Chapter 3
Design and Optimization of Spin-Transfer Torque MRAMs

Xuanyao Fong, Sri Harsha Choday, and Kaushik Roy

Abstract In this chapter, reviews the basics and modeling of spin-transfer torque magnetic RAM (STT-MRAM) for circuit-level failure analysis. A methodology for analyzing failures in STT-MRAM bit-cells is also presented. The optimization of STT-MRAM bit-cells using the presented framework is then discussed, along with several circuit and array architecture-level failure mitigation techniques. We will show that despite the relatively high write energy in STT-MRAM, large capacity last level caches based on STT-MRAM can be more energy efficient than their SRAM counterparts due to the unique characteristics of STT-MRAM.

The cache capacity of high-performance microprocessors is increasing as transistor technology is scaled down. Since the leakage power also increases exponentially with the scaling down of transistor technology, the power dissipation of on-chip caches is an increasingly dominant component of power dissipation in high-performance microprocessors. Non-volatile memories have been proposed as a solution for mitigating the increasing power dissipation in high-performance on-chip caches. Among the currently available non-volatile memory technologies, only spin-transfer torque magnetic random access memory (STT-MRAM) has the desired characteristics for high-performance on-chip cache applications [1]. In this chapter, we discuss the design optimization and modeling of STT-MRAMs, and its potential application in high-performance on-chip caches.

3.1 MRAM Storage Device: The Magnetic Tunnel Junction

The storage device in MRAM is the magnetic tunnel junction or MTJ. An MTJ, as shown in Fig. 3.1, consists of a soft ferromagnetic layer which stores the information (also called the “free” layer), a tunneling layer (usually AlO_x or more commonly,

X. Fong • S.H. Choday • K. Roy (✉)
School of Electrical and Computer Engineering, Purdue University,
West Lafayette, IN 47907, USA
e-mail: xfong@purdue.edu; schoday@purdue.edu; kaushik@purdue.edu

© Springer Science+Business Media New York 2015

R.O. Topaloglu (ed.), *More than Moore Technologies for Next Generation Computer Design*, DOI 10.1007/978-1-4939-2163-8_3

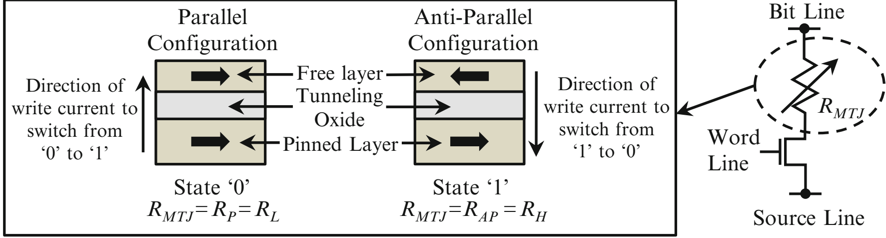

Fig. 3.1 The storage device in the MRAM memory cell is the magnetic tunnel junction (illustrated inset). The memory cell consists of an access transistor and the storage device connected as shown. The current direction for programming the cell using spin-transfer torque is also shown

MgO), and a reference ferromagnetic layer (also called the "fixed" or "pinned" layer). The MTJ can be switched between two stable states. When both the free and the pinned layers are magnetically aligned, the configuration is called the "parallel" state (P), and when the free and the pinned layers are anti-aligned magnetically, the configuration is called the "anti-parallel" state (AP). A metric for MTJ as shown in [2] is its resistance–area (*RA*) product. The *RA* product of the MTJ depends exponentially on the tunnel oxide thickness (t_{MgO}) since the mechanism for electron transport is tunneling. At the same t_{MgO}, the MTJ resistance, R_{MTJ}, depends linearly on the cross-sectional area of the MTJ (A_{MTJ}), similar to an Ohmic conductor. 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 is due to the difference in density of states around the Fermi energy, E_F, in the ferromagnetic layers [3]. When the MTJ is in the P state, 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 state. Thus, R_{MTJ} is low in the P state ($R_{MTJ} = R_P = R_L$) and high in the AP state ($R_{MTJ} = R_{AP} = R_L$). This difference in R_{MTJ}, termed the "tunneling magneto-resistance ratio" (or *TMR*), is given by

$$TMR = \frac{R_{AP} - R_P}{R_P} \times 100\ \% \tag{3.1}$$

and is an important metric for the performance of MTJs as memory elements. Since binary data are represented by and stored as the resistance state of the MTJ, a larger *TMR* also means that the MTJ states can be distinguished more easily. A constant voltage or constant current scheme can be used to sense R_{MTJ} and hence, the MTJ state [4, 5]. In the constant voltage scheme, a fixed voltage is applied across the MTJ and the resulting current through the MTJ is compared to a reference current. The current flowing through the MTJ can be either higher or lower than the reference current, depending on the resistance state of the MTJ. The advantage of the constant voltage scheme is that the current flowing through the MTJ during read operations may be amplified in the sense amplifier to improve sensing speed. However, the disadvantage is that the result of the sensing needs to be converted into an output

voltage. In case of constant current scheme, a fixed current is passed through the MTJ and the voltage developed across the bit-line and the source-line is compared with a reference voltage. The constant current scheme has the advantage that the result of the sensing is already in the voltage domain and hence, no conversion is required. However, the current required to generate sufficient voltage signal for sensing may be large enough to cause *disturb failures*, which will be discussed in detail later.

The magnetic layers are stabilized against thermal effects by engineering them with anisotropies during fabrication. The most common form of anisotropy engineered into the magnetic layers 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. When the volume of the magnet is reduced, the *uniaxial anisotropy energy* must be proportionally increased to maintain the same stability. We will discuss this in more detail in the later sections.

Nano-scale MTJs may be switched using the spin-transfer torque phenomenon which was theoretically predicted by Slonczewski and Berger independently in 1996 [6, 7]. Since then many experiments have observed spin-transfer torque (STT) switching [8–10]. STT exists because magnetism in ferromagnetic metals arises due to the spin property of electrons. The magnetization of the ferromagnet points in a particular direction when the majority of electron spins in it are aligned in that direction. Hence, when current flows through the MTJ, the ferromagnetic layers act as spin filters that polarizes the flowing electrons. Electrons in a spin polarized current flowing into a ferromagnetic layer are able to transfer their spin momentum to it. The spin momentum transferred exerts a torque on the magnetization of the ferromagnetic layer. The magnetization of the ferromagnetic layer is switched if the torque is large enough to overcome all other energies in the ferromagnetic layer. 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.

In an MTJ, the pinned layer is magnetically pinned whereas the free layer is not. Hence, it is easier for spin-transfer torque to switch the free layer than to switch the pinned layer. Let us consider what happens when electrons are flowing from the pinned layer to the free layer in an MTJ. The pinned layer polarizes the incoming electrons which then flow into the free layer. These electrons are polarized in the spin direction of the pinned layer and transfer their spin momentum to the free layer. Hence, a spin-transfer torque is exerted on the free layer to align its magnetization parallel with the pinned layer. Consider instead when electrons flow from the free layer to the pinned layer. Electrons entering the free layer from the metallic interconnect are not polarized and can have any spin direction. Electrons with same spin direction as the pinned layer are able to tunnel across the oxide easily. However, electrons 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 free layer

magnetization anti-parallel with the pinned layer. When the electrons transfer their spin angular momentum to the free layer, their spin directions become aligned with the spin polarization of the pinned layer. They may then tunnel across the oxide easily. From this discussion, we can see that the process of parallelizing the free and pinned layers is more efficient than the anti-parallelizing process, resulting in asymmetry in I_C and J_C [3, 11]. It has been reported that J_C when anti-parallelizing the MTJ can be 10–200 % larger than for parallelizing the MTJ [11, 12].

3.2 Modeling Magnetic Tunnel Junctions

The transient behavior of an MTJ can be modeled only if the essential physics in it are captured. The *I–V* characteristic of the MTJ depends on physical parameters of the MTJ, such as the thickness of the tunneling oxide and the cross-sectional area of the MTJ, and on the magnetization directions of the free and the pinned layers. Since the magnetization of the free layer does not change instantaneously during switching, R_{MTJ} also transition smoothly during MTJ switching. Accurate modeling of the transient behavior of MTJs must model the transient behavior of the free layer and relate it to the *I–V* characteristic of the MTJ. The transient behavior of the free layer magnetization may be modeled using the *Landau–Lifshitz–Gilbert* (LLG) [13] equation, and the *I–V* characteristic of the MTJ may be modeled using the *Non-Equilibrium Green's Function* (NEGF) [3] approach.

3.2.1 The Non-Equilibrium Green's Function (NEGF) Approach

The Non-Equilibrium Green's Function (NEGF) approach may be used to simulate electronic transport through an MTJ [3]. The approach requires the effective mass Hamiltonian representing the MTJ and the MTJ biasing conditions, to be written first. The *I–V* characteristics may be calculated by solving Non-Equilibrium Green's Function (NEGF) equations. Details of the approach are published in [3, 14] and are beyond the scope of this chapter.

The NEGF approach to modeling the *I–V* characteristic of the MTJ has the advantage that model parameters correspond to material parameters and may be obtained from experimental measurements. The model may then be used to predict MTJ characteristics and then validated experimentally. Figure 3.2 shows the successful calibration of the NEGF model to experimentally measured data published in [15, 16].

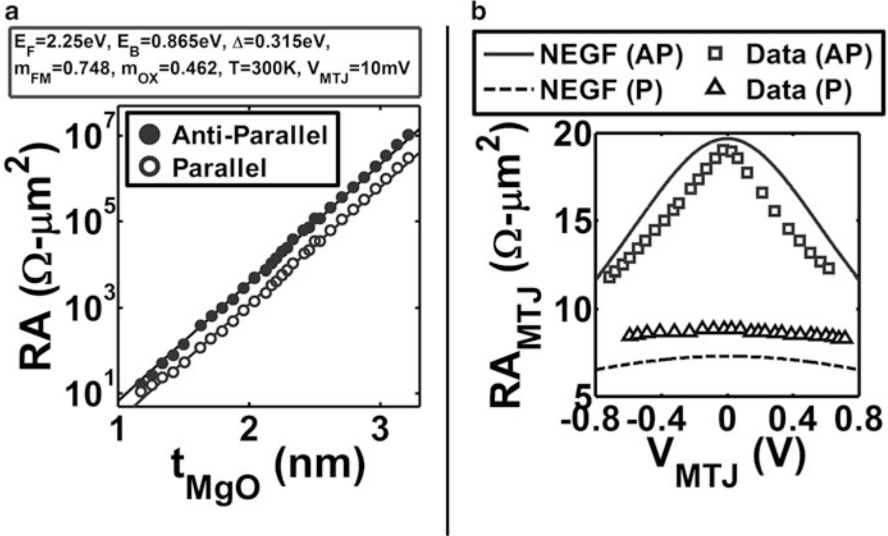

Fig. 3.2 The successful calibration of the NEGF model to experimentally measured data reported in the literature. (**a**) and (**b**) show results for calibration to data from [15] and from [16], respectively

3.2.2 The Landau–Lifshitz–Gilbert (LLG) Equation

The typical approach to simulating the magnetization dynamics in an MTJ is the micromagnetic approach. In this approach, the free magnetic layer in the MTJ is discretized into a 3-D grid of ferromagnetic mono-domains. Since micromagnetic simulations solve the Landau–Lifshitz–Gilbert (LLG) equation numerically, they need to be repeated such that the solutions converge when parameters that should not affect them are varied. For example, the magnetization dynamics are independent of the discretization resolution and the discretization resolution is increased until the numerical solutions of micromagnetic simulations converge.

The LLG equation describes magnetization dynamics of each ferromagnetic mono-domain, and is given by [13]

$$\frac{\partial \widehat{m}}{\partial t} = -|\gamma| \widehat{m} \times \overrightarrow{H}_{EFF} + \alpha \widehat{m} \times \frac{\partial \widehat{m}}{\partial t} \tag{3.2}$$

where $\widehat{m}$ is the unit vector describing the magnetization direction of the mono-domain, γ is the electron gyromagnetic ratio (17.5 MHz/Oe or 2.21×10^5 m/A s), and α is the Gilbert damping factor [13]. An effective magnetic field, $\overrightarrow{H}_{EFF}$, models the forces acting on the mono-domain. In an MTJ, $\overrightarrow{H}_{EFF}$ may be written as

$$\overrightarrow{H}_{EFF} = \overrightarrow{H}_{Ani} + \overrightarrow{H}_{Dip} + \overrightarrow{H}_{Demag} + \overrightarrow{H}_{Ex} + \overrightarrow{H}_{Ext} + \overrightarrow{H}_{TH} + \overrightarrow{H}_{STT} \tag{3.3}$$

$\overrightarrow{H}_{Ani}$, $\overrightarrow{H}_{Dip}$, $\overrightarrow{H}_{Demag}$, $\overrightarrow{H}_{Ex}$, $\overrightarrow{H}_{Ext}$, $\overrightarrow{H}_{TH}$, and $\overrightarrow{H}_{STT}$ describe the effective magnetic fields due to magnetic anisotropies (including uniaxial anisotropy), dipolar coupling of the mono-domain to other magnetic dipoles, the demagnetization field due to the arrangement of the magnetic ensemble, the exchange coupling between mono-domains, any externally applied magnetic field, effects due to temperature, and spin-transfer torque, respectively. The first term in the right-hand side of Eq. (3.2) describes the precession of the magnetization around the axis of the effective magnetic field. On the other hand, the remaining term in the right-hand side of Eq. (3.2) describes the dampening of the precession which forces the magnetization to align with the effective magnetic field.

The free layer in the MTJ is stabilized against thermal effects using shape anisotropy, crystalline anisotropy, etc. Uniaxial anisotropy result in the free layer magnetization to preferentially align itself along a single axis, $\hat{u}$, and the effective anisotropy field may be calculated using

$$\overrightarrow{H}_{Ani} = 2K_{u2}\left(\widehat{m}\cdot\widehat{u}\right)\ \widehat{u} \tag{3.4}$$

where K_{u2} is the second order uniaxial anisotropy constant.

When an ensemble of mono-domains is considered, the demagnetization field due to the geometry of the ensemble needs to be considered. Since $\overrightarrow{\nabla}\times\overrightarrow{H}_{Demag} = 0$ and $\overrightarrow{\nabla}\cdot\overrightarrow{B}_{Demag} = 0$ in a uniformly magnetized mono-domain, the demagnetization field can be written as the gradient of a scalar potential

$$\overrightarrow{H}_{Demag} = -\overrightarrow{\nabla}\Phi M \tag{3.5}$$

where

$$\Phi_M\left(\mathbf{r}\right) = \frac{1}{4\pi}\int M\left(\mathbf{r}'\right)\cdot\overrightarrow{\nabla}\left(\frac{1}{\left|\mathbf{r}-\mathbf{r}'\right|}\right)\ d^3\mathbf{r}' \tag{3.6}$$

and $M(\mathbf{r}')$ is the magnetization of the whole ensemble relative to the origin. Details of the calculation of the demagnetization field in numerical solvers are beyond the scope of this chapter and may be found in [17, 18].

Mono-domains that are far apart may appear to be magnetic dipoles to each other. The magnetic field on a mono-domain due to a magnetic dipole is given by

$$\overrightarrow{H}_{DIP} = \frac{3\left(\overrightarrow{M}\cdot\overrightarrow{r}\right)\ \overrightarrow{r} - \left|\overrightarrow{r}\right|^2\overrightarrow{M}}{4\pi\left|\overrightarrow{r}\right|^5} \tag{3.7}$$

where $\overrightarrow{M}$ is the magnetic moment of the dipole (or $\overrightarrow{M} = M_S\widehat{m}$ if a mono-domain with magnetization direction $\widehat{m}$ is approximated as a point dipole) and $\overrightarrow{r}$ is the vector pointing from the magnetic dipole to the mono-domain.

Thermal energy may also perturb the spin interaction between electrons in a mono-domain and needs to be modeled as well. The formulation of the effect thermal energy has on a mono-domain was presented by Brown in [19]. This effect is captured in Eq. (3.3) using the effective thermal field $\overrightarrow{H}_{TH}$. The thermal field is related to the mono-domain properties by

$$\overrightarrow{H}_{TH} = \overrightarrow{\xi}\sqrt{\frac{2k_B T}{|\gamma|\,\mu_0 M_S V_{Domain}\Delta t}} \tag{3.8}$$

where $\overrightarrow{\xi}$ is a vector with components that are independent standard Gaussian random variables, k_B is the Boltzmann constant, T is the temperature of the magnetic ensemble, μ_0 is the permeability of free space, Δt is the constant time step used in the numerical simulation, M_S and V_{Domain} are the saturation magnetization and the volume of the mono-domain, respectively. The statistics of $\overrightarrow{H}_{TH}$ are such that

$$\langle H_{TH,u}\rangle = 0 \text{ where } u = x, y, z \tag{3.9}$$

$$\langle H_{TH,u}(t)H_{TH,v}(t+\tau)\rangle = \frac{2k_B T}{|\gamma|\,\mu_0 M_S V_{Domain}}\delta(\tau)\,\delta_{uv} \tag{3.10}$$

where u and v denote the component of $\overrightarrow{H}_{TH}$.

Slonczewski and Berger independently showed that when a spin-polarized electron current (spins of every electron in the current are aligned in one direction) flows into a ferromagnetic layer, the electrons transfer their spin momentum to the ferromagnetic layer, exerting a torque on the magnetization of the ferromagnetic layer [6, 7]. The spin-transfer torque effect can be written as

$$-|\gamma|\,\widehat{m}\times\overrightarrow{H}_{STT} = \beta\,(\widehat{m}\times(\widehat{m}\times\widehat{m}_P)) + \beta'\widehat{m}\times\widehat{m}_P \tag{3.11}$$

where β and β' depend on the current, and $\widehat{m}_P$ is the unit vector describing the spin direction of the electrons entering the ferromagnetic layer. In the case of spin valves and of MTJs, $\widehat{m}_P$ corresponds to the magnetization direction of the pinned ferromagnetic layer. It may be convenient to write the spin-transfer torque in Eq. (3.11) as an effective field instead, which is given by

$$\overrightarrow{H}_{STT} = \frac{\beta}{|\gamma|}\,(\widehat{m}_P\times\widehat{m}) - \frac{\beta'}{|\gamma|}\widehat{m}_P \tag{3.12}$$

In Eqs. (3.11) and (3.12),

$$\beta = a_J\frac{|\gamma|}{\mu_0 M_S V_{Domain}}\,\frac{\hbar}{2}\,\frac{I_{Curr}}{e} \tag{3.13}$$

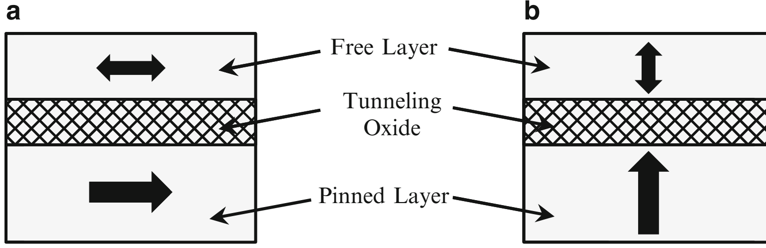

Fig. 3.3 The (**a**) in-plane magnetic anisotropy (IMA) MTJ has magnetizations which are in the plane of the thin film ferromagnetic layers whereas the (**b**) perpendicular-magnetic anisotropy (PMA) MTJ has magnetizations that are perpendicular to the plane of the thin film ferromagnetic layers

where e is the electronic charge, $\hbar$ is the reduced Planck constant, I_{Curr} is the electronic current flowing from the mono-domain into the polarizing ferromagnetic layer, and a_J is dimensionless. β' has the same form as β except a_J is replaced by a_J'. The vector direction of the effective magnetic flux density is the spin direction, $\widehat{m}_P$, of the spin-carrying particles. a_J and a_J' are fitting functions that describe the in-plane and perpendicular-to-plane torques, respectively, relative to the plane containing $\widehat{m}$ and $\widehat{m}_P$. They may be interpreted as the effectiveness of spin-transfer (i.e. the proportion of total available spin-angular momentum that is transferred to the mono-domain).

MTJs with *perpendicular magnetic anisotropy* (PMA) are currently the technology of choice for STT-MRAM application. The magnetic layers in MTJs with PMA have magnetizations that are perpendicular to the plane of the magnetic layers. Previously, MTJs have *in-plane anisotropy* (IMA) in which the magnetic layers have magnetizations that are in-plane to the magnetic layers. The difference between MTJs with IMA and with PMA is illustrated in Fig. 3.3. In IMA, the STT has to overcome both $\overrightarrow{H}_{Ani}$ and $\overrightarrow{H}_{Demag}$. The strength of the effective field that STT needs to overcome is approximately $4\pi M_S$. Furthermore, it is difficult to increase the retention time as the MTJ with IMA is scaled down. These two issues are absent in MTJs with PMA. Since $\overrightarrow{H}_{Ani}$ and $\overrightarrow{H}_{Demag}$ are collinear in MTJs with PMA, the MTJ free layer can be modeled with only uniaxial anisotropy. The relationship between switching the energy barrier, E_A, and the critical switching field is then given by

$$\overrightarrow{H}_C = \frac{2E_A}{\mu_0 M_S V_{FL}} \tag{3.14}$$

where V_{FL} is the volume of the free layer. Also, $E_A = K_{u2} V_{FL}$.

In conventional MRAM, the MTJ free layer magnetization is switched using magnetic fields generated by current carrying wires as shown in Fig. 3.4. The required current for switching the MTJ is

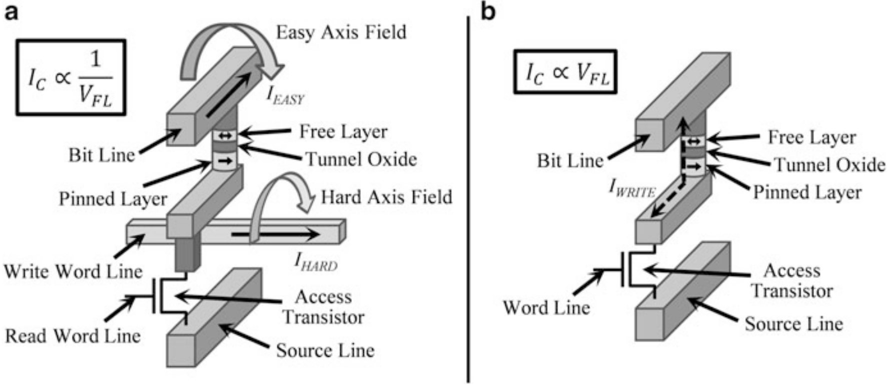

Fig. 3.4 Structures of (**a**) the field-switched MRAM, and (**b**) the spin-transfer torque MRAM

$$I_C = \frac{4\pi r E_A}{\mu_0 M_S V_{FL}} \tag{3.15}$$

where r is the spacing between the wire and the center of the free layer. When the MTJ is scaled down, I_C increases and hence MRAM is not scalable. On the other hand, if the MTJ free layer is approximated as a mono-domain, the effective switching field due to spin-transfer torque, which may be written as

$$\vec{H}_{STT} = \frac{\hbar I_{Curr}}{2e\mu_0 M_S V_{FL}} \left(a_J \left(\widehat{m}_P \times \widehat{m} \right) - a'_J \widehat{m}_P \right) \tag{3.16}$$

scales up at the same rate as $\vec{H}_C$ when the MTJ is scaled down. Hence, spin-transfer torque MRAM overcomes the scalability issue in MRAM.

3.2.3 SPICE Compatible Model of Magnetic Tunnel Junctions

The interaction between device dynamics within the MTJ and the external circuit needs to be considered in the design of STT-MRAM memory cells. Hence, a SPICE compatible model for the MTJ needs to be developed to include MTJ physics during circuit simulations in SPICE. Figure 3.5 shows how an SPICE compatible model for an MTJ with a mono-domain free layer may be implemented. This model captures the magnetization dynamics of the MTJ free layer, and the dependence of the I–V characteristics of the MTJ on the MTJ biasing conditions.

The LLG equation for the free layer may be solved by rewriting Eq. (3.2) in spherical coordinates and noting that the radial component of $\widehat{m}$ is constant. A circuit block consisting of current sources driving a capacitor may then be used to implement a differential equation solver in SPICE by noting that

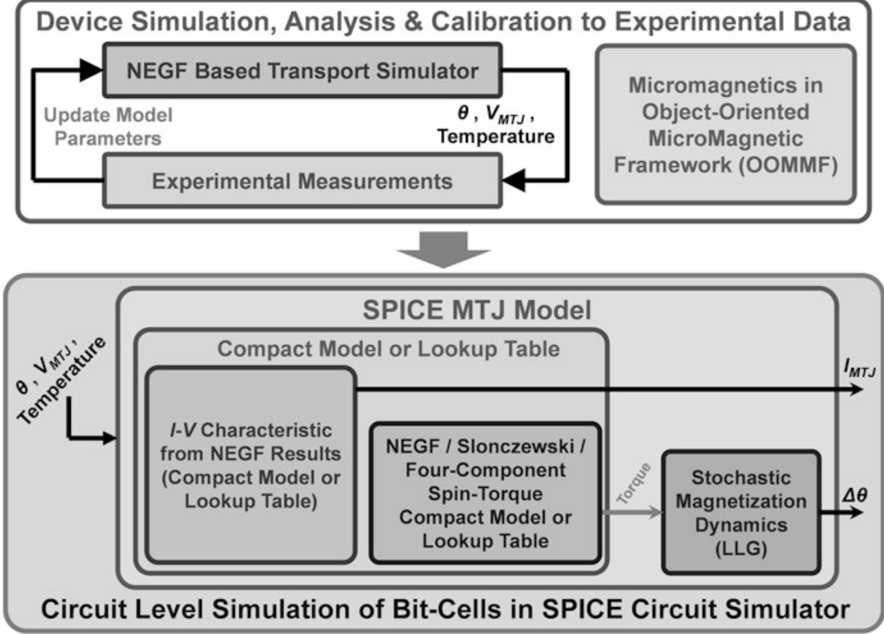

Fig. 3.5 Device/circuit simulation framework used to evaluate STT-MRAM. Device level simulation results are validated using experimental data before parameters are imported into the SPICE model for circuit level simulation of STT-MRAM bit-cells

$$\frac{dv_C}{dt} = \frac{i_C}{C} \tag{3.17}$$

where v_C and i_C are the voltage across and current through the capacitor with capacitance C, respectively. A pair of such circuit blocks can be used to solve the angular components of Eq. (3.2) by representing the right-hand side of Eq. (3.2) as a sum of currents.

The I–V characteristics of the MTJ may be stored in a lookup table by noting that the current flowing through the MTJ depends on both the voltage across the MTJ, as well as the magnetizations $\widehat{m}$ and $\widehat{m}_P$ of the free and pinned layers, respectively. Such a lookup table may consume a lot of memory and is impractical to implement. An alternate method is to note that the dependence of MTJ current on MTJ voltage and on the magnetizations may be decoupled by

$$I_{MTJ}\left(V_{MTJ}\right) = I_{AP}\left(V_{MTJ}\right)\ sin^2\left(\frac{\theta}{2}\right) + I_P\left(V_{MTJ}\right)\ cos^2\left(\frac{\theta}{2}\right) \tag{3.18}$$

where $\widehat{m} \cdot \widehat{m}_P = cos\ \theta$, and $I_{AP}(V_{MTJ})$ and $I_P(V_{MTJ})$ are the MTJ currents in the anti-parallel and parallel configurations, respectively, when the voltage applied across the MTJ is V_{MTJ}. Hence, the I–V characteristics of the MTJ may be implemented using

lookup tables or equations for I_{AP} and for I_P. The lookup tables or equations need to capture the dependence of I_{AP} and I_P on V_{MTJ}, MTJ cross-sectional area, and MTJ tunneling oxide thickness also [20].

3.3 Design of STT-MRAM Memory Cells

The STT-MRAM memory cell may be thought of as a programmable resistor connected with an access transistor as shown in Fig. 3.1. In an on-chip cache array, the gates of the access transistors in each row of memory cells are connected together so that they may be accessed in parallel. The bit and source lines are shared along the column of the array so that individual memory cells along the row being accessed may be written to or read from in parallel. When a memory cell is being accessed, the word line connected to the cell is charged to the supply voltage, V_{DD}, to enable the access transistor. Write operations are performed by charging the bit line and source line to the required voltages so that current will flow through the MTJ to program it. The directionality of the current determines the data being stored in the memory cell. Read operations may be performed either by passing a fixed current through the cell and sensing the voltage developed across the bit and source lines (also called *voltage sensing scheme*), or by clamping the voltages of the bit and source lines and sensing the current flowing through the memory cell (also called *current sensing scheme*). Figure 3.6 shows the biasing conditions of the STT-MRAM memory cell for different operations.

Under process variations, failures may occur during the operation of STT-MRAM memory cells. Variations in MTJ tunnel oxide thickness, t_{MgO}, and MTJ cross-sectional area affect R_{MTJ}, which in turn affect the ability to write into the memory cell, the ability to correctly sense R_{MTJ} of the memory cell, and the ability of the MTJ to retain its configuration when the bit-cell is being read. *Write failures* occur when the MTJ cannot be switched between anti-parallel and parallel

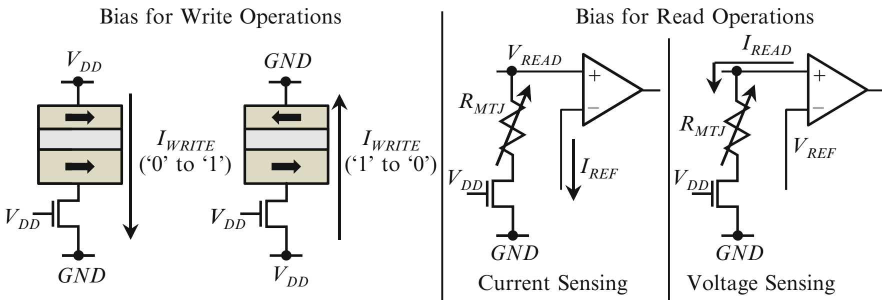

Fig. 3.6 Biasing conditions for read and for write operations of STT-MRAM. In current sensing read operation, the bit-line is clamped at V_{READ}, and the bit-cell current is compared to the reference current, I_{REF}. In voltage sensing, a read current (I_{READ}) is passed through the bit-cell and the voltage on the bit-line is compared to the reference voltage, V_{REF}

configurations. This occurs when the current through the MTJ falls below I_C during write. Read failures occur when R_{MTJ} is incorrectly determined (*decision* failure) or when the MTJ configuration is accidentally switched during read (*disturb* failure). The failure probability of each type of STT-MRAM failure may be calculated using D.C. load line analyses, discussed in the following sections.

3.3.1 Modeling STT-MRAM Failures

The common approach to calculating STT-MRAM failure probabilities assumes distributions for R_{MTJ} and the *TMR* of the MTJ [21], which may be physically incorrect. We now show how STT-MRAM failure probabilities may be calculated without the need to assume distributions for R_{MTJ} and *TMR* of the MTJ.

Write failure occurs when data cannot be written into a STT-MRAM bit-cell within the write cycle. This occurs when R_{MTJ} is too large for the access transistor to provide the required I_C. Write failure may occur when t_{MgO} is too thick, when the access transistor has a threshold voltage (V_T) that is too high, when access transistor width is too small, or when other factors or a combination of factors that results in a write current smaller than I_C flowing through the MTJ occur. The write failure probability ($P_{WR,i}$) for a particular bit-cell may be calculated using D.C. load line analysis as shown in Fig. 3.7a. Consider a bit-cell having an MTJ with cross-sectional area $A_{MTJ,j}$, and I_{MTJ} is exactly I_C for parallel-to-anti-parallel (P-to-AP) switching corresponding to $A_{MTJ,j}$. Further, consider that the MTJ is in parallel (P) configuration with $t_{MgO} = t_{WR,MAX}$ and resistance R_P. I_{MTJ} falls below I_C if $t_{MgO} > t_{WR,MAX}$, and hence data cannot be written into this bit-cell within one write cycle. The same argument holds for an MTJ in AP configuration. Since $t_{WR,MAX}$ depends on $A_{MTJ,j}$, $P_{WR,i}$ for this particular bit-cell can be written as

$$P_{WR,i} = \lim_{\delta \to 0} \sum_{all \quad j} P\left(X - \delta \le X \le X + \delta\right) \cdot P\left(t_{MgO} \ge t_{WR-MAX,j}\right) \quad (3.19)$$

where $X = A_{MTJ,j}$. Since $t_{WR,MAX,j}$ depends on A_{MTJ} and A_{MTJ} is allowed to vary, A_{MTJ} is divided into bins (indexed as j) for numerical calculation of $P_{WR,i}$. The write failure probability of the array (P_{WR}) may be calculated by first using Monte Carlo simulation to generate N access transistor I–V characteristic and calculating $P_{WR,i}$ for each I–V characteristic. P_{WR} may then be calculated as

$$P_{WR} = \sum_{i=1}^{N} P_{WR,i} \quad (3.20)$$

The *disturb failure* probability for a STT-MRAM cell ($P_{RD,i}$) may also be calculated in a similar way by noting that disturb failure occurs when data is accidentally written into the cell during read operations. The D.C. load line used

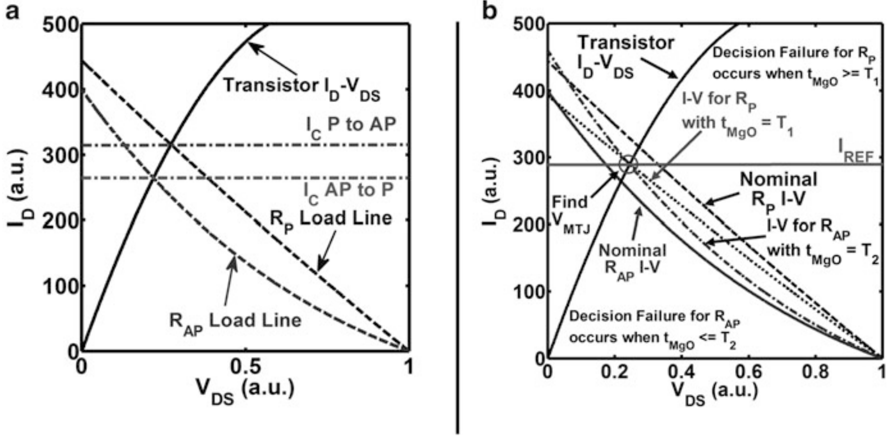

Fig. 3.7 Load lines used for analyzing (**a**) write and disturb failures, and (**b**) decision failures

for calculating $P_{RD,i}$ is the same as that in Fig. 3.7a except that the I–V curve of the MTJ intersects the horizontal axis at $V = V_{READ}$. Consider a memory cell having an MTJ with cross-sectional area $A_{MTJ,j}$, and I_{MTJ} is exactly I_C for P-to-AP switching corresponding to $A_{MTJ,j}$. Further, consider that the MTJ in the bit-cell is in P configuration with $t_{MgO} = t_{RD-MIN}$ and resistance R_P. If $t_{MgO} < t_{RD-MIN}$, I_{MTJ} will rise above I_C, and hence the data gets written into this memory cell within one read cycle, causing a disturb failure. The same argument holds for an MTJ in the AP configuration. However, the read operation involves only one direction of current flow, and for a specific direction of read current flow, either P-to-AP disturbs or AP-to-P disturbs will occur but not both. t_{RD-MIN} depends on $A_{MTJ,j}$, and $P_{RD,i}$ for this particular bit-cell is

$$P_{RD,i} = \lim_{\delta \to 0} \sum_{all\quad j} P\left(X - \delta \le X \le X + \delta\right) \cdot P\left(t_{MgO} \le t_{RD-MIN,j}\right) \qquad (3.21)$$

where $X = A_{MTJ,j}$. Since $t_{RD,MIN,j}$ depends on A_{MTJ} and A_{MTJ} is allowed to vary, A_{MTJ} is divided into bins (indexed as j) for numerical calculation of $P_{RD,i}$. The disturb failure probability of the array (P_{RD}) may be calculated by first using Monte Carlo simulation to generate N access transistor I–V characteristic and calculating $P_{RD,i}$ for each I–V characteristic. P_{RD} may then be calculated as

$$P_{RD} = \sum_{i=1}^{N} P_{RD,i} \qquad (3.22)$$

The calculation for the decision failure probability of a STT-MRAM memory cell (P_{DEC}) depends on the sensing scheme and sense amplifier used. Consider the current sensing scheme where during STT-MRAM read operation, the voltage

across the bit line and the source line is clamped at V_{READ} and a current sense amplifier (SA) compares the current flowing through the memory cell (I_{Cell}) with a reference current, I_{REF}. If $I_{Cell} < I_{REF}$, the MTJ in the memory cell is in the anti-parallel configuration (AP) or $R_{MTJ} = R_{AP}$ and the SA outputs logic '1'. If $I_{Cell} > I_{REF}$, the MTJ in the memory cell is in the parallel configuration (P) or $R_{MTJ} = R_P$ and the amplifier outputs logic '0'. However, due to process variations, I_{Cell} may be higher than I_{REF} when the MTJ is in AP, or lower than I_{REF} when the MTJ is in P. When this occurs, the SA outputs logic '0' when $R_{MTJ} = R_{AP}$ or logic '1' when $R_{MTJ} = R_P$. Such a failure is called a *decision* failure. I_{REF} needs to be carefully chosen to minimize decision failures.

Figure 3.7b illustrates the D.C. load lines used to calculate the decision probability for a particular memory cell ($P_{DEC,i}$) with a particular I_{REF}. For an MTJ in AP at the nominal t_{MgO} and cross-sectional area $A_{MTJ,j}$, its resistance is R_{AP} and the load line is the solid red line. $I_{MTJ} = I_{REF}$ when $t_{MgO} = T_2$. If $t_{MgO} < T_2$, I_{MTJ} will be more than I_{REF} and the SA incorrectly outputs logic '0'. Similarly, $I_{MTJ} = I_{REF}$ if the MTJ is in P and has cross-sectional area $A_{MTJ,j}$, and $t_{MgO} = T_1$. If $t_{MgO} > T_1$, I_{MTJ} will be less than I_{REF} and the SA incorrectly outputs logic '1'. Thus, for this particular STT-MRAM memory cell

$$P_{DEC,i} = \lim_{\delta \to 0} \sum_{all \quad j} P\left(X - \delta \le X \le X + \delta\right) \cdot P\left(T_1 \le t_{MgO} \le T_2\right) \tag{3.23}$$

where $X = A_{MTJ,j}$. Since T_1 and T_2 depend on A_{MTJ} and A_{MTJ} is allowed to vary, A_{MTJ} is divided into bins (indexed as *j*) for numerical calculation of $P_{DEC,i}$. The decision failure probability of the array (P_{DEC}) may be calculated by first using Monte Carlo simulation to generate *N* access transistor *I–V* characteristic and calculating $P_{DEC,i}$ for each *I–V* characteristic. P_{DEC} may then be calculated as

$$P_{DEC} = \sum_{i=1}^{N} P_{DEC,i} \tag{3.24}$$

Because P_{DEC} depends on I_{REF}, I_{REF} may be used as a design parameter to minimize P_{DEC}. To determine the optimum I_{REF} ($I_{REF-OPT}$) that minimizes P_{DEC}, the nominal read currents through the bit-cell when the MTJ is in AP (I_{R-AP}) and when the MTJ is in P (I_{R-P}) are determined first. $I_{REF-OPT}$ is determined by minimizing P_{DEC} in the interval $[I_{R-AP}, I_{R-P}]$. A similar approach may be used to determine the decision failure probability with a voltage sensing scheme.

Finally, the total failure probability of the each memory cell ($P_{FAIL,i}$), may be calculated using

$$P_{FAIL,i} = \lim_{\delta \to 0} \sum_{all \quad j} P\left(X - \delta \le X \le X + \delta\right) \cdot \left[1 - P\left(T_3 \le t_{MgO} \le T_4\right)\right] \tag{3.25}$$

$$T_3 = max \left(T_1, t_{RD-MIN,j}\right) \tag{3.26}$$

$$T_4 = min \left(T_2, t_{WR-MAX,j}\right) \tag{3.27}$$

where $X = A_{MTJ,j}$. Since T_3 and T_4 depend on A_{MTJ} and A_{MTJ} is allowed to vary, A_{MTJ} is divided into bins (indexed as j) for numerical calculation of $P_{FAIL,i}$. The total failure probability of the array (P_{FAIL}) may be calculated by first using Monte Carlo simulation to generate N access transistor I–V characteristic and calculating $P_{FAIL,i}$ for each I–V characteristic. P_{FAIL} may then be calculated as

$$P_{FAIL} = \sum_{i=1}^{N} P_{FAIL,i} \tag{3.28}$$

3.3.2 Optimization of STT-MRAM Memory Cells

Several STT-MRAM bit-cell designs have been published in the literature [16, 22]. STT-MRAM bit-cells can have two configurations as shown in Fig. 3.8: the "standard" connection (SC, Fig. 3.8a) and the "reversed" connection (RC, Fig. 3.8b). Furthermore, there are two possible configurations for sensing the data stored in the cell. Figure 3.6 shows one configuration where sensing is done by connecting the bit-line to the input of the sense amplifier. Note that sensing may also be done by connecting the source-line to the input of the sense amplifier instead.

Figures 3.9 and 3.10 shows the results of the failure analysis (using the methodology presented in the earlier sections) performed on SC and RC STT-MRAM bit-cells. It is clearly shown that the configurations for read and for write operations need to be carefully chosen to optimize the failure probabilities of the cell. Read failures for sensing through bit-line or through source-line may be significantly different, as Fig. 3.9a shows. For the SC bit-cell, sensing from the bit-line only has disturb failures that flip '1' to '0' (SC, P), whereas sensing from the source-line only has disturb failures that flip '0' to '1' (SC, AP). For the RC bit-

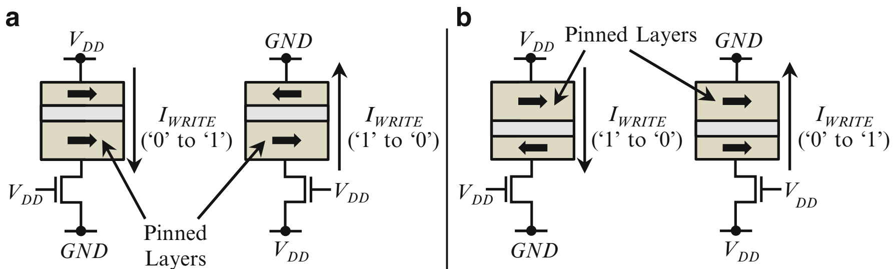

Fig. 3.8 The (**a**) standard, and (**b**) reversed connection 1T-1MTJ STT-MRAM bit-cell structures

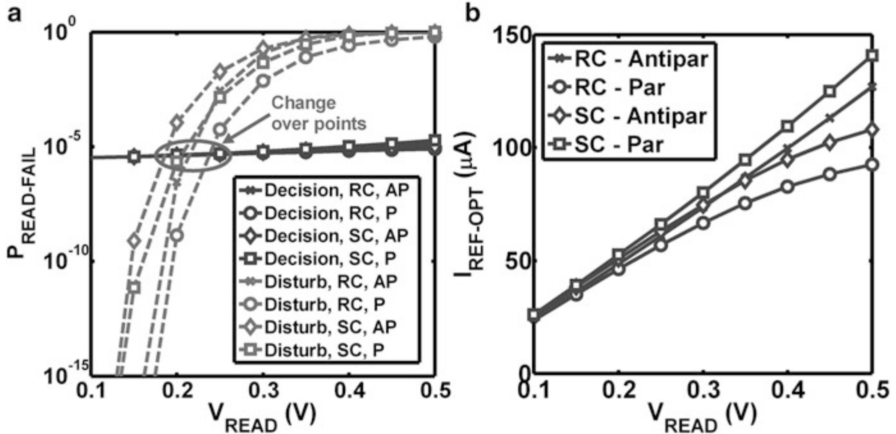

Fig. 3.9 (**a**) Decision and disturb failure probabilities were plotted with varying V_{READ} at constant ATx width. (**b**) The $I_{REF-OPT}$ corresponding to the decision failure in (**a**). V_{READ} was fixed at 0.1 V so that disturb failures are negligible

cell, sensing from the bit-line only has disturb failures that flip ‘0’ to ‘1’ (RC, AP), whereas sensing from the source-line only has disturb failures that flip ‘1’ to ‘0’ (RC, P). Interestingly, decision failures do not change significantly when V_{READ} is sufficiently small. However, the decision failure probability becomes increasingly sensitive to $I_{REF-OPT}$ (shown in Fig. 3.9b) as V_{READ} is reduced. The three failure probabilities are then plotted in the same graph, as shown in Fig. 3.10, to determine the optimum ATx width of the bit-cell. The optimum ATx width depends on whether read failures are decision dominated or disturb dominated.

3.3.3 The 2T-1MTJ STT-MRAM Bit-cell

Note that when read failures are decision dominated, the decision failure probability is minimized when ATx width is 908 nm (Fig. 3.10). However, the ATx width needs to be increased to reduce write failures. 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 just multiple transistors connected in such a way that their gate, source, and drain terminals are shared. When multi-finger transistors are used in the bit-cell design, the effective access transistor width may be varied using two word-lines instead of one (Fig. 3.11), and is called the *2T-1MTJ design* [21]. Word line 1 is used during read operations to switch M1 ON and OFF, while word line 2 keeps M2 OFF. During write operations, both word lines are turned ON and OFF simultaneously.

Let us analyze the 2T-1MTJ design using the failure characteristics in Fig. 3.10 as an example. The write operation of the 2T-1MTJ bit-cell requires both M1 and

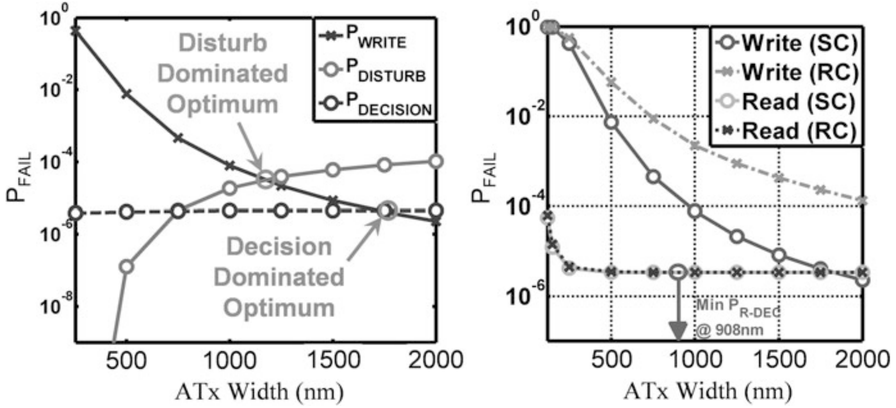

Fig. 3.10 Generally, all three failure graphs are plotted together to determine the optimum ATx width to use as shown on the *left*. The optimum point depends on whether the bit-cell failures are disturb dominated or decision dominated. However, if decision failures are the dominant read failure, then we only have to look at decision failures and write failures to determine the optimum ATx width. As shown on the *right*, decision failures are minimized at a particular ATx width while write failures keep decreasing with increasing ATx width

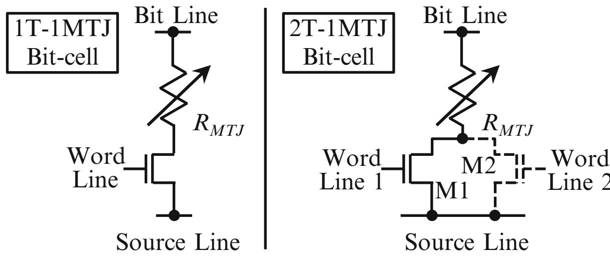

Fig. 3.11 The 2T-1MTJ bit-cell uses two access transistors with separate word lines to optimize for read failures and write failures without the need to tradeoff one for the other

M2 to be turned on. On the other hand, the read operation requires only M1 to be turned on. The size of M1 is optimized for decision failures (908 nm), while the size of M2 is as large as required to meet the write failure, array area, and array capacity requirements. Hence, the decision and the write failure probabilities of the 2T-1MTJ bit-cell may be optimized simultaneously without the need to tradeoff one for the other.

3.3.4 Stretched Write Cycle

The *stretched write cycle* (*SWC*) [23] is another optimization strategy that may be used in STT-MRAM design. SWC takes advantage of the fact that write operations

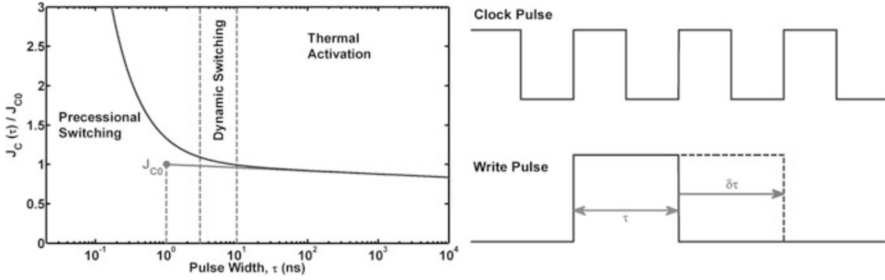

Fig. 3.12 The typical dependence of J_C on the switching time, τ, is shown on the *left*. The write pulse may be stretched by $\delta\tau$ as shown on the *right* (relative to the clock pulse) to reduce J_C

do not occur as frequently as read operations in last level caches. The critical current required for writing into the STT-MRAM bit-cell may then be lowered by allowing a longer time for write operations to complete, as shown in Fig. 3.12.

The write energy comparisons of the optimization techniques presented are shown in Fig. 3.13. The worst case design that mitigates write failures by write-voltage boosting has 18 % higher power dissipation as compared to the nominal design without process variations. The 1T-1MTJ bit-cell energy overhead is reduced to 11 % after optimization, resulting in an area overhead of 5.4 %. However, if an optimized 2T-1MTJ design is used, the energy overhead is reduced to 9 % while the area overhead is increased to 9 %. Finally, the energy dissipation becomes 3 % lower than the nominal case when SWC is used with an optimized 1T-1MTJ bit-cell. This is because the critical write current needed is significantly lower in SWC. Although the write frequency is reduced by 50 % in SWC, the throughput penalty is only 3 %. Hence, we conclude that circuit/architecture co-design can lead to ultralow power last level caches based on STT-MRAMs.

3.4 Comparisons of Cache Arrays Based on SRAM and STT-MRAM

A cache comprises of multiple arrays for storing tags and data bits. In conventional on-chip caches, both the tag and data arrays are implemented using SRAM. Since the tag array requires frequent and fast updates of status bits and history bits, the write latency of STT-MRAM may significantly impact the performance STT-MRAM based tag arrays [24]. Hence, the STT-MRAM cache we will be discussing is a hybrid cache where the tag arrays are implemented using SRAM and the data arrays are implemented using STT-MRAM. In order to estimate the overall cache latency, area and energy consumption of the STT-MRAM cache, the CACTI 6.5 simulator [25] needs to be modified to consider (a) analog read circuits in STT-

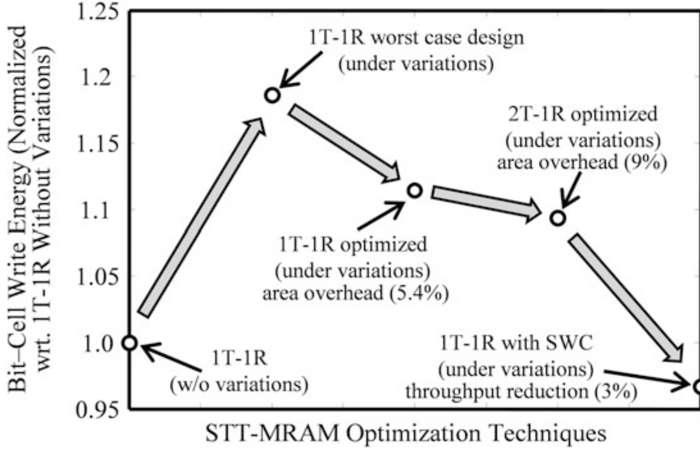

Fig. 3.13 Write energy comparison of the bit-cell optimization techniques and the overhead associated with each optimization technique

MRAM data arrays, (b) SRAM-based tag arrays along with STT-MRAM data arrays, and (c) the bit-cell layout geometries to optimize the array aspect ratio.

A comparison of caches designed with SRAM and STT-MRAM is shown in Fig. 3.14. Note that the capacity of the cache array, and not the cache area, has a more significant impact on whether caches designed with STT-MRAM outperform caches designed with SRAM. As the cache capacity increases, the wire delays in SRAM based caches increases much faster than that in STT-MRAM based caches due to the larger bit-cell footprint. Hence, high capacity caches designed with STT-MRAM have faster access time and are smaller than SRAM based caches. As Fig. 3.14 shows, an 8 MB cache designed with STT-MRAM has lower read latency than an iso-capacity cache designed with SRAM. Similarly, the write latency gap between STT-MRAM based and SRAM based caches reduces with increasing cache capacity.

A similar trend is observed in the dynamic energy consumption of the caches (Fig. 3.14). The energy dissipated in read operations in STT-MRAM based caches is higher than that of SRAM based caches due to power dissipation in the analog read circuits, despite 75 % smaller total cache area. However, the energy dissipation due to interconnects becomes dominant when cache capacity is 1 MB and higher. Therefore, read operation dynamic energy is significantly lower in STT-MRAM based caches. During write operations, STT-MRAM caches dissipate significantly larger energy than SRAM based caches. Finally, the leakage in STT-MRAM based caches is significantly lower than that in SRAM based cache because STT-MRAM bit-cells are non-volatile and have zero standby power. Only the SRAM based tag arrays and periphery dissipate leakage power in caches designed with STT-MRAM.

The total energy dissipation in a cache also depends on factors such as cache access patterns (number of read and write operations) and cache utilization (number

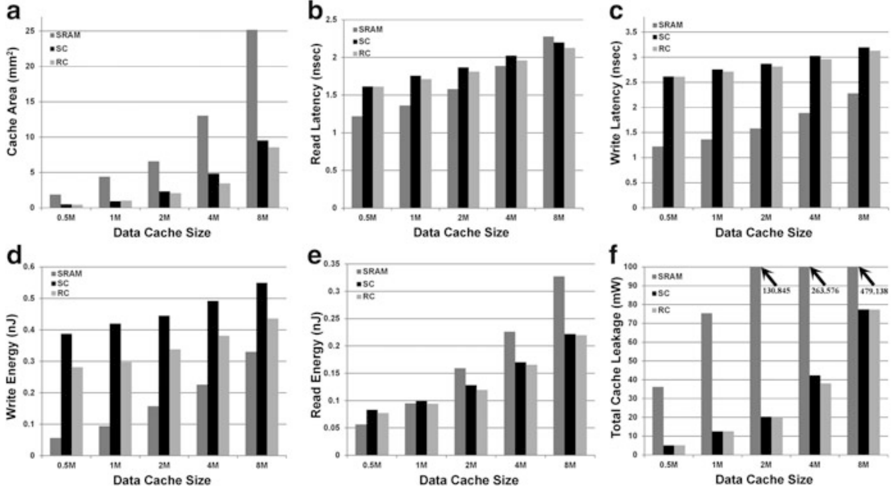

Fig. 3.14 (**a**) Array area of SRAM and STT-MRAM based caches (4-way, 64 B cache line, *B* byte, *M* mega byte), (**b**) read latency and (**c**) write latency, (**d**) read energy per operation and (**e**) write energy per operation, and (**f**) total leakage power

of times a processor accesses the cache per unit cycle). The cache utilization is lower than 30 % in today's processors [26]. Moreover, for lower levels of the cache hierarchy, the cache utilization is significantly lower than 30 %. We have measured L2 cache utilizations for various SPEC2000 benchmarks based on the Simplescalar framework [27] with a 32 KB L1 cache configuration. For a majority of the benchmarks, L2 cache utilization is lower than 3 %. The highest utilization, observed for the AMMP benchmark, is about 13 %, and the average utilization across 16 benchmarks is only 2.2 %.

As shown in Fig. 3.15, a 2 MB STT-MRAM cache shows similar or lower energy consumption than a 0.5 MB SRAM cache when the utilization is lower than 10 %. Although the STT-MRAM cache has significantly lower energy consumption at 0 % utilization (leakage only), the energy dissipation increases drastically due to excessive write energy as the utilization increases. The results are obtained using the following conditions: read and write operation ratio of 2:1, 2 GHz processor speed, and total simulation time of 1 billion processor cycles. Therefore, an STT-MRAM cache can achieve high energy-efficiency along with high capacity in comparison to an SRAM cache, especially in lower levels of the cache hierarchy due to the low cache utilization.

In a conventional SRAM array, column selection is required for storing multiple words in a single row [28]. Since set associativity is common in modern caches, column selection in SRAM arrays is imperative. Furthermore, bit-interleaving can only be achieved by employing column selection. Bit-interleaving is a commonly adopted technique in SRAM arrays (1) to mitigate soft errors [28], and (2) to increase array density by bit-line multiplexing [25]. In the column selection

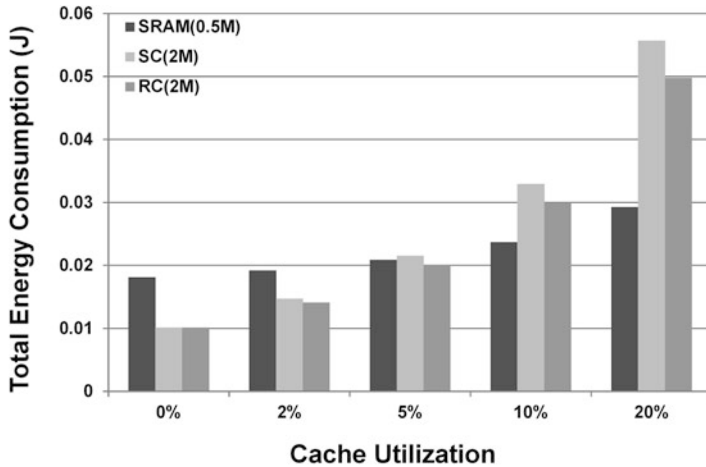

Fig. 3.15 Total energy consumption versus cache utilization for SRAM and for STT-MRAM based caches shows that when 0.1 M data is stored in cache, STT-MRAM dissipation is much lower

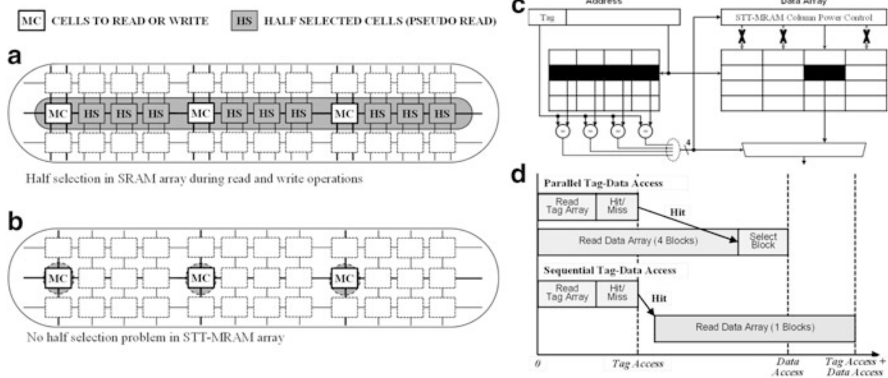

Fig. 3.16 (**a**) SRAM cache access dissipates additional power in the bit-lines of unselected cells, whereas (**b**) STT-MRAM based cache do not have the half-select problem. (**c**) Tag-data access needs to be sequential to take advantage of the lack of half-select problem in STT-MRAM. (**d**) Sequential tag-data access incurs additional read latency since the cache hit needs to occur before reading data

operation of an SRAM array, all unselected bit-cells in the accessed row have to be under read mode to prevent unexpected bit flips, when a word-line is asserted. This phenomenon is commonly known as pseudo-read or half-selection [28]. Note that, in an STT-MRAM array, the non-volatility of bit-cells can eliminate the half selection problem. As presented in Fig. 3.16, the unselected bit-cells can remain in standby mode, and hence, consume no energy during both read and write column selection operations. However, a sequential tag-data access is needed in order to determine which of the columns need to be the selected prior to actual access, which increases

the read latency since a cache hit must occur prior to reading the data array. Based on our simulation parameters, the average read latency penalty is about 500 ps for the 2 MB STT-MRAM based caches. However, the read energy savings is about 40–50 %.

3.5 Conclusion

Based on the simulation results presented in this chapter, we may conclude that spin-transfer torque MRAM is becoming more viable as a technology for on-chip last-level caches. Significant energy savings are achieved due to the large cache capacities enabled by the small footprint of STT-MRAM memory cells. Further reduction in the critical switching current of STT-MRAM will increase the achievable energy savings [29]. The non-volatility of STT-MRAM may also be exploited to enable a new "normally-off" computing paradigm [30]. However, crucial design issues need to be overcome for STT-MRAM to be viable for caches next to the processor and become a truly universal memory technology. For example, the lack of a self-referenced differential sensing scheme in STT-MRAMs limits the performance of its read operations and also its robustness against process variations. Hence, there is a need to explore alternative MTJ structures to improve STT-MRAM performance, and it may take some time before suitable structures become a reality. Even so, STT-MRAM offers exciting possibilities in integrating new functionality into on-chip caches in its current form [31]. This ability to integrate new functionality on-chip to complement the CMOS circuitry may be key in driving the future adoption of on-chip STT-MRAM technology.

References

1. ITRS Roadmap 2014. http://www.itrs.net.
2. Huai Y. Spin-transfer torque MRAM (STT-MRAM): challenges and prospects. AAPPS Bull. 2008;18(6):33–40.
3. Datta D, Behin-Aein B, Datta S, Salahuddin S. Voltage asymmetry of spin-transfer torques. IEEE Trans Nanotechnol. 2012;11(2):261–72.
4. Dorrance R, Ren F, Toriyama Y, Hafez AA, Yang CK, Markovic D. Scalability and design-space analysis of a 1T-1MTJ memory cell for STT-RAMs. IEEE Trans Electron Devices. 2012;59(4):878–87.
5. Fong X, Choday SH, Roy K. Bit-cell level optimization for non-volatile memories using magnetic tunnel junctions and spin-transfer torque switching. IEEE Trans Nanotechnol. 2012;11(1):172–81.
6. Slonczewski JC. Current-driven excitation of magnetic multilayers. J Magn Magn Mater. 1996;159(1–2):L1–7.
7. Berger L. Emission of spin waves by a magnetic multilayer traversed by a current. Phys Rev B. 1996;54(13):9353–8.
8. Myers EB. Current-induced switching of domains in magnetic multilayer devices. Science. 1999;285(5429):867–70.

9. Katine J, Albert F, Buhrman R, Myers E, Ralph D. Current-driven magnetization reversal and spin-wave excitations in Co/Cu/Co pillars. Phys Rev Lett. 2000;84(14):3149–52.
10. Huai Y, Albert F, Nguyen P, Pakala M, Valet T. Observation of spin-transfer switching in deep submicron-sized and low-resistance magnetic tunnel junctions. Appl Phys Lett. 2004;84(16):3118.
11. Ikeda S, Miura K, Yamamoto H, Mizunuma K, Gan HD, Endo M, Kanai S, Hayakawa J, Matsukura F, Ohno H. A perpendicular-anisotropy CoFeB-MgO magnetic tunnel junction. Nat Mater. 2010;9(9):721–4.
12. Kishi T, Yoda H, Kai T, Nagase T, Kitagawa E, Yoshikawa M, Nishiyama K, Daibou T, Nagamine M, Amano M, Takahashi S, Nakayama M, Shimomura N, Aikawa H, Ikegawa S, Yuasa S, Yakushiji K, Kubota H, Fukushima A, Oogane M, Miyazaki T, Ando K. Lower-current and fast switching of a perpendicular TMR for high speed and high density spin-transfer-torque MRAM. In: 2008 IEEE International Electron Devices Meeting, 2008, p. 1–4.
13. Sun J. Spin-current interaction with a monodomain magnetic body: a model study. Phys Rev B. 2000;62(1):570–8.
14. Salahuddin S, Datta S. Self-consistent simulation of quantum transport and magnetization dynamics in spin-torque based devices. Appl Phys Lett. 2006;89(15):153504.
15. Yuasa S, Nagahama T, Fukushima A, Suzuki Y, Ando K. Giant room-temperature magnetoresistance in single-crystal Fe/MgO/Fe magnetic tunnel junctions. Nat Mater. 2004;3(12):868–71.
16. Lin CJ, Kang SH, Wang YJ, Lee K, Zhu X, Chen WC, Li X, Hsu WN, Kao YC, Liu MT, Nowak M, Yu N. 45nm low power CMOS logic compatible embedded STT MRAM utilizing a reverse-connection 1T/1MTJ cell. In: 2009 IEEE International Electron Devices Meeting (IEDM), 2009, p. 1–4.
17. Mansuripur M, Giles R. Demagnetizing field computation for dynamic simulation of the magnetization reversal process. IEEE Trans Magn. 1988;24(6):2326–8.
18. Newell AJ, Williams W, Dunlop DJ. A generalization of the demagnetizing tensor for nonuniform magnetization. J Geophys Res. 1993;98(B6):9551–5.
19. Brown W. Thermal fluctuations of a single-domain particle. Phys Rev. 1963;130(5):1677–86.
20. Fong X, Gupta SK, Mojumder NN, Choday SH, Augustine C, Roy K. KNACK: a hybrid spin-charge mixed-mode simulator for evaluating different genres of spin-transfer torque MRAM bit-cells. In: 2011 International Conference on Simulation of Semiconductor Processes and Devices, 2011, p. 51–4.
21. Li J, Ndai P, Goel A, Salahuddin S, Roy K. Design paradigm for robust spin-torque transfer magnetic RAM (STT MRAM) from circuit/architecture perspective. IEEE Trans Very Large Scale Integr Syst. 2010;18(12):1710–23.
22. Jeong G, Cho W, Ahn S, Jeong H, Koh G, Hwang Y. A 0.24-μm 2.0-V 1T1MTTJ 16-kb nonvolatile magnetoresistance RAM with self-reference sensing scheme. IEEE J Solid-State Circuits. 2003;38(11):1906–10.
23. Augustine C, Mojumder NN, Fong X, Choday SH, Park SP, Roy K. Spin-transfer torque MRAMs for low power memories: perspective and prospective. IEEE Sens J. 2012;12(4):756–66.
24. Rasquinha M, Choudhary D, Chatterjee S, Mukhopadhyay S, Ram TSTT, Yalamanchili S. An energy efficient cache design using Spin Torque Transfer (STT) RAM. In: 2010 IEEE/ACM International Symposium on Low-Power Electronics and Design (ISLPED), 2010, p. 389–94.
25. Muralimanohar N, Balasubramonian R, Jouppi N. Optimizing NUCA organizations and wiring alternatives for large caches with CACTI 6.0. In: 40th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO 2007), 2007, p. 3–14.
26. Ramaswamy S, Yalamanchili S. An utilization driven framework for energy efficient caches. In: Proceedings of the 15th International Conference on High Performance Computing (HiPC'08), 2008, p. 583–94.
27. Simplescalar. 2011 LLC. http://www.simplescalar.com.

28. Park SP, Kim SY, Lee D, Kim J-J, Griffin WP, Roy K. Column-selection-enabled 8T SRAM array with ∼1R/1W multi-port operation for DVFS-enabled processors. In: IEEE/ACM International Symposium on Low Power Electronics and Design, 2011, p. 303–8.
29. Yoda H, Fujita S, Shimomura N, Kitagawa E, Abe K, Nomura K, Noguchi H, Ito J. Progress of STT-MRAM technology and the effect on normally-off computing systems. In: 2012 International Electron Devices Meeting, 2012, p. 11.3.1–4.
30. Kawahara T. Challenges toward gigabit-scale spin-transfer torque random access memory and beyond for normally off, green information technology infrastructure (Invited). J Appl Phys. 2011;109(7):07D325.
31. Lee D, Fong X, Roy K. R-MRAM: a ROM-embedded STT MRAM cache. IEEE Electron Device Lett. 2013;34(10):1256–8.