

Current-Induced Magnetic Switching for High-Performance Computing

Yue Zhang, Weisheng Zhao, Wang Kang, Eyra Deng, Jacques-Olivier Klein, and Dafiné Revelosona

1 Introduction

The shrinking of complementary metal oxide semiconductor (CMOS) fabrication node below 90 nm leads to high static power in memories and logic circuits due to the increasing leakage currents [1]. This power issue limits greatly the miniaturization and improvement of electronic devices. For example, the design of multicore microprocessors for CPU in computer is a proof for this point. In this background, novel technologies to replace the mainstream charge-based electronics are hot topics for both academics and industries. Beyond the electrical charge, the devices based on the spintronics attract a broad attention and show the performance advantages in many aspects [2].

Magnetic tunnel junction (MTJ), one of the most important spintronic devices, is the basic element of magnetoresistance random access memory (MRAM) which becomes a most promising candidate for the next generation of universal non-volatile memory. Among its various features, the magnetization switching in MTJ is a crucial point. Much of the academic and industrial research efforts are presently focused on developing efficient switching strategies. One promising method relies on using spin transfer torque (STT) [3, 4], which involves low threshold currents and well-understood mechanisms. Furthermore, only a bi-directional current is

Y. Zhang • W. Kang

Spintronics Interdisciplinary Center, Beihang University, Beijing 100191, China

IEF, Univ. Paris-Sud 11, UMR8622, CNRS, Orsay 91405, France

W. Zhao (✉)

Spintronics Interdisciplinary Center, Beihang University, Beijing 100191, China

e-mail: weisheng.zhao@buaa.edu.cn

E. Deng • J.-O. Klein • D. Revelosona

IEF, Univ. Paris-Sud 11, UMR8622, CNRS, Orsay 91405, France

needed in this approach, which simplifies greatly the CMOS switching circuits and thereby allows for higher density than the other approaches.

On the other hand, racetrack memory is an emerging spintronic concept based on current-induced domain wall (CIDW) motion in magnetic nanowires [5, 6]. Combining with MTJs as write and read heads, CMOS integrability and fast data access speed can be achieved. In this concept, the data are stored via the magnetizations of magnetic domains separated by domain walls (DWs). Due to STT mechanism, the DWs can be propagated consecutively in a direction by a spin-polarized current, which makes the racetrack memory possible to be widely applied for logic and memory designs.

However, some unexpected effects have been discovered using current-induced approach in the devices with small size (e.g., lateral size of 40 nm), such as erroneous state switching with reading currents and short retention times. These problems are mainly related to the in-plane magnetic anisotropy, which cannot provide a sufficiently high energy barrier to ensure thermal stability [7]. This issue limits greatly the potential for future miniaturization of spintronic devices. One compelling solution addressing this issue involves the perpendicular magnetic anisotropy (PMA) in certain materials (e.g., CoFeB/MgO), because it allows high energy barrier to be attained for small-size structures (<40 nm) while maintaining the possibility of fast-speed operations, high TMR ratios and low threshold currents [8, 9].

Thanks to the diverse advantages demonstrated by spintronics as well as various milestone breakthroughs of its related materials and techniques, hybrid spintronics/CMOS logic and memory circuits open a novel route to manipulate information more efficiently. Taking advantages of spintronic devices, the emerging circuits or systems can also realize low power, high density and high speed. For the past decade, many spintronics based logic and memory circuits and their prototypes have been designed and presented. From the relatively mature spin valve for HDDs [10–12] to recently commercialized STT-MRAM [13–17], from magnetic full adder (MFA) and Magnetic Flip-Flop (MFF) for magnetic processors [18–20] to magnetic content addressable memory (CAM) for internet router and search engines [21, 22], spintronics or concepts based on it has seeped into a majority of the advanced logic and memory systems.

One of the beneficial applications is the computing: the structural and technological limitations of conventional computing systems prevent them from reaching high frequency (~4 GHz) and limit power efficiency [23]. In this chapter, we describe an overview of the devices and circuits for high-performance computing, which are particularly based on current-induced magnetic switching. From the compact modeling to the circuit design and the optimization, the contributions of this chapter have been made at a series of levels.

Firstly we introduce the concepts of PMA STT MTJ and racetrack memory. Their fundamental physics, structures and performances, promising to achieve high performance computing, will be involved. In the following, we present spintronics based computing designs under intense R&D. MFA (1-bit one based on PMA STT MTJ and multi-bit one based on racetrack memory) and CAM are particularly investigated. They demonstrate the performance advantages in terms of area and/or speed and/or energy, compared with the CMOS based conventional ones.

2 Current-Induced Magnetic Switching Spintronic Devices

2.1 Perpendicular Magnetic Anisotropy Spin Transfer Torque Magnetic Tunnel Junction (PMA STT MTJ)

The MTJ nanopillar, as shown in Fig. 1a, is one of the important devices for current spintronics based integrated circuits. Particularly, it is the basic element of MRAM. According to the different switching mechanisms, MTJs can be categorized into certain generations. Field-induced magnetic switching (FIMS) [24, 25] and thermally-assisted switching (TAS) [26, 27] are two achievable and mainstream approaches. Some prototypes or even commercialized products are based on these mechanisms. However, the mandatory utilization of magnetic field in these approaches leads to drawbacks on speed, density and power consumption, which hinder the integration of MRAM for advanced computing or memory applications.

In this context, the current-induced magnetic switching comes into view. STT was proposed independently by Berger and Slonczewski in 1996 [3, 4]. They found that a spin-polarized current injected perpendicularly to the plane could equally influence the magnetizations. This interaction is attributed to angular momentum transferred from the polarized electrons to the local magnetization of the ferromagnetic (FM) layer. Once the amount of electrons exceeding the threshold value (often represented by critical current or critical current density), the STT exerted by the current will switch the magnetization of the free layer of MTJ [13, 28]. The STT switching approach was initially researched on giant magnetoresistance (GMR) effect based spin valve [29, 30], and then focused on the MTJ providing a significantly higher magnetoresistance [31, 32]. In MTJ, one FM layer acts as a polarizer for an electric current, which then transfers angular momentum by exerting a torque on the magnetization of the other FM layer. This current-only approach simplifies

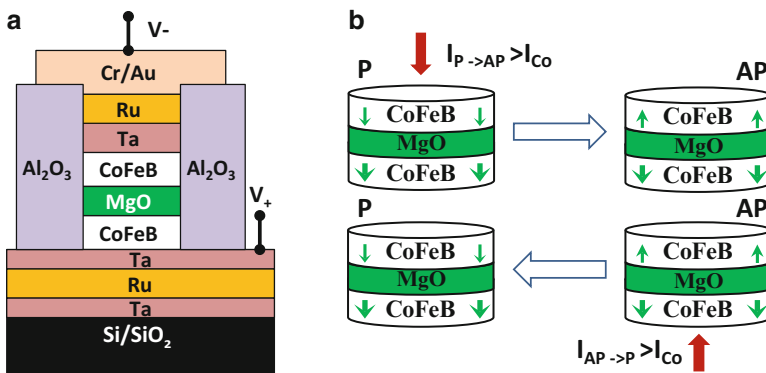


Fig. 1 (a) Vertical structure of an MTJ nanopillar composed of CoFeB/MgO/CoFeB thin films. (b) STT switching mechanism: the MTJ state changes from parallel (P) to anti-parallel (AP) as the positive direction current $I_{P \rightarrow AP} > I_{Co}$, on the contrast, its state will return to P state with the negative direction current $I_{AP \rightarrow P} > I_{Co}$

greatly the switching process as it only requires a bi-directional current (see Fig. 1b). Moreover, the magnitude of current for STT is normally less by an order than that for generating a large magnetic field. As a consequence, STT switching approach is widely considered the most promising one to be applied in the future MRAM applications [33].

The effect of STT on the free layer of MTJ can be described by the extra STT term in the Landau-Lifshitz-Gilbert (LLG) equation [34, 35] given by

$$\frac{d\vec{m}}{dt} = -\gamma\vec{m} \times \vec{H}_{eff} + \alpha\vec{m} \times \frac{d\vec{m}}{dt} - \beta J (\vec{m} \times \vec{m} \times \vec{M}) \quad (1)$$

where \vec{m} and \vec{M} are the unit vectors of the free and pinned layers' magnetizations, α is the damping constant, γ is the gyromagnetic ratio, β is the STT coefficient depending on both the spin polarization and the geometric configuration of the spin torque efficiency. \vec{H}_{eff} is the effective field that includes the external field, the anisotropy field, the magnetostatic field, the Oersted field and the exchange coupling field.

In this equation, the first term on the right is to describe the precession of the field-induced magnetization. The second term describes the intrinsic damping process that results in a decrease of the precessional angle as a function of time. The last term on the right is the STT term whose vector direction is opposite to the damping direction. In the current-induced system, the magnetization switching on the free layer can be considered the competition between the damping term and the STT term (see Fig. 2). When the current density is small, the STT term is weaker than the damping term, then the magnetization dynamics maintain in an equilibrium state. In contrast, if the current density is high enough to make STT term stronger than damping term, the magnetization can be excited to larger precessional angles and further be switched. The critical current is defined as the threshold current to distinguish these two regimes, which is described by STT switching static model presented as follows.

The threshold for excitations driven by STT is given by the critical current. The static behavior to describe STT switching in PMA MTJ is mainly based on the calculation of threshold or critical current I_{C0} , which can be expressed by the Eqs. (2) and (3) [8].

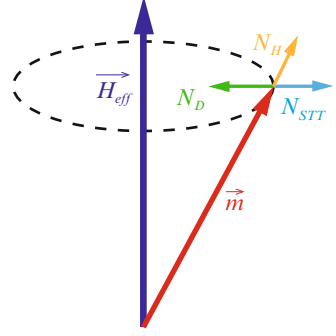
$$E_p = \frac{\mu_0 M_S \times Vol \times H_K}{2} \quad (2)$$

$$I_{C0p} = \alpha \frac{\gamma e}{\mu_B g} (\mu_0 M_S) H_K Vol \quad (3)$$

where H_K is the perpendicular magnetic anisotropy field.

Note that the spin accumulation effects are neglected and the spin polarization efficiency factor g is firstly obtained with the following equation to describe the

Fig. 2 Diagram of the LLG equation. N_D is the damping term, N_{STT} is the STT term and N_H is the field precession term



asymmetric current case [36]. It provides the best agreement with the experimental results illustrated in [8],

$$g = g_{SV} \pm g_{Tunnel} \quad (4)$$

where the sign depends on the free-layer alignment. g_{SV} and g_{Tunnel} are respectively the spin polarization efficiency in a spin valve and tunnel junction nanopillars. They are both predicted by Slonczewski,

$$g_{SV} = \left[-4 + \left(P^{-1/2} + P^{1/2} \right)^3 (3 + \cos \theta) / 4 \right]^{-1} \quad (5)$$

$$g_{Tunnel} = (P/2) / (1 + P^2 \cos \theta) \quad (6)$$

where P is the spin polarization percentage of the tunnel current, θ is the angle between the magnetization of the free and the pinned layers [3, 37].

The good agreement between the physical model and experimental measurement has been verified. Figure 3 shows the verification of static model with the measured data reproduced by Ohno group [8]. The blue and red solid lines represent the STT switching static model for parallel to anti-parallel process and anti-parallel to parallel process, respectively. The blue squares and red points represent the experimental results. From Fig. 3, the overlaps between the lines and the squares (or points) show the good agreement and the feasibility of this physical model to describe the STT switching static behavior.

On the other hand, more recent experimental progress of IBM shows that an MTJ involving symmetric electrodes provides a single spin polarization efficiency factor g for both state change processes (anti-parallel state to parallel state process or parallel state to anti-parallel state process) of MTJ [38], which allows the same critical current for both parallel and anti-parallel states. In this mechanism, g is only related to TMR ratio and described as follows:

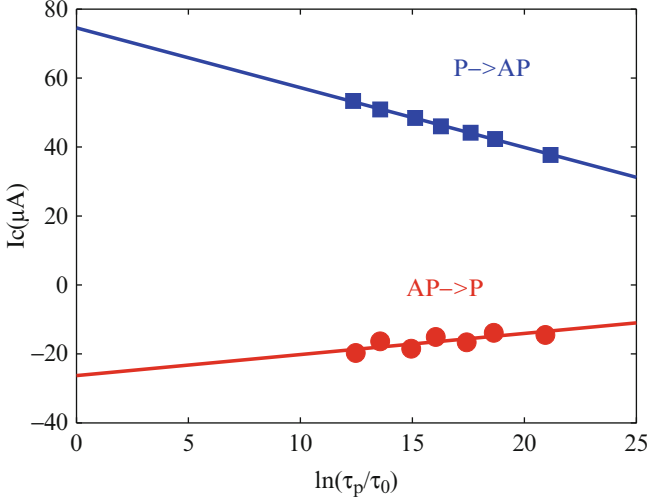


Fig. 3 Verification of the static model with measured data reproduced by [8]. Solid lines represent the STT static model, the red points and blue squares represent the experimental results

$$g = (TMR(TMR + 2))^{1/2} / 2(TMR + 1) \quad (7)$$

The dynamic switching behavior of STT in PMA MTJ shows the dependence of switching current on switching duration. It is considered to be a complex process as it can be categorized into two regimes depending on the relative magnitude between switching current (I) and critical current (I_{C0} , calculated by Eq. (3) for static behavior): thermally assisted ($I < 0.8 I_{C0}$) and precessional ($I > I_{C0}$) switching regimes. Thermally assisted regime can be described by Néel-Brown model and precessional regime can be described by Sun model [39–41]. Note that there are no clear experimental results and theories related to the range from $0.8I_{C0}$ to I_{C0} , we thus neglect this range and consider no effect occurs in this range.

For each model, the relationship between current and duration follows different laws. For practical applications, the two regimes have their own specific interest: the thermally assisted regime corresponds to low current density but slower switching, which is usually used for the sensing operation; the precessional regime corresponds to fast switching (sub 3 ns) but high current density, which is usually used for the writing operation.

In the sub-threshold condition where the current remains below the critical current ($I < 0.8I_{C0}$), the switching can still occur thanks to thermal activation above the voltage/current-dependent barrier. In this case, the switching behavior can be described by Néel-Brown model [42]:

$$\frac{d\text{Pr}(t)}{(1 - \text{Pr}(t))dt} = \frac{1}{\tau_1} \quad (8)$$

$$\tau_1 = \tau_0 \exp\left(\frac{E}{k_B T} \left(1 - \frac{I}{I_{c0}}\right)\right) \quad (9)$$

where τ_0 is the attempt period, $\text{Pr}(t)$ is the switching probability. Eq. (8) can be transformed to a simple formula:

$$t = -\tau_1 \ln(1 - \text{Pr}(t)) \quad (10)$$

These equations demonstrate that the STT dynamic switching behavior is probabilistic or stochastic. However, from Eq. (10), it can convert this stochastic behavior to be deterministic by determining the switching probability. That means ones should apply a specifically long current pulse to get the determined switching probability. This assumption would greatly simplify the description and analyses of the thermally assisted regime. Meanwhile, the stochastic effect is still the key point for this regime, which will be described and integrated in the following part.

In the case that the switching current is near or exceeding the critical one, the STT excitation becomes more obvious and deterministic. The high current pulse drives the magnetization to process, then after reaching the switching time, a magnetization reversal occurs suddenly and quickly [9]. Considering a small thermal fluctuation in this regime with a relatively high thermal stability, the average switching time is given by

$$\frac{1}{\langle \tau \rangle} = \left[\frac{2}{C + \ln\left(\frac{\pi^2 \xi}{4}\right)} \right] \frac{\mu_B P_{pin}}{em_m (1 + P_{pin} P_{free})} (I - I_{c0}) \quad (11)$$

where C is the Euler's constant, $\xi = E/k_{BT}$ is the activation energy in units of k_{BT} , P_{pin} , P_{free} are the tunneling spin polarizations of the pinned and free layers, we assume here that $P_{pin} = P_{free} = P$, m_m is the magnetic moment of free layer. Figure 4 shows the good agreement of this dynamic model with the experimental data extracted from [9]. From this figure, the increase of I and decrease of I_{c0} both contribute to scale down the switching latency. Considering the high currents are always ensured by the large-size transistors, this physical model also implies the alternatives to optimize the tradeoff between the overall area and the speed of hybrid spintronic/CMOS circuits.

Recently, a lot of experimental and theoretical results have shown that, although STT switching may allow sub-nanosecond switching duration, the switching process of STT is intrinsically stochastic, which results from the unavoidable thermal fluctuations of magnetization (see Fig. 5) [43–46]. They are responsible for large fluctuation in the switching duration, which can be proven by the Eqs. (8)–(11) describing the dynamic behavior. Moreover, the stochastic behavior can also be

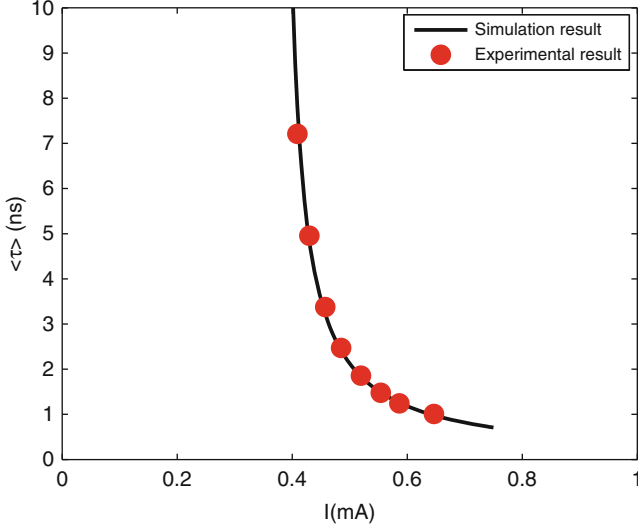
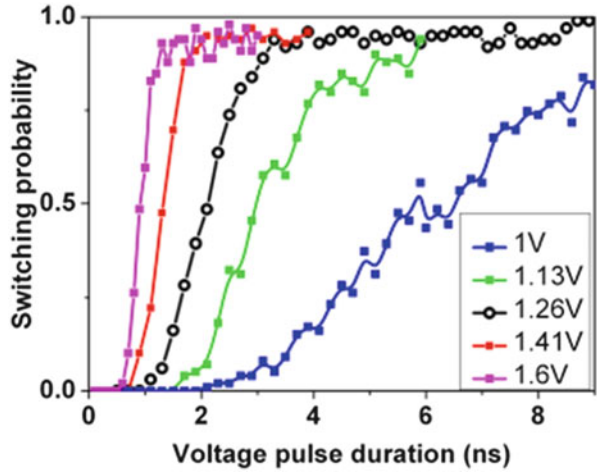


Fig. 4 Comparison of the dynamic model with measured data

Fig. 5 Experimental measurements of STT stochastic switching behaviors, high writing current drives faster speed and higher switching probability [44]



divided into two regimes: thermally assisted ($I < 0.8 I_{C0}$) and precessional ($I > I_{C0}$) switching regimes.

For the thermally assisted regime, we can transform Eq. (10) to another form:

$$\text{Pr}(t) = 1 - \exp(-t/\tau_1) \quad (12)$$

It describes the probability density function (PDF) of the switching duration for this regime, which follows an exponential distribution with characteristic time τ_1 decreasing with the current density.

In the super-threshold region described as precessional regime, the stochastic switching is triggered by a thermal fluctuation which creates an initial angle between the current spin-polarization and the magnetization of magnetic layer. The switching duration then follows a specific exponential-like distribution centered on the average switching delay time calculated by Eq. (11) [47].

From the above expressions, it shows that, in both regions, increasing the switching probability requires to increase either the write current or the current pulse duration. It could also be of great benefit for tolerating the high mismatch and process variations [48, 49].

To address the requirement of high-performance MTJ for the future logic and memory applications, there are usually five criteria to evaluate: small area, high TMR ratio, low STT switching current, capacity to withstand the standard semiconductor processing and high thermal stability. With the shrinking of size, the conventional MTJ with in-plane magnetic anisotropy becomes more and more difficult to satisfy these criteria. Recent material progress showed that the MTJ with PMA could offer lower switching critical current, higher switching speed and higher thermal stability compared with that with in-plane magnetic anisotropy [7]. These can be explained by the following theories.

The barrier energy and critical current of STT switching in the materials with in-plane magnetic anisotropy can be expressed as:

$$E_i = \frac{\mu_0 M_S \times Vol \times H_C}{2} \quad (13)$$

$$I_{C0i} = \alpha \frac{\gamma e}{\mu_B g} (\mu_0 M_S) \left(H_{ext} \pm H_{ani} \pm \frac{H_d}{2} \right) Vol \quad (14)$$

where H_C is the coercive field, H_{ext} is the external field, H_{ani} is the in-plane uniaxial magnetic anisotropy field, H_d is the out-of-plane magnetic anisotropy induced by the demagnetization field, μ_0 is the permeability in the free space, M_S is the saturation magnetization, Vol is the volume of the free layer, μ_B is the Bohr magneton, γ is the gyromagnetic ratio, e is the electron charge, m is the electron mass.

By comparing Eqs. (2) and (13), as H_K is higher than H_C , PMA allows obtaining relatively high barrier energy with a small size. By comparing Eqs. (3) and (14), as H_K is much lower than H_d , the critical current for PMA materials can be significantly reduced.

From 2002, when the first MTJ with PMA was reported, this advantageous structure attracts a great deal of attentions from academics and industries [50]. A variety of material systems has been attempted, for example, rare-earth/transition metal alloys, multilayers and other alloy materials. However, they have not been able to truly realize low critical current and high thermal stability at the same time. This situation didn't change until the Ta/CoFeB/MgO structure was revealed in 2010 [8, 9]. Figure 6 demonstrates the excellent performances of this structure. It takes advantages of CoFeB-MgO interface anisotropy to provide a good tradeoff

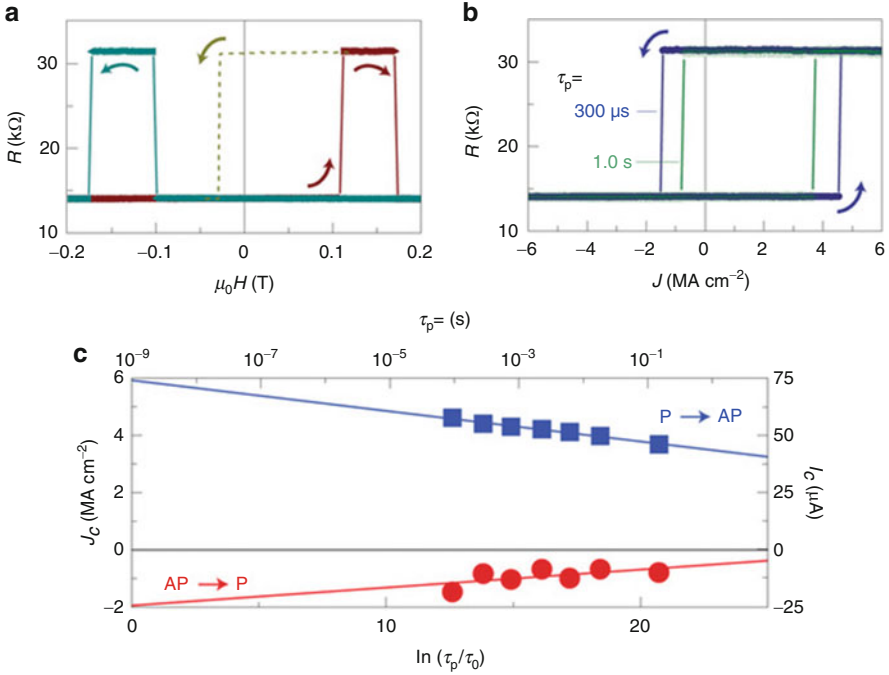


Fig. 6 TMR and current-induced magnetization switching for Ta/CoFeB/MgO structure MTJ with PMA. (a) Perpendicular R-H curve. (b) Typical results of current-induced magnetization switching at current pulse duration of 300 μs and 1.0 s. (c) Critical current density as a function of pulse duration [8]

among the area (40 nm), critical current ($\sim 50 \mu\text{A}$), thermal stability (40 $k_B T$) and TMR ratio ($> 100\%$). Thanks to the material and technical improvement of MTJ, especially MgO based PMA STT MTJ, a lot of persistent and intensive efforts have been made for the past years to develop the high-performance spintronic systems [51, 52].

2.2 Racetrack Memory

The observation of electrical CIDW motion in magnetic nanowires promises numerous perspectives [2, 53, 54] and the most interesting one is to build a novel ultra-dense non-volatile storage device, called “racetrack memory” (see Fig. 7).

The term “racetrack memory” was firstly proposed by Parkin in 2008 [5, 6]. In the concept that he proposed, write head nucleates a local domain in the magnetic nanowire and a current pulse drives the domain to move sequentially from write head to read head. Data or magnetization direction is stored between two artificial potentials or constrictions, which pin the DW as no current pulse is applied. The distance between two constrictions can be extremely small to some nanometers and

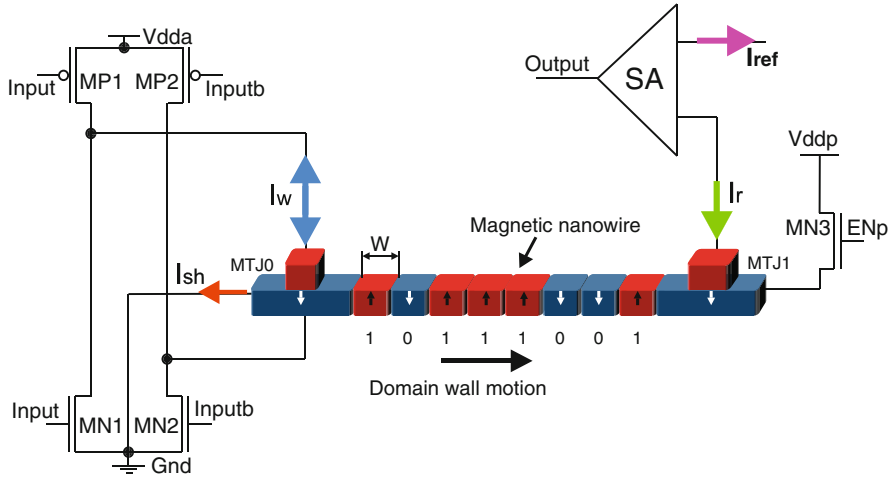


Fig. 7 Racetrack memory based on CIDW motion, which is composed of one write head (MTJ0), one read head (MTJ1) and one magnetic nano-stripe. I_w nucleates data or magnetic domain in the magnetic stripe through STT approach, I_{sh} induces DW motion along the magnetic stripe and I_r detects the magnetization direction through TMR effect

this allows an enormous storage ($>GB$) in a small die area. Compared with other non-volatile memory candidates, the scalability potential of racetrack memory is evident. By using MTJ as write and read heads, its operations, such as DW motion, domain nucleation and detection, can be addressed directly by CMOS circuits [55]. This hybrid integration makes racetrack memory promise high performance like high speed (>100 MHz) and low power beyond classical STT-MRAM. The nanowire can be built in 3D or 2D, the latter one is easier to be fabricated and become the mainstream solution for the current research on this topic. Based on in-plane magnetic anisotropy, the first racetrack memory prototype was presented in 2011 by IBM despite of its small capacity 256 bits [56]. However the intrinsic low energy barrier separating the two in-plane magnetization directions of storage layer leads to short data retention in advanced technology node (e.g., 22 nm) [51]. This drawback limits its use for high-density racetrack memory. PMA in some structures (e.g., CoFeB/MgO) providing a high energy barrier [8, 57, 58] were demonstrated and PMA MTJ become one of the most promising candidates to realize a read head. Advantageous domain wall nucleation current and speed with PMA MTJ were also observed recently [9] and this makes it be a better write head than in-plane MTJ.

The Cross-section structure of racetrack memory is shown in Fig. 8, which includes mainly three parts: a magnetic stripe separated by constrictions to store data, two MTJs as write and read heads. The number of constrictions equals to the number of stored bits. It is noteworthy that the CMOS circuits dominate the whole area of this racetrack memory as the magnetic stripe is implemented at the back-end through 3D integration as MRAM.

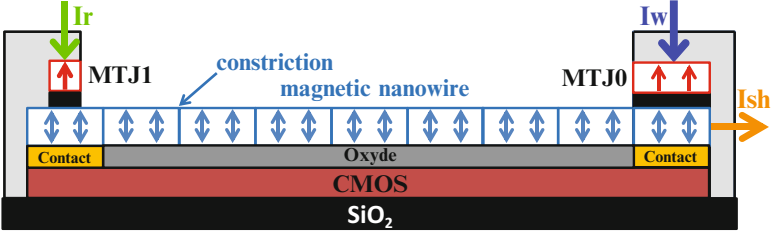


Fig. 8 Cross-section structure of racetrack memory. At the back-end process, the magnetic nanowire is implemented above the CMOS/MTJ interfacing circuits, which generate I_r for reading, I_w for DW nucleation and I_p for DW propagation

Figure 7 shows simultaneously one example of CMOS circuits to generate I_w and I_{sh} , which are respectively bi-directional and uni-directional at the side of write head. I_r is driven by a sense amplifier [59] and it can convert the stored data from different magnetization directions to digital signal “0” or “1”. In order to achieve the best write and read reliability, the width of write and read heads are different. For writing, a lower resistance of MTJ0 with larger width can reduce the rate of oxide barrier breakdown, which is one of the most significant constraints of the high-speed STT switching mechanism. On the contrary, high resistance of the MTJ1 with smaller width for reading can greatly improve the sensing performance.

For the racetrack memory, the speed performance is governed by the velocity of DW motion. The physical model to calculate DW velocity is indispensable for the compact modeling of racetrack memory. According to the previous literatures, the dependence of DW velocities on current and magnetic field can be described by the one dimensional (1D) model. This 1D model is deduced from the LLG equation in a 1D system [60], which can be described as:

$$\dot{\phi}_0 + \alpha \dot{X}/\lambda = \gamma H + \beta u/\lambda + f_{pin} \quad (15)$$

$$\dot{X} - \alpha \lambda \dot{\phi}_0 = v_{\perp} \sin 2\phi_0 + u \quad (16)$$

where X is the position of a DW, and ϕ_0 is the angle that the DW magnetization forms with the easy plane. λ is the width of DW, α is the Gilbert damping constant, β is the dissipative correction to the STT, H is the external field, γ is the gyromagnetic ratio, f_{pin} is the pinning force. The velocity constant v_{\perp} comes from the hard-axis magnetic anisotropy K_{\perp} ($\sim K_{\perp}\lambda/\hbar$). u is spin current velocity. These two equations can describe a lot of qualitative features of DW motion driven by the field and the current. The field acts as a “force” to drive ϕ_0 , the current acts as a “torque” to drive X . In addition, as the “torque” is also contributed from the hard-axis magnetic anisotropy, the state of ϕ_0 can determine whether there is intrinsic pinning or pure STT. Considering only the process after depinning, Eqs. (15) and (16) can be solved analytically and described in the forms of the influence of field and current on the velocity:

$$V = V_H + V_j \quad (17)$$

The velocity is the vector sum of field-induced (V_H) and current-induced velocities (V_j). Above the Walker breakdown field, the field-induced velocity contribution is given by

$$V_H = \alpha^2 \mu H \left\{ 1 - \frac{1}{1 + \alpha^2} \sqrt{1 - \left(\frac{H_W}{H} \right)^2} \right\} \quad (18)$$

where the mobility $\mu = \gamma\lambda/\alpha$, H_w is the Walker breakdown field.

The general racetrack memory is based on CIDW motion, which means there is normally no magnetic field. Hence, the dependence of DW velocity on current is the key point. Regarding the relationship between α , the damping constant, and β , the nonadiabatic coefficient, the dependence can be categorized into three cases. Before introducing these three cases, we should indicate the definition of the spin current velocity [60], which is given by Eq. (19).

$$u = \frac{\mu_B P j_p}{e M_S} \quad (19)$$

where j_p is the propagation current density. Figure 9 shows the dependence of DW velocity on current according to different configurations of α and β , which depends on the material of the magnetic nanowire.

When $\beta > \alpha$,

$$u_{WB} = \frac{1}{2} \gamma H_K \Delta \frac{\alpha}{\beta - \alpha} \quad (20)$$

$$\langle v \rangle = \frac{\beta}{\alpha} u \quad (u < u_{WB}) \quad (21)$$

$$\langle v \rangle = \frac{\beta}{\alpha} u - \frac{\sqrt{\left(1 - \frac{\beta}{\alpha}\right)^2 u^2 - \left(\frac{1}{2} \gamma \Delta H_K\right)^2}}{1 + \alpha^2} \quad (u > u_{WB}) \quad (22)$$

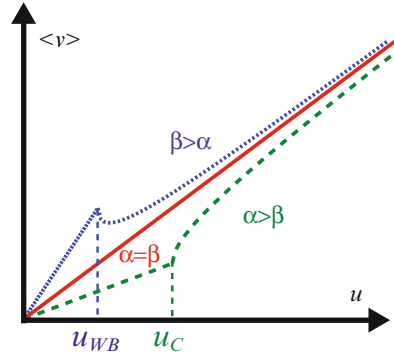


Fig. 9 Dependence of DW velocity on current described by 1D model

When $\alpha > \beta$,

$$u_C = \frac{1}{2}\gamma H_K \Delta \frac{\alpha}{\alpha - \beta} \quad (23)$$

$$\langle v \rangle = \frac{\beta}{\alpha} u \quad (u < u_C) \quad (24)$$

$$\langle v \rangle = \frac{\beta}{\alpha} u + \frac{\sqrt{(1 - \frac{\beta}{\alpha})^2 u^2 - (\frac{1}{2}\gamma \Delta H_K)^2}}{1 + \alpha^2} \quad (u > u_C) \quad (25)$$

When $\alpha = \beta$,

$$\langle v \rangle = u \quad (26)$$

where u_{WB} is the Walker breakdown velocity, u_C is the critical velocity corresponding to the critical current density of DW motion. In order to achieve a high speed racetrack memory, the current density should be more or far more than the critical one. In both cases, when applying a much higher current, the DW velocity approaches to spin current velocity. Therefore, we take this assumption into account, which means we use the spin current velocity to directly represent DW velocity. Thus,

$$V_j = u = \frac{\mu_B P j_p}{e M_S} \quad (27)$$

We verified this physical model by comparing with the micromagnetic simulations done by Ohno group (see Fig. 10) [61]. In this case, we suppose that the DWs are definitely pinned when the current density is lower than the critical one, the velocity is thus kept to zero in this condition. From the figure, a current density of

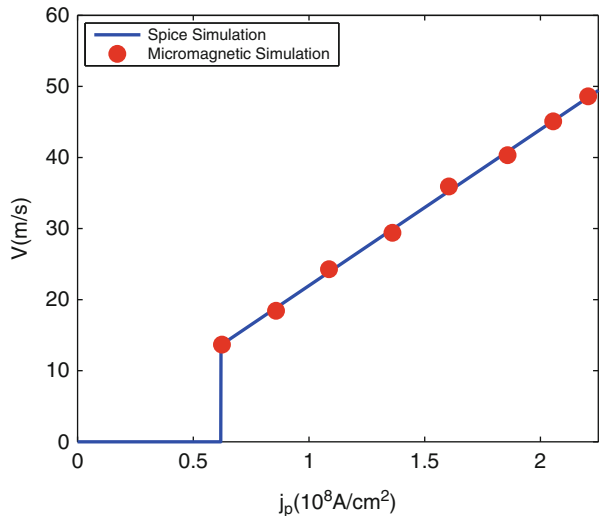


Fig. 10 Good agreements with micromagnetic simulation for DW motion velocity as a function of current density

$\sim 2 \times 10^8$ A/cm² can trigger a DW motion in 50 m/s, which is beneficial for the logic and embedded memory circuits.

By considering the distance W between two adjacent constrictions (see Fig. 7), we can calculate the necessary pulse duration for current to move one storage element by the Eq. (28). For example, when W is 40 nm, the DW velocity is 50 m/s, the pulse duration can be as small as 0.8 ns. If neglecting the nucleation process, the frequency of racetrack memory can thus be as high as 1 GHz. If considering the nucleation process time (e.g., 1–2 ns), the frequency can still be 500 MHz.

$$D = W/V_j \quad (28)$$

Caused by the thermal activation, stochastic nature has been found for DW motion in diverse structures and materials. With the reduction of the applied current or field by optimizing the techniques and the materials, the stochasticity of DW motion will be further enhanced [62]. DW velocity and displacement are susceptible to stochastic effect, which exerts a considerable influence to the feasibility and reliability of DW-based devices, not only racetrack memory. However, as there have not been some coherent experimental results or physical theories concerning the pure CIDW in PMA materials, we refer to the measurements of DW motions in spin valve induced simultaneously by current and field. Under this condition, after depinning, the DW motion velocity is found to follow a Gaussian-like specific distribution centered with the value calculated by Eq. (27) [63]. We analyze the dependence of cumulative probability of DW motion versus different current pulse durations and magnitudes in Fig. 11. It illustrates a coherent functionality of stochastic behavior where we can also find that higher and longer current pulse yields a more probable DW motion.

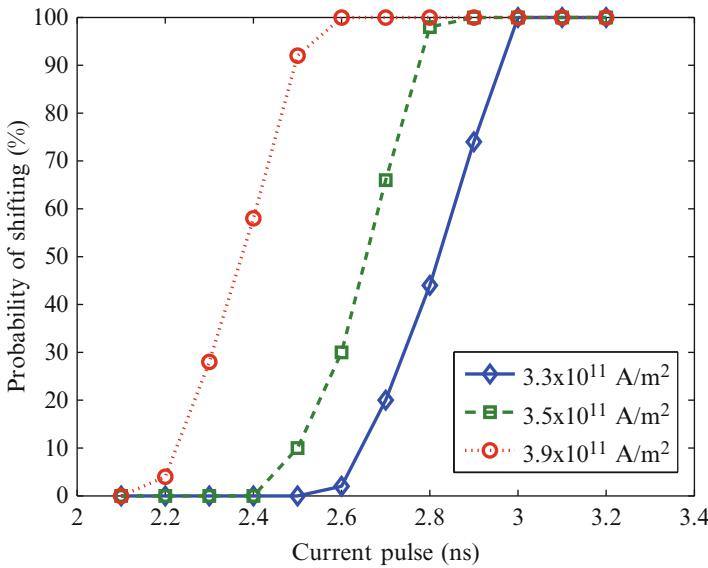


Fig. 11 Dependence of cumulative probability of DW motion versus shifting current pulse duration and magnitude

3 Current-Induced Magnetic Switching Based Hybrid Spintronics/CMOS Circuits for High-Performance Computing

3.1 Sensing Circuit

Due to the TMR effect, MTJ presents the property of resistance difference for different states. This resistance property allows MTJ to be compatible with CMOS sense amplifier circuit that detects the MTJ's configuration and amplifies them to logic level. Among various sense amplifiers [64–66], pre-charge sense amplifier (PCSA) is proposed to provide not only the best tradeoff between sensing reliability and power efficiency, but also high-speed performance [59]. Thereby we focus on PCSA and apply it for the hybrid logic circuits involved in this chapter.

The PCSA circuit (see Fig. 12) consists of a pre-charge sub-circuit (MP2-3), a discharge sub-circuit (MN2) and a pair of inverters (MN0-1 and MP0-1), which act as an amplifier. Its two branches are normally connected to a couple of MTJs with complementary states. It operates in 2 phases: “Pre-charge” and “Evaluation”. During the first phase, “CLK” is set to “0” and the outputs (“Qm” and “/Qm”) are pulled-up to “Vdd” or logic “1” through MP2-3 while MN2 remains off. During the second phase, “CLK” becomes “1”, MP2-3 are turned off and MN2 on. Due to the resistance difference between the two branches, discharge currents are different. The lower resistance branch will be pulled-down to reach more quickly the threshold voltage of the transistor (MP0 or MP1), at that time, the other branch will be pulled up to “Vdd” or logic “1” and this low-resistance branch will continue to drop to “Gnd” or logic “0”.

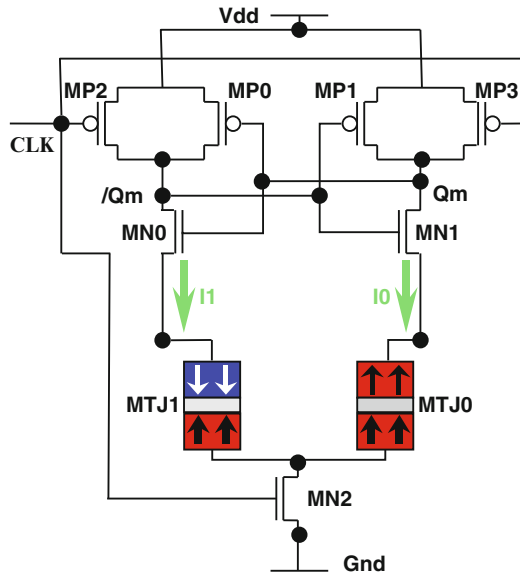


Fig. 12 Pre-charge sense amplifier (PCSA) for MTJ state detection and amplification to logic level

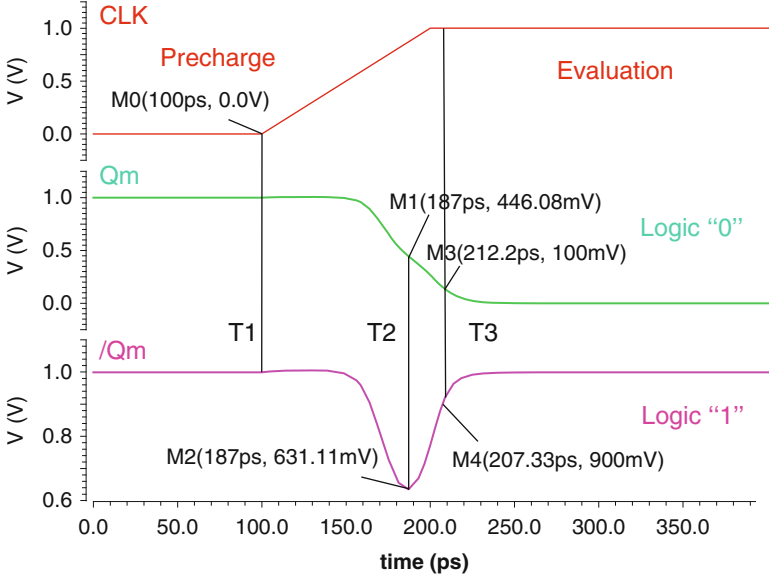


Fig. 13 PCSA sensing operation in the case of MTJ0 with “parallel” state and MTJ1 with “anti-parallel” state

Figure 13 shows a sensing operation of PCSA in the case of MTJ0 with “parallel” state and MTJ1 with “anti-parallel” state. Before the moment “T1”, it is “Precharge” phase. Both outputs are pulled up to 1 V. Then the “Evaluation” phase starts from “T1”, two branches begin to discharge after a small delay considering the rising time of “CLK” signal. At the time “T2”, the branch “Qm” reaches the threshold firstly and this branch will continue to decrease to “0”. At the same time, MP0 begin to work and recharge the complementary branch “/Qm” back to “1”. This sensing operation is so speedy. From the figure, we can find the whole process costs a sensing delay less than 100 ps. From the point of view of consumption energy, a sensing operation can only cost as low as 10 fJ. This high-speed and low-power feature makes PCSA suitable for the logic applications.

We use PCSA circuit in the hybrid MTJ/CMOS design for the other reason: the read disturbance induced by sensing operations can be significantly decreased. It is important for embedded STT-MRAM as it is an intrinsic nature and difficult to correct in logic circuit where complex error correction circuit (ECC) is prevent to ensure fast computing speed (e.g., 1 GHz). The read disturbance can be regarded as the unexpected switches during the sensing operation. As the sensing current is usually much lower than the critical current, the switch probability can be described by Néel-Brown model. If there are N bits of MTJs in the chip, the chip failure rate F_{chip} can be calculated by Eq. (29).

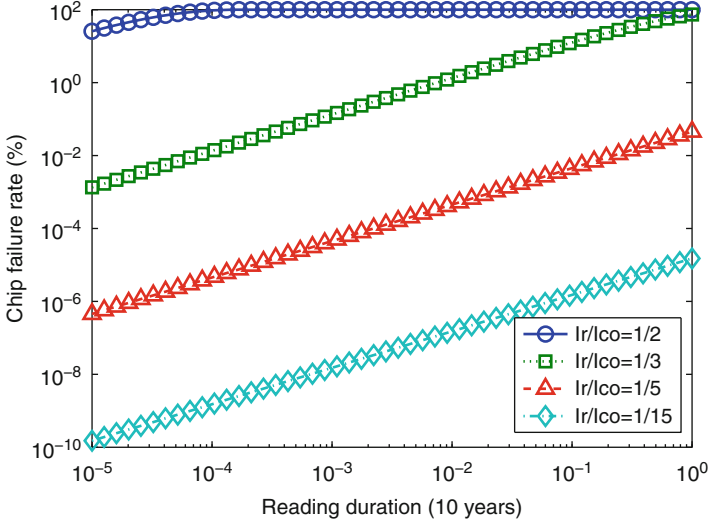


Fig. 14 Dependence of chip failure rate on reading duration for different reading current

$$F_{chip} = 1 - \exp \left[-N \frac{\tau}{\tau_0} \exp \left(-\Delta \left(1 - \frac{I_r}{I_{CO}} \right) \right) \right] \quad (29)$$

where N is the number of bits per word, I_r is sensing current, I_{CO} is the critical current, τ is the read duration and τ_0 is the attempt period. As shown in Fig. 14, lower I_r and shorter τ can reduce greatly the chip failure rate for the STT-MRAM with the same thermal stability factor $\Delta = 40$.

In reality, numbers of words of memories (e.g., 1 k) normally share a sense amplifier. As shown in Fig. 15, a 16 k-bit (1 k words of 16 bits) PCSA sensing circuit has been studied. This enormous parallel structure leads to a huge capacitance, which drives the current pulse through the MTJ. As a result, an evaluation phase lasts almost 10 ns. By taking the effect of stochastic behavior into account, Monte-Carlo simulations after 1 μ s of sensing duration (i.e., 1,000-time sensing operations) has been performed (see Fig. 16). We found that the 33 errors occurred among 100 simulations. They are caused by either mismatch and process variations of CMOS part or STT stochastic behaviors of MTJ, or sometimes by both of them.

To identify the impact from each of them, we also performed Monte-Carlo simulations for sensing circuit with only mismatch and process variations. We found that the read disturbance was $\sim 11\%$. Compared with the result presented in Fig. 16, we can conclude that the stochastic behavior of MTJ greatly increases the error probability for a long-pulse current, and that this PCSA is not suitable for very large memory systems.

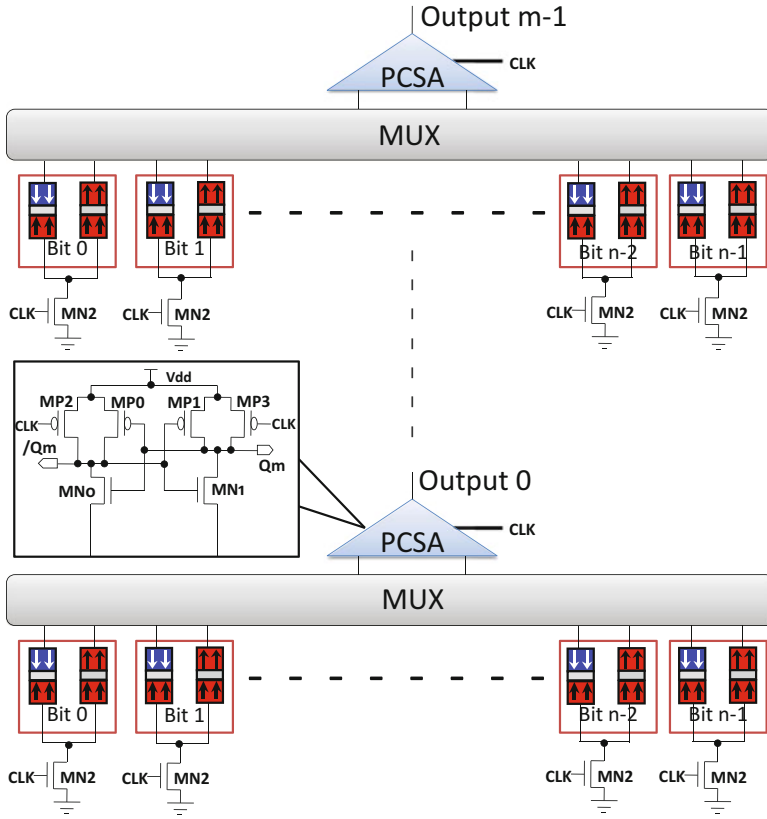


Fig. 15 Schematic for 16 k-bits PCSA sensing circuit ($m = 16$, $n = 1$ k)

To improve further the sensing reliability, a separated PCSA (SPCSA) was proposed, as shown in Fig. 17 [67]. The basic structure as well as the operation of SPCSA is similar to that of the PCSA. The main difference is that SPCSA separates the discharging and evaluation stages with two different paths, which alleviate greatly the voltage headroom problem, enabling it to operate at a relatively lower supply voltage. Meanwhile, thanks to the separated discharge and evaluation stages, we can amplify the input signals before entering the evaluation stage so as to tolerate the input-offset. In addition, two inverters (IV1 and IV2) and two NMOS transistors (MN2 and MN3) are added connecting between the discharging and evaluation stages, to amplify the limited current or voltage difference (due to the limited TMR ratio) between the two discharging paths flowing through MTJ0 and MTJ1, thus tolerating significantly the process variations and increasing greatly the sensing margin. Figure 18 shows the statistical sensing error rate of SPCSA compared to PCSA. As can be seen, SPCSA provides a much higher sensing reliability with the same hardware.

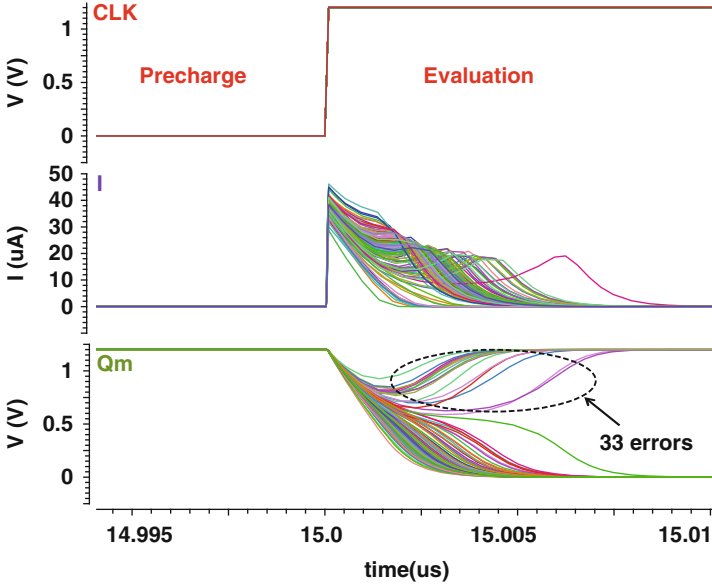


Fig. 16 Monte-Carlo simulation of a 16 k-bit PCSA circuit

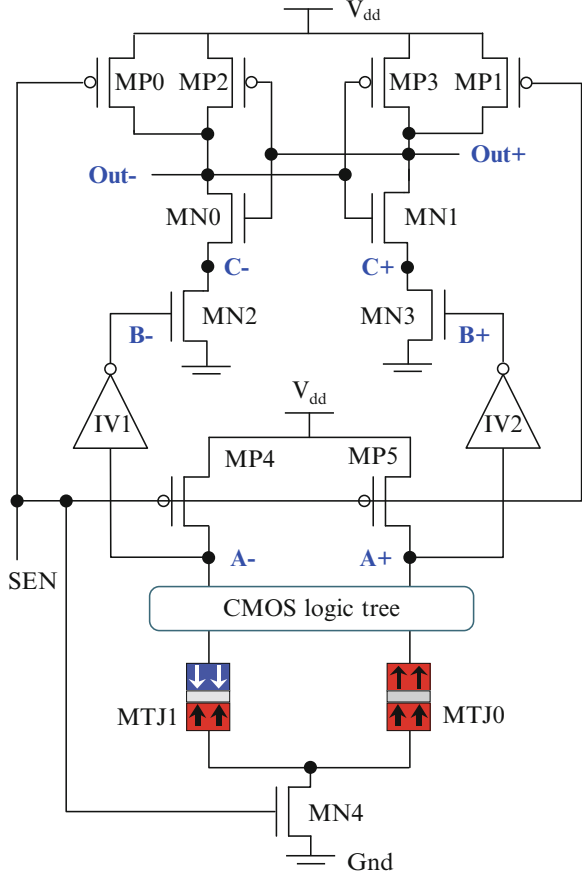
3.2 Writing Circuit

According to STT switching mechanism, a bi-directional current is required to switch the magnetization in free layer of MTJ. In order to achieve high-speed logic design, high current is required to ensure the speed. In contrast to the low power and low area of the sensing circuit, the writing circuit for PMA STT MTJ occupies the main area and power of the whole circuit. As a result, the study on writing circuit is of importance to hybrid MTJ/CMOS circuit design.

In concert with the sensing circuit for a couple of MTJs with complementary states, a writing circuit to generate the bi-directional current for switching a couple of MTJs is designed as Fig. 19. Two NMOS (MN0-1) and two PMOS (MP0-1) transistors construct the main circuit. Each time one NMOS and one PMOS are always left open and the others closed, which creates a path to make the current pass from “Vdda” to “Gnd”. Through two NOR and three NOT logic gates, the signals “Input” and “EN” control respectively the current direction and activation. Normally, it requires a “Vdda” higher than “Vdd” for logic operations to avoid the area overhead in the write circuit.

In order to generate the maximum current flowing through the couple of MTJs, both the transistors (one PMOS and one NMOS) should operate in their linear region above the threshold voltage V_{TH} to obtain the relatively lower resistances. In this case, they should satisfy the conditions: $V_{DS} \ll 2(V_{GS} - V_{TH})$ for NMOS and $V_{DS} \gg 2(V_{GS} - V_{TH})$ for PMOS. Their resistances, R_{on} and R_{op} , can be

Fig. 17 The schematic of the separated pre-charge sense amplifier (SPCSA)



approximately expressed by Eqs. (30) and (31), and the generated current can be obtained through the Eq. (32),

$$R_{on} = \frac{1}{\mu_n C_{ox} \frac{W}{L} (V_{GS} - V_{TH})} \quad (30)$$

$$R_{op} = \frac{1}{\mu_p C_{ox} \frac{W}{L} (V_{SG} - |V_{TH}|)} \quad (31)$$

$$I_{write} = \frac{V_{dda}}{R_p + R_{ap} + R_{on} + R_{op}} \quad (32)$$

where μ_n is the electron mobility, μ_p is the hole mobility, C_{ox} is the gate oxide capacitance per unit area, W is the channel width, L is the channel length, V_{GS} is the gate-source voltage.

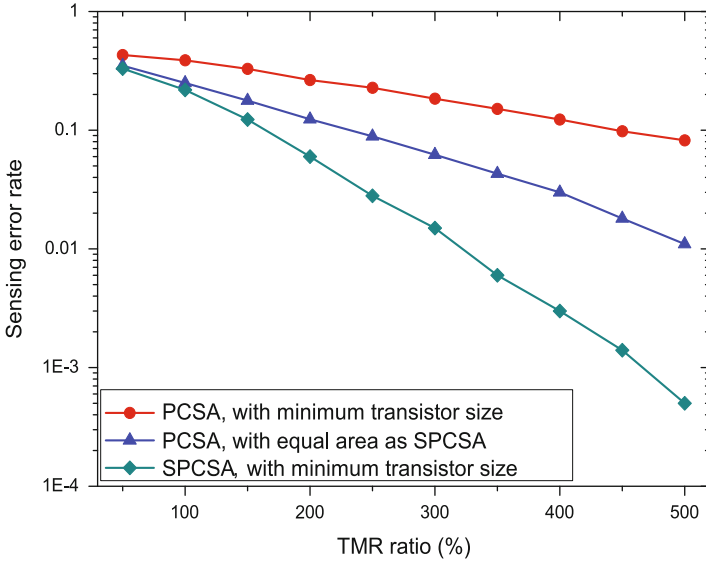


Fig. 18 Sensing error rate of the SPCSA circuit with Monte-Carlo simulations

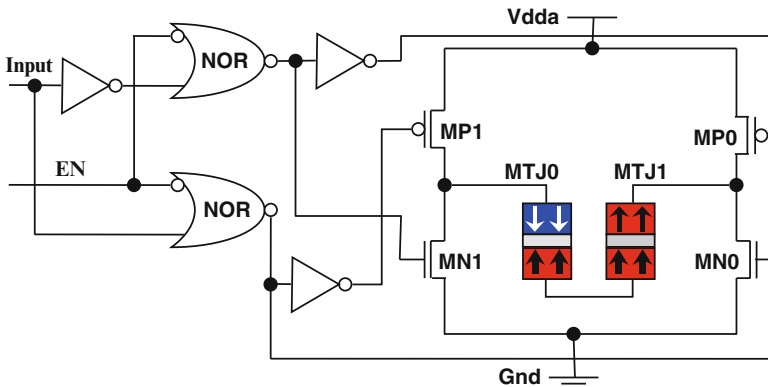


Fig. 19 Full writing schematic for STT writing approach, which is composed of two modified inverters and logic control circuits

By simulating a writing operation including anti-parallel to parallel switching and parallel to anti-parallel switching (see Fig. 20). We can find that the writing operation is not activated until the signal “EN” is set to “1”. The states of the couple of MTJs remain always opposite and the switching direction follows the signal “Input”.

From Eqs. (30)–(32), we find that the most efficient method to improve the current value is by increasing W , but this leads to significant area overhead. Figure 21 shows a study of area, speed and energy performance for this circuit. Here, only the area of four transistors (MN0-1, MP0-1) has been taken into account as the area of logic control circuit is the same for different simulation and is often in

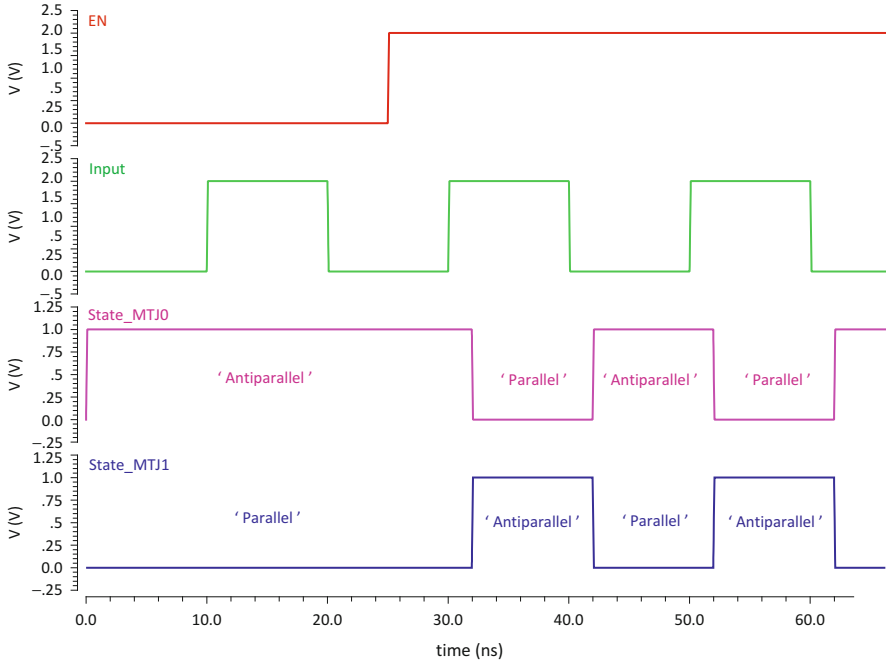


Fig. 20 Writing operation for a couple of MTJs with complementary states. The signals “Input” and “EN” control respectively the current direction and activation

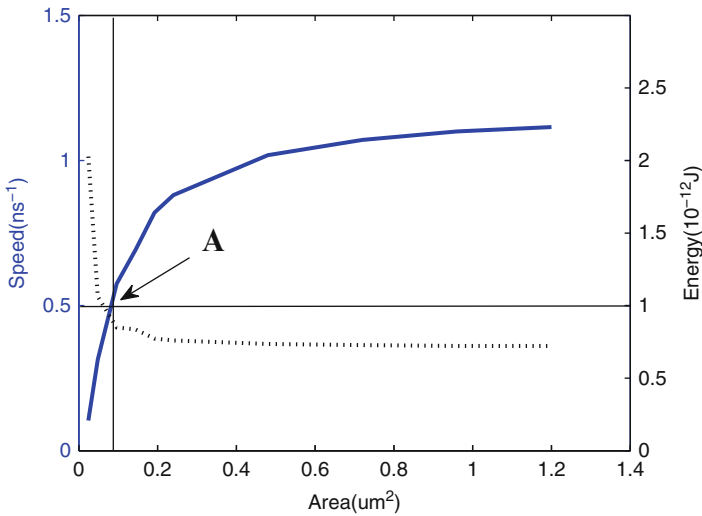


Fig. 21 Dependence of circuit switching speed (solid line) and energy dissipation (dotted line) versus die area with four transistors (MN0-1, MP0-1)

the minimal size. A strong dependence between area and speed can be found, especially when the area is smaller than $0.2 \mu\text{m}^2$. The speed improvement becomes less significant for larger areas and saturates at ~ 1.1 GHz, which is different from the 2 GHz obtained with a single cell. There are two reasons for explaining this: first, “V_{dda}” is set to 2 V as 2.2 V is the breakdown limit for CMOS technology node [68]; second, there are a couple of MTJs, so the bias-voltage for each one cannot be larger than 1 V as there is also bias on the transistors in the circuit.

The energy of each switching operation has been calculated with Eq. (33). We also find a turning point, $\sim 0.1 \mu\text{m}^2$, below which the energy will be increased rapidly with a smaller area due to the extremely long switching duration as the current I_{write} approaching to the threshold I_{CO} . Contrarily, the energy is nearly the same for whatever the size larger than $\sim 0.1 \mu\text{m}^2$. This is firstly because that the writing current and speed approach to be saturated. Even if the writing current can increase continuously, from Eq. (11), the current is inversely proportional to the switching duration when the current is much higher than the critical one. Therefore the energy will inevitably be saturated for a high writing current.

$$E_{\text{operation}} = V_{\text{dda}} \times I_{\text{write}} \times \text{Duration} \quad (33)$$

The region around the crossing point of the two curves (point “A” in Fig. 21) can be localized. It can be considered as a good tradeoff among the area ($\sim 0.096 \mu\text{m}^2$ or 30 F^2), power (1 pJ) and speed (~ 500 MHz) performance of this switching circuit, and be suitable to build up both logic chip and memory. This analysis can also help to investigate the circuits with special requirements like 800 MHz operating frequency.

For the advanced node below 90 nm, high reliability is becoming more and more crucial for the IC design [69–72]. Thanks to the integration of STT stochastic behavior into this model, an overall reliability investigation becomes possible. Figure 22 shows the statistical Monte-Carlo simulations of 100 complete writing

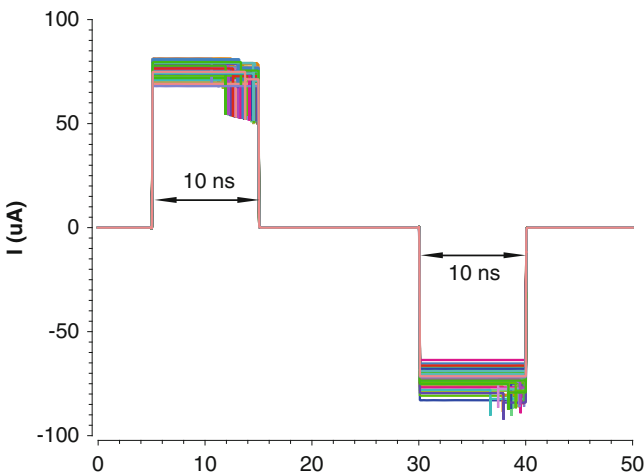


Fig. 22 Monte-Carlo simulation of a whole writing operation implemented by the writing circuit

operations using the writing circuit shown in Fig. 19 (however, instead of a couple of MTJs, there is only one MTJ connected in the circuit for this part of study). The complete writing operation includes the switching from parallel to anti-parallel and from anti-parallel back to parallel. Similarly to the case of the sensing circuit, the writing current at each write event is different because of the mismatch and process variation of CMOS part. The switching delay times vary randomly due to the stochastic behavior of MTJ cell. Since writing current is normally larger than sensing current, the stochastic effect in writing operation is relatively weaker than that in sensing operation. This can be proven by Fig. 22, the variation of every event is not so enormous.

Writing current magnitude and pulse duration are two key factors for the writing operation. As mentioned above, the writing current magnitude is dependent on the die area of writing circuit. We then perform the Monte-Carlo simulations for different writing pulse durations (5, 10 and 20 ns) to observe the dependence of writing Bit Error Rate (BER_W) versus die area of writing circuit (four main transistors: MP0-1 and MN0-1). The simulation results shown by Fig. 23 demonstrate their tradeoff relation: the increase of area can improve the BER performance. The reason is that a larger circuit allows larger write current, which in average reduces the time required to switch. For a given pulse duration, this increases the switching probability. Correlatively, it is observed that a longer pulse can also increase the reliability, which confirms the explanation mentioned above.

On the other hand, in order to overcome both power and reliability issues of conventional switching circuits due to the STT stochastic behavior, a self-enabled “error-free” strategy was proposed [73]. The corresponding circuit schematic is shown in Fig. 24. A sense amplifier (S.A) associated to the MTJ detects its state and outputs the data in logic level. The “self-enable” signal depends on the comparison

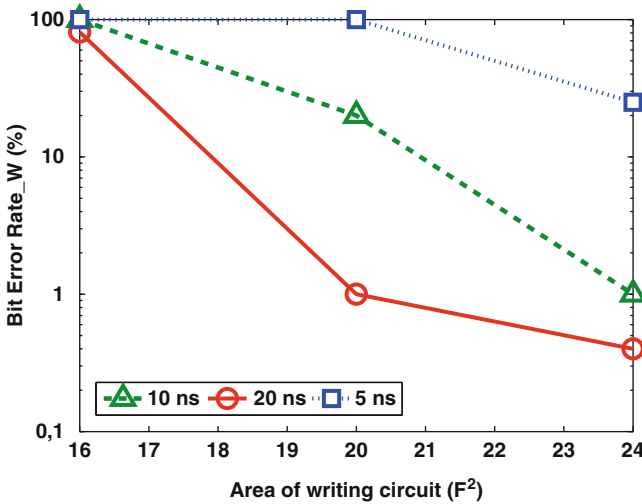
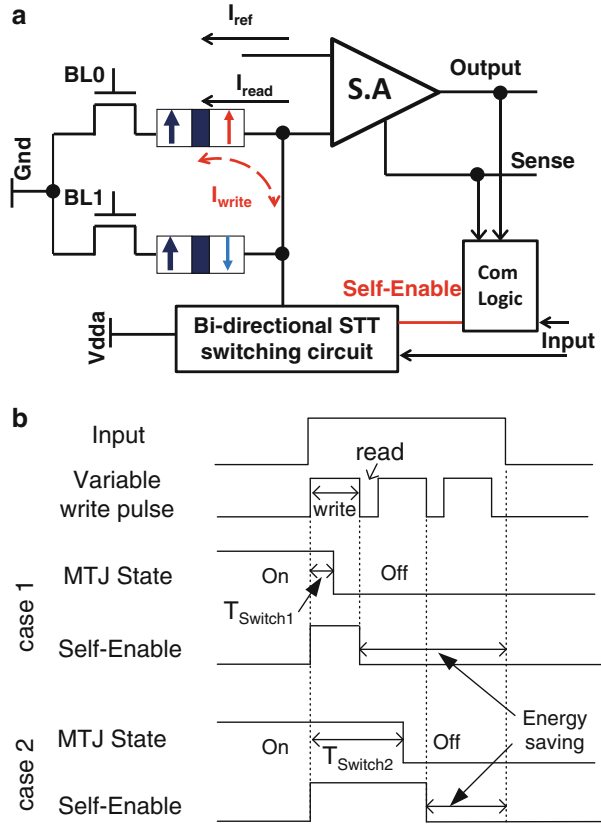


Fig. 23 Dependence of writing Bit Error Rate (BER_W) versus die area of writing circuit

Fig. 24 Self-enabled “error-free” switching circuits and strategy with adaptive driver pulse duration (a) Scheme of proposed switching circuit (b) Varied duration for “self-enable” signal



result between output and “Input” data. For instance, it becomes “ON” as output is different from “Input” data. The fixed long writing pulse is replaced by a sequence of short duration T_{SS} including both switching and sensing operations. Thanks to the stochastic behaviors of STT magnetic switching, the state of MTJ can be changed just after one short write pulse, as shown in Fig. 24b. After that, “self-enable” is set to “OFF” and no current flows through the MTJ. Different from a self-adaptive write circuit designed for memristor, the proposed circuit takes benefits from the stochastic behaviors of STT switching. Moreover, periodic sensing is used to obtain the STT-MRAM storage in logic level for the comparison with “Input” data. This is due to the relatively low TMR or R_{Off}/R_{on} ratio of MTJ (e.g., 150–250 %). The frequency of read operations equals normally to the global clock (e.g., 500 MHz).

This switching circuit with self-enable mechanism presents a number of advantages. Firstly, it allows “error-free” as the switching operation becomes fully deterministic instead of stochastic behaviors caused by the intrinsic STT and PVT variations. As the write pulse duration is shortened and the number of switching operation is also reduced, the lifetime of oxide barrier can be greatly improved.

As mentioned above, the state of MTJ may be erroneously changed by a read current, “self-enable” becomes automatically “ON” to correct this error. Thereby, this proposed circuit provides evident high reliability.

Secondly, high power efficiency can be achieved by eliminating completely the additional power to tolerate the process voltage temperature (PVT) variations and stochastic behaviors. Another power saving comes from the reduced switching numbers as the “self-enable” signal is activated only while the stored data is different from “Input” data. On average, half of the switching operations can be economized, but exact power saving depends greatly on applications. Note that, for asynchronous applications, in addition to power saving, better operating speed could also be expected.

3.3 Magnetic Full Adder

Aiming to overcome the issue of rising standby and dynamic power, magnetic processor based on spintronic devices is thus expected. Since addition is the basic operation of the arithmetic/logic unit of any processors, MFAs attract a lot of attention and several designs based on diverse technologies are proposed in the last years [18, 19]. Here, we present a 1-bit MFA based on PMA STT MTJ (STT-MFA) [74] and a multi-bit MFA based on PMA racetrack memory [75].

3.3.1 1-Bit MFA Based on PMA STT MTJ

Figure 25 shows a 1-bit STT-MFA circuit, which is based on the generic logic-in-memory structure [74]. To evaluate the logic function, PCSA circuit is used. The inputs are “ A ”, “ C_i ” and “ B ”, and the outputs are “ SUM ” and “ C_o ”. Among them, the input “ B ” relates to non-volatile storage PMA STT MTJ. The MOS tree is designed according to Eqs. (34)–(37) and the truth table shown in Table 1.

$$SUM = A \oplus B \oplus C_i = ABC_i + \overline{ABC}_i + \overline{ABC}_i + \overline{ABC}_i \quad (34)$$

$$\overline{SUM} = \overline{ABC}_i + \overline{ABC}_i + \overline{ABC}_i + \overline{ABC}_i \quad (35)$$

$$C_o = AB + AC_i + BC_i \quad (36)$$

$$\overline{C_o} = \overline{AB} + \overline{AC}_i + \overline{BC}_i \quad (37)$$

For “ SUM ” logic, the MOS tree corresponds directly to the logic relationship among the inputs “ A ”, “ C_i ” and “ B ”, we can simply adapt it to the general structure with a couple of complementary PMA STT MTJ. However, it is a little difficult for “ C_o ” logic as there is the term AC_i in the logic function Eq. (36) and we cannot adapt the schematic to the general “logic-in-memory” structure. It can be inferred that the impact of the term AC_i on the resistance is equivalent to a sub-branch

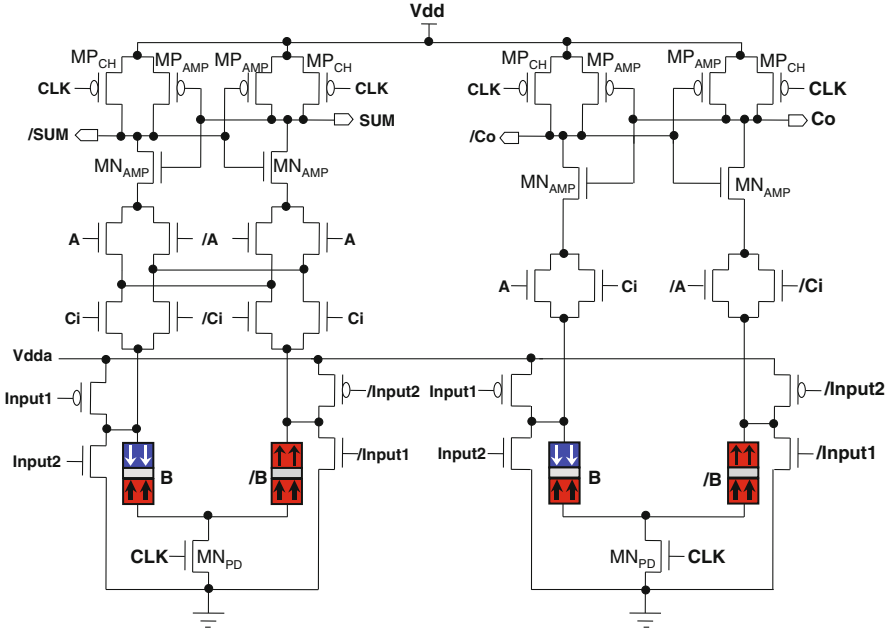


Fig. 25 STT-MFA architecture with “SUM” (left) and output carry “Co” (right) sub-circuits, “A” is volatile data for computing, “B” is non-volatile data using as quasi-constant [74]

Table 1 Truth table of “SUM” and “Co” logic gate for MFA

A	B	C	SUM	Co
0	0	0	0	0
0	0	1	1	0
0	1	0	1	0
0	1	1	0	0
1	0	0	1	0
1	0	1	0	1
1	1	0	0	1
1	1	1	1	1

connecting PCSA and the discharging transistor ($MN2$ in Fig. 12). Table 2 exhibits the true table and the resistance configuration of “Co” logic. R_{OFF} and R_{ON} are respectively the close and open resistances of MOS transistor. R_L and R_R are respectively the whole resistance of the left and right branch of PCSA. We can find that whatever the value of “A” and “Ci”, the sub-branches AC_i and $\overline{AC_i}$ have no impact on the output. If “A” and “Ci” are different, the resistances of the two sub-branches are the same. If they are the same, their comparison corresponds to that of R_L and R_R in the condition of $R_{ON} > R_{AP}$, which is always true for PMA STT MTJ under present technology condition. This allows the term AC_i to be deleted from Eq. (36) and we can obtain the “Co” logic circuit shown in Fig. 25.

Table 2 Truth table and resistance configuration of “ C_o ” for MFA

A	B	C	Resistance comparison	C_o	Sub-branch AC_i	Sub-branch \overline{AC}_i
0	0	0	$R_L > R_R$	0	$2R_{OFF}$	$2R_{ON}$
0	0	1	$R_L > R_R$	0	$R_{OFF} + R_{ON}$	$R_{OFF} + R_{ON}$
0	1	0	$R_L > R_R$	0	$2R_{OFF}$	$2R_{ON}$
0	1	1	$R_L < R_R$	0	$R_{OFF} + R_{ON}$	$R_{OFF} + R_{ON}$
1	0	0	$R_L > R_R$	0	$R_{OFF} + R_{ON}$	$R_{OFF} + R_{ON}$
1	0	1	$R_L < R_R$	1	$2R_{ON}$	$2R_{OFF}$
1	1	0	$R_L < R_R$	1	$R_{OFF} + R_{ON}$	$R_{OFF} + R_{ON}$
1	1	1	$R_L < R_R$	1	$2R_{ON}$	$2R_{OFF}$

The PMA STT MTJs connect serially with a common central point. In order to program MTJs, we use a writing circuit composed of pass transistors, which are connected respectively to the bottom and top electrodes of the serial branch. In such a manner, as a control signal (“ $Input1$ ” or “ $Input2$ ”) is activated, the first PMA STT MTJ noted “ B ” is put in high resistance state (R_{AP}) or low resistance state (R_P) while the second PMA STT MTJ noted “ $/B$ ” is put in the complementary state R_P or R_{AP} .

It is noteworthy that there is neither capacitance for the data sensing and nor magnetic field for data programming in this new structure beyond the previous structures [18, 19]. Therefore, this design allows efficient area minimization and is suitable for advanced fabrication nodes below 65 nm.

Figure 26 illustrates the transient simulation of 1-bit STT-MFA shown in Fig. 25. It is performed by using PMA STT MTJ compact models introduced above and CMOS 40 nm design kit. The time-dependent behaviors of outputs (“ SUM ” and “ C_o ”) confirm the logic functionality of full addition. For instance, for the operation “ A ” = “1”, “ B ” = “0”, “ C_i ” = “0”, the result is “1” and no carry yields; for the operation “ A ” = “1”, “ B ” = “0”, “ C_i ” = “1”, the result is “0” and the carry is “1”.

Figure 27 emphasizes one sensing operation of this STT-MFA and shows the analog behaviors. It confirms the pre-charge, evaluation and amplification process described previously. Moreover, we find that the sensing delay of “ $Output_{C_o}$ ” (~127 ps) is shorter than that of “ $Output_{SUM}$ ” (~147 ps). This is due to the higher resistance of the branch associated with “ $Output_{SUM}$ ”, leading to lower current and slower amplification.

The delay time and dynamic energy are generally two crucial parameters to evaluate the performance of computation system. We have studied the effects of three possible factors: the size of discharge transistor (MN_{PD} in Fig. 25), PMA STT MTJ resistance-area product (RA) and TMR ratio. Figure 28 demonstrates the performance dependence of this STT-MFA in terms of delay time and dynamic power on the size of discharge transistor. We can find a tradeoff between the speed and power performance by varying the die area. A larger discharge transistor can drive a higher sensing current and faster amplification of PCSA circuit, but cost more energy.

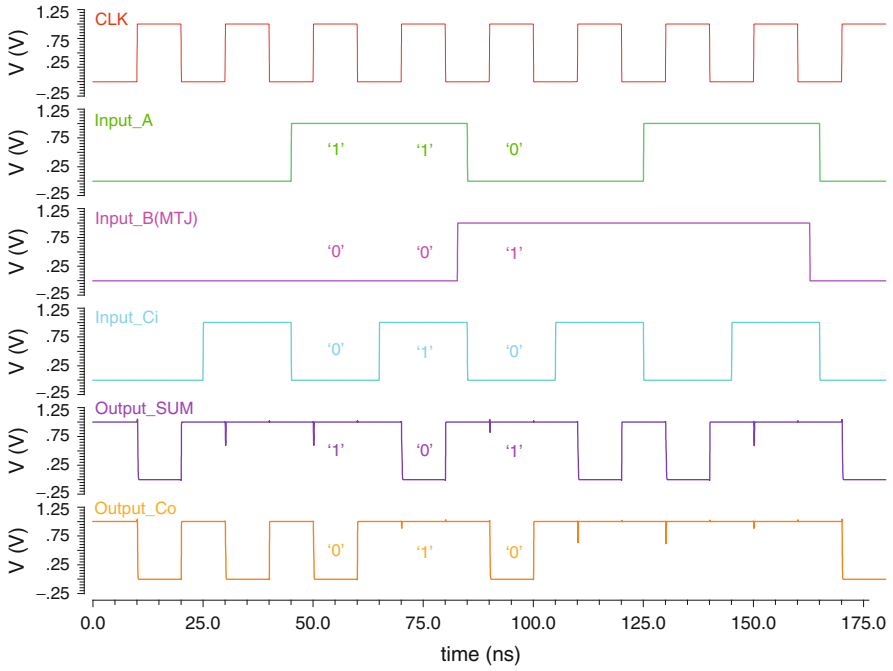


Fig. 26 Transient simulations of 1-bit STT-MFA in 40 nm node

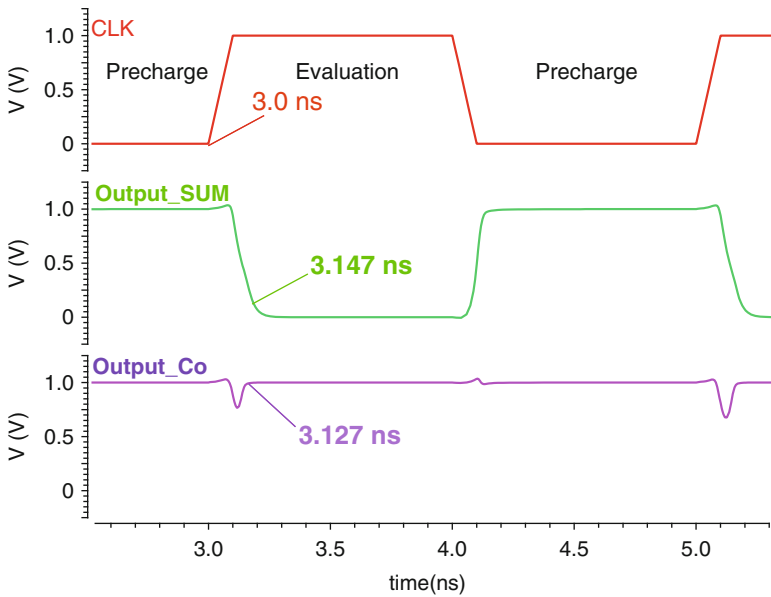


Fig. 27 One sensing operation of the PCSA based STT-MFA: outputs are pre-charged as CLK is set to “0” and are evaluated as CLK is set to “1”

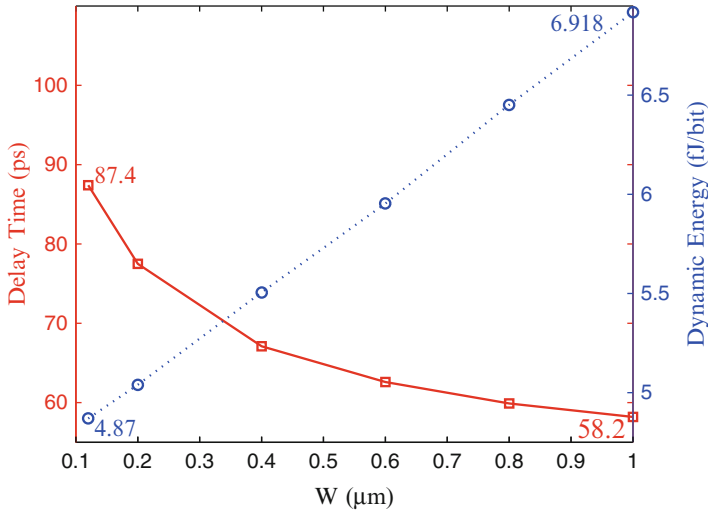


Fig. 28 Dependence of delay time (*red solid line*) and dynamic energy (*blue dotted line*) on the width of discharge transistor for STT-MFA

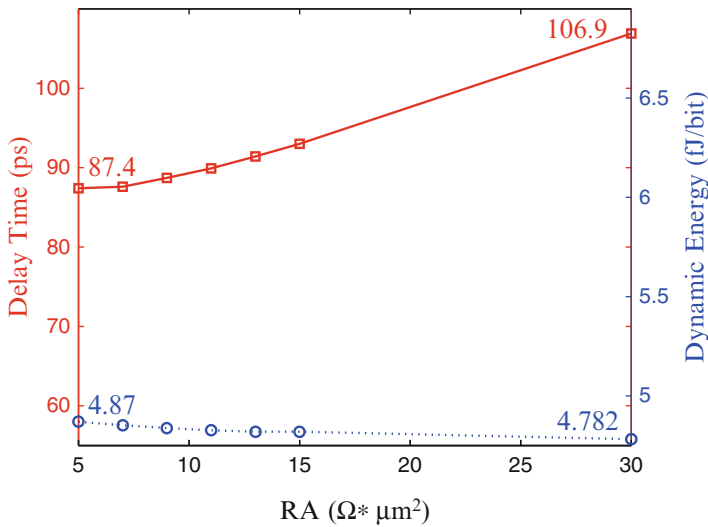


Fig. 29 Dependence of delay time (*red solid line*) and dynamic energy (*blue dotted line*) on the resistance-area product (*RA*) of PMA STT MTJ

Figure 29 shows the *RA* dependence for this STT-MFA. By decreasing *RA*, the delay time becomes shorter while keeping a relatively steady dynamic power performance. This confirms that the speed advantage of using low *RA*.

We also investigate the dependence between TMR ratio of PMA STT MTJ and STT-MFA performance. Figure 30 shows that faster speed is possible by increasing the TMR ratio while the dynamic energy changes slightly.

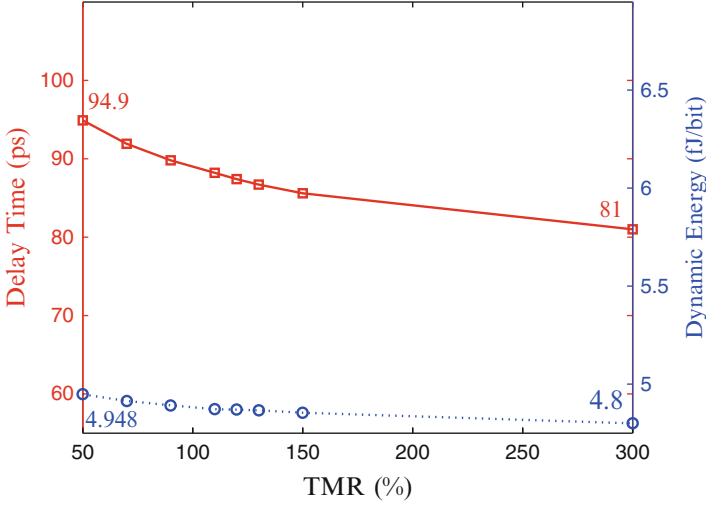


Fig. 30 Dependence of delay time (red solid line) and dynamic energy (blue dotted line) on PMA STT MTJ TMR ratio for 1-bit STT-MFA

Table 3 Comparison of 1-bit STT-MFA with CMOS only full adder

Performance	CMOS full adder (40 nm)	STT-MFA
Delay time	75 ps	87.4 ps
Dynamic power @500 MHz	2.17 μ W	1.98 μ W
Standby power	71 nW	<1 nW [31]
Data transfer energy	>1 pJ/bit	<1 fJ/bit
Die area	46 MOS	38 MOS + 4 MTJs

According to the above analyses, a PMA STT MTJ with lower RA and higher TMR ratio is expected to perform fast computation while keeping nearly the same dynamic energy. In the recent experimental demonstration of the MTJ, a low RA (e.g., $5 \Omega\mu\text{m}^2$) and high TMR ratio (e.g., 200 %) can be achieved in PMA STT MTJ [51, 76, 77].

We compare the STT-MFA with conventional CMOS only full adder in terms of delay time, dynamic power, standby power, data transfer energy and die area (see Table 3). The CMOS-only full adder is taken from the standard cell library of STMicroelectronics 40 nm design kit. Two full latches are added to synchronize the outputs with clock signal.

In conventional computing architectures, logic and memory are completely separated [78]. In order to perform a logic operation, both the instruction and data need to be read from memory units (i.e., cache and main memories), and then moved to logic unit. The results are transferred back to the memory units after the computing. In the STT-MFA circuit based on “logic-in-memory” architecture, logic operations are processed directly with the magnetic data stored in MTJs and

the addition result is written to other MTJs for the next operations. Long latency and high dynamic power due to data moving can be significantly economized. For example, the data transfer energy (~ 1 pJ/mm/bit @22 nm [1]) becomes much lower thanks to the shorter distance between memory and computing unit, which is about some μm or below in STT-MFA instead of some mm for CMOS only logic circuit.

Furthermore, thanks to the 3D integration of STT-MRAM, the die area of this design (38 MOS + 4 MTJs) is advantageous compared to those of the CMOS full-adder (46 MOS). However, its energy-delay product (EDP) exceeds that of a CMOS full-adder by approximately 10 % since it takes more time for PCSA amplification process. Due to the non-volatility of PMA STT MTJ, the new chip can be powered off completely and this allows the standby power to be reduced significantly down to 0.75 nW [77]. Thereby, the STT-MFA can greatly reduce the consumption in a full computing system, especially for those normally in OFF state.

Another critical idea of this design is to use a programming frequency (e.g., 1 kHz) of STT-MRAM much lower than the computing frequency. Thereby, the switching power for non-volatile storage becomes insignificant to other power consumption in a full system. We can continue to reduce it by shortening the non-volatile data retention (e.g., 1 day). Moreover, the programming energy for the non-volatile data (bit “B” in Fig. 25) can be reduced, following the area minimization [79] and new material development for MTJs (e.g., ~ 0.1 pJ/bit).

3.4 Multi-Bit MFA Based on Racetrack Memory

PMA racetrack memory is distinguished as it can store and shift multiple bits of data through CIDW motion along a magnetic nanowire. This advantageous feature makes it possible to design a high speed and compact multi-bit serial MFA.

Figure 31 shows the detailed schematic of CARRY circuit of the multi-bit MFA based on PMA racetrack memory including MTJ writing circuit [75]. “A” and “B” are multi-bit input data stored in different nanowires. Each data is designed to be stored in dual magnetic nanowires with exactly opposite configuration to minimize the variation between two complementary data (e.g., “A” and “ \bar{A} ”) as the same I_{shift} is used in the dual nanowires to move the DWs [3]. At each rising edge of CLK, “ C_o ” and “ \bar{C}_o ” are evaluated through the PCSA circuit and become inputs of a writing circuit, which generates writing current I_{write} to reverse or just conserve the state of nucleation MTJs (“ $C_{nucleation}$ ” and “ $\bar{C}_{nucleation}$ ”). At each falling edge of CLK, propagating current I_{shift} induces the DW motion of all magnetic nanowires (“A”, “ \bar{A} ”, “B”, “ \bar{B} ”, “C”, “ \bar{C} ”) simultaneously, moves next magnetic domains under the read MTJ for next adding operation.

The operation of SUM circuit is similar to that of CARRY: the SUM output and its complement are evaluated through the SUM PCSA circuit and become inputs of SUM writing circuit, which generates the writing current to write these values into

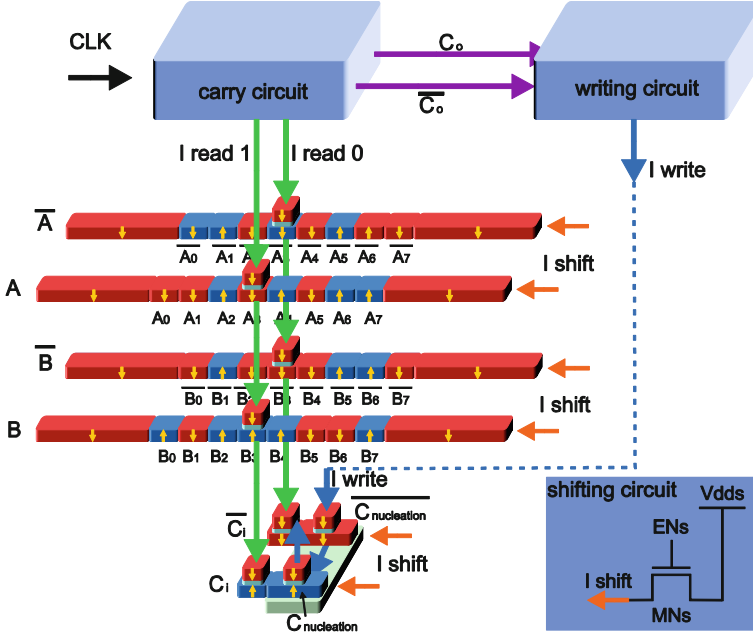


Fig. 31 CARRY circuit of multi-bit MFA based on PMA racetrack memory including MTJ writing circuit

the most significant bits of SUM nanowire. These bits are then shifted in the way that at the end of all additions, the sooner the bit is calculated, the greater its weight is.

The transient simulation of this multi-bit MFA shows the addition operation of two random 8-bit words: “A” = “01110011” (Fig. 32c) and “B” = “01011010” (Fig. 32d). “CLK” (Fig. 32a) drives PCSA circuit and “ I_{shift} ” (Fig. 32b) induces DW motion in the magnetic nanowire. The outputs “SUM” (Fig. 32e) and “ C_o ” (Fig. 32f) are firstly pre-charged to logic “1” when “CLK” = “0” and are evaluated when “CLK” is set to “1”. The DW motions are implemented in the pre-charge phase in order to avoid the disturbance to the output evaluation. The serial addition is performed from the least significant bit and the simulation result “SUM” = “11001101” and “ C_o ” = “01110010” confirms the correct operation of MFA.

Between two addition evaluations, there is a data transition process to achieve multiple bits operation. Figure 33 demonstrates the CARRY transition including DW nucleation and motion. Carry-out “ C_o ” (Fig. 33b) is firstly pre-charged to “Vdd” before the time “M0”, after the rising edge of “CLK” (Fig. 33a), “ C_o ” is evaluated by the PCSA and becomes the input signal of writing circuit (Fig. 33b). I_{write} is generated to nucleate DW in the magnetic nanowire (Fig. 33c, d). I_{shift} is in the following activated to propagate the DW and replace the value of carry-in “ C_i ” with “ C_o ” at the time “M3” for next cycle of addition (Fig. 33e).

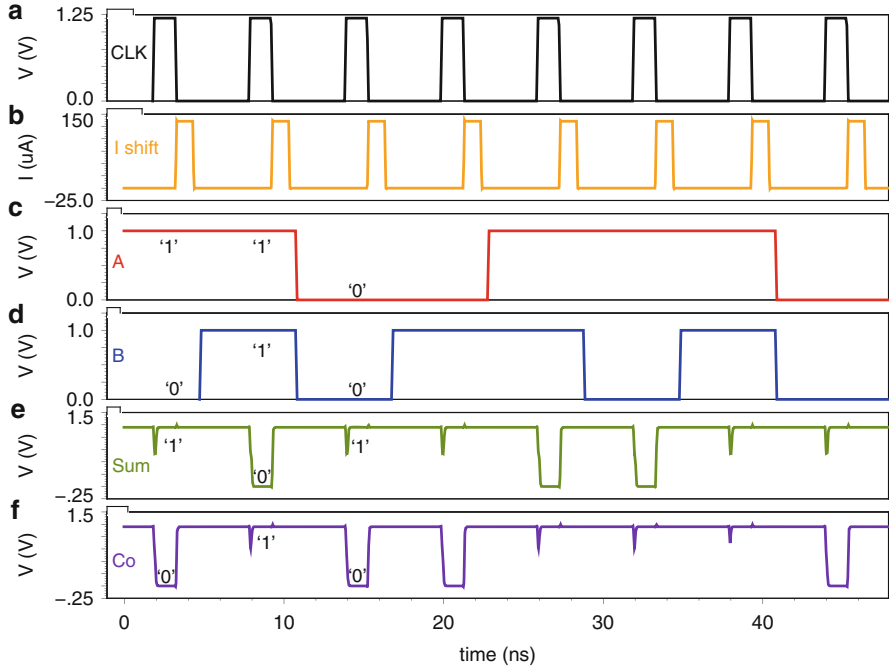


Fig. 32 Transient simulation of the multi-bit MFA. (a) CLK signal (b) Data shifting current pulse I_{shift} (c) Input data “A” (d) Input data “B” (e) “SUM” (f) “ C_o ”

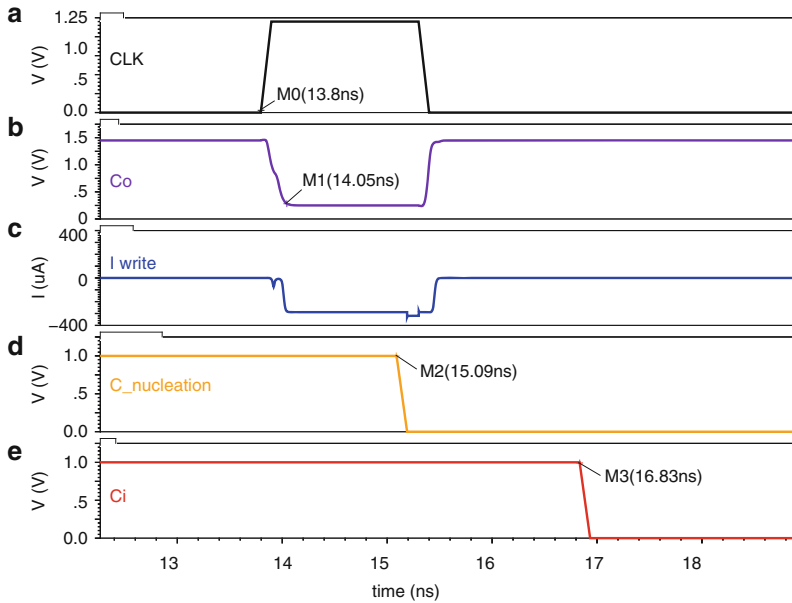


Fig. 33 Transient simulation of CARRY storage (a) CLK signal (b) “ C_o ” (c) DW nucleation current I_{write} (d) State of DW nucleation MTJ (e) Carry in for next adding operation

We then analyze the performance of this multi-bit MFA in terms of delay and power dissipation. Indeed, several parameters, such as the size of transistors and “Vdd”, can affect greatly on them. A first look at the repartition of consumed energy in this MFA shows that the energy for nucleation and propagation is of the same order of magnitude, and higher than that of data sensing or logic computing.

In order to propagate the magnetic domains to their next positions (shift 1 bit), we must supply a current I_{shift} in a period t_{pulse} . The period t_{pulse} is the necessary time for all magnetic domains to move from their current positions to their next positions. It corresponds to the propagation delay, which is inversely proportional to I_{shift} . Consequently, the propagation energy does not vary much (seeing that this energy is the integral of the product $V_{pulse} \times I_{shift} \times t_{pulse}$ and the power supply voltage V_{pulse} is kept invariable). Simulations show that energy needed for shifting all racetrack memories 8 bits is about 29 pJ.

Since the energy needed for propagation is almost invariable, we can reduce the propagation delay by increasing I_{shift} . Normally, one transistor based current source is used to generate the DW propagation current, thus the size of transistor determines the generated current in propagation circuit. Figure 34 shows the tradeoff dependence of propagation delay on the width of transistor MN2: the reduction of propagation delay at the cost of satisfying the area.

The writing circuit nucleates domain walls under the MTJ write head (e.g., MTJ0 in Fig. 7) by passing through a bi-directional current I_{write} . I_{write} is proportional to both supply voltage V_{write} and transistors’ size. V_{write} will be set as high as possible in order to minimize the size of transistors (MN0-1 and MP0-1) while keeping the switching current at fixed value. In this setup, V_{write} is set to 2 V to avoid the breakdown of oxide barrier at 65 nm technology node. A higher I_{write} can reduce the switching delay, but increase the power consumption.

A study of the tradeoff among the width of transistors, switching speed and power dissipation have been made to find out optimal operation point (see Fig. 35). In this analysis, the width of transistors W is started at 0.35 μm because I_{write} is not

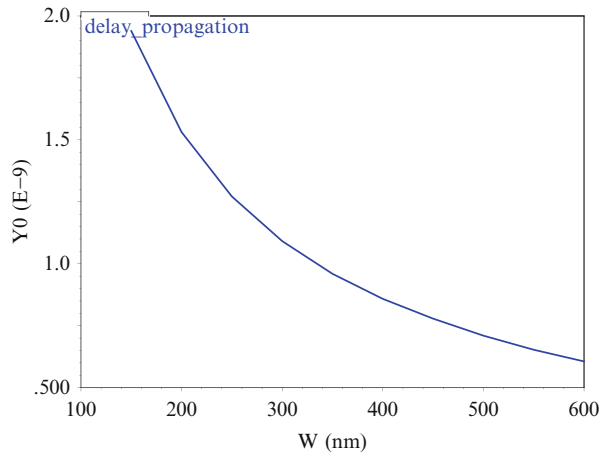


Fig. 34 Dependence of domain wall propagation delay on the transistor width of propagation circuit

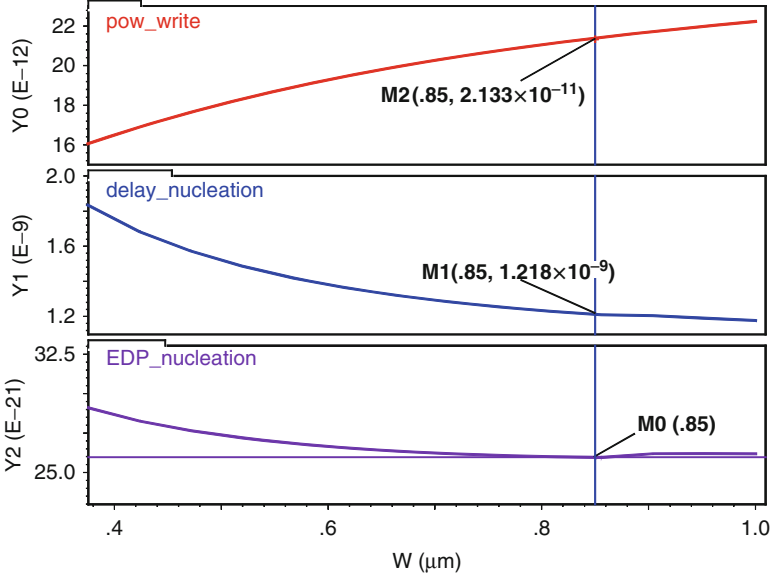


Fig. 35 Dependence of energy consumption (*pow_write*), switching speed (*delay_nucleation*) and the energy-delay product (*EDP_nucleation*) versus the width of four transistors using in writing circuit

Table 4 Comparison of 8-bit MFA based on racetrack memory with CMOS only full adder included transferring and writing data

Performance	CMOS full adder (65 nm)	MFA
Write time	200 ps	2 ns
Write energy	16 fJ/8 bits	(21.39 + 29) pJ/8 bits
Transfer time	~ns	0
Transfer energy	8 pJ/mm (for 8 bits)	~0
Die area	310 MOS	23 MOS + 18 MTJs

high enough to switch the state of MTJ below this value. This curve shows that one can increase W to reduce the factor of merit EDP until the point “M0” ($W = 0.85 \mu\text{m}$) and then it slightly goes up. Thereby the operating points should be chosen around the optimum (e.g., $W = 0.75\text{--}1.0 \mu\text{m}$) to address different applications. For instance, the two markers “M1” and “M2” show that when $W = 0.85 \mu\text{m}$, the switching power and latency are 21.33 pJ and 1.22 ns, respectively. The switching current I_{write} equals to 291 μA in this case.

In order to understand the advantages and disadvantages of this multi-bit MFA based on PMA racetrack memory, we compare its performance with that of a CMOS-only series adder (see Table 4), which uses a full adder taken from the library of STMicroelectronics 65 nm design kit.

For the comparison with CMOS only multiple bits full adder regarding writing and transferring data, we see that the chip area of the MFA based on PMA racetrack memory is significantly reduced. The 8-bit MFA uses only 23 MOS transistors, 18 MTJs and 8 magnetic nanowires instead of 22 MOS plus 8×3 Flip-Flops (310 MOS transistors totally) for an 8-bit series CMOS full adder. Although the number of transistors decrease 13 times, the area reduction is about 4.5 times since the writing circuit and propagating circuit requires the transistors with 6.3 times and 3.3 times minimum width ($0.135 \mu\text{m}@65 \text{ nm}$ technology node), respectively. The total delay of one operation of the new MFA is $\sim 2.1 \text{ ns}$, composed of DW nucleation ($\sim 1.2 \text{ ns}$), motion ($\sim 0.7 \text{ ns}$) and detection ($\sim 180 \text{ ps}$). It can be thus driven by a CLK frequency up to 470 MHz, which can be further increased with the feature size shrinking. This latency is of the same order with that of CMOS circuit (read time + transfer time + operate time + transfer time + write time). On the contrary, the MFA consume six times dynamic energy more than the CMOS only full adder since energy needed for nucleation and propagation is still too large with current technology. However, we have not yet addressed the static energy in this comparison. Regarding that power must be supplied in order to maintain stored data in CMOS-only storage circuit, the MFA does not require energy to conserve information thanks to its total non-volatility. This allows the circuit to be turned off safely in “idle” mode without data backup. All the operations can be retrieved instantly after power-on. This instant on/off capability promises to overcome completely the rising standby power issue due to leakage currents and could be very useful for normally-off systems [77].

It is important to note that for this non-volatile MFA, operations are performed directly with the data (“A”, “ \bar{A} ”, “B”, “ \bar{B} ”) stored in magnetic nanowires, which plays the role of shift registers. We do not take into account the writing circuits of “A” and “B” to keep the same comparison condition as the writing circuits of data are considered in the CMOS shift register part, not in the adder. The number of writing circuit is then reduced to two for respectively SUM and CARRY circuits, which are shared by the eight bits.

3.5 *Content Addressable Memory (CAM) Based on Racetrack Memory*

CAM is a computer memory that can output the address of search data. It compares search data with stored data and returns the match location with its high-speed fully-parallel manner. Therefore it is widely used in mobile, internet routers and processors to provide fast data access and ultra-high density [21]. The mainstream CAMs are composed of large-capacity volatile SRAM blocks (see Fig. 36a), which lead to high static power and large die area [80]. These become the key challenges for the future R&D of CAM. Replacing volatile memories by non-volatile memories or applying hybrid non-volatile logic-in-memory circuits is a promising

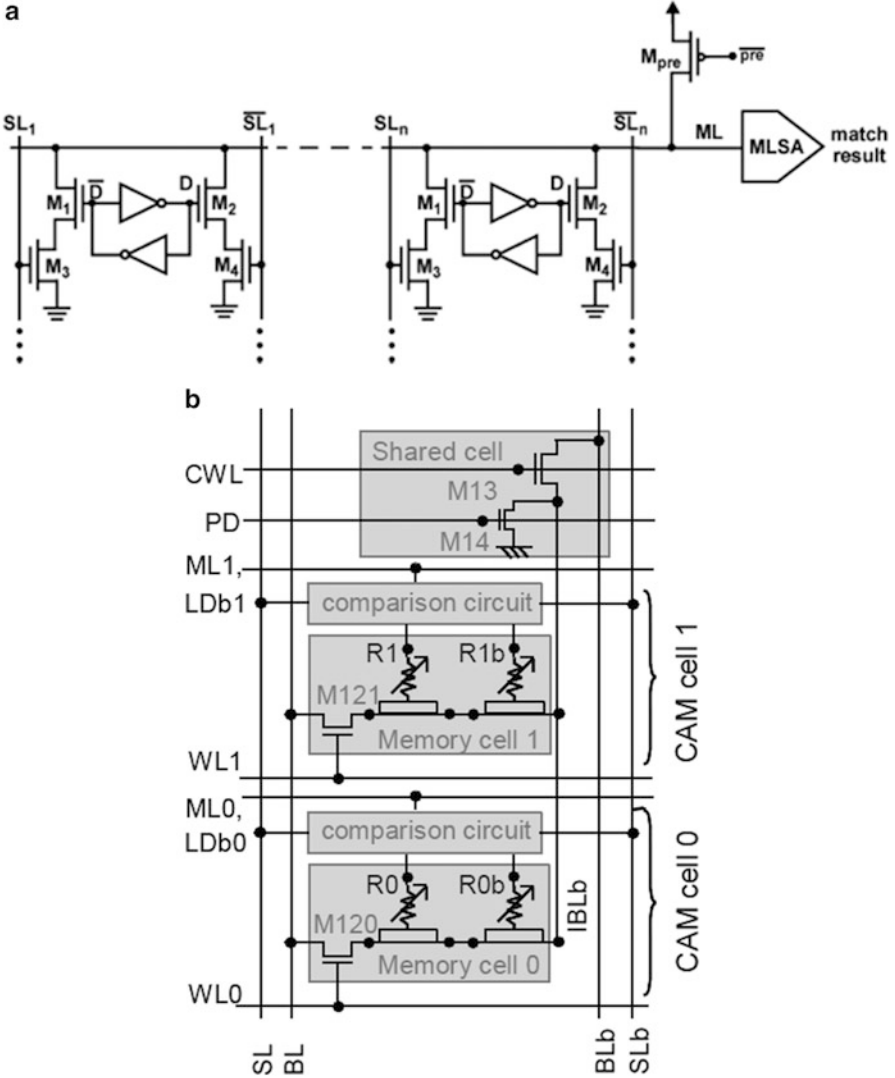


Fig. 36 Conventional CAM. (a) SRAM based CAM. (b) DW-CAM [21, 22]

solution to build non-volatile CAM and overcome both these drawbacks. This topic is currently under intense investigation. For instance, a DW motion MRAM based CAM (DW-CAM) was prototyped recently (see Fig. 36b), which demonstrated important progress in terms of power and density [22]. However, this DW-CAM used a three-terminal MTJ as storage element and every memory cell had one comparison circuit and one selected transistor, which lead to a high bit-cell cost and cannot allow the expected ultra-high density.

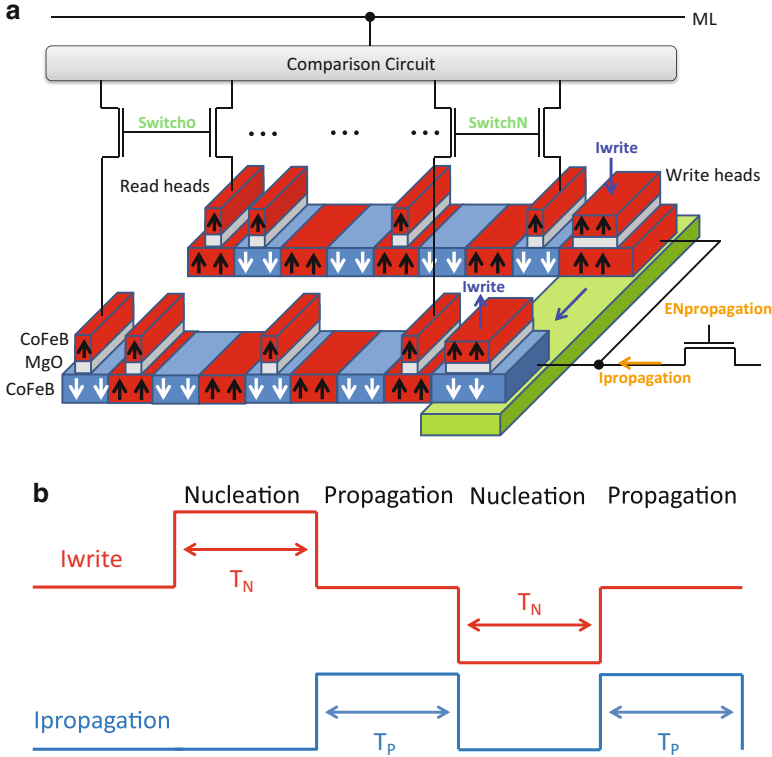


Fig. 37 (a) Structure of dual nanowires based RM-CAM. One writing current pulse nucleates a couple of MTJs with complementary configurations. A propagation current pulse drives the dual nanowires synchronously. Every dual wires share a comparison circuit. (b) One example of current pulse configuration for I_{write} and $I_{propagation}$. T_N and T_P are respectively their pulse durations

In this situation, a design of CAM based on complementary dual PMA racetrack memories (RM-CAM) was proposed [81]. Its non-volatile feature can reduce the static power due to leakage currents. The CMOS based DW nucleation and sensing circuits are globally shared to scale down the cell area. The complementary dual nanowires structure allows the local sensing and fast data search operation.

The RM-CAM is composed of comparison circuits, PMA racetrack memories and DW nucleation/propagation circuits. A couple of complementary magnetic nanowires are used to present one word (see Fig. 37a) in order to obtain the most reliable and fast access operation for CAM applications as this solution benefits the maximum TMR value instead of $TMR/2$ for conventional single nanowire structures. We design the comparison circuit based on PCSA, which allows minimum power and sensing errors. This RM-CAM includes a couple of PMA MTJs connected together as the write heads. Due to the different directions of the writing current pulse I_{write} through these two MTJs, they can nucleate the complementary configurations through STT switching mechanism under the same I_{write} pulse.

One of the critical challenges for complementary magnetic nanowires is to synchronize precisely the domain wall positions. Here, the same current pulse $I_{propagation}$ propagates domains in the dual nanowires and we implement the DW pinning constrictions with the same distance in the magnetic nanowires [82]. To avoid the interference between the DW nucleation and the previous data, write heads do not hold the data storage and there is always a $I_{propagation}$ pulse following each DW nucleation (see Fig. 37b). There are also a couple of PMA MTJs at each bit of storage elements as read heads. Since lower resistance can reduce the rate of breakdown and higher resistance can improve the sensing performance, the size of the read heads should therefore be smaller than that of the write heads to obtain the best switching and sensing reliability.

The comparison circuit (see Fig. 38) consists of two parts: a PCSA detects the complementary magnetizations of the read heads by two reading current pulses (I_{read} and I_{readb}) and outputs a logic value; the transistors MN3-MN6 build a classical NOR-type CAM. The signal “MLpre” is used to pre-charge the match line (ML). In case that the search line “SL” (“SLb” is its complementary signal) matches the stored data, there is no path to discharge and ML will thus be asserted. In contrast, ML will be discharged.

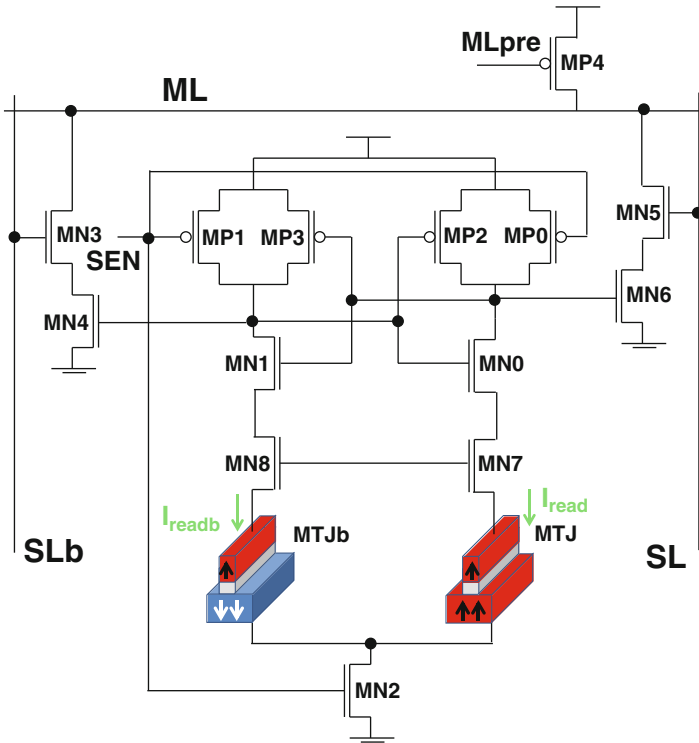


Fig. 38 Schematic of the comparison circuit. It outputs the logic value “1” or “0” according to the configuration of complementary MTJs. MN3-6 transistors build up a NOR-type CAM

The fast search operation as shown in [22] can be also expected in the RM-CAM. At first, we program the magnetic nanowires, and the switch signals then select each bit of magnetic nanowires to be loaded in the comparison circuit. By sequentially triggering the switch signals, all the words can be explored. If there is no match case, DW nucleation and propagation will be carried out to enter new words for the next search. The programming speed of magnetic nanowires depends on T_N and T_P , which are respectively the pulse durations of I_{write} and $I_{propagation}$. They can be both sped up to ~ 1 ns. According to the current pulse configuration shown in Fig. 37b, the worst case of programming duration is $N \times (T_N + T_P)$, where N is the number of pinning potentials in the magnetic nanowire. We can benefit a higher speed for the repeated bits such as “111” and “000” when only one DW nucleation is required for three bits.

In order to improve the area efficiency, every couple of dual nanowires shares the comparison circuit in this RM-CAM (see Fig. 37a). Unlike the DW-CAM where there is a large transistor for nucleation for every storage cell, the same write head is shared for one magnetic nanowire in RM-CAM, and the CMOS area dedicated for each storage cell becomes ignored for a long track with numerous pinning constrictions. This structure thus allows an ultra-high density.

An 8-bits-width-8-words-depth PMA RM-CAM shown in Fig. 39 has been designed. Firstly, we implement the transient simulation for the search operation without DW propagation (see Fig. 40a). The clock signal “CLK” involves the

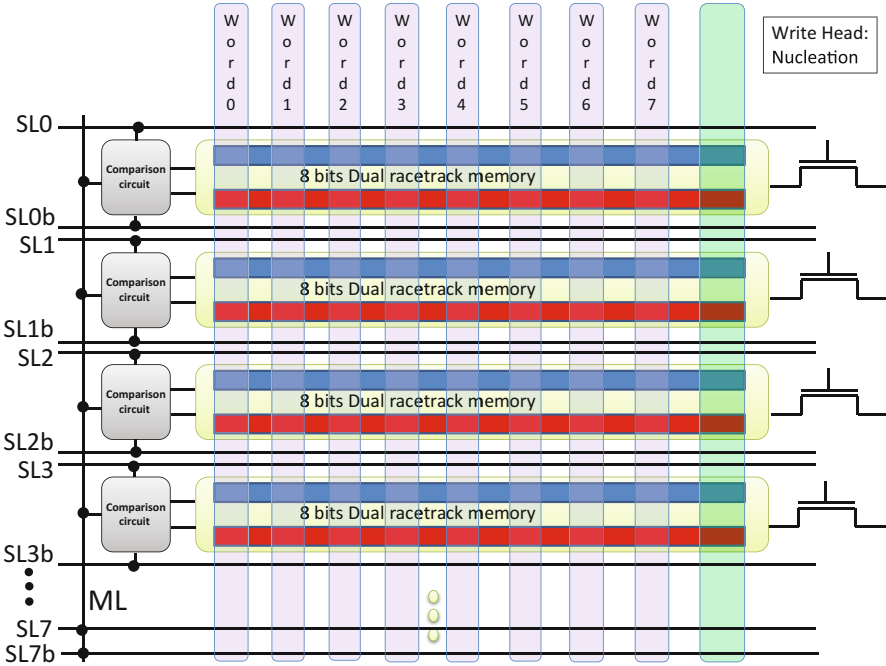


Fig. 39 Schematic of an 8 × 8 bits RM-CAM. Each word is composed of the bits at the same positions in 8 different dual nanowires; they can be driven to move simultaneously by the propagation currents

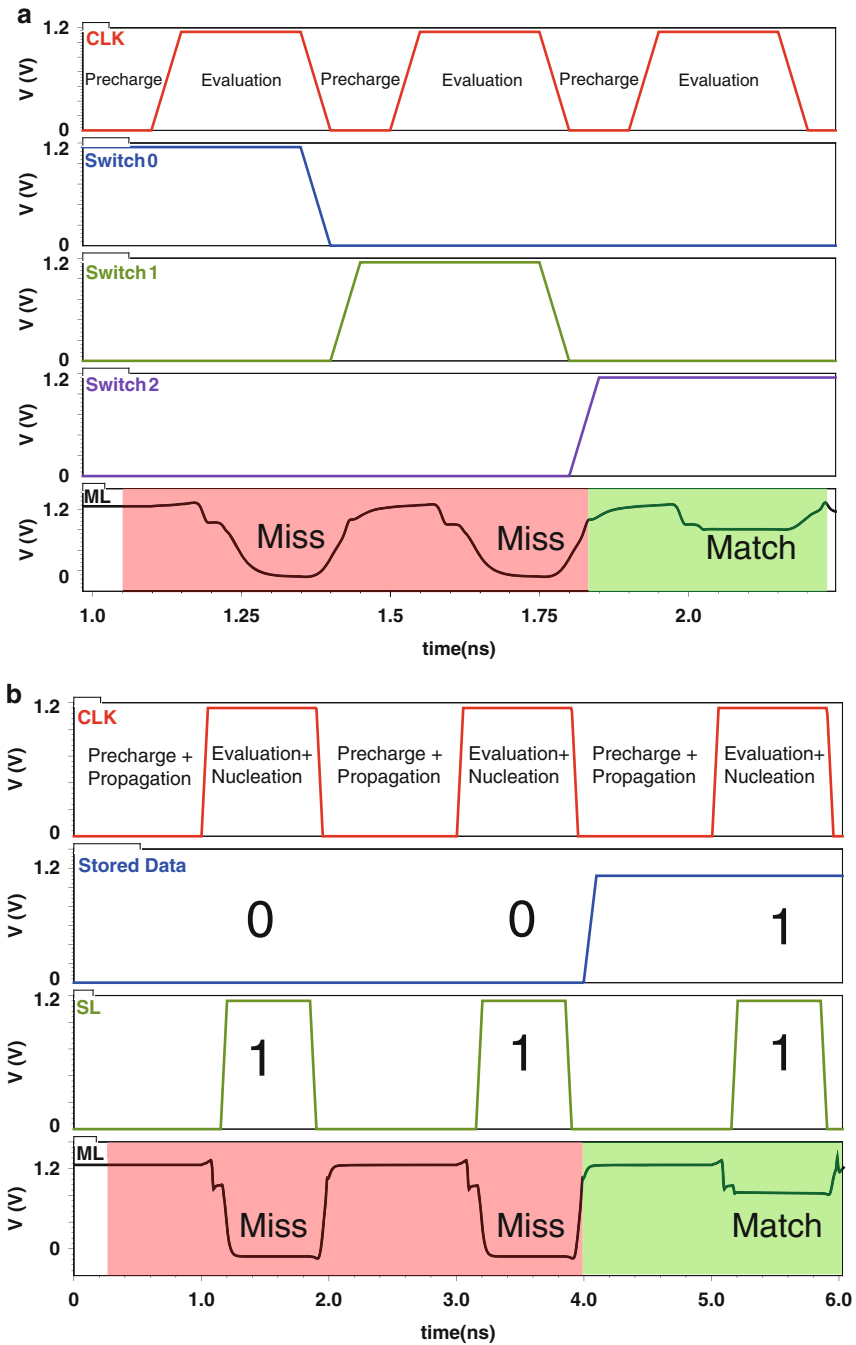


Fig. 40 Transient simulations of the RM-CAM: (a) Without DW nucleation and propagation. (b) With DW nucleation and propagation

“Pre-charge” phase and the “Evaluation” phase. During the “Pre-charge” phase, both of the signals “SEN” and “MLpre” (see Fig. 38) are set low to pre-charge the PCSA circuit and the match line “ML”. The first word “Word0” has been loaded by enabling the signal “Switch0”. With the response of the signal “Miss”, “Switch1” will be then activated and so on. This process doesn’t stop until the appearance of the match case. We find that this search operation needs only ~ 0.45 ns, which is faster than that of conventional SRAM-based CAM and DW-CAM. In addition, the energy consumption of searching is as low as ~ 12 fJ/bit/search, which can be further reduced by the decrease of activity rate thanks to the segmentation of the match line [83].

In case that no storage data can match the search word, a new word will be nucleated and propagated into the magnetic nanowire for the next round of search. Figure 40b shows the transient simulation result of the worst case: 1-bit miss process. It means that the rest 7 bits of the search word match the stored data, only one bit is different from the stored data. As shown in Fig. 40b, the search bit is “1”, if no match is found, the propagation current pulse will start to drive the DW propagation, until “SL” and “Stored data” match each other. We can find the whole operation, consisting of “Pre-charge”, “Propagation” and “Evaluation” phases, only requires ~ 2 ns. This suggests a high operating frequency up to 500 MHz, comparable to that of traditional CAM [84].

We estimate the cell area for RM-CAM with Eq. (38):

$$A_C = \frac{A_{CO} + A_{NU} + A_{PR} + N \times \text{MAX}(A_{BT} + A_{LS})}{N} \quad (38)$$

where A_{CO} denotes the area of a comparison circuit, which is ~ 50 F², A_{NU} denotes the area of a DW nucleation circuit, which is ~ 48 F², A_{PR} denotes the area of a propagation current generating circuit, which is ~ 7 F², A_{BT} is the area of every bit in racetrack memory, A_{LS} is the area of two load selecting transistors for every bit and N is the number of bits per word.

Due to the 3D integration of MTJs above CMOS circuit, only the larger one between the MTJs’ area and the selecting transistors’ area will be involved for calculating the full area. For this design, A_{BT} is ~ 6 F² considering 2 F between two adjacent constrictions. Coincidentally, A_{LS} is also ~ 6 F² with the minimum size. If the distance between two adjacent constrictions exceeds 2 F, only A_{BT} would be taken into account in Eq. (38). As $N=8$, the cell area per bit is therefore ~ 19 F², which is much lower than that of SRAM-based CAM or DW-CAM [21, 22]. Meanwhile, with the increase of the bit number per word, the area of shared CMOS circuits for data comparison, DW nucleation and motion would become negligible (see Fig. 41). The cell area per bit will approach to $\text{MAX}(A_{BT} + A_{LS})$ (e.g., ~ 6 F² for this design).

With the performance analyses above, the comparison of CAMs based on different technologies is summarized. From Table 5, we can find that non-volatility of racetrack memory allows RM-CAM to eliminate the static power. DW propagation in the racetrack memory benefits for improving the search speed. Most importantly,

Fig. 41 Dependence of full area versus number of bits per word

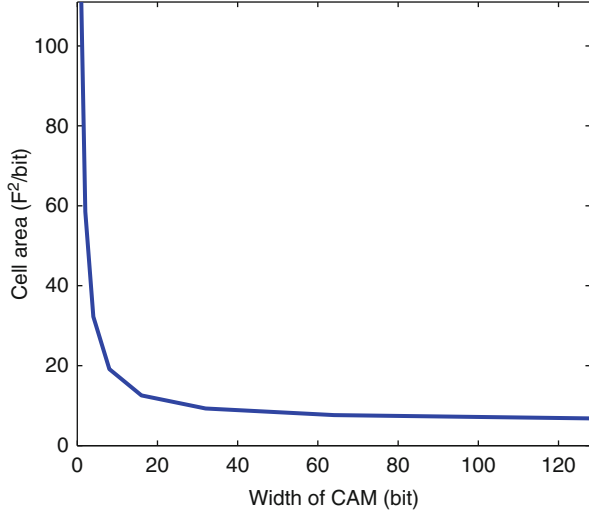


Table 5 Comparison of CAMs based on different technologies

Type	SRAM based CAM [21]	DW-CAM [22]	RM-CAM
Cell area (F ² /bit)	540	~815	~19
Cycle time (ns)	2	5	~0.45
Energy (fJ/bit/search)	9.5	~30	~12
Static power	Yes	No	No

RM-CAM shows a great advantage in terms of density thanks to 3D integration and sharing of CMOS circuits (comparison circuit, DW propagation circuit and DW nucleation circuit). However, we have to mention that the cell area of DW-CAM shown in Table 5 does not consider the 3D integration. If it applies 3D integration, its cell area could reach N times that of RM-CAM. For example, if $N = 8$ for RM-CAM, the cell area of DW-CAM would be about 160 F²/bit.

4 Conclusions and Perspectives

This chapter focused on two current-induced magnetic switching technologies for high-performance computing: the PMA STT MTJ and the racetrack memory. The work covers from theoretical study to hybrid circuit design and performance analyses. Through this work, the integration functionality of the current-induced spintronic devices based on PMA materials has been proven. The performance analyses of related hybrid logic circuits demonstrated that spintronic devices could provide various advantages compared with conventional systems, such as scalability, low switching current and high operation speed. Thanks to these, the application

potential of PMA spintronic devices to achieve future low-power high-density high-speed electronic systems can be confirmed.

In details, non-volatility allows the hybrid systems to be powered off while saving the data, and then to eliminate the static power consumption. This feature can reduce greatly the overall power consumption, especially for normally-off systems. 3D integration technology can improve the system's density efficiency. Moreover, it can shorten the distance between logic and memory, which helps to save considerably the transfer energy and time. Although the switching speed of MTJ doesn't show an evident advantage compared to conventional CMOS, it is still sufficient for logic and memory application. To overcome this challenge, using of the CIDW motion is an alternative solution. Considering that the distance between two adjacent DW is 40 nm and the propagation speed can be as high as 100 m/s, switching a state by propagating DW can be as fast as 400 ps. That is why we believe the CIDW motion based racetrack memory design has a great potential for the future high-speed low-power systems.

The emergence of spintronics is to achieve more efficient and reliable applications, which could overcome the issues of mainstream charge-based electronics. The term "efficient" here concerns many factors, which involve power, density and frequency, etc. This aim is the "beacon", which indicates the direction of the progress of spintronics. On this route, the innovative technologies are appearing ceaselessly, and an emerging mechanism would be replaced by a more emerging alternative.

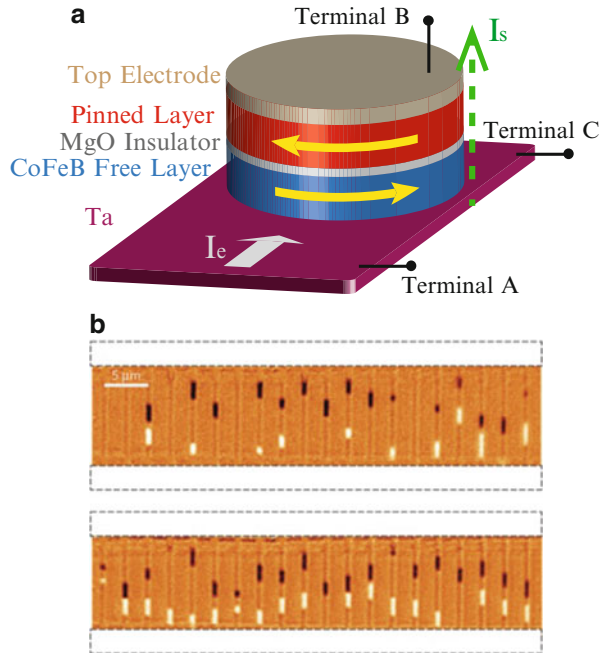
Along with the downscaling pace of MTJs beyond sub-volume limit (~ 40 nm), MTJ displays a relatively high thermal stability factor and low STT critical current. This so-called "high spin torque efficiency" is a strong stimulus for high density MTJ application. Spin torque efficiency is defined here as [85]:

$$\kappa = \frac{E}{I_{c0}} \quad (39)$$

where E is barrier height (or thermal stability factor), and I_{c0} is the average critical current. Spin torque efficiency reflects the capability of spin polarized current to reverse the barrier height. Practically, when the lateral size of MTJ scaling down to the sub-volume limit, sub-volume activation effects make the leading term guiding the magnetization switch in devices, which is negligible when lateral diameter is larger than the limit. Thanks to this effect, the scaling gain (faster operation, higher density and improved spin torque efficiency) can be further continued, which benefits greatly for the miniaturization of MTJ. As a result, the high spin torque efficiency should be considered in the future work involving the small-size MTJs and hybrid circuits.

Beyond STT, spin orbit torque (SOT) is demonstrated to be able to switch magnetization and nucleate DWs. Two main effects referred to SOT have been observed: spin hall effect (SHE) [86, 87] and Rashba effect [88] (see Fig. 42). Compared with the STT switching mechanism, these effects are exhibited with assets in terms of power, speed and reliability. For example, three-terminal devices

Fig. 42 (a) Three-terminal device based on the giant spin Hall effect in β -Ta/CoFeB. (b) Kerr micrographs showing the DW motion induced by Rashba effect [87, 88]



based on SHE can yield a more efficient spin torque which means to require a lower current [89, 90]. As a consequence, the power consumption can be further reduced. In addition, the current is not applied through the tunnel junction, which avoids the damage caused by the high current density. Furthermore, the separation of writing path and reading path can improve greatly the reliability performance.

References

1. International roadmap for semiconductor (ITRS) (2010)
2. C. Chappert, A. Fert, F. Nguyen Van Dau, The emergence of spin electronics in data storage. *Nat. Mater.* **6**, 813–823 (2007)
3. J.C. Slonczewski, Current-driven excitation of magnetic multilayers. *J. Magn. Magn. Mater.* **159**, L1–L7 (1996)
4. L. Berger, Emission of spin waves by a magnetic multilayer traversed by a current. *Phys. Rev. B* **54**, 9353–9358 (1996)
5. M. Hayashi, L. Thomas, R. Moriya, C. Rettner, S.S.P. Parkin, Current-controlled magnetic domain-wall nanowire shift register. *Science* **320**, 209–211 (2008)
6. S.S.P. Parkin, M. Hayashi, L. Thomas, Magnetic domain-wall racetrack memory. *Science* **320**, 190–194 (2008)
7. S. Ikeda et al., Recent progress of perpendicular anisotropy magnetic tunnel junctions for nonvolatile VLSI. *SPIN* **02**, 1240003 (2012)
8. S. Ikeda et al., A perpendicular-anisotropy CoFeB–MgO magnetic tunnel junction. *Nat. Mater.* **9**, 721–724 (2010)

9. D.C. Worledge et al., Spin torque switching of perpendicular Ta/CoFeB/MgO-based magnetic tunnel junctions. *Appl. Phys. Lett.* **98**, 022501 (2011)
10. B. Dieny et al., Giant magnetoresistive in soft ferromagnetic multilayers. *Phys. Rev. B* **43**, 1297–1300 (1991)
11. S. Parkin et al., Magnetically engineered spintronic sensors and memory. *Proc. IEEE* **91**, 661–680 (2003)
12. A. Barthélémy et al., Magnetoresistance and spin electronics. *J. Magn. Magn. Mater.* **242–245** (Part 1), 68–76 (2002)
13. Z. Diao et al., Spin-transfer torque switching in magnetic tunnel junctions and spin-transfer torque random access memory. *J. Phys. Condens. Matter* **19**, 165209 (2007)
14. D. Apalkov et al., Spin-transfer torque magnetic random access memory (STT-MRAM). *J. Emerg. Technol. Comput. Syst.* **9**, 13:1–13:35 (2013)
15. Y. Chen et al., A 130 nm 1.2 V/3.3 V 16 kb spin-transfer torque random access memory with nondestructive self-reference sensing scheme. *IEEE J. Solid-State Circuits* **47**, 560–573 (2012)
16. Y. Chen, X. Wang, H. Li, H. Xi, Y. Yan, W. Zhu, Design margin exploration of spin-transfer torque RAM (STT-RAM) in scaled technologies. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **18**, 1724–1734 (2010)
17. W. Zhao, S. Chaudhuri, C. Accoto, J. Klein, C. Chappert, P. Mazoyer, Cross-point architecture for spin-transfer torque magnetic random access memory. *IEEE Trans. Nanotechnol.* **11**, 907–917 (2012)
18. S. Matsunaga et al., Fabrication of a nonvolatile full adder based on logic-in-memory architecture using magnetic tunnel junctions. *Appl. Phys. Exp.* **1**, 091301 (2008)
19. Y. Gang, W. Zhao, J.-O. Klein, C. Chappert, P. Mazoyer, A high-reliability, low-power magnetic full adder. *IEEE Trans. Magn.* **47**, 4611–4616 (2011)
20. Y. Lakys, W. Zhao, J. Klein, C. Chappert, Low power, high reliability magnetic flip-flop. *Electron. Lett.* **46**, 1493–1494 (2010)
21. K. Pagiamtzis, A. Sheikholeslami, Content-addressable memory (CAM) circuits and architectures: a tutorial and survey. *IEEE J. Solid-State Circuits* **41**, 712–727 (2006)
22. R. Nebashi et al., A content addressable memory using magnetic domain wall motion cells. in *2011 Symposium on VLSI Circuits (VLSIC)* (2011), pp. 300–301
23. W. Zhao et al., High performance SOC design using magnetic logic and memory. in *VLSI-SOC: Advanced Research for Systems on Chip* (2012), pp. 10–33.
24. W.J. Gallagher, S.S.P. Parkin, Development of the magnetic tunnel junction MRAM at IBM: from first junctions to a 16-Mb MRAM demonstrator chip. *IBM J. Res. Develop.* **50**, 5–23 (2006)
25. S.A. Wolf et al., Spintronics: a spin-based electronics vision for the future. *Science* **294**, 1488–1495 (2001)
26. I.L. Prejbeanu et al., Thermally assisted MRAM. *J. Phys. Condens. Matter* **19**, 165218 (2007)
27. I.L. Prejbeanu et al., Thermally assisted MRAMs: ultimate scalability and logic functionalities. *J. Phys. D Appl. Phys.* **46**, 074002 (2013)
28. D.C. Ralph, M.D. Stiles, Spin transfer torques. *J. Magn. Magn. Mater* **320**, 1190–1216 (2008)
29. E.B. Myers, D.C. Ralph, J.A. Katine, R.N. Louie, R.A. Buhrman, Current-induced switching of domains in magnetic multilayer devices. *Science* **285**, 867–870 (1999)
30. J.A. Katine, F.J. Albert, R.A. Buhrman, E.B. Myers, D.C. Ralph, Current-driven magnetization reversal and spin-wave excitations in Co/Cu/Co pillars. *Phys. Rev. Lett.* **84**, 3149–3152 (2000)
31. J.Z. Sun, D.C. Ralph, Magnetoresistance and spin-transfer torque in magnetic tunnel junctions. *J. Magn. Magn. Mater.* **320**, 1227–1237 (2008)
32. A. Kalitsov, M. Chshiev, I. Theodonis, N. Kioussis, W.H. Butler, Spin-transfer torque in magnetic tunnel junctions. *Phys. Rev. B* **79**, 174416 (2009)
33. T. Kawahara, K. Ito, R. Takemura, H. Ohno, Spin-transfer torque RAM technology: review and prospect. *Microelectron. Reliab.* **52**, 613–627 (2012)
34. J.Z. Sun, Spin-current interaction with a monodomain magnetic body: a model study. *Phys. Rev. B* **62**, 570–578 (2000)

35. J. Xiao, A. Zangwill, M.D. Stiles, Macrospin models of spin transfer dynamics. *Phys. Rev. B* **72**, 014446 (2005)
36. G.D. Fuchs et al., Adjustable spin torque in magnetic tunnel junctions with two fixed layers. *Appl. Phys. Lett.* **86**, 152509 (2005)
37. J.C. Slonczewski, Currents, torques, and polarization factors in magnetic tunnel junctions. *Phys. Rev. B* **71**, 024411 (2005)
38. J.Z. Sun et al., Effect of subvolume excitation and spin-torque efficiency on magnetic switching. *Phys. Rev. B* **84**, 064413 (2011)
39. Y. Zhang et al., Compact modeling of perpendicular-anisotropy CoFeB/MgO magnetic tunnel junctions. *IEEE Trans. Electron Devices* **59**, 819–826 (2012)
40. R.H. Koch, J.A. Katine, J.Z. Sun, Time-resolved reversal of spin-transfer switching in a nanomagnet. *Phys. Rev. Lett.* **92**, 088302 (2004)
41. R. Heindl, W.H. Rippard, S.E. Russek, M.R. Pufall, A.B. Kos, Validity of the thermal activation model for spin-transfer torque switching in magnetic tunnel junctions. *J. Appl. Phys.* **109**, 073910 (2011)
42. L.-B. Faber, W. Zhao, J.-O. Klein, T. Devolder, C. Chappert, Dynamic compact model of spin-transfer torque based magnetic tunnel junction (MTJ). in *4th International Conference on Design Technology of Integrated Systems in Nanoscal Era*, 2009. DTIS'09 130–135 (2009)
43. Y. Zhang et al., Electrical modeling of stochastic spin transfer torque writing in magnetic tunnel junctions for memory and logic applications. *IEEE Trans. Magn.* **49**, 4375–4378 (2013)
44. T. Devolder et al., Single-shot time-resolved measurements of nanosecond-scale spin-transfer induced switching: stochastic versus deterministic aspects. *Phys. Rev. Lett.* **100**, 057206 (2008)
45. J.J. Nowak et al., Demonstration of ultralow bit error rates for spin-torque magnetic random-access memory with perpendicular magnetic anisotropy. *IEEE Magn. Lett.* **2**, 3000204 (2011)
46. Z. Wang, Y. Zhou, J. Zhang, Y. Huai, Bit error rate investigation of spin-transfer-switched magnetic tunnel junctions. *Appl. Phys. Lett.* **101**, 142406 (2012)
47. H. Tomita et al., Unified understanding of both thermally assisted and precessional spin-transfer switching in perpendicularly magnetized giant magnetoresistive nanopillars. *Appl. Phys. Lett.* **102**, 042409 (2013)
48. K. Lee, S.H. Kang, Development of embedded STT-MRAM for mobile system-on-chips. *IEEE Trans. Magn.* **47**, 131–136 (2011)
49. Y. Kim et al., Integration of 28nm MJT for 8~16Gb level MRAM with full investigation of thermal stability. in *2011 Symposium on VLSI Technology (VLSIT)* (2011), pp. 210–211
50. N. Nishimura et al., Magnetic tunnel junction device with perpendicular magnetization films for high-density magnetic random access memory. *J. Appl. Phys.* **91**, 5246–5249 (2002)
51. M. Gajek et al., Spin torque switching of 20 nm magnetic tunnel junctions with perpendicular anisotropy. *Appl. Phys. Lett.* **100**, 132408 (2012)
52. D.C. Worledge et al., Recent advances in spin torque MRAM. in *Memory Workshop (IMW)*, 2012 4th IEEE International (2012), pp. 1–3
53. D. Ravelosona et al., Domain wall creation in nanostructures driven by a spin-polarized current. *Phys. Rev. Lett.* **96**, 186604 (2006)
54. S. Mangin et al., Current-induced magnetization reversal in nanopillars with perpendicular anisotropy. *Nat. Mater.* **5**, 210–215 (2006)
55. Y. Zhang et al., Perpendicular-magnetic-anisotropy CoFeB racetrack memory. *J. Appl. Phys.* **111**, 093925 (2012)
56. A.J. Annunziata et al., Racetrack memory cell array with integrated magnetic tunnel junction readout. in *Electron Devices Meeting (IEDM)*, 2011 I.E. International (2011), pp. 24.3.1–24.3.4
57. Y. Zhang et al., Current induced perpendicular-magnetic-anisotropy racetrack memory with magnetic field assistance. *Appl. Phys. Lett.* **104**, 032409 (2014)
58. Y. Zhang et al., Implementation of magnetic field assistance to current-induced perpendicular-magnetic-anisotropy racetrack memory. *J. Appl. Phys.* **115**, 17D509 (2014)

59. W. Zhao, C. Chappert, V. Javerliac, J.-P. Noziere, High speed, high stability and low power sensing amplifier for MTJ/CMOS hybrid logic circuits. *IEEE Trans. Magn.* **45**, 3784–3787 (2009)
60. A. Thiaville, Y. Nakatani, J. Miltat, Y. Suzuki, Micromagnetic understanding of current-driven domain wall motion in patterned nanowires. *EPL* **69**, 990 (2005)
61. S. Fukami et al., Current-induced domain wall motion in perpendicularly magnetized CoFeB nanowire. *Appl. Phys. Lett.* **98**, 082504 (2011)
62. H. Tanigawa et al., Domain wall motion induced by electric current in a perpendicularly magnetized Co/Ni nano-wire. *Appl. Phys. Exp.* **2**, 053002 (2009)
63. X. Jiang et al., Enhanced stochasticity of domain wall motion in magnetic racetracks due to dynamic pinning. *Nat. Commun.* **1**, 25 (2010)
64. W. Kang et al., High reliability sensing circuit for deep submicron spin transfer torque magnetic random access memory. *Electron. Lett.* **49**, 1283–1285 (2013)
65. W. Kang et al., Variation-tolerant and disturbance-free sensing circuit for deep nanometer STT-MRAM. *IEEE Trans. Nanotechnol.* **13**, 1088–1092 (2014)
66. W. Kang et al., Variation-tolerant high-reliability sensing scheme for deep submicrometer STT-MRAM. *IEEE Trans. Magn.* **50**, 1–4 (2014)
67. W. Kang et al., Separated pre-charge sensing amplifier for deep submicron MTJ/CMOS hybrid logic circuits. *IEEE Trans. Magn.* **6**, 3400305–5 (2014)
68. CMOS065 Design Rule Manual, STMicroelectronics, Geneva, Switzerland (2010)
69. W. Zhao et al., Failure and reliability analysis of STT-MRAM. *Microelectron. Reliab.* **52**, 1848–1852 (2012)
70. W. Kang et al., A low-cost built-in error correction circuit design for STT-MRAM reliability improvement. *Microelectron. Reliab.* **53**, 1224–1229 (2013)
71. W. Kang et al., DFSTT-MRAM: dual functional STT-MRAM cell structure for reliability enhancement and 3D MLC functionality. *IEEE Trans. Magn.* **6**, 3400207 (2014)
72. W. Kang et al., A radiation hardened hybrid spintronic/CMOS nonvolatile unit using magnetic tunnel junctions. *J. Phys. D Appl. Phys.* **47**, 405003 (2014)
73. Y. Lakys, W. Zhao, T. Devolder, Y. Zhang, J. Klein, D. Ravelosona, C. Chappert, Self-enabled “error-free” switching circuit for spin transfer torque MRAM and logic. *IEEE Trans. Magn.* **48**, 2403–2406 (2012)
74. E. Deng et al., Low power magnetic full-adder based on spin transfer torque MRAM. *IEEE Trans. Magn.* **49**, 4982–4987 (2013)
75. H.-P. Trinh et al., Magnetic adder based on racetrack memory. *IEEE Trans. Circuits Syst. I: Regular Papers* **60**, 1469–1477 (2013)
76. W. Kim et al., Extended scalability of perpendicular STT-MRAM towards sub-20 nm MTJ node. in *Electron Devices Meeting (IEDM)*, 2011 I.E. International (2011), 24.1.1–24.1.4
77. H. Yoda et al., Progress of STT-MRAM technology and the effect on normally-off computing systems. in *Electron Devices Meeting (IEDM)*, 2012 I.E. International (2012), pp. 11.3.1–11.3.4
78. D.A. Patterson, J.L. Hennessy, *Computer Organization and Design: The Hardware/Software Interface* (Elsevier, Amsterdam, 2012)
79. F. Ren, D. Markovic, True energy-performance analysis of the MTJ-based logic-in-memory architecture (1-bit full adder). *IEEE Trans. Electron Devices* **57**, 1023–1028 (2010)
80. N.S. Kim et al., Leakage current: Moore’s law meets static power. *Computer* **36**, 68–75 (2003)
81. Y. Zhang, W. Zhao, J.-O. Klein, D. Ravelosona, C. Chappert, Ultra-high density content addressable memory based on current induced domain wall motion in magnetic track. *IEEE Trans. Magn.* **48**, 3219–3222 (2012)
82. W. Zhao, D. Ravelosona, J. Klein, C. Chappert, Domain wall shift register-based reconfigurable logic. *IEEE Trans. Magn.* **47**, 2966–2969 (2011)
83. S. Matsunaga et al., Fully parallel 6T-2MTJ nonvolatile TCAM with single-transistor-based self match-line discharge control. in *2011 Symposium on VLSI Circuits (VLSIC)* (2011), pp. 298–299.

84. H. Kadota, J. Miyake, Y. Nishimichi, H. Kudoh, K. Kagawa, An 8-kbit content-addressable and reentrant memory. *IEEE J. Solid-State Circuits* **20**, 951–957 (1985)
85. J.Z. Sun et al., Spin-torque switching efficiency in CoFeB-MgO based tunnel junctions. *Phys. Rev. B* **88**, 104426 (2013)
86. J.E. Hirsch, Spin hall effect. *Phys. Rev. Lett.* **83**, 1834–1837 (1999)
87. L. Liu et al., Spin-torque switching with the giant spin hall effect of tantalum. *Science* **336**, 555–558 (2012)
88. I.M. Miron et al., Fast current-induced domain-wall motion controlled by the Rashba effect. *Nat. Mater.* **10**, 419–423 (2011)
89. W. Kang et al., An overview of spin-based integrated circuits. in *Asia and South Pacific Design Automation Conference (ASP-DAC)* (2014), pp. 676–683
90. Z.H. Wang et al., Perpendicular-anisotropy magnetic tunnel junction switched by spin-hall-assisted spin-transfer torque. *J. Phys. D Appl. Phys.* **48**, 065001 (2015)