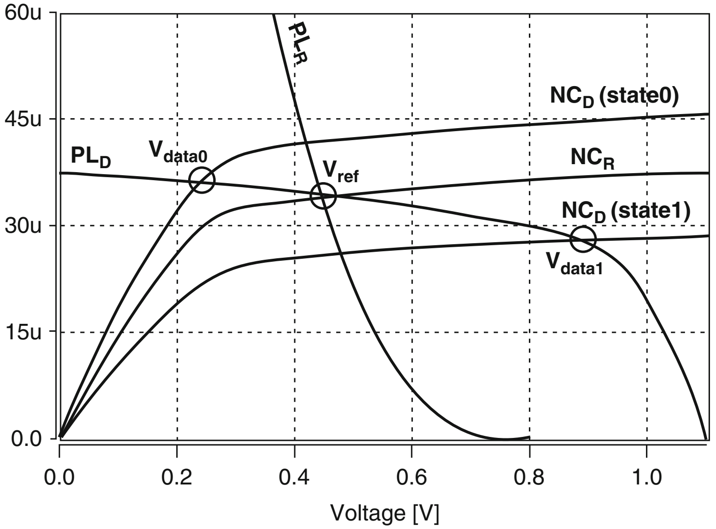


Fig. 4.11 I–V characteristics of clamp NMOS and load PMOS in a conventional sensing circuit

the saturation current of NC_D is larger than that of PL_D for 0 and accordingly V_{data0} is low, whereas the saturation current of NC_D is smaller than that of PL_D for 1 and V_{data1} is high.

The sense amplifier determines 0 or 1 of the data MTJ by comparing output voltages (V_{data0}, V_{data1}, V_{ref}) of the sensing circuit. Read is successful when the difference between V_{data} and V_{ref} ($\Delta V_0 = V_{ref} - V_{data0}$, $\Delta V_1 = V_{data1} - V_{ref}$) is larger than the offset voltage of the sense amplifier (V_{SA_OS}). Note that V_{data} is susceptible to PVT variations. Thus, it is important to design V_{ref} in a way to trace PVT variations of V_{data}.

4.4.2.2 Read Yield

A circuit designer needs to prevent two types of functional failure during read operation. Sensing failure occurs when $\varDelta V_0$ or $\varDelta V_1$ is smaller than V_{SA_OS}. Read disturbance failure is possible when I_s exceeds I_c (i.e. unintentional switching during sensing). Considering these two, a statistical read yield model can be built in the following way.

The statistical distributions of $\varDelta V_0$, $\varDelta V_1$, and V_{SA_OS} can be modeled by Gaussian distributions. For a single cell, the read access pass yield for 0 or 1 (RAPY$_{\mathrm{Cell0}}$ or RAPY$_{\mathrm{Cell1}}$) is given by [24]

$$RAPY_{Cell0,1} = \frac{\mu_{\Delta V_{0,1}} - \mu_{SAOS}}{\sqrt{\sigma^2_{\Delta V_{0,1}} + \sigma^2_{SAOS}}} \tag{4.8}$$

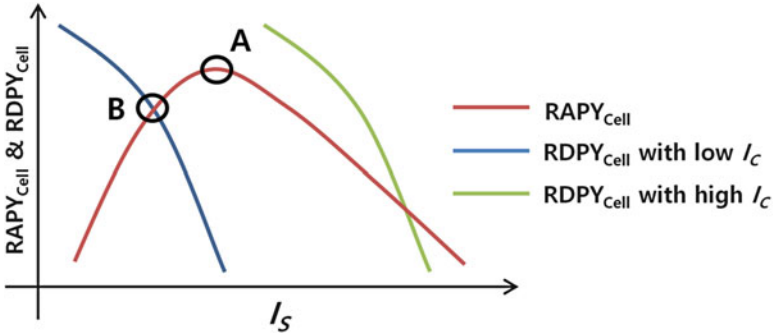


Fig. 4.12 Read yield ($\text{RAPY}_{\text{Cell}}$ and $\text{RDPY}_{\text{Cell}}$) trends as a function of I_s and I_c

where $\mu_{\Delta V_{0,1}}$ and μ_{SA_OS} are the mean of $\varDelta V_{0,1}$ and V_{SA_OS}, respectively, and $\sigma_{\Delta V_{0,1}}$ and σ_{SA_OS} are the standard deviation of $\varDelta V_{0,1}$ and V_{SA_OS}, respectively. $\text{RAPY}_{\text{Cell}}$ is then defined as the smaller of $\text{RAPY}_{\text{Cell0}}$ and $\text{RAPY}_{\text{Cell1}}$.

$$RAPY_{Cell} = \min\left(RAPY_{Cell0},\ RAPY_{Cell1}\right) \tag{4.9}$$

The criterion for read disturbance failure is $I_s \geq I_c$. Thus, the read disturbance pass yield (RDPY) is given by:

$$RDPY_{Cell} = \frac{\mu_{I_c} - \mu_{I_s}}{\sqrt{\sigma_{I_c}^2 - \sigma_{I_s}^2}} \tag{4.10}$$

I_s has large effects both on RAPY and on RDPY. As illustrated in Fig. 4.12, $\text{RDPY}_{\text{Cell}}$ decreases as I_s increases. In addition, $\text{RDPY}_{\text{Cell}}$ is lower when I_c is lower. On the other hand, there is an optimum I_s which maximizes $\text{RAPY}_{\text{Cell}}$ because it is difficult to achieve small $\sigma_{\Delta V_{0,1}}$ and large $\mu_{\Delta V_{0,1}}$ when I_s is too low and too high, respectively. Therefore, depending on I_c, different design strategies are applicable to maximize read yield. For high I_c, $\text{RDPY}_{\text{Cell}}$ is also high, so that I_s is tuned to maximize $\text{RAPY}_{\text{Cell}}$ (Point A). For low I_c, it is desired to find I_s to make $\text{RDPY}_{\text{Cell}}$ and $\text{RAPY}_{\text{Cell}}$ equal (Point B). In general, I_c continually scales down as the feature size shrinks, which means that controlling read disturb yield is becoming of great significance.

4.4.3 Advanced Sensing Circuits

Assuring adequate sensing margin ($\varDelta V_0$ and $\varDelta V_1$) for STT-MRAM at deeply scaled nodes necessitates extensive design efforts owing to the decrease in supply voltage and the increase in process variation. Further, STT-MRAM must be designed to avoid potential read disturbance, desiring a low-current sensing method. To solve these challenges, various types of advanced sensing circuits have been developed.

4.4.3.1 Source Degeneration PMOS

The load PMOS and the clamp NMOS shown in Fig. 4.10 can become a significant source of process variation for the sensing circuit. The clamp NMOS has a large source resistance, and it operates as a source degeneration resistance to keep the current through the clamp NMOS as constant as possible. On the other hand, the source of the load PMOS is directly connected to a voltage supply (V_{DD}), leading to large current variation. To reduce the variation effect of the load PMOS, a source degeneration scheme can be adopted by inserting a degeneration PMOS between the source of the load PMOS and the voltage supply, as shown in Fig. 4.13 [24]. This is relatively a simple method to increase read yield by minimizing the variation of ΔV_0 and ΔV_1 caused by the process variation of the load PMOS.

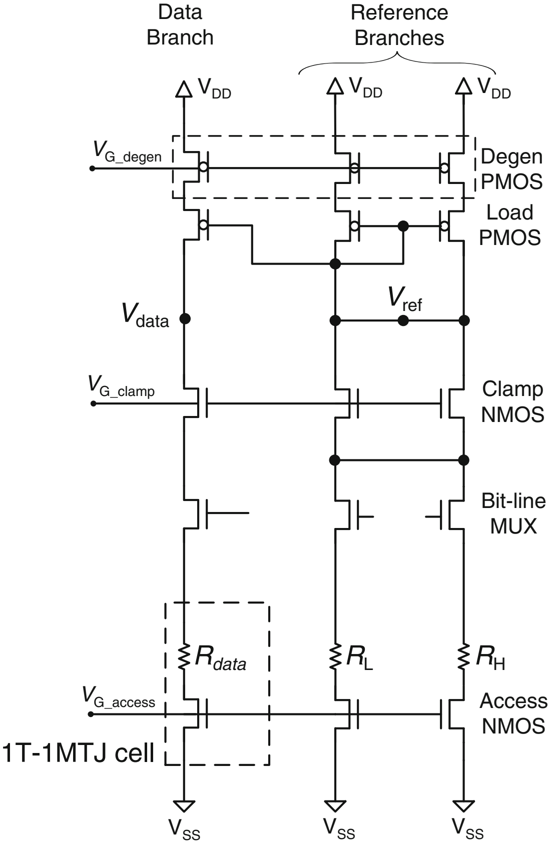


Fig. 4.13 A sensing circuit that employs degeneration PMOS [24]

4.4.3.2 Self Body Biasing

Assuring sensing margin (ΔV_0 and ΔV_1) is more challenging at lower V_{DD} and higher V_{TH} (threshold voltage), i.e., when the voltage headroom (V_{DD}–V_{TH}) is smaller. Note that high V_{TH} transistors are widely adopted for low standby power applications. Figure 4.14 describes a sensing circuit that can mitigate this challenge by utilizing self body biasing [25]. The body bias can decrease V_{TH} of the load PMOS when the sensing circuit is active. This helps secure the sensing margin during read operation while not causing high leakage current at the standby mode. In addition, it internally generates a body voltage without utilizing a body voltage generator, so that the area overhead is minimal compared with the conventional sensing circuit (Fig. 4.10). This scheme can be coupled with the degeneration PMOS described in Sect. 4.4.3.1.

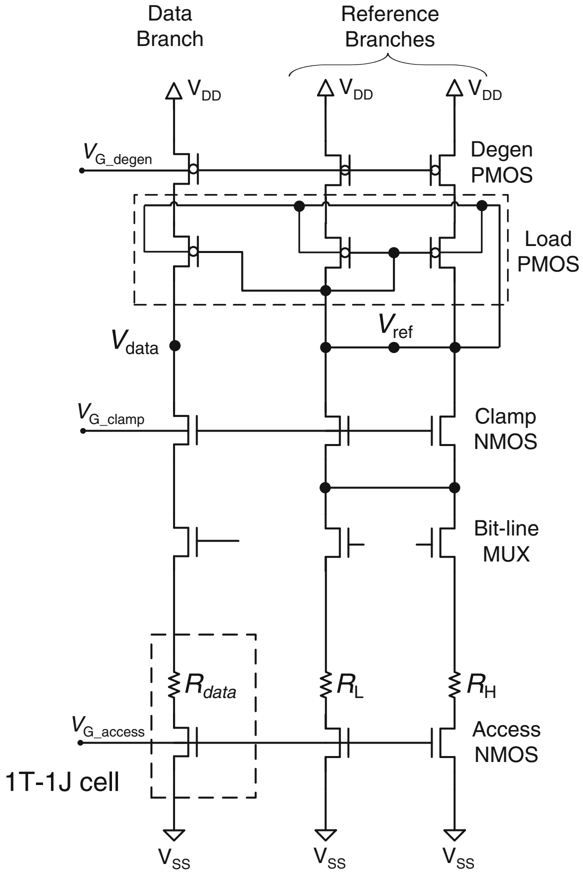


Fig. 4.14 A sensing circuit that employs self-body biasing [25]

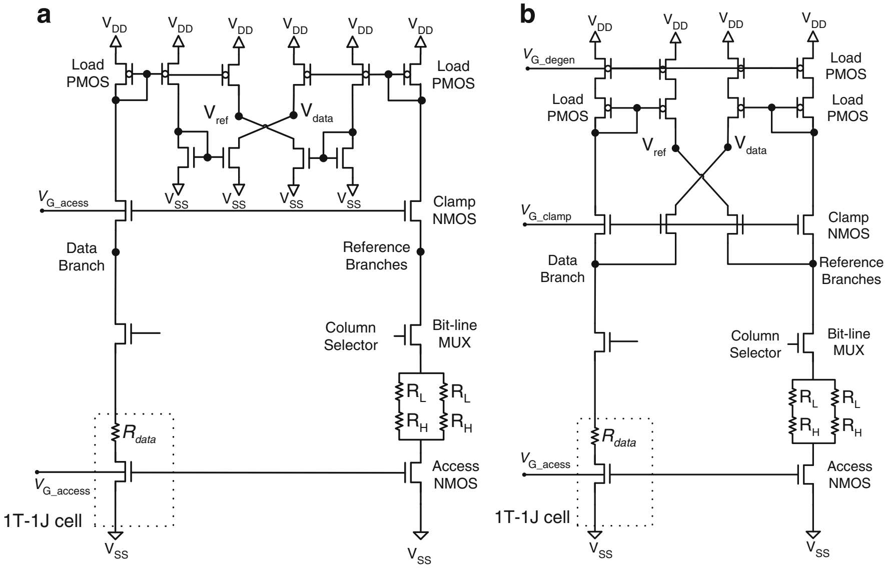


Fig. 4.15 (**a**) A sensing circuit with a symmetric cross-coupled current mirror; (**b**) a split-path sensing circuit [26]

4.4.3.3 Split-Path Sensing

A sensing circuit can adopts variable V_{ref} to increase ΔV_0 and ΔV_1. V_{ref} is modulated according to the MTJ state. At 0, V_{ref} increases, so does ΔV_0. Whereas at 1, V_{ref} decreases, hence ΔV_1 increases. Figure 4.15a shows a conventional sensing circuit with variable V_{ref} with a symmetric cross-coupled current mirror. A drawback of this circuit is that the mismatch occurred during such current mirroring increases the standard deviation of ΔV_0 and ΔV_1. An alternative scheme has been proposed by adopting a split-path sensing circuit, shown in Fig. 4.15b [26]. The split path enables variable V_{ref} while minimizing the number of current mirrors. The variable V_{ref} enhances $\mu_{\Delta V_{0,1}}$ by doubling it. The minimized number of current mirrors reduces $\sigma_{\Delta V_{0,1}}$.

4.4.3.4 Offset-Canceling Triple-Stage Sensing

One of emerging challenges for STT-MRAM sensing circuits is to reduce I_s while maintaining sensing margins. This is attributed to rapid reduction in I_c with p-MTJ (owing to high STT efficiency addressed in Sect. 4.2) and also with a demand for smaller MTJ (e.g. diameter <20 nm). Reduced I_s results in larger $\sigma_{\Delta V_{0,1}}$, which directly reduces sensing yields. The impact of $\sigma_{\Delta V_{0,1}}$ is often greater than $\mu_{\Delta V_{0,1}}$. Figure 4.16 illustrates an offset-canceling triple-stage (OCTS) sensing circuit [27].

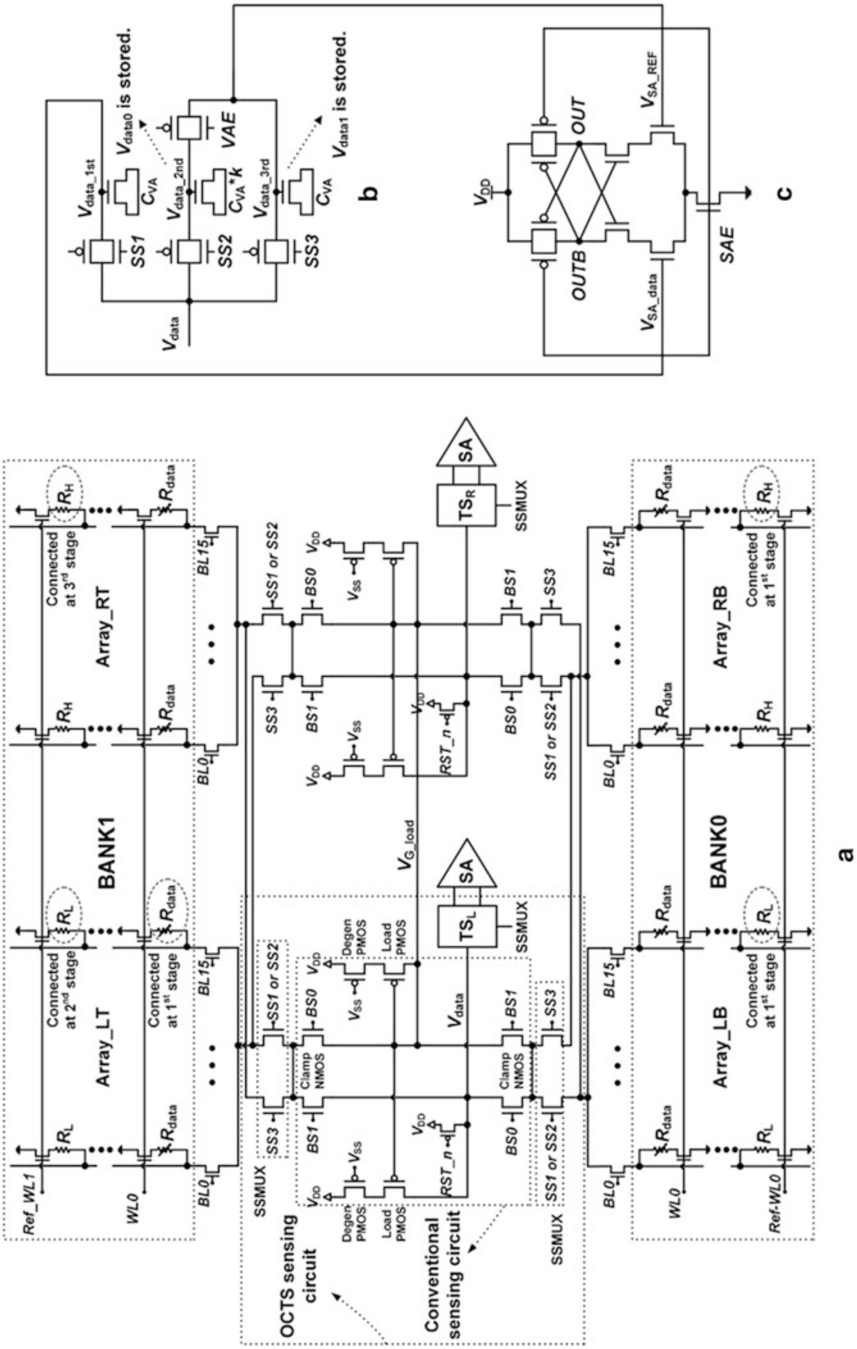


Fig. 4.16 An offset-canceling triple-stage (OCTS) sensing circuit. Diagrams of (**a**) an OCTS sensing circuit in a simplified array architecture; (**b**) the TS_L block of (**a**); (**c**) the SA block of (**a**) [27]

This is designed for reducing $\sigma_{\Delta V_{0,1}}$ by canceling the offsets of the sensing circuit caused by process variations. The principle of OCTS is to sense progressively three cells that are the data cell and the two reference cells with 0 and 1 through one sensing circuit. The output voltages are then added or subtracted to cancel out the offsets.

A drawback of OCTS is such that there is only one sensing circuit sequentially to read data and reference cells. Hence it is difficult to avoid a read speed penalty, though this may be tolerable for most applications. In addition, gate capacitors are required to store the sensed value at each stage, which increases the array size.

4.4.3.5 Self-Reference Sensing

OCTS can effectively cancel out the offsets of the sensing circuit, but cannot improve the sensing yield related to MTJ process variation. Self-reference circuits, shown in Fig. 4.17, generate V_0, V_1, and V_{ref} with only one MTJ. Such sensing circuits minimize the offsets caused not only by the sensing circuit but also by MTJ process variation. Figure 4.17a is a relatively simple self-reference scheme. It reads the MTJ cell and store V_0 and V_1 in capacitance at the first stage. At the second stage, the MTJ cell is written to 0 by a larger write current than that of the first stage. The sensing circuit reads this cell to generate V_{ref}. Because the stored information of the MTJ is removed during the read operation, this method is destructive, so that the readout value must be written back to the MTJ at the last step. This degrades the read speed and consumes more energy. In contrast, an alternative scheme, which is nondestructive, is shown in Fig. 4.17b [28]. This scheme allows maintaining the MTJ state after the first sensing. When the MTJ is at 0, the resistance of the cell is nearly constant regardless of the current flowing through the MTJ cell. The resistance change is detected by current when the MTJ turns into 1. Accordingly, this nondestructive self-reference scheme generates V_{ref} without write and write-back processes, overcoming the drawbacks of the circuit in Fig. 4.17a. However, it is difficult to secure the resistance difference when the current difference is subtle, which may cause a challenge for ever decreasing operating current requirement with scaling.

4.4.4 *Array Architecture*

In general, memory array architecture is an essential design parameter that influences performance, power consumption, yield, and chip size. Determining an optimal array architecture is therefore dependent on bitcell specification, chip specification, target yield, and even reliability. The array efficiency, a ratio of bitcell array area over total memory area including peripheral circuits, becomes higher as the array size increases. But, this leads to degradation in performance because parasitic resistances and capacitances increase owing to the increase in the number

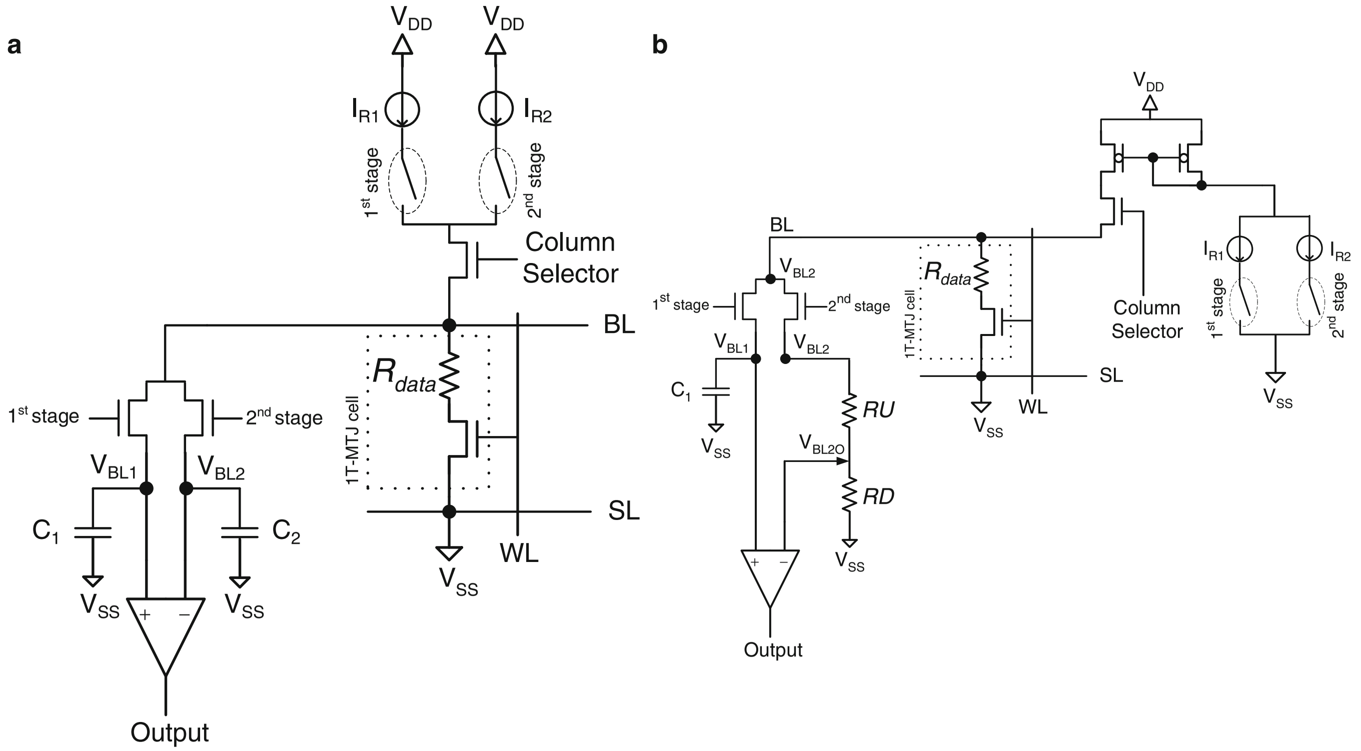


Fig. 4.17 Self-reference sensing circuits: (**a**) conventional scheme; (**b**) non-destructive scheme [28]

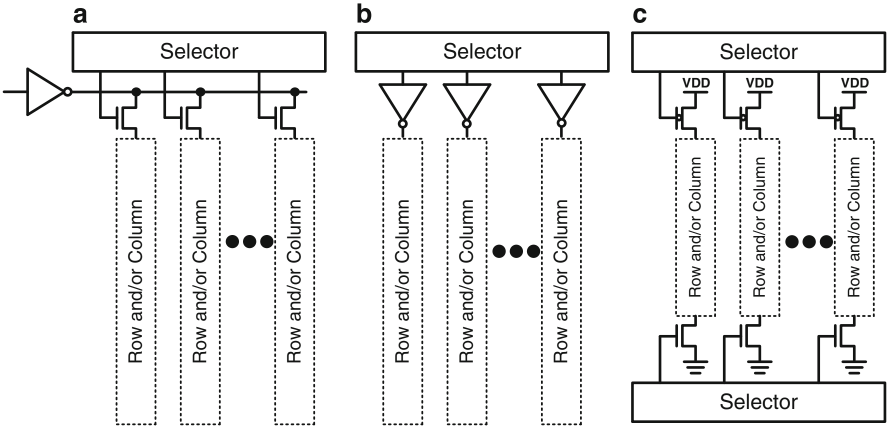


Fig. 4.18 Multiplexer architectures: (**a**) single write driver; (**b**) merged writer drivers; (**c**) separated write driver

of cells connected to BL. As shown above, read circuit design is more challenging for STT-MRAM, so that an effort to increase the array size must be considered cautiously in a way not to degrade read performance.

4.4.4.1 Multiplexer (MUX) Architecture

STT-MRAM employs a different MUX structure compared with conventional memories. For the read operation of STT-MRAM, selected BL is connected to the sensing circuit and SL to the ground (*GND*). For the write operation of 0, the write driver connected to selected BL drives V_{DD} and the driver connected to SL drives *GND*, and vice versa for 1. As shown in Fig. 4.18, these read and write operations can be enabled by different types of MUX structures. Figure 4.18a is a common architecture to achieve a small footprint since only one write driver is required for all BL and SL. However, the current originated from the write driver must pass through the MUX, hence, reducing I_w. Figure 4.18b shows a merged write driver for which selection control utilizes an independent write driver for each BL and SL. This results in a larger area, but provides higher I_w owing to the absence of MUX along the write path. Figure 4.18c illustrates selectors to drive V_{DD} and *GND* from both terminals of BL and SL by separating a write driver. The area becomes even larger, however, this MUX can realize higher yield by controlling parasitic mismatches because the lengths of BL and SL are the same for all the cells.

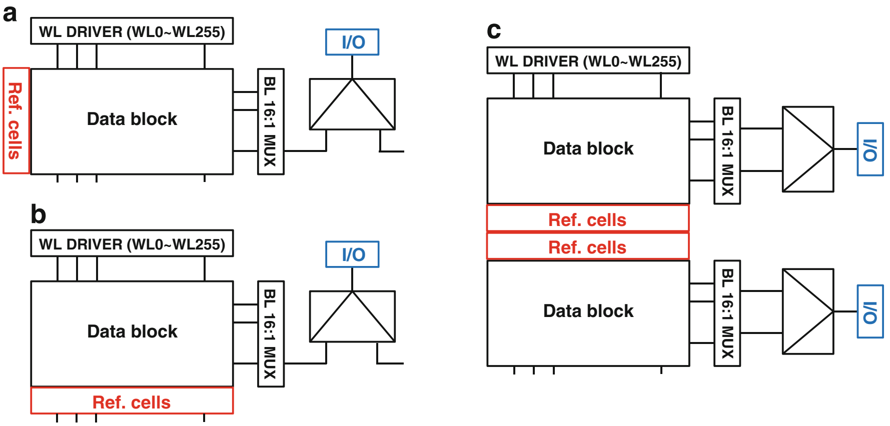


Fig. 4.19 Reference cell array architectures: (**a**) row reference; (**b**) column reference; (**c**) shared column reference

4.4.4.2 Reference Cell Architecture

STT-MRAM reference cell architectures are generally categorized according to the position of reference cells, as shown in Fig. 4.19. An array structure which places all reference cells into one WL (Fig. 4.19a) has an advantage in achieving small footprint. But, this architecture is not preferable for high-capacity STT-MRAM owing to parasitic mismatches between reference cells and data cells. In comparison, the reference architecture in Fig. 4.19b, which places all reference cells along one BL, has an advantage of achieving higher read yield. There is an area penalty associated with this, though tolerable. While each data block ordinarily has its own reference cell array, it is also possible to design a shared reference scheme for which two data blocks share two adjacent reference cells (Fig. 4.19c).

4.5 Co-design of MTJ and Logic

Designing an IC at an advanced technology node with embedded STT-MRAM requires a proven statistical circuit model which addresses systematic and random variations of MTJ and logic [29]. It is important to understand the challenges imposed by deep scaling of the logic technology. Recent work addressed a first-of-its-kind statistical circuit model and its application for designing a STT-MRAM building block and its array [30]. Figure 4.20 illustrates key components of this co-design methodology. The model can be seamlessly integrated into a common CMOS circuit simulation environment. The statistical variability-aware model fits Si data by covering PVT variations. The model is also combined with a micromagnetic physical model, hence, allowing co-optimization of MTJ physical parameters and

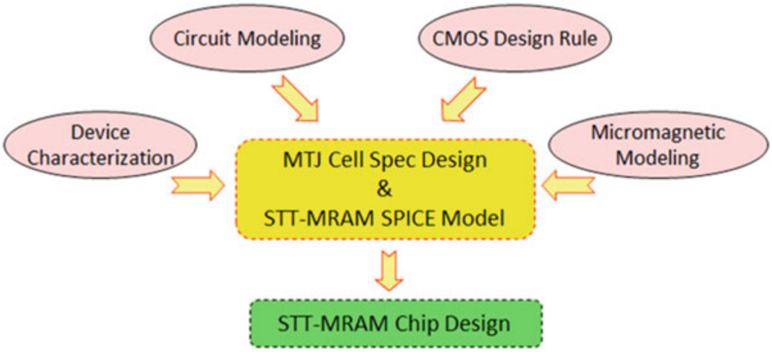


Fig. 4.20 Conceptual illustration of key components of a CMOS-MTJ co-design methodology [30]

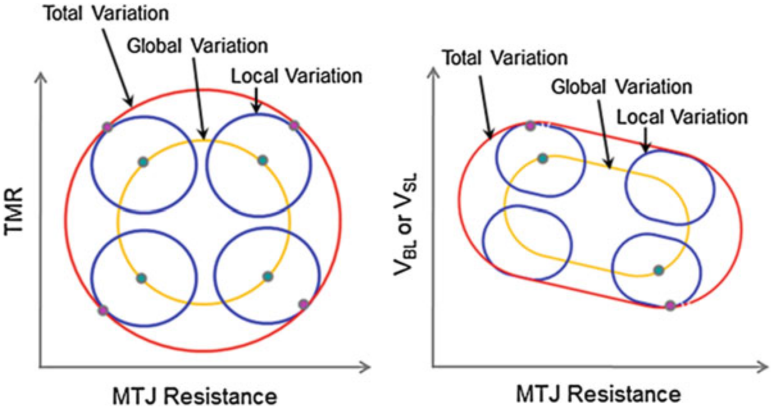


Fig. 4.21 Illustration of STT-MRAM statistical variation models which correlate key device parameters such as *R*, TMR, and V_c [29, 30]

cell circuit parameters. Hence, the model is applied to tune device parameter specifications required to meet target chip performance, yield, and reliability. Correlations among critical design parameters are systematically examined by the model. An example is shown in Fig. 4.21. By performing statistical Monte Carlo simulations, the model is capable of predicting an array functional yield. The model correlates various functional failure modes to physical cell defects or circuit-design errors. The accuracy of the model has been validated by chip-level functionality and yield data [31].

4.6 Perspective

Modern SOC memory subsystems are diverse or complicated, so that difficult to be served by one prevalent memory. It is desirable to optimize each SOC platform by a different combination of memory attributes such as speed, power consumption, reliability, and cost. In this aspect, STT-MRAM is attractively positioned since its building block MTJ can be tuned for a broad range of memory attributes which can serve largely different types of SOC applications. For example, low-power STT-MRAM can become an ideal embedded nonvolatile memory for battery-powered wireless connectivity networks pertaining to Internet-of-Things and wearable electronics, not only by storing nonvolatile codes, but also by storing and executing fast data [32, 33]. This type of STT-MRAM simplifies the conventional memory subsystem and also extends battery life. In addition, its logic-friendly design and process compatibility can realize such benefits at advanced logic nodes for which it is difficult to employ conventional embedded nonvolatile memory technology. On the other hand, high-performance ($<\sim$5 ns) STT-MRAM can serve as an alternative to SRAM. Despite the fact that STT-MRAM is slower than SRAM at a discrete circuit level, the memory subsystem can be architected in a way that the performance can be comparable or even better at a system level. Furthermore, there is a significant range of custom SRAM for which its leakage power and cost (chip area) are critical drawbacks. One emerging case is Level-3 cache for mobile CPU. Moreover, even higher performance STT-MRAM potentially realized in custom-designed bitcells and circuits may move up to a higher level of memory hierarchy (Level-2 cache). In addition, high-throughput embedded STT-MRAM can become an attractive memory for GPU by providing higher on-chip memory density at lower energy consumption. Finally, the MTJ applied for STT-MRAM can be utilized for security and anti-tampering applications. Examples include one-time programmable memory, random number generator, and physically unclonable function.

References

1. Kang SH, Lee K. Emerging materials and devices in spintronic integrated circuits for energy-smart mobile computing and connectivity. Acta Mater. 2013;61:952–73.
2. Hosomi M, Yamagishi H, Yamamoto T, et al. A novel nonvolatile memory with spin torque transfer magnetization switching: spin-RAM. IEDM Tech Dig. 2005;2005:459–62.
3. Lin CJ, Kang SH, Wang YJ, et al. 45 nm low power CMOS logic compatible embedded STT MRAM utilizing a reverse-connection 1T/1MTJ cell. IEDM Tech Dig. 2009;2009:279–82.
4. Ikeda S, Miura K, Yamamoto H, et al. A perpendicular-anisotropy CoFeB-MgO magnetic tunnel junction. Nat Mater. 2010;9:721–4.
5. Rizzo ND, Houssameddine D, Janesky J, et al. A fully functional 64 Mb DDR3 ST-MRAM built on 90 nm CMOS technology. IEEE Trans Magn. 2013;49(7):4441–6.
6. Thomas L, Jan G, Zhu J, et al. Perpendicular STT-MRAM with high spin-torque efficiency and thermal stability for embedded memory applications. J Appl Phys. 2014;155(17):172615
7. Sekikawa M, Kiyoyama K, Hasegawa H, et al. A novel SPRAM-based reconfigurable logic block for 3D-stacked reconfigurable spin processor. IEDM Tech Dig. 2008;2008:1–3.
8. Ohno H. A hybrid CMOS/magnetic tunnel junction approach for nonvolatile integrated circuits. In: VLSI Technology Symposium, 2009, p. 122–23.

9. Ohno H, Endoh T, Hanyu T, et al. Magnetic tunnel junction for nonvolatile CMOS logic. IEDM Tech Dig. 2010;2010:9.4.1–4.
10. Ando K. Nonvolatile magnetic memory. J Fed. 2001;12:89–95.
11. Ando K, Ikegawa S, Abe K, et al. Roles of non-volatile devices in future computer systems: normally-off computers. In: Energy-aware systems and networking for sustainable initiatives. Hershey: IGI Global; 2012. p. 83–907.
12. Kawahara T. Scalable spin-transfer torque RAM technology for normally-off computing. IEEE Des Test Comput. 2011;28(1):52–63.
13. Jullière M. Tunneling between ferromagnetic films. Phys Lett A. 1975;54(3):225–6.
14. Jaffrès H, Lacour D, Nguyen Van Dau F, et al. Angular dependence of the tunnel magnetoresistance in transition-metal-based junctions. Phys Rev B. 2001;64:064427.
15. Ikeda S, Hayakawa J, Ashizawa Y, et al. Tunnel magnetoresistance of 604% at 300 K by suppression of Ta diffusion in CoFeB/MgO/CoFeB pseudo-spin-valves annealed at high temperature. Appl Phys Lett. 2008;93:082508.
16. Choi YS, Tsunematsu H, Yamagata S, et al. Novel stack structure of magnetic tunnel junction with MgO tunnel barrier prepared by oxidation methods: preferred grain growth promotion seed layers and bi-layered pinned layer. Jpn J Appl Phys. 2009;48:120214.
17. Maehara H, Nishimura K, Nagamine Y, et al. Tunnel magnetoresistance above 170% and resistance–area product of 1 Ω-μm^2 attained by in-situ annealing of ultra-thin MgO tunnel barrier. Appl Phys Express. 2011;4:033002.
18. Slonczewski JC. Current-driven excitation of magnetic multilayers. J Magn Magn Mater. 1996;159:L1–7.
19. Berger. Emission of spin waves by a magnetic multilayer traversed by a current. Phys Rev B. 1996;54:9353–8.
20. Diao Z, Panchula A, Ding Y, et al. Spin transfer switching in dual MgO magnetic tunnel junctions. Appl Phys Lett. 2007;90:132508.
21. Lee YM, Yoshida C, Tsunoda K et al. Highly scalable STT-MRAM with MTJs of top-pinned structure in 1T/1MTJ cell. In: VLSI Technology Symposium, 2010, p. 49–50.
22. Kawahara T. 2 Mb SPRAM with bit-by-bit bi-directional current write and parallelizing-direction current read. IEEE J Solid-State Circuits. 2008;43(1):109.
23. Maffitt TM. Design considerations for MRAM. IBM J Res Dev. 2006;50(1):25.
24. Kim J, Ryu K, Kang SH, et al. A novel sensing circuit for deep submicron spin transfer torque MRAM. IEEE Trans Very Large Scale Integr Syst. 2012;20(1):181–6.
25. Kim J, Ryu K, Kim JP et al. An STT-MRAM sensing circuit with self-body biasing in deep submicron technologies. IEEE Trans Very Large Scale Integr Syst. 2014;22(7):1630-4 doi:10.1109/TVLSI.2013.2272587.
26. Kim J, Na T, Kim JP. A split-path sensing circuit for spin-torque transfer MRAM. IEEE Trans Circuits Syst, 2014. doi:10.1109/TCSII.2013.2296136.
27. Na T, Kim J, Kim JP et al. An offset-canceling triple-stage sensing circuit for deep submicrometer STT-RAM. IEEE Trans Very Large Scale Integr Syst. 2014;22(7):1620-4. doi:10.1109/TVLSI.2013.2294095.
28. Chen Y. A nondestructive self-reference scheme for spin-transfer torque random access memory. In: Design, Automation and Test in Europe Conference and Exhibition (DATE), 8–12 March 2010, p. 148–53.
29. Zhu X, Kang SH. Spin-transfer-torque MRAM: device architecture and modeling. In: Wang X, editor. Metallic spintronics devices. CRC, 2014 p. 21–70.
30. Zhu X, Kang SH. Variation-aware device modeling and design for embedded STT-MRAM array. In: 55th MMM Conference HC-13, 2010
31. Kim JP, Kim T, Hao W et al. A 45nm 1Mb embedded STT-MRAM with design techniques to minimize read-disturbance. In: VLSI Circuits Symposium, 2011, p. 296–97.
32. Kang SH. Embedded STT-MRAM for energy-efficient and cost-effective mobile systems. In: VLSI Technology Symposium, 2014, p. 36–7.
33. Lee K, Kan JJ, Kang SH. Unified embedded non-volatile memory for emerging mobile markets. In: ISLPED, 2014, p. 131–6.