7 On-Chip Non-volatile STT-MRAM for Zero-Standby Power

Xuanyao Fong and Kaushik Roy

In this Chapter, we present spin-transfer torque magnetic random access memory (STT-MRAM) suitable for IoT applications. Its ability to operate at low supply voltages, non-volatility, good endurance, and small bit-cell footprint are especially attractive for IoT applications in which low energy consumption is crucial. We will present the fundamentals of STT-MRAM. The design of the STT-MRAM storage device, memory bit-cell and memory array architecture are also discussed to highlight the benefits STT-MRAM brings to IoT applications, as well as the design issues that need to be considered. We then present a device/circuit/architecture co-design approach for STT-MRAM. Finally, we will discuss the trends in STT-MRAM and give some perspectives on the future of STT-MRAM design.

7.1 Introduction

As discussed in Chap. 6, several non-volatile memory technologies such as embedded flash (eFlash) and resistive RAM (ReRAM) are available for implementing memories with truly random access for IoT applications (Yamauchi et al. 2015). However, STT-MRAM is especially attractive as compared to ReRAM and eFlash due to its relative ease of integration into the back-end-of-line (BEOL) in the CMOS fabrication process, ability to operate at <1.2 V supply voltages, <100 ns read and write delays, good endurance ($>10^{14}$ cycles) and bit-cell footprint as small as 40 F^2 (F is the smallest feature size of the CMOS technology) (ITRS Roadmap 2014). In this chapter, we discuss the modeling, design, and optimization of STT-MRAMs and present some of the potential benefits in relation to IoT applications. As we will see later in this chapter, a highly desirable trait of STT-MRAM for IoT applications is that they only need to be powered when they are being accessed, which results in zero-standby power.

7.2 The Magnetic Tunnel Junction and STT-MRAM

The storage device in STT-MRAM is the magnetic tunnel junction or MTJ. The structure of a typical MTJ with in-plane magnetic anisotropy is shown in Fig. 7.1. The MTJ may be visualized as a stack consisting of two nano-magnets sandwiching a tunneling oxide barrier (usually AlO_x or more commonly MgO). One nano-magnet is a soft ferromagnetic layer used to store the information (also called the "*free*"

X. Fong (✉)
National University of Singapore, Singapore 119077, Singapore
e-mail: elefongx@nus.edu.sg

K. Roy
Purdue University, West Lafayette, IN 47907, USA

© Springer International Publishing AG 2017
M. Alioto (ed.), *Enabling the Internet of Things*, DOI 10.1007/978-3-319-51482-6_7

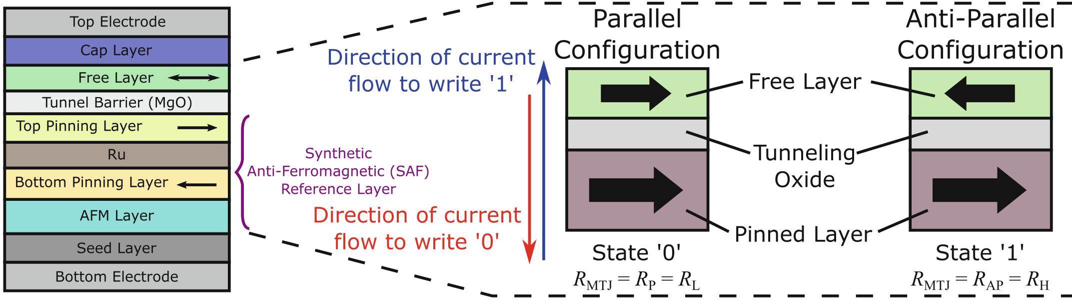


Fig. 7.1 The storage device in the STT-MRAM is the magnetic tunnel junction (MTJ). The typical stack structure of an MTJ with in-plane anisotropy is shown. It is easier to understand the operation of an MTJ by only considering the simple tri-layer stack illustrating the parallel and anti-parallel MTJ configurations. The directions of programming current flow through the bit-cell are shown using colored arrows whereas the black arrows indicate possible magnetization directions of magnetic layers in the MTJ

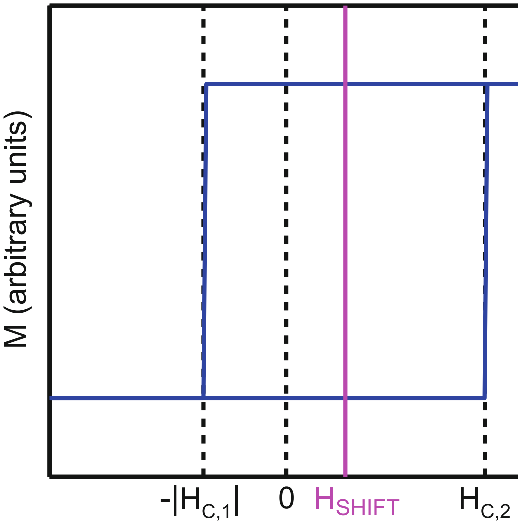


Fig. 7.2 The *M-H* loop of an MTJ without SAF based pinned layer exhibits a shifted hysteresis loop as illustrated here, which corresponds to a degraded retention time

layer) whereas the other nano-magnet is a hard ferromagnetic layer for use as a reference layer (also called the "*fixed*" or "*pinned*" layer).

In many MTJ stacks, a synthetic anti-ferromagnetic (SAF) layer is used for the pinned layer to reduce the stray magnetic field. The stray magnetic field may shift the *M-H* loop of the MTJ as illustrated in Fig. 7.2. The *M-H* loop (*M*: magnetization, *H*: applied magnetic field) exhibits a horizontal shift, H_{SHIFT}. $H_{C,1}$ and $H_{C,2}$ are symmetric about H_{SHIFT} (i.e., $H_{C,2} - H_{SHIFT} = H_{SHIFT} - |H_{C,1}|$). However, the thermal stability and retention time of the MTJ is determined by *min* $(|H_{C,1}|, |H_{C,2}|)$. Thus, non-zero H_{SHIFT} reduces the thermal stability and retention time of the MTJ.

An example of a SAF layer is shown in Fig. 7.1. The anti-ferromagnetic (AFM) layer pins the magnetization of the bottom pinning layer via exchange bias effect (Nogués et al. 2005). The top pinning layer is anti-ferromagnetically coupled to the bottom pinning layer via interlayer exchange coupling with a non-magnetic spacer such as *Ru* (Parkin et al. 1990). Exchange coupling increases the magnetic field needed to switch the magnetizations of the coupled pinning layers and effectively pins the magnetizations of both ferromagnetic layers. This allows the magnetization of the top pinning layer to be used as a reference.

The MTJ is designed to be switchable between two stable states. When the magnetization directions of both the free and the pinned layers point in the same direction, the MTJ configuration is called the "*parallel*" state (P). When instead the magnetization directions of the free layer and the pinned layer point in opposite directions, the MTJ configuration is called the "*anti-parallel*" state (AP). An important metric for the MTJ is its resistance-area (*RA*) product (Huai 2008). The *RA* product of the MTJ varies exponentially with the thickness of the tunneling oxide barrier (t_{MgO}) since the mechanism for electron transport is tunneling. The MTJ

resistance, R_{MTJ}, depends linearly on the cross-sectional area of the MTJ (A_{MTJ}) similar to an Ohmic conductor when t_{MgO} is constant. R_{MTJ} also depends on the relative magnetic polarization of the free layer with respect to the pinned layer. The dependence of R_{MTJ} on magnetic polarization arises due to the difference in density of states around the Fermi energy, E_F, in the ferromagnetic layers (Datta et al. 2012). When the MTJ is in the P configuration, the density of states of like-spins around E_F is very high in the ferromagnetic layers. Conversely, the density of states of like-spins around E_F in the ferromagnetic layers is very low when the MTJ is in AP configuration. Thus, R_{MTJ} is low in the P configuration ($R_{MTJ} = R_P = R_L$) and high in the AP configuration ($R_{MTJ} = R_{AP} = R_H$). This difference in R_{MTJ}, termed the "*tunneling magneto-resistance ratio*" (or *TMR*), is quantified as

$$TMR = \frac{R_{AP} - R_P}{R_P} \times 100\% \qquad (7.1)$$

and is an important metric for the performance of MTJs as memory elements. A larger *TMR* also means that MTJ states can be distinguished more easily. Binary data may then be represented and stored as the resistance state of the MTJ.

The magnetic layers in an MTJ, which may be considered as nano-magnets, are engineered with anisotropy energies to satisfy thermal stability and data retention requirements. The required energy barrier height (E_B, in Joules) and retention time (τ, in seconds) must satisfy:

$$\Delta = \frac{E_B}{k_B T} > ln\left(\frac{m\tau}{\tau_0\, ln\, 2}\right) \qquad (7.2)$$

where k_B is the Boltzmann constant, T is the temperature in Kelvin, m is the number of bits in the memory and τ_0 is the characteristic time in seconds ($\tau_0 \approx 1$ ns). $\Delta = 40.66$ corresponds to a retention time of about 10 years for $m = 1$. In real STT-MRAM arrays, $\Delta > 70$ may be required (Naeimi et al. 2013). The minimum magnetic field for switching a nano-magnet, H_k, and the energy barrier are related by (Sun 2000):

$$H_k = \frac{2E_B}{\mu_0 M_{sat} V_{FL}} \qquad (7.3)$$

where μ_0 is the permeability of vacuum, and M_{sat} and V_{FL} are the saturation magnetization and volume of the nano-magnet, respectively. E_B of the pinned layer in the MTJ is engineered to be much larger than that of the free layer in the MTJ so that the pinned layer magnetization direction is fixed. As such, only the magnetization direction of the free layer can change during operation. The most common form of anisotropy engineered into the free layer of an MTJ is the *uniaxial anisotropy*. This causes the magnetization of the magnetic layers to have a preferential alignment axis—the magnetization will align along this axis when no external stimulus is present. The *uniaxial anisotropy energy density*, K_{u2}, and E_B of the nano-magnet are related by

$$K_{u2} V_{FL} = E_B \qquad (7.4)$$

Hence, $K_{u2} V_{FL}$ must be kept constant to maintain the same thermal stability when the volume of the nano-magnet is reduced. We will discuss this in more detail in the later sections. In the presence of a stray magnetic field that shifts the *M-H* loop of the MTJ by H_{SHIFT} (as in Fig. 7.2), the effective barrier height becomes

$$E_B = 0.5\mu_0 M_{sat} V_{FL} (H_k - H_{SHIFT}) \qquad (7.5)$$

Hence, the free layer of the MTJ needs to be engineered with a larger K_{u2} to compensate for the barrier height degradation due to H_{SHIFT}.

7.2.1 Spin-Transfer Torque

Nano-scale MTJs may be switched using the spin-transfer torque (STT) phenomenon shown in Fig. 7.3a, which was theoretically predicted by Slonczewski and Berger independently in 1996 (Berger 1996; Slonczewski 1996). Since then many experiments have observed STT switching (Huai et al. 2004; Katine et al. 2000; Myers 1999). STT arises due to the spin property of electrons. When the majority of electron spins in a nano-magnet are aligned in a particular direction, the magnetization of the nano-magnet also points in that direction. This is illustrated by the different density of states for different electron

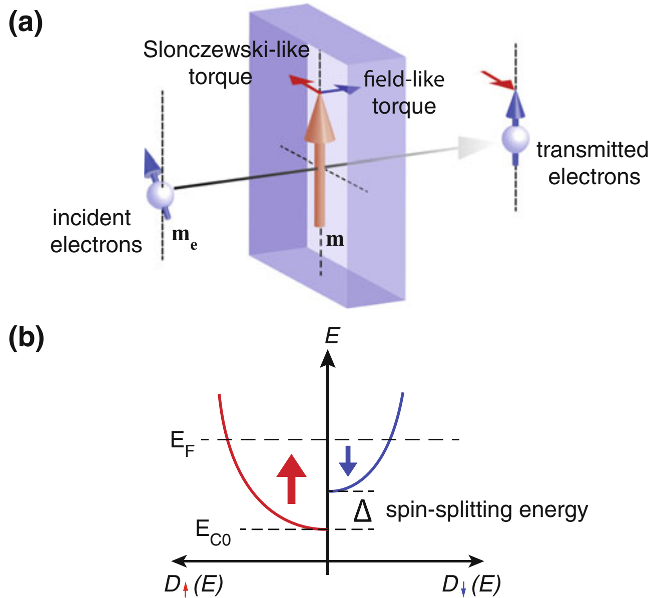


Fig. 7.3 (**a**) When an electron carries spin polarization that is non-collinear with the magnetization of the nano-magnet it is incident on, it exchanges spin angular momentum with the nano-magnet. Consequently, spin-transfer torque (decomposed into Slonczewski-like and field-like components) is exerted on the magnetization of the nano-magnet to align it into the spin polarization direction of the electron. This occurs because magnetism is a result of unequal spin density of states in the nano-magnet as illustrated in (**b**)—the flow of electrons with other spin directions perturb the total spin populations in the nano-magnet, which manifests as a torque on the magnetization of the nano-magnet

spins in Fig. 7.3b. When an electron is incident on a nano-magnet, it experiences an exchange field—the same field that aligns the spin directions of all electrons in the ferromagnet—if its spin polarization direction is non-collinear with the magnetization direction of the nano-magnet. The exchange field exerts a torque on the spin polarization of the electron, aligning it with the magnetization direction of the nano-magnet. Hence, when current flows through the MTJ, the ferromagnetic layers also act as spin filters that polarize the spin direction of electrons constituting the current flow. Whereas the exchange field exerts torque on the spin polarization of the electron, due to conservation of spin angular momentum, an equal and opposite torque (spin-transfer torque) is exerted on the magnetization of the nano-magnet to align it with the spin polarization direction of the electron. Hence, the electrons in a spin-polarized current (i.e., a current in which the majority of electrons are spin polarized in a particular direction) transfer their spin momentum to the nano-magnet and in the process exerts a torque on the magnetization of the nano-magnet that tries to align the magnetization direction of the nano-magnet with the spin polarization direction of the spin-polarized current. The magnetization direction of the nano-magnet may be switched if the torque exerted is large enough to overcome all other energies in the nano-magnet. Furthermore, the rate of spin momentum transfer and the torque exerted are proportional to the rate of electron flow or the current, and determine the switching time. The current or current density needed to achieve a specific switching time is the critical current, I_C, or critical current density, J_C. It is often easier to analyze the MTJ in terms of the *intrinsic switching current density*, J_{C0}, given as (Huai 2008; Sun 2000)

$$J_{C0} = \frac{2e\alpha M_{sat} t_{FL} H_{eff}}{\hbar\eta} \quad (7.6)$$

Here, e is the elementary charge, α and t_{FL} are the Gilbert damping constant and thickness of the free layer, respectively. $\hbar$ is the reduced Planck constant, η is the spin polarization efficiency factor, and H_{eff} is the effective magnetic field spin-transfer torque must overcome when switching the free layer magnetization.

In an MTJ, it is easier for spin-transfer torque to switch the free layer than to switch the pinned layer because the E_B of the pinned layer is much higher than that of the free layer. Let us consider what happens when electrons are flowing from the pinned layer to the free layer in an MTJ. The pinned layer spin-polarizes the incoming electrons which then flow into the free layer. These electrons are spin-polarized in the direction of the pinned layer magnetization and transfer their spin momentum to the free layer. The spin-transfer torque exerted on the free layer tries to align the free layer magnetization with that of the pinned layer (i.e., MTJ is switched into the P configuration). Consider when electrons flow from the free layer to the pinned layer instead. Electrons entering the free layer from the non-magnetic metal interconnect are not spin-polarized and can have any spin direction. Electrons spin-polarized in direction of the pinned layer magnetization are able to tunnel across the oxide easily. Those with the opposite spin polarization may not tunnel across the oxide easily and accumulate in the free layer. These electrons transfer their spin angular momentum to the free layer and exert a torque that aligns the direction of free layer magnetization opposite to that of the pinned layer (i.e., MTJ is being switched into the AP configuration). Consequently, the spin directions of the electrons become aligned with the magnetization direction of the pinned layer and they may then easily tunnel across the oxide. From this discussion, we can see that the process of parallelizing the MTJ configuration is more efficient than that for anti-parallelizing the MTJ configuration (i.e., I_C and J_C are asymmetric and depends on switching direction (Datta et al. 2012; Ikeda et al. 2010). It has been reported that J_C for anti-parallelizing the MTJ can be 10%–200% larger than for parallelizing the MTJ (Ikeda et al. 2010; Kishi et al. 2008).

As just mentioned, spin-transfer torque, $\boldsymbol{\tau}_{STT}$, is only exerted when the spin polarization direction, $\boldsymbol{p}$, of the electrons incident on the nano-magnet is non-collinear with the magnetization direction of the nano-magnet, $\boldsymbol{m}$. It can be shown that (Salahuddin et al. 2008)

$$\boldsymbol{\tau}_{STT} = \beta[\varepsilon(\mathbf{m} \times \mathbf{p} \times \mathbf{m}) + \varepsilon'(\mathbf{p} \times \mathbf{m})] \quad (7.7)$$

where β is a scalar factor proportional to the current flowing through the MTJ. ε and ε' are scalar factors proportional to the strength of Slonczewski-like and field-like torques, respectively. In an MTJ, the spin polarization direction of the electrons is pointing in the magnetization direction of the pinned layer or that of the free layer. The magnetization directions of the two magnetic layers are collinear to maximize the *TMR* and the distinguishability between the P and AP configurations of the MTJ. Thus, $\boldsymbol{\tau}_{STT}$ should be negligible. However, thermal effects perturb the magnetization of the nano-magnets in the MTJ, causing them to be non-collinear and $\boldsymbol{\tau}_{STT}$ can be large enough to overcome all other energies in the free layer of the MTJ. Hence, spin-transfer torque switching of the MTJ is a stochastic process since thermal effects are random in nature. The thermal effect may be modeled as a fluctuating magnetic field written as (Brown 1963)

$$\boldsymbol{H}_{EFF} = \xi\sqrt{\frac{\alpha k_B T}{\gamma\mu_0 M_{sat} V_{FL} \delta t}} \quad (7.8)$$

$\boldsymbol{\xi}$ is a 3-vector whose components are independent Gaussian random variables with zero mean and unit variance. γ is the gyromagnetic ratio

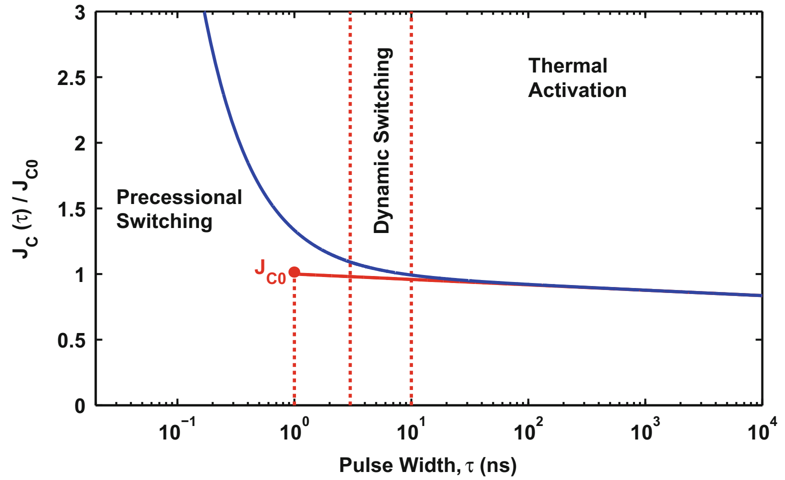


Fig. 7.4 This graph illustrates an example of the pulse width, τ, dependence of J_C needed to program an MTJ with 0.5 success probability. The spin torque generated at low current densities is still able to switch the MTJ state due to heating of the MTJ, which increases the thermal effects

(17.6 MHz/Oe) and δt is the constant time step used in numerical simulation of the MTJ dynamics.

Also, it was found that J_C depends on the pulse width, τ (Huai 2008). An example of this is illustrated in Fig. 7.4. Three switching regimes can be observed: the *precessional*, *thermal* and *dynamic* regimes. In the precessional regime, $\boldsymbol{\tau}_{STT}$ completely overcomes all other energies in the free layer to switch the MTJ. Thermal effects only affect the magnetization direction of the free layer in the MTJ just prior to onset of J_C. Thus, the dependence of programming failure on J_C is determined by the distribution of free layer magnetization direction just prior to applying the switching current pulse. In the thermal regime, $\boldsymbol{\tau}_{STT}$ alone is unable to overcome all other energies in the free layer. However, thermal effects, which are also increased due to Joule heating by the current flowing through the MTJ, assists $\boldsymbol{\tau}_{STT}$ in switching the MTJ configuration. The dependence of programming failure on J_C in the thermal regime is determined by the random time needed for thermal effects to sufficiently reduce the effective barrier height such that $\boldsymbol{\tau}_{STT}$ can switch the MTJ configuration. In the dynamic regime, the dependence of programming failure on J_C is determined by both the distribution of free layer magnetization prior to applying J_C and the random time needed for thermal effects to sufficiently reduce the effective barrier height.

7.2.2 Integrating MTJ with CMOS Technology

When designing STT-MRAM arrays, the fabrication steps need to be understood in order to understand the impact of design choices on the characteristics of the MTJ, the impact of the bit-cell topology on the bit-cell footprint, layout of the memory array in terms of area overhead, and performance and energy overheads due to increased parasitics. Figure 7.5 illustrates one method of integrating MTJs into the back end of the CMOS fabrication process (back-end-of-line, BEOL) developed by Qualcomm (Kang et al. 2014). The MTJs are placed in between metal layers that form the interconnects of the integrated circuit (IC). After the chemical mechanical polishing (CMP) step (step 1) to expose the metal layer on which the MTJs are to be placed, the layers constituting the MTJ stack are deposited (steps 2 and 3). The MTJ stack consists of many thin layers including those needed to form the top and bottom electrodes (TE and BE, respectively), the seed layer for growing high quality magnetic thin films, the magnetic layers needed to form the

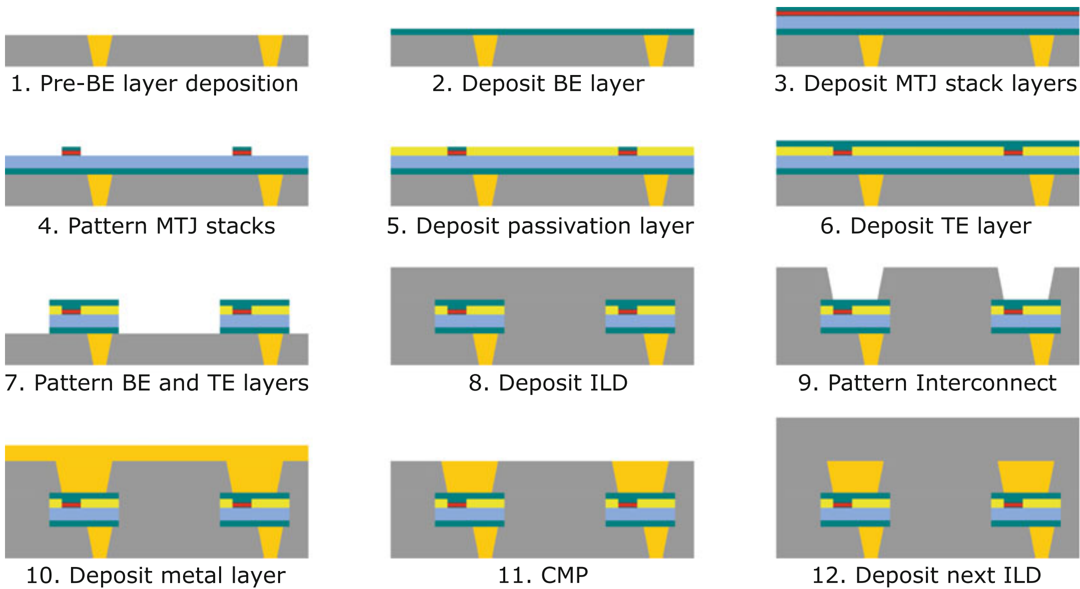


Fig. 7.5 An example flow for fabricating magnetic tunnel junctions in the back-end-of-line (BEOL) of the CMOS fabrication process. BE: bottom electrode, TE: top electrode, ILD: interlayer dielectric, CMP: chemical mechanical polishing

pinned and free layers of the MTJs, and also the tunnel barrier. In Fig. 7.5, the seed layer and layers constituting the pinned layer of the MTJ are deposited on the bottom electrode (BE) before the tunneling oxide layer. The first lithographic step is then applied to pattern the MTJ pillars using the BE as the etch stop layer (step 4). Since the thin film layers constituting the MTJ stack are very sensitive to particle contaminants, a dielectric passivation layer is immediately deposited after MTJ patterning (step 5) to protect them. Particle contamination can be detrimental to MTJ characteristics. A CMP step is performed to expose the capping layer of the MTJs after the dielectric passivation layer is deposited. Thereafter, the TE layer is deposited to contact the capping layer of the MTJs (step 6). The TE and BE of individual MTJ pillars are then defined by lithography and etch (step 7). The next interlayer dielectric (ILD) layer is then deposited (step 8). Lithography and etching are then performed to define the next layer of interconnect (step 9). This layer also serves as an electrical connection to the MTJs. Note that the material for the TE layer have been chosen so that it may be used as an etch stop. The next metal layer is then deposited (step 10) and a CMP step (step 11) is performed to complete the definition of the metal interconnect layer. The next ILD layer is then deposited (step 12) and the rest of the BEOL fabrication process continues.

7.2.3 Impact of STT-MRAM Integration on IC Design

The STT-MRAM array may be fabricated together with other CMOS circuits in an system-on-chip (SoC) and hence the placement of the MTJ layers needs to account for the interconnect wiring in the rest of the silicon wafer. For example, increasing the separation between the lower metal layers to accommodate the MTJ may increase parasitics in the interconnects in other parts of the IC, which can negatively affect the overall performance of the IC. The MTJs may be formed between the higher interconnect metal layers (such as between the last 1× metal layer and the first 2× interconnect metal layer) where the separation between layers across the IC is large enough accommodate the MTJs. As a

result, the minimum pitch between MTJs and hence the integration density of STT-MRAM, may be limited by the minimum pitch between metal interconnects. Furthermore, the via parasitics between the MTJ and the transistors below may be significantly increased if the MTJ is placed too high in the interconnect layers. Hence, process for integrating MTJs may become an important factor not only in determining the characteristics of the MTJ, but also in the overall performance of the IC.

The order of stack deposition may also impact the footprint of the STT-MRAM bit-cell, as we shall see in the next section. The MTJ pinned layer stack may be deposited first followed by the MgO and then the free layer stack. Consider if the pinned layer of the MTJ is connected to the transistor layer below as shown in Fig. 7.6a. This may be achieved by placing the MTJ on top of a stack of vias. If the free layer is to be connected to the transistor below instead as Fig. 7.6b illustrates, the metal layer on top of the TE needs to be extended to one side and then connected to the transistor below through a stack of vias. The additional area needed to accommodate this extension may increase the footprint of the STT-MRAM bit-cell that requires such an MTJ connection. An alternative scheme is to swap the order of pinned layer and free layer deposition but the impact on the MTJ characteristics depends on the fabrication process.

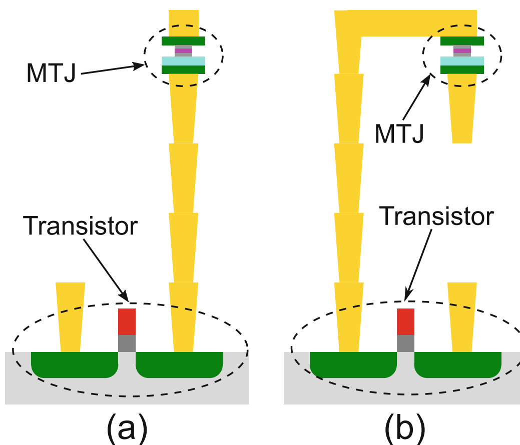
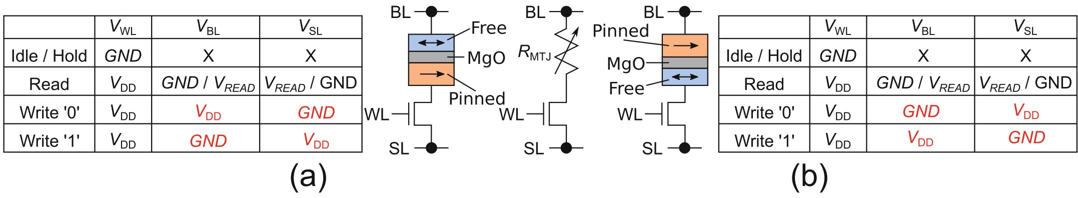


Fig. 7.6 The bottom pinned MTJ may have its pinned layer directly connected to the transistor below as shown in (**a**). If the free layer is to be connected instead and the order of layer deposition cannot be altered, the metal layer above TE has to be extended, as shown in (**b**), and connected to the transistor below through a stack of vias

7.3 Design of the STT-MRAM Bit-Cell

Figure 7.7 shows the topology of the basic one-transistor one-MTJ (1T-1M) STT-MRAM bit-cell. One electrode of the MTJ is connected to the *bit line* (BL) while the other electrode is connected to the access transistor (ATx). ATx is also connected to the *source line* (SL) as shown, and the *word line* (WL) is connected to the gate of ATx. WL is used to turn ATx ON and OFF. The 1T-1M STT-MRAM bit-cells can have two configurations (Kishi et al. 2008; Lin et al. 2009): the “standard” connection (SC) as shown in Fig. 7.7a, and the “reversed” connection (RC) shown in Fig. 7.7b. The bit-cell is accessed for read and for write operations by charging WL to V_{DD} to turn ATx ON. Read operation may then be performed by sensing the resistance between BL and SL. Write operations are performed by setting the voltages on BL and SL to the values shown in Fig. 7.7. Let us consider the SC configuration for example. A bit-cell having MTJ in P configuration stores ‘0’ whereas a bit-cell having MTJ in AP configuration stores ‘1’. The bit-cell is programmed with ‘0’ by applying V_{DD} and *GND* to BL and SL, respectively. The electrons constituting the write current flow from the pinned layer into the free layer of the MTJ to parallelize the MTJ configuration. A ‘1’ is programmed into the bit-cell by applying V_{DD} and *GND* to SL and BL, respectively. The electrons flow from the free layer to the pinned layer in this configuration, and exert torque that anti-parallelizes the MTJ configuration. Note that the size of ATx, the value of V_{DD}, and the write current pulse width are all designed such that the current flowing through MTJ during write operations is larger than I_C. Compared to the SC bit-cell configuration, the connections of



	V_{WL}	V_{BL}	V_{SL}
Idle / Hold	*GND*	X	X
Read	V_{DD}	*GND* / V_{READ}	V_{READ} / GND
Write '0'	V_{DD}	V_{DD}	*GND*
Write '1'	V_{DD}	*GND*	V_{DD}

	V_{WL}	V_{BL}	V_{SL}
Idle / Hold	*GND*	X	X
Read	V_{DD}	*GND* / V_{READ}	V_{READ} / GND
Write '0'	V_{DD}	*GND*	V_{DD}
Write '1'	V_{DD}	V_{DD}	*GND*

Fig. 7.7 The topology of the 1T-1M STT-MRAM bit-cell is shown. The (**a**) "standard" and (**b**) "reversed" connection differs in the way the MTJ is connected to the access transistor. The voltages on the control lines of the bit-cell corresponding to various operations of each bit-cell configuration are shown in the tables

the MTJ are swapped in the RC bit-cell configuration. Hence, the voltages on BL and SL for write operations of the RC bit-cell configuration are also swapped as compared to the SC bit-cell configuration.

During read operations, a sense amplifier is used to sense the MTJ state in the STT-MRAM bit-cell through BL or through SL. Also, a constant voltage or constant current scheme may be used to sense R_{MTJ} (Dorrance et al. 2011; Fong et al. 2012). In the constant voltage scheme, a fixed voltage, V_{RD}, is applied across the bit line and the source line of the STT-MRAM bit-cell and the resulting current flowing through the MTJ, I_{MTJ}, is compared to a reference current, I_{REF}. I_{MTJ} can be either higher or lower than I_{REF}, depending on the resistance state of the MTJ. The advantage of the constant voltage scheme is that I_{MTJ} during read operations may be amplified in the sense amplifier to improve sensing speed. The disadvantage is that the result of the sensing needs to be converted into an output voltage. In the constant current scheme, a fixed current, I_{RD} is passed through the MTJ and the voltage developed across the bit line and the source line, V_{BC}, is compared with a reference voltage, V_{REF}. The constant current scheme has the advantage that the result of the sensing is the voltage domain does not need to be converted. Furthermore, since the MTJ is programmed by passing current through it as we will see later, I_{RD} may be limited to prevent accidental programming of the MTJ during read operations or read-disturb failures. However, the $|V_{BC}\text{-}V_{REF}|$ signal generated by I_{RD} may be too small to be easily sensed.

7.3.1 Design Issues of the 1T-1M STT-MRAM Bit-Cell

Let us now discuss the major design issues of 1T-1M STT-MRAM bit-cells. The source degeneration of ATx during write operations, shared read and write current paths, and single-ended sensing scheme are the three major issues in 1T-1M STT-MRAM bit-cell. As we shall see, these design issues have conflicting design requirements, which constrain the design space of 1T-1M STT-MRAM bit-cell. If not carefully taken care of, these design issues may lead to excessive energy consumption, and degraded STT-MRAM performance and reliability.

7.3.1.1 Source Degeneration of ATx

In order to program the 1T-1M STT-MRAM bit-cell, bi-directional write current flow is required. Figure 7.8 shows the voltages on the circuit nodes of the bit-cell during write operations. When current flows from BL to SL as illustrated in Fig. 7.8a, the overdrive voltage of ATx (V_{GS}) is at V_{DD}. When the direction of current flow is reversed as shown in Fig. 7.8b, the voltage on the source terminal of ATx is at *GND* $< V_S < V_{DD}$. As such, $V_{GS} < V_{DD}$ and the drive strength of ATx is weakened. Furthermore, the asymmetry in ATx drive strength may be exacerbated by the asymmetry in I_C of the MTJ. The size of ATx may need to be enlarged to ensure successful write operations when passing current from SL to BL. This increases the write current supplied when write current is flowing from BL to SL, which increases energy consumption and may degrade the reliability of the MTJ tunneling oxide barrier. The reliability of

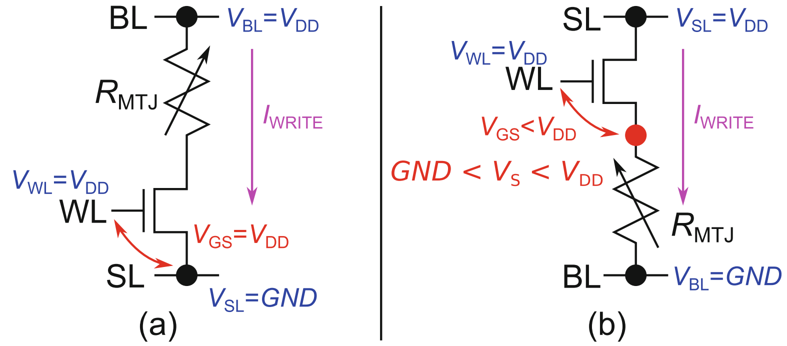


Fig. 7.8 (**a**) The overdrive voltage, V_{GS}, of the access transistor is at V_{DD} when write current flows from BL to SL. (**b**) $V_{GS} < V_{DD}$ when write current flows from SL to BL, and the drive strength of the access transistor is reduced

the MTJ tunneling oxide barrier is crucial for maintaining the large *TMR* required for reliably distinguishing between R_P and R_{AP}.

7.3.1.2 Shared Read and Write Current Paths

During write operations, current needs to be passed through the MTJ to exert STT so as to switch the MTJ configuration. During read operations, current flows through the MTJ regardless of whether the sensing scheme used is the constant voltage scheme or constant current scheme. Hence, the current flowing through the MTJ needs to be limited to avoid *disturb failures*. A disturb failure occurs when the MTJ is accidentally programmed during a read operation. Since the circuit nodes in the bit-cell need to be charged or discharged during sensing operations, limiting the amount of current flow through the bit-cell limits the rate at which these circuit nodes are charged or discharged. As a consequence, the sensing speed is reduced. Furthermore, limiting the amount of current flowing through the bit-cell during sensing may also reduce the signal margin available for the sense amplifier to determine the MTJ state. As a result, the sense amplifier may not be able to reliably distinguish between R_P and R_{AP}.

Note that it may be more difficult to limit the current flowing through the MTJ during read operation if the ATx is enlarged to mitigate the source degeneration problem discussed previously. The conflicting requirements for mitigating source degeneration and disturb failure may significantly constrain the design space. For example, E_B may be increased to reduce disturb failures. Doing so increases I_C, resulting in increased write currents. Hence, design choices must be made carefully to achieve an optimum 1T-1M STT-MRAM design.

7.3.1.3 Single-Ended Sensing Scheme

The sensing scheme for 1T-1M STT-MRAM bit-cells described in the earlier sections uses *single-ended sensing*—the voltage across or current through the bit-cell is compared with a common reference during sensing. Under process variations, the resistance of MTJs will deviate from the nominal value as depicted by the scatter plot in Fig. 7.9. Some bit-cells having MTJ in P configuration may have read current smaller than the reference current, or read voltage larger than the reference voltage. Also, some bit-cells having MTJ in AP configuration may have read current larger than the reference current, or read voltage smaller than the reference voltage. The sense amplifier will sense the bit-cells as having MTJ in AP configuration in the former case whereas the bit-cells are sensed as having MTJ in P configuration in the latter case. This is also called *decision failure*.

7.3.2 Scalability of STT-MRAM

When designing STT-MRAM, it is crucial for STT-MRAM designers to understand how the size of the access transistor and the MTJ impact

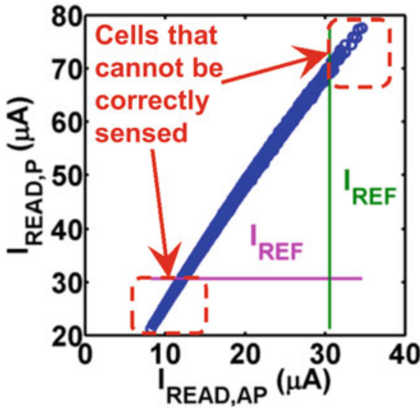


Fig. 7.9 Under process variations, the read currents through the 1T-1M STT-MRAM bit-cells may be distributed as this example plot shows. Each data point on the plot is the read current through a bit-cell when its MTJ is in the P configuration and the AP configuration. The bit-cells falling to the right of and below I_{REF} will not be correctly sensed by the sense amplifier

the characteristics of the STT-MRAM bit-cell. Furthermore, it is desirable to scale down the size of the STT-MRAM bit-cell to achieve high bit density and reduce the cost per bit. The size of the MTJ may need to be varied to achieve this while also ensuring that its resistance is not too large for the access transistor to provide adequate write current.

MTJs may be engineered with two flavors of magnetic anisotropy for satisfying thermal stability and retention time requirements—perpendicular and in-plane magnetic anisotropy (PMA and IMA, respectively). The magnetization of the magnetic layers in the MTJ with IMA points within the plane of the wafer on which the MTJ is deposited. On the other hand, the magnetization of the magnetic layers in the MTJ with PMA points perpendicular to the plane of the wafer. In both flavors of MTJs, the minimum magnetic field to switch the MTJ configuration, H_k, and the energy barrier, E_B, are related by (Sun 2000):

$$H_k = \frac{2E_B}{\mu_0 M_{sat} V_{FL}} \tag{7.9}$$

where μ_0 is the permeability of vacuum, and M_{sat} and V_{FL} are the saturation magnetization and volume of the nano-magnet, respectively.

For MTJs with IMA, the shape of the nano-magnets are engineered to achieve the required energy barrier height (Devolder 2011). In these nano-magnets, the width is defined by the limits of the process technology. The length and thickness of the nano-magnet are varied to engineer its energy barrier. However, the aspect ratio (AR = length/width) is usually in the range of 2–3 to keep the MTJ footprint small and for manufacturability (Apalkov et al. 2010). The thickness of the nano-magnet is then varied to meet thermal stability requirements. On the other hand, the nano-magnets in MTJs with PMA have other anisotropies, such as interfacial perpendicular anisotropy (Ikeda et al. 2010) or magneto-crystalline anisotropy (Apalkov et al. 2010), that overcome the shape anisotropy. The nano-magnets in MTJs with PMA may be modeled as having only uniaxial anisotropy and

$$H_k = \frac{2K_{u2}}{\mu_0 M_{sat}} \tag{7.10}$$

where K_{u2} is the effective uniaxial anisotropy constant.

Scaling analysis of the footprint of IMA and PMA based MTJs conclude that IMA based MTJ has excellent scalability down to 10–20 nm widths for thermal stability factor, $\Delta = 60$ (Apalkov et al. 2010; Devolder 2011). Below this critical width, the geometry of the free layer causes the anisotropy energies in the nano-magnet to favor a PMA configuration. Also, the critical width is increased if the required thermal stability factor is larger. This has motivated a shift in research focus toward PMA based MTJs.

Another crucial characteristic of the PMA based MTJ is that its I_{C0} is constant under size scaling (Apalkov et al. 2010). Thus, the voltage required to switch the MTJ, V_{C0}, scales as

$$V_{C0} = \frac{4 \times RA \times I_{C0}}{\pi w^2} \tag{7.11}$$

where a cylindrical MTJ with diameter w is assumed. The supply voltage required is expected to scale down together with the CMOS technology node (Roadmap 2014),

which is important for IoT and other applications requiring low energy consumption. Hence, the *RA* product of the MTJ needs to be scaled down (by reducing the thickness of the tunneling oxide barrier) so that the access transistor is able to supply sufficient write current to the MTJ. However, it was shown that the *TMR* of the MTJ may be substantially degraded when its *RA* product is reduced (Yuasa et al. 2004).

Another important consideration when scaling down the *RA* product of the MTJ is the available voltage signal for the sense amplifier to distinguish between a bit-cell storing an MTJ with P configuration from one storing an MTJ with AP configuration. Let us now derive a minimum magneto-resistance (MR) ratio of the bit-cell needed to meet sensing requirements when the resistance of the access transistor is neglected. The inclusion of the resistance of the access transistor will increase the MR ratio we calculate. We will assume a constant current sensing scheme so as to limit read-disturb failures. Under this scheme, the current being passed through the bit-cell during sensing is limited to a fraction of I_{C0}. The voltage developed across the bit-cell is compared to a reference voltage by a sense amplifier to determine the configuration of the MTJ. Ordinarily, the difference in voltage that the sense amplifier sees needs to be at least 50 mV. This means the difference in voltage due to the difference in MTJ resistance is $\delta V \geq 0.1$. Hence, for a cylindrical MTJ with diameter w, the MR ratio has to satisfy:

$$\text{MR ratio} \geq \frac{0.1(0.25\pi w^2)}{\kappa I_{C0} RA} \qquad (7.12)$$

κI_{C0} is the current passed through the bit-cell to sense the MTJ configuration. For a cylindrical MTJ with 50 nm diameter and $\kappa I_{C0} = 5$ μA, the MR ratio needs to be larger than 300% for $RA \leq 13$ Ω-μm^2. Fortunately, the required MR ratio decreases rapidly when the MTJ diameter is scaled down. For $w \leq 30$ nm and $RA \geq 6$ Ω-μm^2, the required MR ratio is less than 250%. Note that in the preceding analysis, we have assumed that I_{C0} does not scale down with w.

An alternative scheme that relaxes the requirement to scale the *RA* product of the MTJ is to scale down I_{C0} as the footprint of the MTJ is scaled down. The options for doing so, from analyzing Eq. (7.6), are to: (1) decrease the damping factor, α, (2) decrease M_{sat}, or (3) to improve the polarization efficiency factor, η. Reducing M_{sat} may require a change of material for the free layer of the MTJ, whereas increasing η also requires careful optimization of the interface between the free layer and tunneling oxide barrier in the MTJ. It was found recently that α and the interfacial magnetic anisotropy energy density may be optimized in a PMA based MTJ having an FeB free layer sandwiched between two MgO based tunneling oxide barriers (Tsunegi et al. 2014). Reducing α is promising because both the randomness of STT switching and I_{C0} are reduced. The magnitude of the random thermal field depends on $\sqrt{\alpha}$ (see Eq. (7.8)) and hence, reducing α also reduces the stochasticity of STT based switching in MTJs. Thus, recent research works are focused on reducing α. However, the STT-MRAM designer must ensure that the required MR ratio to meet sensing requirements is achievable when I_{C0} is reduced.

7.3.3 Alternative STT-MRAM Bit-Cell Topologies

As we saw in the preceding sections, one of the most challenging aspect of STT-MRAM design is the conflicting design requirements for read operations and for write operations—improving write operation in the 1T-1M STT-MRAM bit-cell requires degrading the read operation and vice versa. However, alternative bit-cell topologies may open new pathways for optimizing the design of STT-MRAM. In the following sections, we will present several alternative STT-MRAM bit-cell topologies that are able to mitigate the conflicting design requirements in STT-MRAM.

7.3.3.1 The 2T-1M STT-MRAM Bit-Cell

The motivation for using the 2T-1M STT-MRAM bit-cell topology may be understood by first analyzing the 1T-1M STT-MRAM bit-cell. The major failure mechanisms in the

1T-1M STT-MRAM bit-cell are decision, disturb, and write failures (Fong et al. 2012). Decision and disturb failures were briefly described in Sect. 7.3.1. *Write failure* occurs when the MTJ configuration is not successfully programmed during the write operation. The optimization methodology proposed in (Fong et al. 2012) sizes the width of the access transistor (ATx) so as to optimize the total failure probability as shown in Fig. 7.10. In this example, the STT-MRAM bit-cell is designed using a commercial 45 nm CMOS technology with MTJs that have 40 × 100 nm elliptical cross-section area. The read operation may be designed to select the dominant read failure mechanism. For example, disturb failure is the dominant read failure mechanism when the read operation uses a constant voltage sensing scheme that has a large read voltage. In the example shown in Fig. 7.10, the decision failure probability, P_{DECISION}, is minimized when ATx width is 908 nm. However, when the dominant mechanism for read failure is decision failure, the ATx width needs to be increased to reduce the write failure probability (P_{WRITE}) and optimize the total failure probability. As a result, the minimum P_{DECISION} cannot be achieved.

Alternatively, the design constraint can be relaxed by noting that multi-finger transistors are typically used to implement very wide transistors. Multi-finger transistors are multiple transistors connected in such a way that their gate, source, and drain terminals are shared. When multi-finger transistors are used in the STT-MRAM bit-cell design, the effective access transistor width may be varied using two word lines instead of one as illustrated in Fig. 7.11, which results in the 2T-1M STT-MRAM bit-cell topology (Li et al. 2010). During write operations, both word lines are turned ON and OFF in unison to supply maximum current through the MTJ. During read operation, Word Line 2 keeps M2 OFF whereas Word Line 1 switches M1 ON and OFF. The width of M1 may then be optimized for decision failures (908 nm in our example), while the width of M2 is made as large as required to fulfill write failure, array area, and array capacity requirements. Hence, the P_{WRITE} of the 2T-1MTJ bit-cell may be optimized without worsening P_{DECISION}.

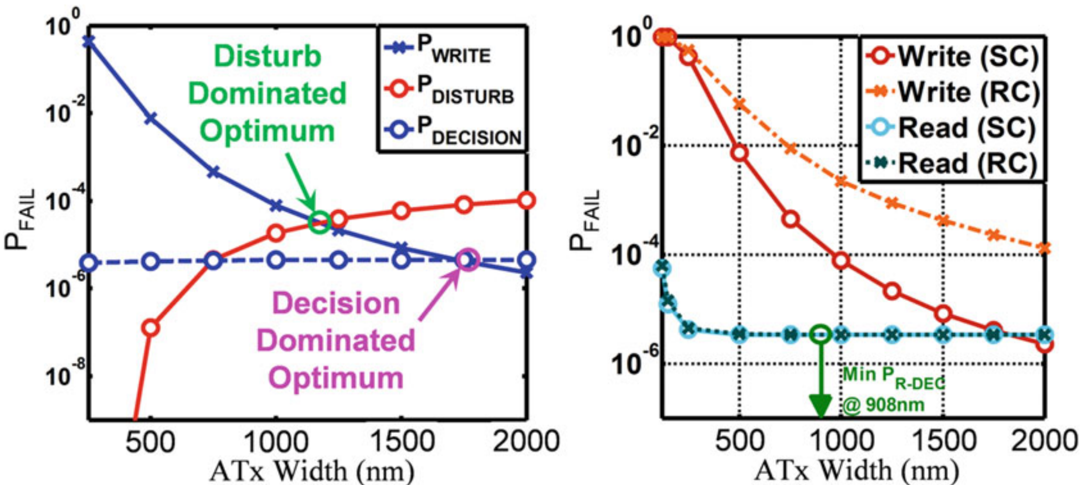


Fig. 7.10 The read-decision, read-disturb, and write failure rates of an example 1T-1M STT-MRAM bit-cell design are shown on the left. The optimum width for the access transistor (ATx) depends on whether the dominant failure mechanism for read operations is decision failure or disturb failure. When the ATx width is small, write failure probability is significantly larger than that for read operations. Based on the optimization methodology proposed in (Fong et al. 2012), the optimum ATx width is 1829 nm, as shown on the right, when decision failure is the dominant read failure mechanism

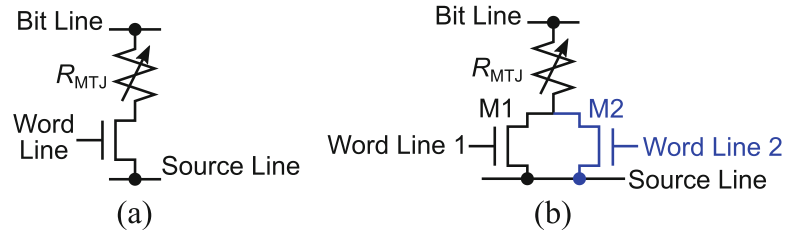


Fig. 7.11 The schematic of the (**a**) 1T-1M and (**b**) 2T-1M STT-MRAM bit-cells. If the 1T-1M STT-MRAM bit-cell requires an ATx with large width, the ATx may be formed by connecting two ATx's with smaller width in parallel. The 2T-1M STT-MRAM topology is implemented by connecting the gates of the ATx's to different word lines, enabling independent control to each ATx

7.3.3.2 Nonvolatile SRAM Using MTJs

The write current requirement of the STT-MRAM bit-cells presented thus far may be relaxed by using a longer write current pulse width. Consequently, the write access delays of these bit-cells may be very long (possibly ≥ 1 μs), and are unsuitable for IoT applications that require faster write access delays. Several alternative STT-MRAM bit-cell structures (see Fig. 7.12) have been proposed in the literature to target IoT applications that require faster access delays (Abe et al. 2004; Ohsawa et al. 2012; Yamamoto and Sugahara 2009). The proposed bit-cell structures are similar to the conventional 6T SRAM bit-cell structure shown in Fig. 7.12. MTJ0 and MTJ1, which store complementary data (i.e., one stores '0' and the other stores '1'), are used to skew the cross-coupled inverters in the SRAM cell to implement a non-volatile SRAM (NV-SRAM) bit-cell. When the NV-SRAM cell is powered off, the MTJs are first programmed with the state of the cell. When the NV-SRAM is powered back on later, the MTJs skew the cross-coupled inverters so that the state the NV-SRAM returns to the state before it was powered off. The advantage of NV-SRAM is that since the SRAM bit-cell topology is preserved, only the array peripheral circuitry for write operations need to be modified to meet write requirements. Furthermore, the differential nature of the bit-cell structure enables fast self-referenced differential read operations for fast read access delays. Another advantage of NV-SRAM is that only the bit-cells that are being accessed need to be powered ON—the rest may be powered OFF to save on standby power, which is highly desirable for IoT applications.

Although NV-SRAMs are able to achieve short access delays, their restore operations depend on the ability of the MTJs to skew the cross-coupled inverters. Hence, the proposed NV-SRAM bit-cells may be very sensitive to mismatches in the characteristics of the MTJ and the CMOS transistors. Furthermore, the sizes of the transistors in the bit-cells shown in Fig. 7.12 may need to be enlarged to meet the write current requirements. The proposed NV-SRAM bit-cells illustrated in Fig. 7.12 also need extra transistors as compared to the 1T-1M STT-MRAM bit-cell. Moreover, both MTJs in the NV-SRAM bit-cells need to be programmed which results in high write energy consumption.

7.4 STT-MRAM Array Architecture and Layout

In the preceding sections, the design of STT-MRAM bit-cells and the MTJ have been discussed. We have seen that at the bit-cell level, it is extremely challenging to design STT-MRAM to meet the performance achievable by 6T SRAMs. The write and read access

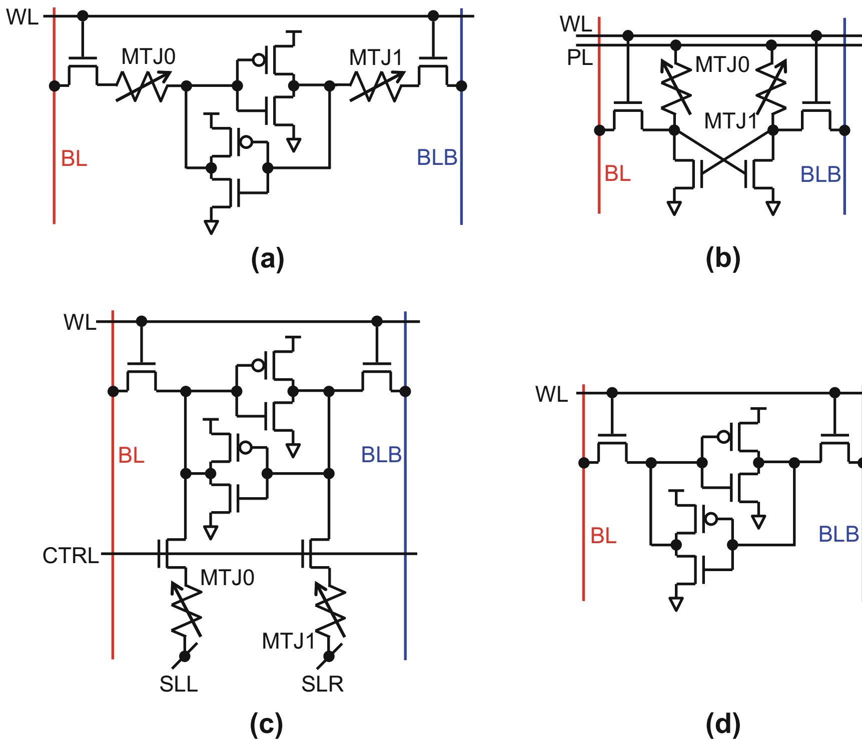


Fig. 7.12 The structures of NV-SRAMs proposed (**a**) in (Abe et al. 2004), (**b**) in (Ohsawa et al. 2012), and (**c**) in (Yamamoto and Sugahara 2009) are shown. The structure of (**d**) a volatile 6T SRAM is also shown for comparison

delays may be limited by large I_C needed and the single-ended nature of the sensing scheme, respectively. Although the STT-MRAM bit-cell does not outperform 6T SRAM, STT-MRAM arrays have several advantages over their 6T SRAM counterparts (Park et al. 2012).

The logical organization of memories stores data words sequentially as shown in Fig. 7.13a. This may result in a memory footprint that is not feasible for physical implementation. Alternatively, several data words are stored in every row as shown in Fig. 7.13b. *Bit-interleaving* (Park et al. 2011) is used to reduce the wiring in the peripheral circuitry that selects the array columns during data access. Note that bit-interleaving is commonly applied to 6T SRAM arrays to mitigate soft errors due to particle strikes. When a data word is being accessed, the word line corresponding to the address of the data word is turned ON. The address of the data word also selects the columns to be accessed in the array.

The schematic and corresponding layout of an array of 1T-1M STT-MRAM bit-cells are shown in Fig. 7.14a, b, respectively. The bit-cells are arranged into rows and columns where a word line turns on the ATx's of bit-cells connected to the same row. The bit-cells connected along the column stores the data bit of the data word stored in the corresponding row. Due to the nonvolatile nature of STT-MRAM, only the accessed bit-cells need to be powered ON. Power is not supplied to the rest of the bit-cells, which saves leakage power consumption in the array. Note that the STT-MRAM array based on NV-SRAM bit-cells may also be organized in a similar manner so that they may be powered OFF when idle.

Now, let us compare the access operation of a 6T SRAM array to that of the STT-MRAM array. Figure 7.15a illustrates the access operation to a

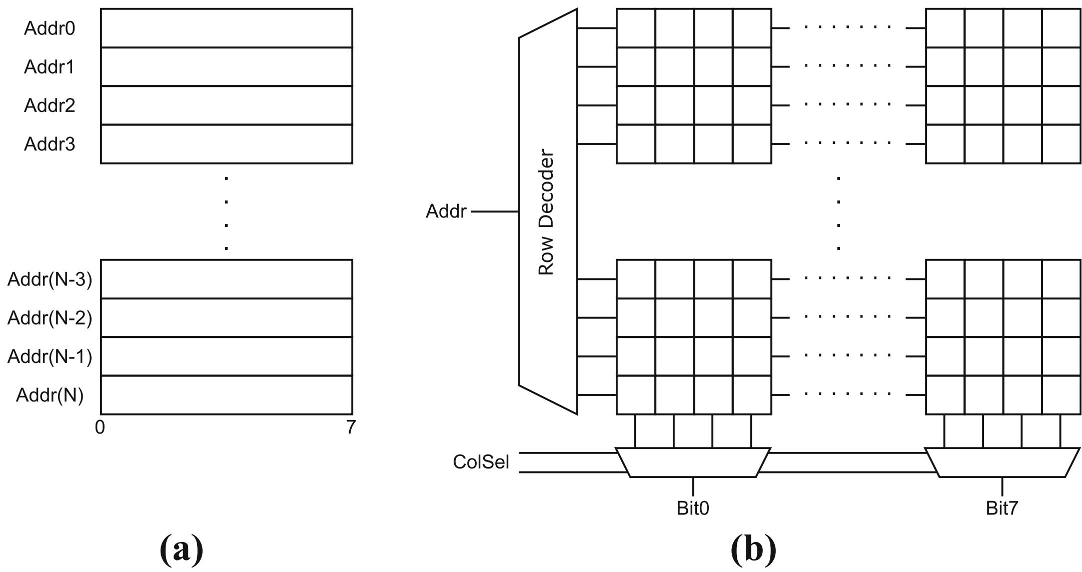


Fig. 7.13 An example showing the (**a**) logical and (**b**) physical organization of a memory consisting of N 8-bit wide data words (Bit0 to Bit7). Each row stores 4 data words. Bit-interleaving (i.e., the bit-cells storing the same bit position of each data word are located together physically as a group) is used to reduce the wiring to the column selection multiplexers shown in (**b**). These multiplexers are two-way multiplexers that allow Bit0 to Bit7 to behave as input and output

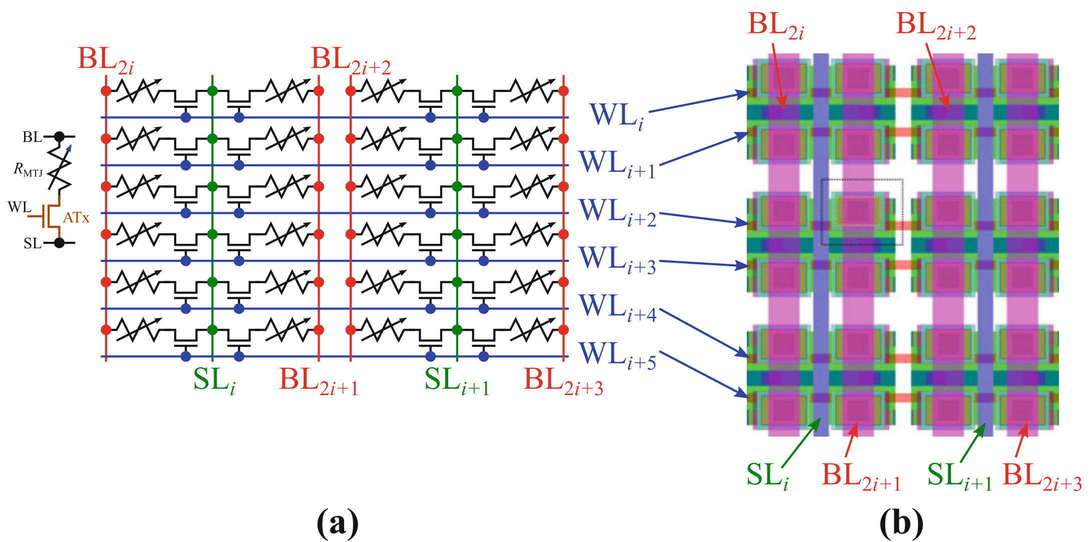


Fig. 7.14 This figure illustrates the (**a**) schematic and (**b**) array layout of an example array of 1T-1M STT-MRAM bit-cells. An individual 1T-1M STT-MRAM bit-cell is shown on the left. Only 6 rows × 4 columns are shown. Note that every SL is shared between two columns to reduce the total array area. In this layout, the pitch between MTJs limits the lowest achievable layout area

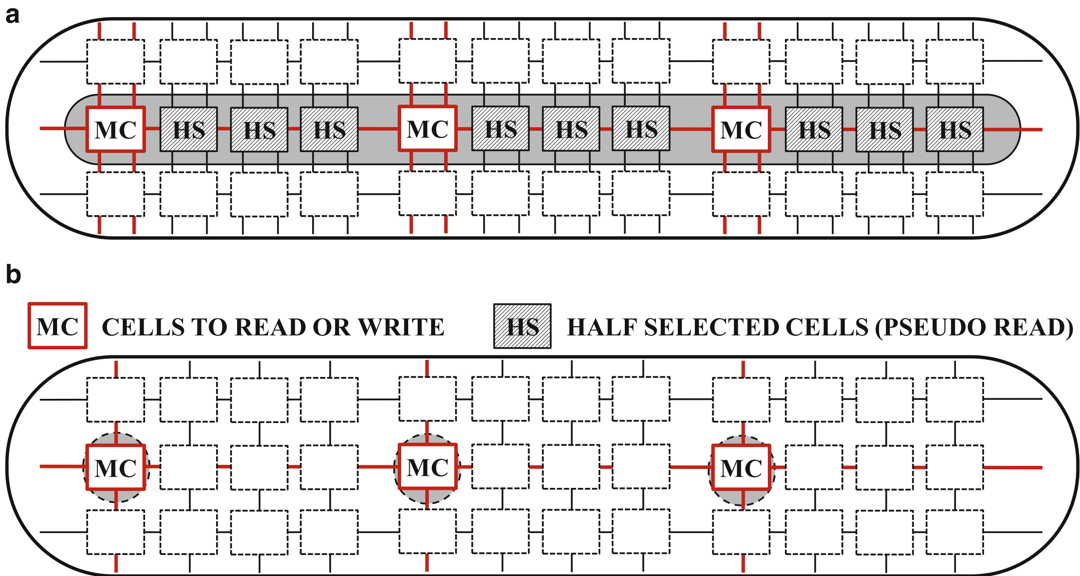


Fig. 7.15 (**a**) 6T SRAM arrays have the half select issue because power must be continuously supplied to all bit-cells to retain data. Thus, unaccessed columns in the selected row need to be biased in the pseudo-read condition to prevent disturb failures. (**b**) Non-volatile STT-MRAM cells do not have the half select issue because power is only supplied to bit-cells that are being accessed

row in a 6T SRAM array. When a row is accessed, only the columns corresponding to the word being selected are accessed. The bit lines connected to the other columns must be precharged to put bit-cells connected to them in the *pseudo-read condition*. Doing so ensures that the data stored in those bit-cells are not accidentally overwritten (i.e., the *half-select issue*). This is required because the SRAM bit-cells need to be continuously powered to retain data. Since STT-MRAM is nonvolatile, power does not need to be supplied to the bit-cells that are not accessed as Fig. 7.15b shows. Hence, the half-select issue present in the 6T SRAM array is absent in the STT-MRAM array. Furthermore, only the peripheral circuitry consumes standby power, leading to significant energy savings that is desirable for IoT application.

Further comparisons show that although the 6T SRAM bit-cell outperforms the 1T-1M STT-MRAM bit-cell, the 1T-1M STT-MRAM based cache may outperform the 6T SRAM based cache (Park et al. 2012). This is illustrated by the cache comparisons reproduced from (Park et al. 2012) in Fig. 7.16. First, the 1T-1M STT-MRAM based cache may have as much as 3× higher capacity than its 6T SRAM counterpart at the same bit-cell array footprint as shown in Fig. 7.16a. As the cache capacity increases, the read and write latencies of the 6T SRAM based cache increases faster than the STT-MRAM based cache as shown in Fig. 7.16b, c. This is because the access latencies are dominated by the delays needed to charge and discharge parasitic capacitances in the cache. The 1T-1M STT-MRAM has a smaller bit-cell footprint than the 6T SRAM bit-cell. Hence, the parasitic line capacitances increase much more quickly in 6T SRAM cache than the STT-MRAM cache when its capacity is increased. Consequently, the read and write energies of 6T SRAM cache increases faster than their 1T-1M STT-MRAM counterpart as plotted in Fig. 7.16d, e. The increase in read and write energies of 6T SRAM cache with increasing cache capacity is exacerbated by the half-select issue discussed earlier. The biggest advantage of the 1T-1M STT-MRAM based cache over its 6T SRAM counterpart is the lower leakage power consumption graphed in Fig. 7.16f.

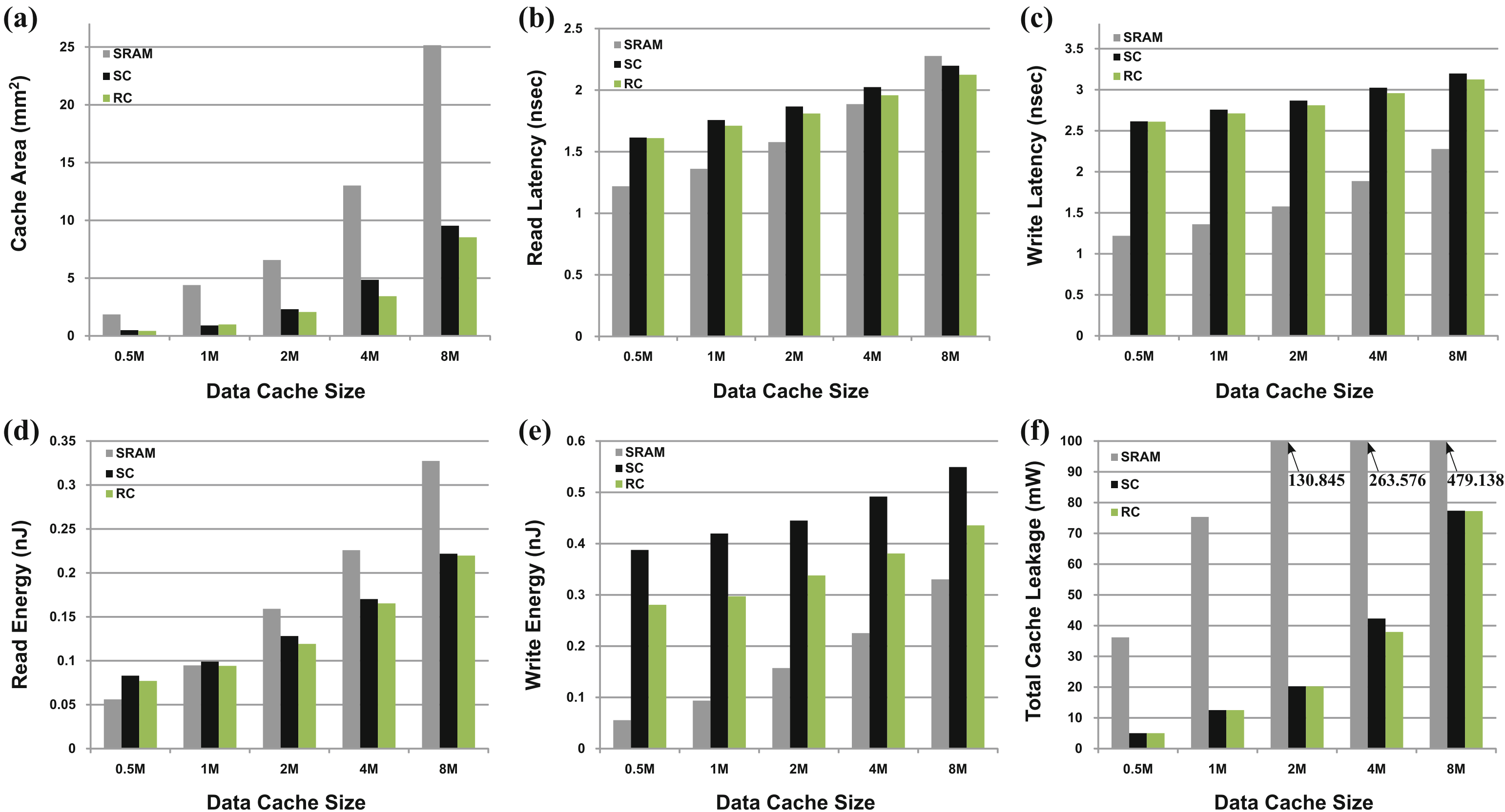


Fig. 7.16 (**a**) Array area of SRAM and STT-MRAM based data caches (4-way, 64B cache line, B = Byte, M = Mega Byte), (**b**) read latency and (**c**) write latency, (**d**) read energy per operation and (**e**) write energy per operation, and (**f**) total leakage power

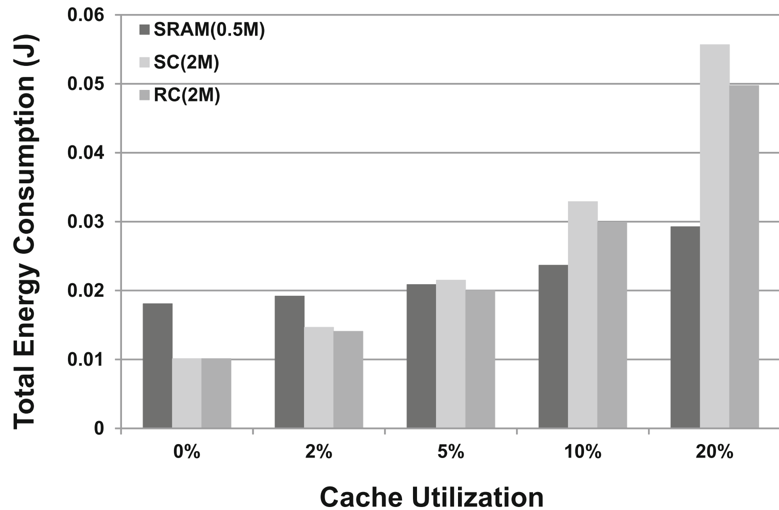


Fig. 7.17 Total energy consumption versus cache utilization for SRAM and for STT-MRAM based caches for various levels of cache utilization. Note that due to the difference in cache capacity, the cache utilization of the STT-MRAM based caches (SC and RC) is only a quarter of that in the 6T SRAM based cache is the data stored is at most 0.5 MB

Real world applications affect how the cache is accessed (number of read and write operations, number of cycles during which the cache is idle, etc.) as well as the cache utilization (amount of data used by the program that is stored in the cache as a fraction of the total cache capacity). The comparison between 1T-1M STT-MRAM based cache and 6T SRAM based cache in (Park et al. 2012) analyzes the performance of the caches using SPEC2000 benchmarks that emulate real world applications. The total energy consumption versus cache utilization for caches based on 6T SRAM, standard connection (SC) and reversed connection (RC) 1T-1M STT-MRAM bit-cells is plotted in Fig. 7.17. Consider when the data that needs to be stored in cache is 0.1 MB. This corresponds to 20% cache utilization in the 6T SRAM based cache whereas the cache utilization is only 5% in the STT-MRAM based caches. The results show that the total energy consumption of the STT-MRAM based caches is 25%–35% lower than that of the 6T SRAM based cache.

At this juncture, we would like to highlight several design techniques proposed in the literature to further reduce the energy consumption of STT-MRAM arrays. The common source line technique was proposed to reduce the bit-cell footprint of the 1T-1M STT-MRAM bit-cell, which also reduces the parasitic line capacitances that need to be charged and discharged during memory accesses (Zhao et al. 2012). Techniques have also been proposed in the literature to improve the energy efficiency of write operations in STT-MRAM, which may also improve the reliability of the tunneling oxide barrier in the MTJ as explained in Sect. 7.3.1. Examples of such techniques include the stretched write cycle (Augustine et al. 2012), the balanced write scheme (Lee et al. 2012), and write optimization techniques discussed in (Kim et al. 2012). The STT-MRAM array architecture may also be modified to implement early write termination (Zhou et al. 2009), partial line update (Park et al. 2012), and write biasing techniques (Ahn et al. 2013; Jung et al. 2013; Mao et al. 2014; Rasquinha et al. 2010; Wang et al. 2013b). The objective of these array architecture modifications is to ensure that write operations into the STT-MRAM array are performed only when they are required, and in an energy efficient manner. Some proposed array architecture design techniques even exploit the asymmetry in write characteristics of STT-MRAM bit-cells to improve the write energy efficiency of the STT-MRAM array (Kwon et al. 2013; Sun et al. 2012).

Although significant improvements to STT-MRAM may be achieved using the design techniques presented earlier, they may not be the

most optimum. In the next section, we present a device/circuit/array architecture co-design technique that optimizes the design of the STT-MRAM array for energy efficiency in the presence of process variations. This technique leverages insights from earlier work that retention time requirements may be relaxed in certain applications (Jog et al. 2012; Li et al. 2013a, b; Sun et al. 2011, 2014; Smullen et al. 2011) and that error-correcting codes (ECC) may be used to mitigate retention time errors as well as write failures and disturb failures that are caused by thermal effects (Xu et al. 2009). This is particularly useful for caches used in some IoT applications.

7.5 ECC and Device/Circuit/Architecture Co-design for Low-Power and Higher Reliability

It is important to note that many IoT applications do not need 10 years of data retention time, which is the requirement of most other memory applications. For example, the IoT system may periodically wake up from sleep mode, sample new data, process the new data with that gathered when it was previously awake, store results and needed data in memory, and discard old data before returning to sleep mode. In such cases, the memory only needs to retain data long enough so that it is still available when the system next wakes up. To save on leakage energy, a nonvolatile memory technology is desirable for implementing caches for these applications. Although dynamic RAM (DRAM) may be used, it may need to be frequently refreshed because its retention time is tens to hundreds of milliseconds, whereas the retention time for IoT applications maybe several days to several weeks. Hence, DRAM based caches may lead to high energy consumption in systems for IoT applications. Another requirement for the memory system is that the any overhead in energy or latency due to transitioning between sleep and active modes must be sufficiently low compared to the leakage energy savings.

STT-MRAM may be designed to fulfill the design requirements of memories for IoT applications. The design strategy proposed in (Pajouhi et al. 2015) exploits the fact that the retention time requirement in STT-MRAM is tunable (either by changing the shape anisotropy or interfacial anisotropy of MTJs with in-plane and perpendicular magnetic anisotropy, respectively). In the proposed approach, STT-MRAM is designed with a retention time target that is significantly relaxed from the conventional target of 10 years so as to save write energy. Note that it is difficult to achieve 10 year retention time target for large STT-MRAM arrays (Kwon et al. 2015; Naeimi et al. 2013). Moreover, the disturb failure probability of STT-MRAM is increased when the retention time target is reduced. The design methodology proposed in (Pajouhi et al. 2015) mitigates the increased disturb failure rate using error correction codes (ECC). ECC may also be used to mitigate failures caused by process variations (Kwon et al. 2015; Xu et al. 2009; Yang et al. 2015).

To exemplify the aforementioned STT-MRAM design approach, let us consider an example of STT-MRAM cache design with 10 years retention time requirement as done in (Pajouhi et al. 2015). The graph of 10 year retention failure probability versus the thermal stability factor for different memory word lengths shown in Fig. 7.18a imply the tradeoff between retention time and I_{C0}, derived as:

$$I_{C0} = \frac{e\alpha}{\hbar\eta}\left(4\Delta k_B T + H_{\mathrm{Demag}}\mu_0 M_{\mathrm{sat}} V\right) \quad (7.13)$$

Here, α, M_{sat}, and V are the Gilbert damping factor, saturation magnetization, and volume, respectively, of the free layer of the MTJ. η is the spin polarization efficiency describing the spin-transfer torque generated per unit current. $\hbar$, e and μ_0 are the reduced Planck's constant, elementary charge and permeability of free space, respectively. H_{Demag} is the demagnetizing field that spin-transfer torque needs to overcome in the free layer of the MTJ. In the MTJ with perpendicular magnetic anisotropy (PMA), $H_{\mathrm{Demag}}\mu_0 M_{\mathrm{sat}} V \ll E_B$ and hence the

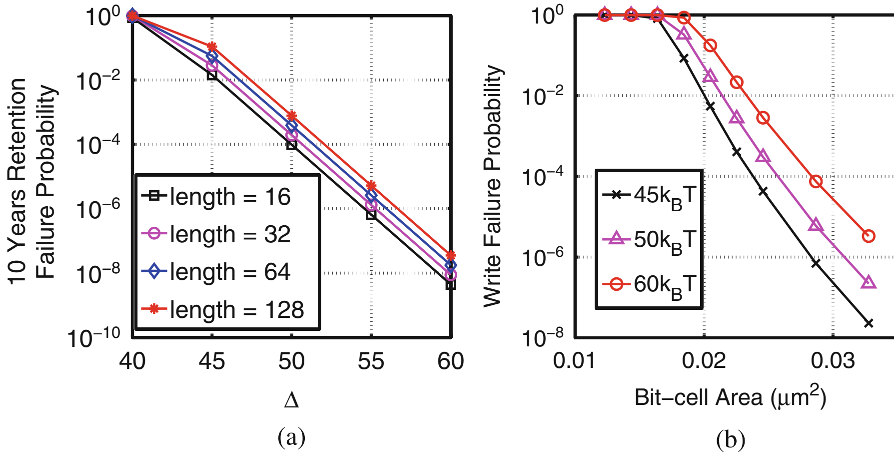


Fig. 7.18 (**a**) The retention failure probability versus thermal stability factor for different word lengths. (**b**) The write failure probability versus bit-cell layout area for different thermal stability factors

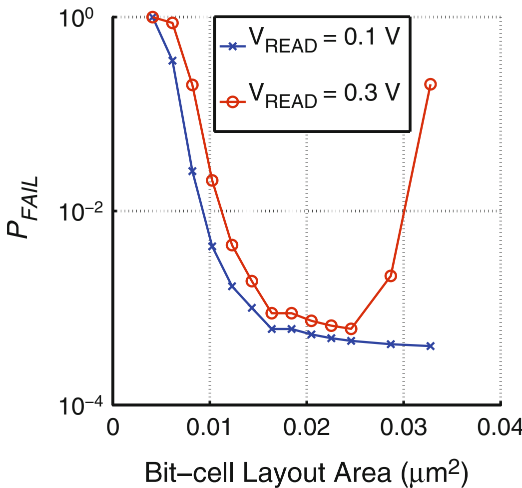


Fig. 7.19 The probability of read failure for constant voltage sensing scheme with $V_{\mathrm{READ}} = 0.1$ and 0.3 V

demagnetizing field term in Eq. (7.13) may be neglected. Due to reduced I_{C0} by reducing Δ, the write failure probability of an STT-MRAM bit-cell decreases under the same bit-cell bias conditions as graphed in Fig. 7.18b. Note that the following analysis assumes that the MTJ resistance is independent of Δ, which may not be generally true.

The biasing conditions of the bit-cell also determine the read failure probability of the STT-MRAM array. Figure 7.19 graphs the read failure probability versus bit-cell layout area for STT-MRAM bit-cells designed with MTJs having $\Delta = 45$. The constant voltage scheme is used, and the read failure probabilities are graphed for bit-cell read voltage of $V_{\mathrm{READ}} = 0.1$ V and $V_{\mathrm{READ}} = 0.3$ V. The word line voltage is $V_{\mathrm{DD}} = 1.0$ V. In the bit-cell layouts used, the bit-cell area is dominated by the access transistor. Thus, the resistance of the access transistor is large compared to that of the MTJ when the bit-cell area is small. This results in significant read failures because the bit-cell resistance when the MTJ is in P configuration is not substantially different than when the MTJ is in AP configuration (also called *decision failures*). When the bit-cell area is large, the resistance of the MTJ is dominant compared to that of the access transistor. If V_{READ} is large, the state of the MTJ may be accidentally switched during read operations (also called *disturb failures*). In Fig. 7.19, $V_{\mathrm{READ}} = 0.3$ V results in significant number of disturb failures when the bit-cell area is larger than ~0.025 μm². This is not the case when $V_{\mathrm{READ}} = 0.1$ V, and the read failure probability approaches the minimum due to decision failure. To simplify the following analysis, we will consider the case when $V_{\mathrm{READ}} = 0.1$ V. Other parameters of the memory array that we analyze are listed in Table 7.1.

When the STT-MRAM is designed without ECC and MTJs of different Δ may be chosen, the optimum Δ depends on the bit-cell layout area (Pajouhi et al. 2015). Figure 7.20a graphs the total failure probability of the STT-MRAM bit-cell versus its layout area for different Δ. When the bit-cell area is small, write failure is dominant because the access transistor is unable to supply enough write current to the MTJ. The total failure probabilities approach the minimum determined by retention failure as the layout area of the bit-cell is increased. For bit-cell area ranging from ~0.0175 to ~0.025 μm^2, bit-cell having MTJs with $\Delta = 60$ has higher write failure probability than those having MTJs with $\Delta = 50$. Hence, write failure is still dominant and a small Δ is preferred. However, MTJ with $\Delta = 45$ is not suitable because the total failure probability of the bit-cells using them is limited by retention failure. If the bit-cell area can be larger than ~0.025 μm^2, retention failure is the dominant failure mechanism and MTJ with higher Δ is preferred. This result implies that at fixed STT-MRAM array area budget, the desired cache capacity and the total array failure probability needs to be jointly considered to select the Δ for the MTJ.

Table 7.1 Parameters of the memory array used in our example

CMOS Technology	32 nm
MTJ cross-section	Equivalent to 64 × 64 nm
Total Capacity	64 Mega-byte
Length of stored data word	64 bits
Hamming Code	71 bit code words, 1 bit error correction
BCH Code	78 bit code words, 2 bit error correction

The graph of total failure probability versus bit-cell layout area of STT-MRAM arrays implemented with different ECC schemes and Δ of the MTJ is shown in Fig. 7.20b. A trend similar to Fig. 7.20a may be observed. The ECC schemes encode data words as code words before writing into the array so as to correct for bit errors due to read, write or retention failures. The length of these code words are longer than that of data words as shown in Table 7.1. Due to stronger error correction capability of the Bose-Chaudhuri-Hocquenghem (BCH) code utilized, the minimum achievable total failure probability for STT-MRAM implemented with BCH code may be lower than that implemented with Hamming code. Note that the minimum achievable total failure probability in the STT-MRAM arrays implemented with ECC schemes is much lower than the STT-MRAM implemented without ECC. As the bit-cell layout area is increased, the point at which the total failure probability of

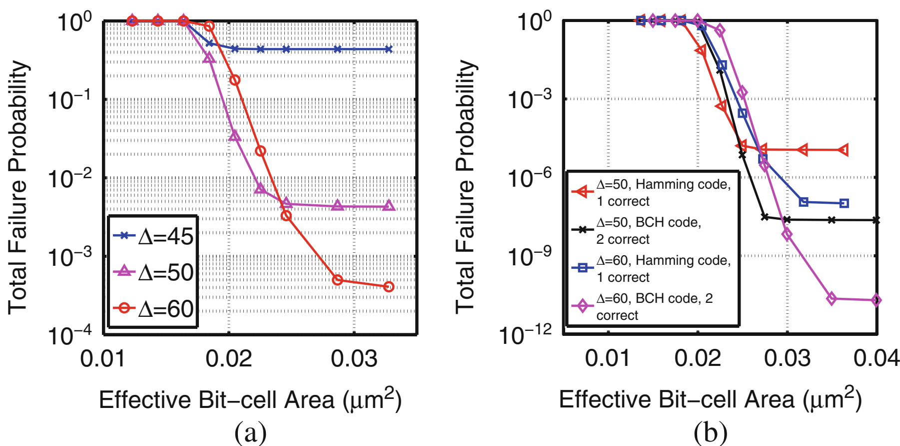


Fig. 7.20 Total failure probability of STT-MRAM arrays implemented (**a**) without ECC, and (**b**) with ECC based on Hamming code and BCH code with single- and double-error correction capability, respectively

Table 7.2 Possible design choices based on device/circuit/array co-design methodology presented

1	Minimize failure probability, no bit-cell area constraint	Use MTJ with $\Delta = 60$, use BCH code ECC
2	Bit-cell layout area $< 0.0225\ \mu m^2$, minimize failure probability	Use MTJ with $\Delta = 50$, use Hamming code ECC
3	$0.025\ \mu m^2 <$ bit-cell layout area $< 0.0285\ \mu m^2$, minimize failure probability	Use MTJ with $\Delta = 50$, use BCH code ECC

the arrays start decreasing depends on the Δ of the MTJ used, as well as the ability to correct for bit errors. For the arrays implemented without ECC, the one using MTJs with $\Delta = 45$ are more severely affected by retention failures than the array using MTJs with $\Delta = 50$. Consequently, it is not as apparent that the total failure probability of the array using MTJs with $\Delta = 45$ decreases earlier than the array using MTJs with $\Delta = 50$.

Figure 7.20b also shows there are differences between the STT-MRAM arrays implemented with Hamming code and those implemented with BCH code. One difference is the point at which the total failure probability starts decreasing. As mentioned earlier, write failure is dominant when the bit-cell layout area is small. Since fewer bits needs to be written into the array implemented with Hamming code for ECC as compared to the array implemented with BCH code for ECC, the probability of write failure is also smaller. Thus, the total failure probability of the array implemented with Hamming code as the ECC scheme decreases earlier than in the array implemented with BCH code as the ECC scheme at the same Δ.

Another difference is the point at which retention failure starts becoming dominant compared to write failures. As Fig. 7.20b shows, in the STT-MRAM arrays using MTJs with $\Delta = 60$, retention failure becomes dominant when the bit-cell layout area is larger than $0.03\ \mu m^2$. Retention failure becomes dominant in the STT-MRAM arrays using MTJs with $\Delta = 50$ when the bit-cell layout area is larger than $\sim 0.025\ \mu m^2$. Note that for the STT-MRAM implemented without ECC, retention failure becomes dominant when the bit-cell area is larger than ~0.0225 and $\sim 0.0275\ \mu m^2$ for the MTJs with $\Delta = 50$ and $\Delta = 60$, respectively. Hence, by considering the array layout area budget, total failure probability, and desired capacity, the optimum STT-MRAM array design can be selected. Table 7.2 lists some possible considerations and the optimum STT-MRAM array design corresponding to them. A similar analysis based on the flow we just described may be applied to design STT-MRAM arrays that fulfill the requirements of IoT systems.

7.6 Other Uses of MTJ for Sensing, Random Number Generation

We have discussed the design and optimization of STT-MRAM for IoT applications and presented several unique characteristics of STT-MRAM earlier. Several works in the literature exploit these unique characteristics to embed additional functionality within the STT-MRAM array while incurring minimal area overhead and near zero performance penalty (Ahn et al. 2013; Fukushima et al. 2014, 2015; Fong et al. 2015; Lee et al. 2013; Zhang et al. 2014a, b, 2015). These additional functionality are particularly interesting for ultralow power IoT systems. In the following sections, we will briefly discuss three design techniques that embed new functionality in STT-MRAM. We discuss how STT-MRAM may be used as truly random number generator (TRNG) and as physically unclonable function (PUF). TRNGs and PUFs may be used as on-chip security hardware, which is needed in IoT applications. Finally, we present a methodology for embedding read-only memory (ROM) in an STT-MRAM array. The embedded ROM stores data which may be different from that stored in RAM, and may be used to accelerate functions that use ROM, such as many digital signal processing (DSP) functions.

7.6.1 STT-MRAM as True Random Number Generators

The STT-MRAM write process is stochastic as we have discussed in Sect. 7.2.1 and may be exploited for random number generation. Since thermal disturbance is used as the entropy source, the STT-MRAM bit-cell may be used as a truly random number generator (TRNG) (Ahn et al. 2013; Fukushima et al. 2014). On-chip security hardware, which is crucial for IoT applications, may use TRNGs for generating high quality encryption keys. Hence, STT-MRAM that can also function as high quality TRNG with little overhead may be a very cost effective solution to enable on-chip security hardware suitable for IoT applications.

The operation concept of the STT-MRAM based TRNG illustrated in Fig. 7.21a is as follows. When a random number is requested, data stored in a row of STT-MRAM bit-cells is read out and stored in a buffer. Next, the random number is generated by performing a special write operation to the same row of bit-cells. During this step, a predetermined data (either all ‘0’ or all ‘1’) is first written to every bit-cell in the row with 100% probability of success. Thereafter, the complementary data is written to every bit-cell in the row (all ‘1’ if all ‘0’ were written previously and vice versa). When complementary data is being written, the current supplied to each bit-cell is such that the state of the bit-cell has 50% probability of switching. After the write operation is completed, the data stored in the bit-cells is read out and supplied to the circuitry requesting the random number. Finally, the data stored in the buffer is written back into the bit-cells to restore them to the previous state.

Note that three write operations are required in the TRNG scheme just described. The energy consumption of this TRNG scheme may be very high due to the high write energy of STT-MRAM bit-cells, which was discussed in Sect. 7.4. Figure 7.21b shows that the overall write energy may be reduced by reducing the number of write operations (Ahn et al. 2013). Once data has been read out of the bit-cells and stored in a buffer, we may immediately generate the random number by overwriting the bit-cells storing ‘0’ with ‘1’, and those storing ‘1’ with ‘0’. The current supplied to each bit-cell during this write operation is such that the probability of write failure in each bit-cell is 50%. After the random number is generated, we compare it to

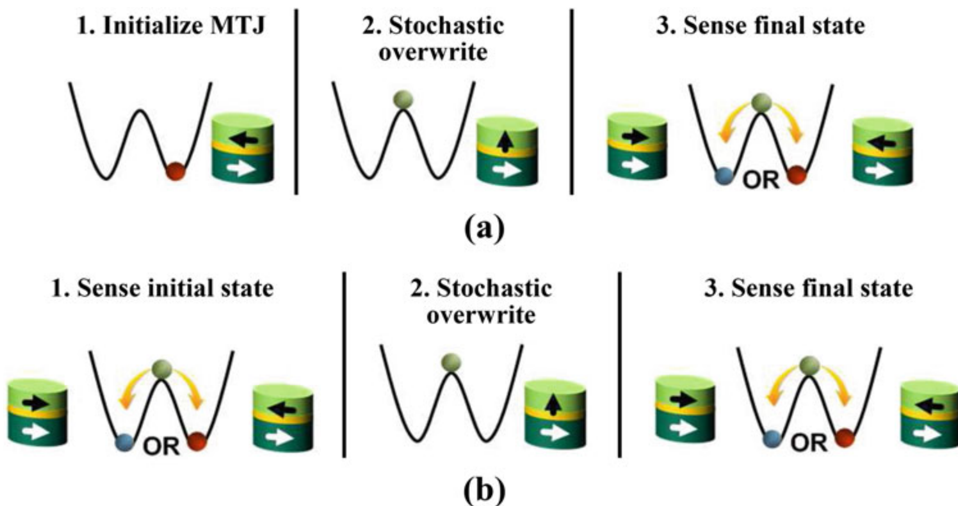


Fig. 7.21 These steps show how STT-MRAM may be used as a truly random number generator by exploiting the stochastic nature of its write operation. (**a**) The MTJ may be initialized to some predetermined state prior to random number generation, which incurs large energy dissipation. (**b**) The initialization step in (**a**) is redundant and may be eliminated to reduce energy consumption

the data stored in the buffer to identify the bit-cells whose data that has been modified by the random number generation process. The data in these bit-cells are then restored from the buffer.

Although the STT-MRAM based TRNG may generate very high quality random numbers (Ahn et al. 2013), it may be very challenging to choose the write current needed to implement a process variation tolerant random number generation process. Due to process variations, the amount of write current supplied to each bit-cell during the random number generation process may result in switching probabilities that are not exactly 50%. This issue may be exacerbated if the TRNG scheme with fewer write operations is used. Depending on the degradation of the switching probability in the random number generation process, strong postprocessing techniques may be required to ensure the quality of random numbers generated by STT-MRAM based TRNGs is sufficiently high (Dichtl 2008; Lacharme 2008; Von Neumann 1951).

7.6.2 Physically Unclonable Functions Using STT-MRAM

Another interesting aspect of STT-MRAM is that the process variations in the bit-cells may be exploited to implement a memory-based physically unclonable function (PUF) (Zhang et al. 2014a, b, 2015). PUFs may be used as on-chip security hardware that generates chip-unique keys for IoT applications. The input and output of the PUF are called the *challenge* and *response*, respectively. Since it is desired that the PUF generate chip-unique keys, the set of challenge-response pairs (CRPs) of a PUF must be unique (the *uniqueness* criterion). During operation, the set of CRPs for each PUF is fixed (the *reliability* criterion). To ensure that the PUF is indeed unclonable, the response of a PUF to any challenge exploits variations in the fabrication process so that it is random and not known at design time (the *randomness* criterion).

A memory-based PUF (MemPUF) may be implemented using 1T-1M STT-MRAM bit-cells by exploiting the single-ended nature of the sensing scheme (Zhang et al. 2014a). Every bit-cell in the STT-MRAM array may be grouped into pairs where one is the *data* cell and the other is the *reference* cell as shown in Fig. 7.22a. The challenge is given as the address to the STT-MRAM array and the response will be generated. In the flow chart shown in Fig. 7.22b, the data stored in the STT-MRAM is first copied to a buffer when the array operates as a PUF. When a challenge is given and a response is required, the data and reference

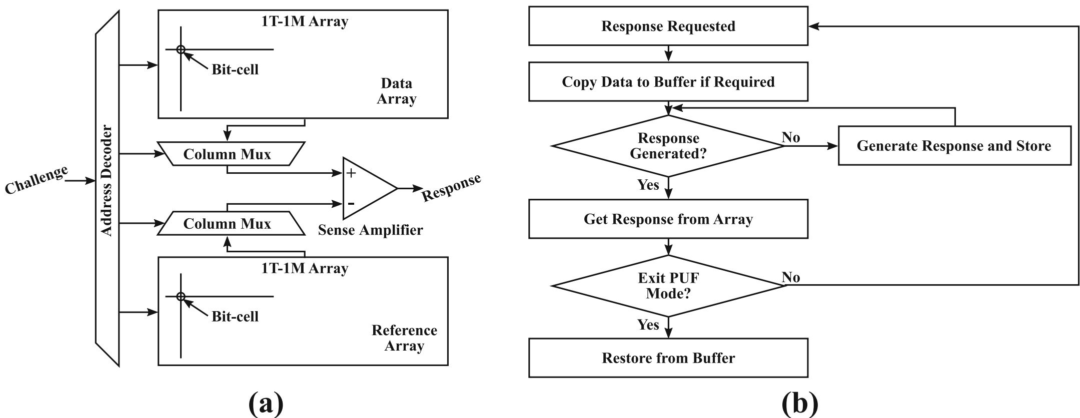


Fig. 7.22 (**a**) The schematic showing how an STT-MRAM array consisting of a data array and a reference array, which are identical 1T-1M bit-cell arrays, may be used as a physically unclonable function (PUF). (**b**) The flow chart shows how the PUF functionality may be used without overwriting the data stored in the STT-MRAM array

cells selected by the challenge are written with '0'. The response is generated by comparing the resistance of the data cell with its companion reference cell. Since the data cell and reference cell are storing the same values, the result of the resistance comparison depends on the process variations in the bit-cells. The corresponding bit in the response is '0' if the resistance of the data cell is smaller and '1' otherwise. When the STT-MRAM resumes operation as RAM, the data stored in the buffer is restored into the array.

The abovementioned scheme may not be reliable due to thermal perturbations. Since the data and reference cells store the same value during response generation, thermal perturbations in the magnetic layers of the MTJs in the bit-cells, which are random by nature, may affect the result of the resistance comparison. As a result, the response to the challenge may be unreliable (i.e., response to the challenge is not deterministic). The proposed STT-MRAM based MemPUF in (Zhang et al. 2014a) overcomes this design issue by using two operational phases: the *enrolment* and the *regeneration* phase. During the enrolment phase, every bit-cell in the array is written with '0'. Then, the resistance of every data cell is compared with its companion reference cell. The cell with the larger resistance is overwritten with a '1' and the enrolment phase ends. During the regeneration phase, the resistance of the data cell selected by the challenge is compared to that of its companion reference cell. The result of the comparison gives the corresponding bit of the response to the challenge. Hence, the operation of the proposed STT-MRAM based MemPUF must begin with an enrolment phase. The reliability of the proposed STT-MRAM based MemPUF is enhanced during the enrolment phase by ensuring that the data cell and reference cell stores complementary data. Furthermore, the CRPs generated may depend on whether '1' or '0' was written during the first step of the enrolment phase. If the *TMR* variation is sufficiently large and random, CRPs generated by writing '1' during the first step in the enrolment phase may be uncorrelated with those generated by writing '0' during the first step in the enrolment phase. This may be exploited to expand the set of CRPs to improve its resilience against attacks.

Although promising as on-chip security hardware for IoT applications, there are several disadvantages of the MemPUF scheme just described. First, the reliability of the MemPUF may be degraded by disturb failures. The currents flowing through the bit-cells during response generation must be sufficiently small that the states of the data cell and the reference cell are not accidentally switched. Next, to utilize the MemPUF as a RAM as well, a small buffer is needed. Prior to enrolment, the data stored in the selected data and reference cells are first copied into the buffer. The data in the buffer is restored to the data and reference cells once the MemPUF response is no longer needed, incurring additional write energy consumption.

7.6.3 Embedding Read-Only Memory in STT-MRAM

As illustrated earlier in Fig. 7.14, every column in the STT-MRAM array consists of a bit line to which bit-cells in the column are connected to. The physical connection of the STT-MRAM bit-cells in the array may be exploited such that every bit-cell stores a RAM bit in their constituent MTJ as well as a read-only memory (ROM) data (Fong et al. 2015; Lee et al. 2013). The embedding of ROM in a column of the array is enabled by an additional bit line as illustrated by Fig. 7.23. The layout of an example STT-MRAM bit-cell array without and with embedded ROM is shown in Fig. 7.24a, b, respectively. Note that when the size of the access transistor is large, the additional bit line may be placed without changing the footprint of the bit-cell array. Each bit-cell in the column is connected to either BL0 or BL1. Bit-cells connected to BL0 store a ROM value of '0' whereas those connected to BL1 store a ROM value of '1'.

In the ROM-embedded STT-MRAM array shown in Fig. 7.23, a pair of pass transistors connects BL0 and BL1 to the RAM mode peripheral circuitry. During RAM mode operations, the pair of pass transistors are turned ON. Hence,

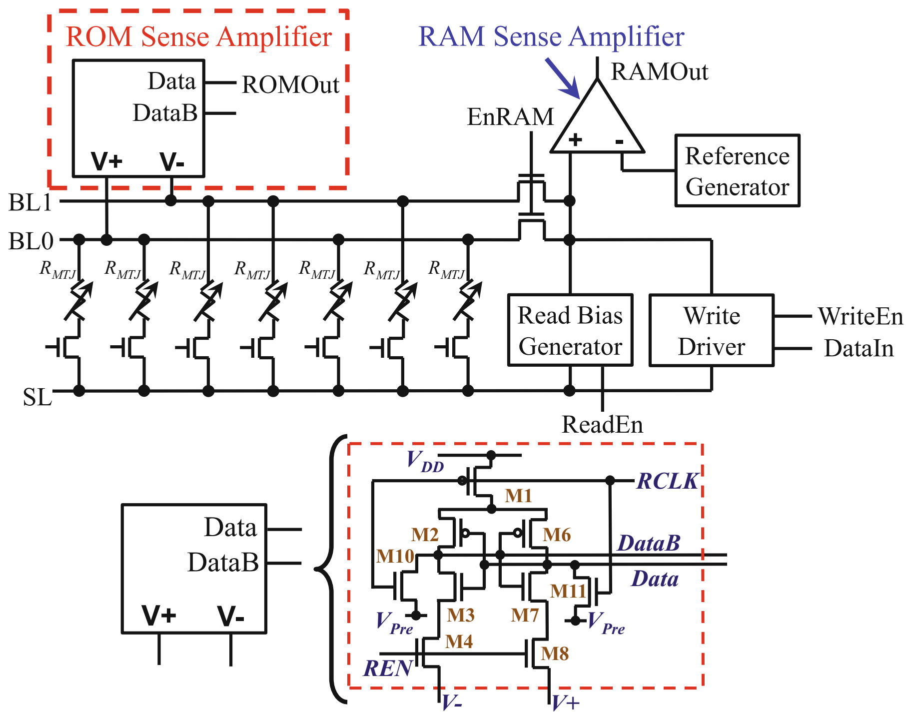


Fig. 7.23 This schematic shows a column of an STT-MRAM array with embedded read-only memory (ROM). The physical connection of each bit-cell to BL0 or BL1 is used to store ROM data in each bit-cell. RAM data is stored using the MTJ in each bit-cell and hence can be different from the ROM data

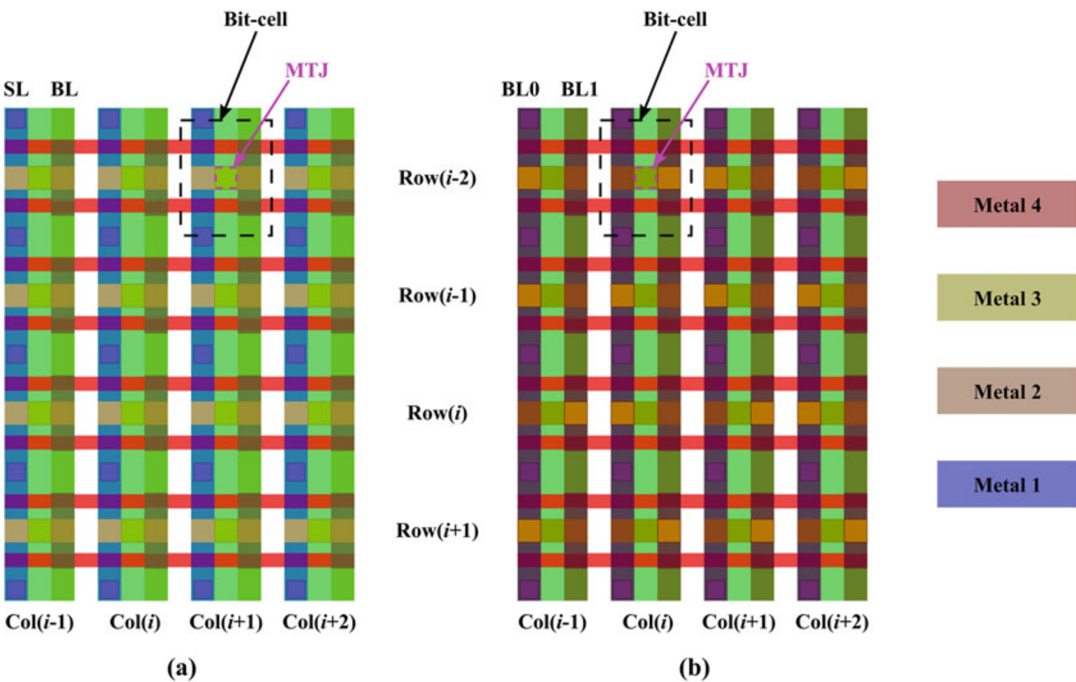


Fig. 7.24 The layout of an STT-MRAM bit-cell array (**a**) without embedded ROM and (**b**) with embedded ROM. ROM data is stored as the via connection from the MTJ to BL0 or BL1

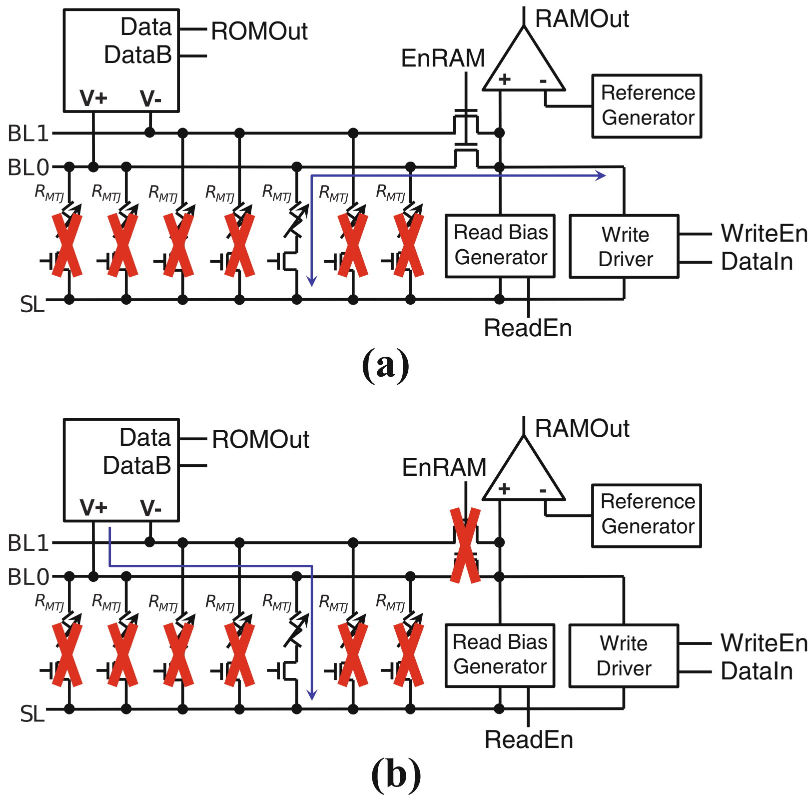


Fig. 7.25 Current path through the ROM-embedded STT-MRAM array during ROM operation when the bit-cell selected stores (**a**) ROM data '0' and (**b**) ROM data '1'

BL0 and BL1 are electrically shorted during RAM mode operations. Consequently, the peripheral circuitry is able to access any selected bit-cell regardless of whether it is connected to BL0 or BL1. During ROM mode operations, the pair of pass transistors is turned OFF. The SL of the column is then grounded. The access transistor of the selected bit-cell is turned ON next, followed by the ROM mode sensing latch. Figure 7.25a, b show the current path in the ROM-embedded STT-MRAM array during ROM operation when a bit-cell storing ROM data '0' and ROM data '1' is selected, respectively. Note that of the two bit lines, the one connected to the selected bit-cell has a much smaller resistance to SL compared to the other bit line. Hence, during ROM mode sensing operation, the output of the sensing latch that senses the bit line connected to the selected bit-cell is easily discharged to *GND* whereas the other is charged to V_{DD} by the cross-coupled inverter action in the latch. Note that because RAM mode operations require BL0 and BL1 to be electrically connected whereas ROM mode operations require BL0 and BL1 to be electrically disconnected, RAM mode and ROM mode operations to the same column of bit-cells cannot occur simultaneously.

The ROM embedded in the STT-MRAM may be used to store instructions and data closer to the processor core, and applications utilizing the ROM may be accelerated even though RAM mode and ROM mode operations cannot occur simultaneously. Note that without the embedded ROM, instructions and data are stored off-chip and need to be fetched on-chip before the processor core may operate on them (Fong et al. 2015; Lee et al. 2013). The cost of accessing off-chip memory may greatly outweigh the cost due to not being able to perform RAM mode and ROM

mode operations simultaneously. This is the case in the benchmark programs analyzed in (Fong et al. 2015; Lee et al. 2013) and it was observed that applications exploiting the ROM embedded in STT-MRAM may be accelerated by as much as 30%.

7.7 Perspectives and Trends

As we saw in this chapter, STT-MRAM may be suitable for a wide array of IoT applications. Comparisons of STT-MRAM with other eNVM technologies in Chap. 6 show that the low voltage and high speed of STT-MRAM is suitable for IoT applications. Thus, STT-MRAM may be useful for implementing memory systems for IoT applications with unstable power supplies. The non-volatility and relatively fast access speed of STT-MRAM may also enable IoT systems that can quickly transition between sleep and active states, which significantly improves energy efficiency. Another attractiveness of STT-MRAM is that its unique characteristics may also be exploited to embed functionality, which is useful for IoT applications, in the memory array with few overhead as discussed in Sect. 7.6. However, such cost savings must be weighed against the increase in STT-MRAM array design complexity as well as the implications on array test methodology and failure rate.

The immediate STT-MRAM design challenges that need to be addressed are the high write energy consumption, and the severely limited design space imposed by the two-terminal nature of the MTJ (Augustine et al. 2010, 2011; Mojumder and Roy 2012). Several works investigated alternative spintronic device structures that are based on the MTJ concept (Braganca et al. 2009; Fong and Roy 2013; Fong et al. 2013, 2014; Huda and Sheikholeslami 2013; Kim et al. 2013; Mojumder et al. 2011a, b, 2013; Shiota et al. 2012; Wang et al. 2012, 2013a; Wang and Chien 2013). Of these device proposals, the two-terminal device based on voltage-controlled magnetic anisotropy (Shiota et al. 2012; Wang et al. 2012, 2013a; Wang and Chien 2013) and the multi-terminal device based on spin Hall effect or spin-orbit torque (Kim et al. 2013; Wang et al. 2013a) have garnered much research interest recently.

The two-terminal device exploiting voltage-controlled magnetic anisotropy has a stack composition that is very similar to that in the conventional MTJ structure we presented earlier (Shiota et al. 2012; Wang et al. 2012, 2013a; Wang and Chien 2013). Hence, the device integration issues are not significantly different from that for the conventional MTJ. It was found that the perpendicular magnetic anisotropy of the CoFeB free layer in the MTJ may be modulated by the voltage-induced electric-field at the CoFeB/MgO interface. When the applied voltage reduces the perpendicular magnetic anisotropy, the hysteresis loop of the CoFeB free layer narrows. Due to the stray field from the pinned layer, the hysteresis loop is not centered about zero applied magnetic field. Hence, the applied voltage may narrow the hysteresis loop to the point that the stray field from the pinned layer aligns the free layer magnetization parallel to that of the pinned layer. The magnetization of the free layer may be aligned anti-parallel to that of the pinned layer using a magnetic field (Wang et al. 2012).

Alternatively, current-induced spin-transfer torque, just like in the conventional MTJ, may be used to switch the state of the MTJ (Wang and Chien 2013). When this is the case, the polarity of the voltage applied across the MTJ must be designed such that the current flowing through the MTJ tries to anti-parallelize the MTJ configuration. The magnitude of the applied voltage is then used to determine whether the stable MTJ state is anti-parallel (for small applied voltage) or parallel (for large applied voltage) (Wang and Chien 2013). A major disadvantage of this scheme is that although the current required to program the MTJ is reduced, there is a significant increase in MTJ resistance. As a result, the voltage needed to program the MTJ may be significantly larger than that available for IoT applications. Furthermore, this may limit the write energy reduction. Hence, the major focus of research on voltage-controlled magnetic

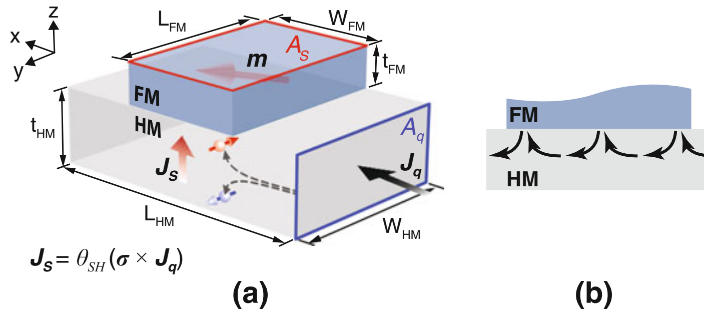


Fig. 7.26 (**a**) Current-induced torque is exerted on a magnetic layer adjacent to a heavy metal as shown here. (**b**) One possible mechanism is the spin Hall effect. After exchanging spin angular momentum with the adjacent magnetic layer, the spin polarization direction of an electron may be realigned by the spin Hall effect in the heavy metal and gets reinjected into the magnetic layer. Hence, an electron may transfer multiple units of spin angular momentum to the magnetic layer

anisotropy is on devising a method and/or device structure that allows the MTJ state to be switched without the need for current flow or magnetic fields.

Unlike the data storage devices using voltage-controlled magnetic anisotropy, those utilizing spin-orbit torque or spin Hall effect are still current-driven by nature (Kim et al. 2013; Wang et al. 2013a). In these devices, the MTJ grown on top of a heavy metal (or HM, such as Pt (Miron et al. 2011), β-Ta (Liu et al. 2012), β-W (Pai et al. 2012), CuBi (Niimi et al. 2012), or CuIr (Niimi et al. 2011)) with its free layer in direct contact with HM. Consider the structure shown in Fig. 7.26a. When charge current flows in HM, a spin current (i.e., a current in which all constituent electrons are identically spin-polarized) that flows perpendicular to the direction of charge current flow is generated via the spin Hall effect. In this spin current, electrons carrying opposite spin polarizations flow in opposite directions. Furthermore, the spin polarization direction of the electron is perpendicular to both charge and spin current flow directions. When the HM is in contact with the free layer of an MTJ, the spin current can get injected into the free layer of the MTJ from HM, which exerts spin-transfer torque on the magnetization of the free layer. After transferring spin angular momentum to the free layer of the MTJ, the spin polarization of the electron may get realigned and injected into the free layer again. Hence, an electron may transfer multiple units of spin angular momentum to the free layer of the MTJ as illustrated by Fig. 7.26b. Note that since the free layer of the MTJ is metallic, some charge current is shunted through it. Due to structural asymmetry and strong spin-orbit interaction, electrons at the interface between HM and the free layer may also experience a Rashba field (Gambardella and Miron 2011). The spin Hall effect and Rashba field contributes to the spin-orbit torque that is exerted on the magnetization of the free layer of the MTJ.

The spin Hall effect is promising for reducing the write energy in STT-MRAM because the spin current generated, $\boldsymbol{I}_\mathbf{s}$, can be controlled by engineering the geometry of the structure in Fig. 7.26a. $\boldsymbol{I}_\mathbf{s}$ is given as:

$$I_s = \theta_{\mathrm{SH}} \frac{A_s}{A_q} I_q \boldsymbol{\sigma} \tag{7.14}$$

where A_q and A_s are as shown in Fig. 7.26a, $\boldsymbol{\sigma}$ is the electron spin polarization direction, and θ_{SH} is the spin Hall angle of the HM. It was also experimentally demonstrated that spin-orbit torque may be used to switch the magnetization of a free layer with PMA (Yu et al. 2014). Furthermore, the current for programming the MTJ is not passed through the tunneling oxide of the MTJ and hence the reliability of the tunnel oxide barrier is improved. This is crucial in maintaining high *TMR*. The resistance of the

HM layer may also be reduced to lower the voltage required to drive the write current, which is desirable for reducing write energy. Because the device structure that enables spin-orbit torque is compatible with that having voltage-controlled magnetic anisotropy, a device structure that utilizes voltage-controlled magnetic anisotropy for assisting spin-orbit torque switching may be the key to achieving STT-MRAM for ultralow power IoT applications.

References

K. Abe, S. Fujita, T. H. Lee, Architecture of three-dimensional circuit using nanoscale memory devices, Eur. Micro Nano Syst., Noisy le Grand, France, 2004. TIMA, Grenoble, France (2004), pp. 225–229

J. Ahn, S. Yoo, K. Choi, Write intensity prediction for energy-efficient non-volatile caches, *Int. Symp. Low Power Electron. Des.* (2013), pp. 223–228

D. Apalkov, S. Watts, A. Driskill-Smith, E. Chen, Z. Diao, V. Nikitin, Comparison of scaling of in-plane and perpendicular spin transfer switching technologies by micromagnetic simulation. IEEE Trans. Magn. **46**(6), 2240–2243 (2010)

C. Augustine, A. Raychowdhury, D. Somasekhar, J. Tschanz, K. Roy, V. K. De, Numerical analysis of typical STT-MTJ stacks for 1T-1R memory arrays, *Int. Electron Devices Meet.* (2010), pp. 22.7.1–22.7.4

C. Augustine, A. Raychowdhury, D. Somasekhar, J. Tschanz, V. De, K. Roy, Design space exploration of typical stt mtj stacks inmemory arrays in the presence of variability and disturbances. IEEE Trans. Electron Devices **58**(12), 4333–4343 (2011)

C. Augustine, N.N. Mojumder, X. Fong, S.H. Choday, S.P. Park, K. Roy, Spin-transfer torque mrams for low power memories: perspective and prospective. IEEE Sens. J. **12**(4), 756–766 (2012)

L. Berger, Emission of spin waves by a magnetic multilayer traversed by a current. Phys. Rev. B **54**(13), 9353–9358 (1996)

P.M. Braganca, J.A. Katine, N.C. Emley, D. Mauri, J.R. Childress, P.M. Rice, E. Delenia, D.C. Ralph, R.A. Buhrman, A three-terminal approach to developing spin-torque written magnetic random access memory cells. IEEE Trans. Nanotechnol. **8**(2), 190–195 (2009)

W.F. Brown, Thermal fluctuations of a single-domain particle. Phys. Rev. **130**(5), 1677–1686 (1963)

D. Datta, B. Behin-Aein, S. Datta, S. Salahuddin, Voltage asymmetry of spin-transfer torques. IEEE Trans. Nanotechnol. **11**(2), 261–272 (2012)

T. Devolder, Scalability of magnetic random access memories based on an in-plane magnetized free layer. Appl. Phys. Exp. **4**(9), 093001 (2011)

M. Dichtl, Bad and good ways of post-processing biased physical random numbers, in *Fast Software Encryption* (Springer, Berlin, 2008), pp. 137–152

R. Dorrance, F. Ren, Y. Toriyama, A. A. Hafez, C. K. Yang, D. Markovic, Scalability and design-space analysis of a 1T-1MTJ memory cell, in *IEEE/ACM Int. Symp. Nanoscale Archit.* (2011), pp. 32–36

X. Fong, K. Roy, Complementary polarizers STT-MRAM (CPSTT) for on-chip caches. IEEE Electron Device Lett. **34**(2), 232–234 (2013)

X. Fong, S.H. Choday, K. Roy, Bit-cell level optimization for non-volatile memories using magnetic tunnel junctions and spin-transfer torque switching. IEEE Trans. Nanotechnol. **11**(1), 172–181 (2012)

X. Fong, K. Roy, Low-power robust complementary polarizer STTMRAM (CPSTT) for on-chip caches, in *5th IEEE Int. Mem. Work.* (2013), pp. 88–91

X. Fong, R. Venkatesan, A. Raghunathan, K. Roy, Non-volatile complementary polarizer spin-transfer torque on-chip caches: a device/circuit/systems perspective. IEEE Trans. Magn. **50**(10), 1–11 (2014)

X. Fong, R. Venkatesan, D. Lee, A. Raghunathan, K. Roy, Embedding read-only memory in spin-transfer torque mram based on-chip caches. IEEE Trans. Very Large Scale Integr. Syst. **24**(3), 992–1002 (2016)

A. Fukushima, T. Seki, K. Yakushiji, H. Kubota, H. Imamura, S. Yuasa, K. Ando, Spin dice: a scalable truly random number generator based on spintronics. Appl. Phys. Exp. **7**(8), 083001 (2014)

A. Fukushima, K. Yakushiji, H. Kubota, S. Yuasa, Spin dice (physical random number generator using spin torque switching) and its thermal response, in *IEEE Magn. Conf.* (2015), pp. 1–1

P. Gambardella, I.M. Miron, Current-induced spin-orbit torques. Philos. Trans. A. Math. Phys. Eng. Sci. **369**, 3175–3197 (2011)

Y. Huai, Spin-transfer torque MRAM (STT-MRAM): challenges and prospects. AAPPS Bull. **18**(6), 33–40 (2008)

Y. Huai, F. Albert, P. Nguyen, M. Pakala, T. Valet, Observation of spin-transfer switching in deep submicron-sized and low-resistance magnetic tunnel junctions. Appl. Phys. Lett. **84**(16), 3118–3120 (2004)

S. Huda, A. Sheikholeslami, A novel STT-MRAM cell with disturbance-free read operation. IEEE Trans. Circ. Syst. I, Reg. Papers **60**(6), 1534–1547 (2013)

S. Ikeda, K. Miura, H. Yamamoto, K. Mizunuma, H.D. Gan, M. Endo, S. Kanai, J. Hayakawa, F. Matsukura, H. Ohno, A perpendicular-anisotropy CoFeB-MgO magnetic tunnel junction. Nat. Mater. **9** (9), 721–724 (2010)

A. Jog, A. K Mishra, C. Xu, Y. Xie, V. Narayanan, R. Iyer, C. R. Das, Cache revive: architecting volatile STT-RAM caches for enhanced performance in CMPs, in *Proceedings of the 49th Annu. Des. Autom. Conf.—DAC'12* (2012), p. 243

J. Jung, Y. Nakata, M. Yoshimoto, H. Kawaguchi, Energy-efficient spin-transfer torque RAM cache exploiting additional all-zero-data flags, in *Int. Symp. Qual. Electron. Des.* (2013), pp. 216–222

S. H. Kang, X. Li, S. Gu, K. Lee, X. Zhu, STT MRAM magnetic tunnel junction architecture and integration (2014)

J. Katine, F. Albert, R. Buhrman, E. Myers, D. Ralph, Current-driven magnetization reversal and spin-wave excitations in Co/Cu/Co pillars. Phys. Rev. Lett. **84** (14), 3149–3152 (2000)

Y. Kim, S. K. Gupta, S. P. Park, G. Panagopoulos, K. Roy, Write-optimized reliable design of STT MRAM, in *Proceedings of the 2012 ACM/IEEE Int. Symp. Low power Electron. Des.—ISLPED' 12* (2012), p. 3

Y. Kim, S.H. Choday, K. Roy, DSH-MRAM: differential spin hall MRAM for on-chip memories. IEEE Electron Device Lett. **34**(10), 1259–1261 (2013)

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, K. Ando, Lower-current and fast switching of a perpendicular TMR for high speed and high density spin-transfer-torque MRAM, in *2008 I.E. Int. Electron Devices Meet.* (2008), pp. 1–4

K. Kwon, S.H. Choday, Y. Kim, K. Roy, AWARE (Asymmetric Write Architecture with REdundant blocks): a high write speed STTMRAM cache architecture. IEEE Trans. Very Large Scale Integr. Syst. **1**, 1–1 (2013)

K. Kwon, X. Fong, P. Wijesinghe, P. Panda, K. Roy, High-density & robust STT-MRAM array through device/circuit/architecture interactions. IEEE Trans. Nanotechnol., **14**(6), 1024–1034 (2015)

P. Lacharme, Post-processing functions for a biased physical random number generator, in *Fast Software Encryption* (Springer, Berlin, 2008), pp. 334–342

D. Lee, S. K. Gupta, K. Roy, High-performance lowenergy STT MRAM based on balanced write scheme, in *Proceedings of the 2012 ACM/IEEE Int. Symp. Low power Electron. Des.—ISLPED' 12* (2012), p. 9

D. Lee, X. Fong, K. Roy, R-MRAM: A ROM-embedded STT MRAM cache. IEEE Electron Device Lett. **34** (10), 1256–1258 (2013)

J. Li, P. Ndai, A. Goel, S. Salahuddin, K. Roy, Design paradigm for robust spin-torque transfer magnetic RAM (STT MRAM) from circuit/architecture perspective. IEEE Trans. Very Large Scale Integr. Syst. **18**(12), 1710–1723 (2010)

Q. Li, J. Li, L. Shi, C. J. Xue, Y. Chen, Y. He, Compiler-assisted refresh minimization for volatile STT-RAM cache, in *2013 18th Asia South Pacific Des. Autom. Conf.* (2013), pp. 273–278

J. Li, L. Shi, Q. Li, C.J. Xue, Y. Chen, Y. Xu, W. Wang, Low-energy volatile STT-RAM cache design using cache-coherence-enabled adaptive refresh. ACM Trans. Des. Autom. Electron. Syst. **19**(1), 1–23 (2013b)

C. J. Lin, S. H. Kang, Y. J. Wang, K. Lee, X. Zhu, W. C. Chen, X. Li, W. N. Hsu, Y. C. Kao, M. T. Liu, M. Nowak, N. Yu, 45 nm low power CMOS logic compatible embedded STT MRAM utilizing a reverse-connection 1T/1MTJ cell, in *2009 I.E. Int. Electron Devices Meet.* (2009), pp. 1–4

L. Liu, C.-F. Pai, Y. Li, H.W. Tseng, D.C. Ralph, R.A. Buhrman, Spin-torque switching with the giant spin hall effect of tantalum. Science **336**(6081), 555–558 (2012)

M. Mao, G. Sun, Y. Li, A. K. Jones, Y. Chen, Prefetching techniques for STT-RAM based last-level cache in CMP systems, in *2014 19th Asia South Pacific Des. Autom. Conf.* (2014), pp. 67–72

I.M. Miron, K. Garello, G. Gaudin, P.-J. Zermatten, M.V. Costache, S. Auffret, S. Bandiera, B. Rodmacq, A. Schuhl, P. Gambardella, Perpendicular switching of a single ferromagnetic layer induced by in-plane current injection. Nature **476**(7359), 189–193 (2011)

N.N. Mojumder, K. Roy, Proposal for switching current reduction using reference layer with tilted magnetic anisotropy in magnetic tunnel junctions for spin-transfer torque (STT) MRAM. IEEE Trans. Electron Devices **59**(11), 3054–3060 (2012)

N.N. Mojumder, S.K. Gupta, S.H. Choday, D.E. Nikonov, K. Roy, A three-terminal dual-pillar STT-MRAM for high-performance robust memory applications. IEEE Trans. Electr. Devices **58**(5), 1508–1516 (2011a)

N. N. Mojumder, S. K. Gupta, K. Roy, Dual pillar spin transfer torque MRAM with tilted magnetic anisotropy for fast and error-free switching and near-disturb-free read operations, in *69th Device Res. Conf.* (2011), pp. 67–68

N.N. Mojumder, X. Fong, C. Augustine, S.K. Gupta, S.H. Choday, K. Roy, Dual pillar spin-transfer torque mrams for low power applications. ACM J. Emerg. Technol. Comput. Syst. **9**(2), 1–17 (2013)

E.B. Myers, Current-induced switching of domains in magnetic multilayer devices. Science **285**(5429), 867–870 (1999)

H. Naeimi, C. Augustine, A. Raychowdhury, S. Lu, J. Tschanz, STTRAM scaling and retention failure. Intel Technol. J. **17**(1), 54–75 (2013)

Y. Niimi, M. Morota, D.H. Wei, C. Deranlot, M. Basletic, A. Hamzic, A. Fert, Y. Otani, Extrinsic spin Hall effect induced by iridium impurities in copper. Phys. Rev. Lett. **106**(12), 126601 (2011)

Y. Niimi, Y. Kawanishi, D.H. Wei, C. Deranlot, H.X. Yang, M. Chshiev, T. Valet, A. Fert, Y. Otani, Giant spin Hall effect induced by skew scattering from bismuth impurities inside thin film CuBi alloys. Phys. Rev. Lett. **109**, 156602 (2012)

J. Nogués, J. Sort, V. Langlais, V. Skumryev, S. Suriñach, J.S. Muñoz, M.D. Baró, Exchange bias in nanostructures. Phys. Rep. **422**(3), 65–117 (2005)

T. Ohsawa, H. Koike, S. Miura, H. Honjo, K. Tokutome, S. Ikeda, T. Hanyu, H. Ohno, T. Endoh, 1Mb 4T-2MTJ nonvolatile STT-RAM for embedded memories using 32b fine-grained power gating technique with 1.0ns/200ps wake-up/power-off times, in *2012 Symp. VLSI Circuits* (2012), pp. 46–47

C.F. Pai, L. Liu, Y. Li, H.W. Tseng, D.C. Ralph, R.A. Buhrman, Spin transfer torque devices utilizing the giant spin Hall effect of tungsten. Appl. Phys. Lett. **101**(12), 1–5 (2012)

Z. Pajouhi, X. Fong, K. Roy, Device/Circuit/Architecture Co-Design of Reliable STT-MRAM, in *Proc. 2015 Des. Autom. Test Eur. Conf. Exhib.* (2015), pp. 1437–1442

S. P. Park, S. Y. Kim, D. Lee, J.-J. Kim, W. P. Griffin, K. Roy, Column-selection-enabled 8T SRAM array with 1R/1W multi-port operation for DVFS-enabled processors, in *IEEE/ACM Int. Symp. Low Power Electron. Des.* (2011), pp. 303–308

S. P. Park, S. Gupta, N. Mojumder, A. Raghunathan, K. Roy, Future cache design using STT MRAMs for improved energy efficiency, in *Proceedings of the 49th Annu. Des. Autom. Conf.—DAC'12* (2012), p. 492

S.S.P. Parkin, N. More, K.P. Roche, Oscillations in exchange coupling and magnetoresistance in metallic superlattice structures: Co/Ru, Co/Cr, and Fe/Cr. Phys. Rev. Lett. **64**(19), 2304–2307 (1990)

M. Rasquinha, D. Choudhary, S. Chatterjee, S. Mukhopadhyay, S. Yalamanchili, An energy efficient cache design using spin torque transfer (STT) RAM, in *Proceedings of the 16th ACM/IEEE Int. Symp. Low power Electron. Des.—ISLPED'10* (2010), p. 389

ITRS Roadmap (2014)

S. Salahuddin, D. Datta, S. Datta, Key role of non equilibrium spin density in determining spin torque, in *2008 Device Res. Conf.* (2008), pp. 161–162

Y. Shiota, T. Nozaki, F. Bonell, S. Murakami, T. Shinjo, Y. Suzuki, Induction of coherent magnetization switching in a few atomic layers of FeCo using voltage pulses. Nat. Mater. **11**(1), 39–43 (2012)

J.C. Slonczewski, Current-driven excitation of magnetic multilayers. J. Magn. Magn. Mater. **159**, L1–L7 (1996)

C. W. Smullen, V. Mohan, A. Nigam, S. Gurumurthi, M. R. Stan, Relaxing non-volatility for fast and energy-efficient STT-RAM caches, in *2011 I.E. 17th Int. Symp. High Perform. Comput. Archit.* (2011), pp. 50–61

J. Sun, Spin-current interaction with a monodomain magnetic body: a model study. Phys. Rev. B **62**(1), 570–578 (2000)

Z. Sun, X. Bi, H. Li, W.-F. Wong, Z.-L. Ong, X. Zhu, W. Wu, Multi retention level STT-RAM cache designs with a dynamic refresh scheme, in *Proceedings of the 44th Annu. IEEE/ACM Int. Symp. Microarchitecture—MICRO'11* (2011), p. 329

G. Sun, Y. Zhang, Y. Wang, Y. Chen, Improving energy efficiency of write-asymmetric memories by log style write, in *Proceeding of the 2012 ACM/IEEE Int. Symp. Low power Electron. Des., ISLPED'12* (2012), pp. 173–178

Z. Sun, X. Bi, H. Li, W.-F. Wong, X. Zhu, STT-RAM cache hierarchy with multiretention MTJ designs. IEEE Trans. Very Large Scale Integr. Syst. **22**(6), 1281–1293 (2014)

S. Tsunegi, H. Kubota, S. Tamaru, K. Yakushiji, M. Konoto, A. Fukushima, T. Taniguchi, H. Arai, H. Imamura, S. Yuasa, Damping parameter and interfacial perpendicular magnetic anisotropy of FeB nanopillar sandwiched between MgO barrier and cap layers in magnetic tunnel junctions. Appl. Phys. Exp. **7**(3), 033004 (2014)

J. Von Neumann, Various techniques used in connection with random digit. Natl. Bur. Stand. Appl. Math. Ser. **12**, 36–38 (1951)

W.G. Wang, C.L. Chien, Voltage-induced switching in magnetic tunnel junctions with perpendicular magnetic anisotropy. J. Phys. D. Appl. Phys. **46**(7), 74004 (2013)

W.-G. Wang, M. Li, S. Hageman, C.L. Chien, Electric-field assisted switching in magnetic tunnel junctions. Nat. Mater. **11**(1), 64–68 (2012)

K.L. Wang, J.G. Alzate, P. Khalili Amiri, Low-power non-volatile spintronic memory: STT-RAM and beyond. J. Phys. D. Appl. Phys. **46**(7), 074003 (2013a)

J. Wang, X. Dong, Y. Xie, OAP: an obstruction-aware cache management policy for STT-RAM last-level caches, in *Des. Autom. Test Eur. Conf. Exhib.* (2013), pp. 847–852

W. Xu, Y. Chen, X. Wang, T. Zhang, Improving STT MRAM storage density through smaller-than-worst-case transistor sizing, in *Des. Autom. Conf.* (2009), pp. 87–90

S. Yamamoto, S. Sugahara, Nonvolatile static random access memory using magnetic tunnel junctions with current-induced magnetization switching architecture. Jpn. J. Appl. Phys. **48**(4), 043001 (2009)

T. Yamauchi, Prospect of embedded non-volatile memory in the smart society, in *2015 Int. Symp. VLSI Technol. Syst. Appl.* (2015), pp. 1–2

J. Yang, B. Geller, M. Li, T. Zhang, An information theory perspective for the binary STT-MRAM cell operation channel, in *IEEE Trans. Very Large Scale Integr. Syst.* (2015), pp. 1–1

G. Yu, P. Upadhyaya, Y. Fan, J.G. Alzate, W. Jiang, K.L. Wong, S. Takei, S.A. Bender, L.-T. Chang, Y. Jiang, M. Lang, J. Tang, Y. Wang, Y. Tserkovnyak, P.K. Amiri, K.L. Wang, Switching of perpendicular magnetization by spin-orbit torques in the absence of external magnetic fields. Nat. Nanotechnol. **9**, 548–554 (2014)

S. Yuasa, T. Nagahama, A. Fukushima, Y. Suzuki, K. Ando, Giant room-temperature magnetoresistance in single-crystal Fe/MgO/Fe magnetic tunnel junctions. Nat. Mater. **3**(12), 868–871 (2004)

L. Zhang, X. Fong, C.-H. Chang, Z. H. Kong, K. Roy, Feasibility study of emerging non-volatile memory based physical unclonable functions, in *2014 I.E. 6th Int. Mem. Work.* (2014), pp. 1–4

L. Zhang, X. Fong, C.-H. Chang, Z. H. Kong, K. Roy, Highly reliable memory-based physical unclonable function using spin-transfer torque MRAM, in *2014 I.E. Int. Symp. Circuits Syst.* (2014), pp. 2169–2172

L. Zhang, X. Fong, C.-H. Chang, Z.H. Kong, K. Roy, Optimizating emerging nonvolatile memories for dual-mode applications: data storage and key generator. IEEE Trans. Comput. Des. Integr. Circuits Syst. **34**(7), 1176–1187 (2015)

B. Zhao, J. Yang, Y. Zhang, Y. Chen, H. Li, Architecting a common source-line array for bipolar non-volatile memory devices, in *2012 Des. Autom. Test Eur. Conf. Exhib.* (2012), pp. 1451–1454

P. Zhou, B. Zhao, J. Yang, Y. Zhang, Energy reduction for STT-RAM using early write termination, in *IEEE/ACM Int. Conf. Comput. Des.—Dig. Tech. Pap.* (2009), pp. 264–268