Low Power Magnetic Full-Adder based on Spin Transfer Torque MRAM

Erya Deng¹,², Yue Zhang¹,², Jacques-Olivier Klein¹,², Dafiné Ravelsona¹,², Claude Chappert¹,², and Weisheng Zhao¹,²∗

¹IEF, Univ. Paris-Sud, Orsay, 91405, France
²CNRS, UMR 8622, Orsay, 91405, France
∗Contact: weisheng.zhao@u-psud.fr

Power issues have become a major problem of CMOS logic circuits as technology node shrinks below 90 nm. In order to overcome this limitation, emerging logic-in-memory architecture based on non-volatile memories (NVMs) are being investigated. Spin transfer torque (STT) Magnetic RAM (MRAM) is considered one of the most promising NVMs thanks to its high speed, low power, good endurance and 3D back-end integration. This paper presents a novel magnetic full-adder (MFA) design based on perpendicular magnetic anisotropy (PMA) STT-MRAM. It provides advantageous power efficiency and die area compared with conventional CMOS-only full adder (FA). Transient simulations have been performed to validate this design by using an industrial CMOS 40 nm design kit and an accurate STT-MRAM compact model including physical models and experimental measurements.

Index Terms— low-power design, full-adder, resistive switching memories, complementary cells

I. INTRODUCTION

As CMOS technology shrinks below 90 nm node, exponentially increasing standby power is becoming major miniaturization bottleneck for mainstream logic systems [1]. The roadmap of ITRS report mentions static power will dominate the power consumption in the future [2]. Academic and industrial research and development efforts focus recently on hybrid circuits based on non-volatile memories, such as Spin Transfer Torque Magnetic Random Access Memory (STT-MRAM) [3], Conductive-Bridge RAM (CBRAM) [4] and Oxide Resistive RAM (OxRRAM) [5], in order to overcome this power issue. The Magnetic tunnel junction (MTJ) is the basic element of MRAM [6-8], and is composed of a thin insulating barrier (e.g. MgO [9]) separating two ferromagnetic (FM) layers (Fig. 1a). Based on the tunnel magneto-resistance (TMR) effect [3], the nanopillar resistance, \( R_P \) or \( R_{AP} \), depends on the relative orientation, Parallel (P) or Anti-Parallel (AP), of magnetizations in the two FM layers.

Spin Transfer Torque (STT) is a promising switching mechanism that supports high power efficiency and fast writing speed [10]. This basic physical mechanism enables profoundly simplifying CMOS circuitry, as only a bipolar current is required (Fig. 1b). The MTJ switches state as the current exceeds a given critical current denoted as \( I_{C0} \). Meanwhile, recent progress demonstrates the perpendicular magnetic anisotropy (PMA) in CoFeB/MgO structures provides a high-energy barrier, \( E \), to resist the thermal instability of in-plane anisotropy. It also presents the advantages of lower threshold current, higher speed operation and higher TMR [11-12].

Innovative circuits based on hybrid MRAM/CMOS circuits have been presented recently [13-14]. For instance, magnetic look-up-table (MLUT) and non-volatile flip-flop (MFF) were introduced for reconfigurable logic circuits and power down applications [15-16]. In order to build low-power high-density arithmetic/logic unit for processors, magnetic full-adder (MFA) design based on non-volatile memory implementing the addition operation has been proposed and exhibits satisfying properties [17-18]. It could also overcome the communication bottleneck between separated logic module and memory block. However, the use of capacitance for data sensing and magnetic field for data programming limits further miniaturization. As the pre-charge sense amplifier (PCSA) [19] can remarkably improve the reliability of MRAM cell sensing, the non-volatile PCSA based MRAM full-adder could lead to ultra-low power and high density ICs [20]. Even so, the mandatory heating for this structure is not suitable for power saving, and the heating devices consume area. Furthermore, despite resolving all the above problems, a fully non-volatile full-adder based on domain walls has not become mature for applications to date [21-22].

![Figure 1](image-url)
(RA) and TMR ratio on the delay time and dynamic power performances have been analyzed. Finally, we compare the simulated performance of this STT-MFA and a conventional CMOS-only full adder (FA) to confirm the low power advantage of this new design.

II. PCSA BASED STT-MFA ARCHITECTURE

A. General “logic-in-memory” architecture for STT-MRAM

The proposed generic STT-MFA structure is shown in Fig.2, and is composed of three parts: a PCSA circuit evaluates the logic result on the outputs (see Fig.3a), a write logic block programs the STT-MRAM cells (see Fig.3b), and a logic control data block. Considering every bit of STT-MRAM costs a high programming energy (~ from 0.2 to 0.5 pJ/bit@40 nm) and relatively low switching speed (~ns), the logic data block contains a MOS logic tree and STT-MRAM in order to keep an area-power-efficient advantage. In this case, the logic volatile data can be driven by a high processing frequency, \( F_{\text{computing}} \), contrarily to analog non-volatile data, which should be changed with a relatively low frequency i.e. they are more critical data or quasi-constant for computing.

A PCSA circuit (see Fig. 3a) 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. It operates in 2 phases. During the first phase, “CLK” is set to ‘0’ and the outputs (“Qm” and “Qm’”) are pulled-up to “\( V_{DD} \)” 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 (MP1 or MP2), at that time, the other branch will be pulled up to “\( V_{DD} \)” or logic ‘1’ and this low-resistance branch will continue to drop to “Gnd” or logic ‘0’.

We use PCSA circuit in this STT-MFA design (see Fig.3a) for two reasons: firstly, the dynamic sensing allows the amplification from analog data to digital with ultra-low power; secondly, the read disturbance induced by sensing operations can be significantly decreased [25-26]. The latter 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). As shown in Fig.4 and Eq.1, lower sensing current \( I_r \) and shorter duration \( \tau \) can reduce greatly the chip failure rate for the STT-MRAM with the same thermal stability factor \( \Delta = 40 \).

\[
F_{\text{chip}} = 1 - \exp\left\{-N \frac{\tau}{\tau_0} \exp\left\{-\Delta \left(1 - \frac{I_r}{I_{\text{co}}} \right)\right\}\right\}
\]

where \( N \) is the number of bits per word, which equals to 1 for the logic in memory architecture shown in Fig.2 different from memory array; \( F_{\text{chip}} \) is the error switching rate induced by the reading operations.

![Figure 2](image-url) Logic gates composed of PCSA, MOS logic tree for volatile data and STT-MRAM cells for non-volatile data.

![Figure 3](image-url) (a) Pre-Charge Sense Amplifier (PCSA) for MTJ state detection and amplification to logic level (b) Full writing schematic for STT writing approach, which is composed of two modified inverters.

![Figure 4](image-url) Low reading current \( I_r \) and short duration \( \tau \) can reduce greatly the chip failure rate shown in Y axis. X axis presents the ratio of accumulated read duration on 10 years.

A pair of STT-MRAM cells presents non-volatile input data and they are always in opposite state to ensure the necessary high sensing speed [19]. Thanks to the proximate position of the complementary cells, this structure suffers from lower mismatch variation, which is one of the most important limitations for STT-MRAM due to its hybrid fabrication process Magnetic/CMOS.

The bi-directional write circuit used in this STT-MFA (see Fig.3b) is suitable for the programming of the complementary cells. It is composed of two inverters connected with a couple of complementary STT-MRAM cells. Each time one NMOS and PMOS transistors (e.g. MN3 and MP5) are active to generate current, and the others are closed. Only one operation is required to switch both of them instead of individual switching for classical STT-MRAM [17]. However, it requires a Vdda higher than Vdd for logic operations (see Table.III) to avoid the area overhead in the write circuit [23].

B. Detailed Magnetic Full Adder based on STT-MRAM

Fig.5 shows a 1-bit STT-MFA circuit, the inputs are “A”, “Ci” and “B”, and the outputs are “SUM” and “Co”.

Copyright (c) 2013 IEEE. Personal use is permitted. For any other purposes, permission must be obtained from the IEEE by emailing pubs-permissions@ieee.org.
Among them, the input “B” relates to non-volatile storage STT-MRAM. To evaluate the logic function, PCSA circuit provides the best sensing reliability and power efficiency while keeping high-speed performance (~200 ps) compared with other sense amplifiers for fast memory reading [19, 27]. Depending on the MOS state in the logic tree and the STT-MRAM element state, the discharge currents are different in both branches and the current sense amplifier latches opposite logic value on “SUM”, “/SUM” and respectively “Co”, “/Co”.

![Figure 5. STT-MRAM Magnetic Full-Adder (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.](image)

The MOS tree is designed according to Eq. 2-5 and the truth table shown in Table.I. For “SUM” logic, the MOS tree corresponds directly to the logic relationship among the inputs “A”, “B” and “Ci”. We can simply adapt it to the general structure with a couple of complementary STT-MRAM cells (see Fig.2). However, it is a little difficult for “Co” logic as there is the term ACi in the logic function Eq.4 and we cannot adapt the schematic to the general “logic-in-memory” structure. It can be inferred that the impact of the term ACi on the resistance is equivalent to a sub-branch connecting PCSA and the discharging transistor (i.e. MN2 in Fig.3a). Table II exhibits the true table and the resistance configuration of “Co” logic. Roff and Ron are respectively the close and open resistances of MOS transistors. Ri and Rf 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 ACi and A Ci have no impact on the output. If “A” and “Ci” are different, the resistance of the two sub-branches is the same. If they are the same, their comparison corresponds to that of Ri and Rf in the condition of Roff>RF, which is always true for STT-MRAM. This allows the term ACi to be deleted from Eq.4 and we can obtain the “Co” logic circuit shown in Fig.5.

\[
\text{Sum} = A \oplus B \oplus C_i = ABC_i + AB\bar{C}_i + \bar{A}BC_i + \bar{A}\bar{B}C_i
\]

\[
\bar{C}_o = \bar{A}B + AC_i + BC_i
\]

The STT-MRAM devices connect serially with a common central point. In order to program STT-MRAM cells, we use a write logic composed of pass transistors, which are connected respectively to the bottom (BE) and top electrodes (TE) of the serial branch and to the common point. In such a manner, as a control signal (“Input1” or “Input2”) is activated, the first STT-MRAM device noted B is put in high resistance state (Roff) or low resistance state (Rf) while the second STT-MRAM device noted “/B” is put in the complementary state Rf or Ron.

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 [17-18, 20]. Therefore, this design allows efficient area minimization and is suitable for advanced fabrication nodes below 65 nm.

### III. STT-MFA Validation and Analysis

#### A. PMA STT-MRAM compact model

A CoFeB/MgO/CoFeB PMA STT-MTJ compact model was recently presented based on the understanding of fundamental physical mechanisms and experimental measurements [23, 29]. This model takes into account static, dynamic, and stochastic behaviors.

The static behavior of STT switching in PMA MTJ mainly relies on the calculation of the threshold or critical current \( I_{\text{Co}} \), as shown in Eq.6 [13]:

\[
I_{\text{Co}} = \alpha \gamma e \left( \frac{\mu_B g}{\mu_B g} \right) H_B V = 2\alpha \gamma e \frac{E}{\mu_B g} \tag{6}
\]

where \( E \) is the barrier energy, \( \alpha \) is the magnetic damping constant, \( \gamma \) is the gyromagnetic ratio, \( e \) is the electron charge, \( \mu_B \) is the Bohr magnetron, \( V \) is the volume of the free layer and \( k_B \) is the Boltzmann constant.

The switching dynamics of STT in PMA MTJ was described in [14] and the dependence of switching current \( I_{\text{write}} \) and the duration \(<\tau>\) is shown as follows:

\[
\frac{1}{<\tau>} = \frac{2}{C + \ln \left( \frac{2 \xi}{\mu_B P_{\text{ref}} \mu_B P_{\text{ref}}} \right)} \left( \frac{\mu_B P_{\text{ref}}}{Cm} \right) \left( I_{\text{write}} - I_{\text{co}} \right) \tag{7}
\]

where \( C \approx 0.577 \) is the Euler’s constant; \( \xi = E/k_B T \) is the activation energy in units of \( k_B T \); \( P_{\text{ref}} \) and \( P_{\text{free}} \) are the tunneling spin polarizations of the reference and free layers.
respectively; \( m \) is the magnetic moment of free FM layer. We assume that \( P_{\text{ref}} = P_{\text{free}} = P \) for this compact model.

Fig.6 shows the DC and transient simulation of this compact model, which could verify the agreement of the static and dynamic behaviors between physical models and experimental measurements. The switching delay is inversely proportional to the writing current (see Eq. 7).

Moreover, we find that the sensing delay of “Output_Co” (~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” (see also Fig.5), 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_D) in Fig.5, STT-MRAM resistance-area product (RA) and Tunnel magnetoresistance (TMR) ratio. Fig. 9 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.

**B. Transient simulation of STT-MFA & parameter analysis**

Fig.7 illustrates the transient hybrid spice simulation of 1-bit STT-MFA shown in Fig.5. It is performed by using PMA STT-MRAM compact models introduced above and CMOS 40 nm design kit. The time-dependent behaviors of outputs (“SUM” and “Co”) confirm the logic functionality of full addition. For instance, for the operation “A” = ‘1’, “B” = ‘0’, “Ci” = ‘0’, the result is ‘1’ and no carry yields; for the operation “A” = ‘1’, “B” = ‘0’, “Ci” = ‘1’, the result is ‘0’ and the carry is ‘1’.

Moreover, we find that the sensing delay of “Output_Co” (~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” (see also Fig.5), 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_D) in Fig.5, STT-MRAM resistance-area product (RA) and Tunnel magnetoresistance (TMR) ratio. Fig. 9 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.

**B. Transient simulation of STT-MFA & parameter analysis**

Fig.7 illustrates the transient hybrid spice simulation of 1-bit STT-MFA shown in Fig.5. It is performed by using PMA STT-MRAM compact models introduced above and CMOS 40 nm design kit. The time-dependent behaviors of outputs (“SUM” and “Co”) confirm the logic functionality of full addition. For instance, for the operation “A” = ‘1’, “B” = ‘0’, “Ci” = ‘0’, the result is ‘1’ and no carry yields; for the operation “A” = ‘1’, “B” = ‘0’, “Ci” = ‘1’, the result is ‘0’ and the carry is ‘1’.

Moreover, we find that the sensing delay of “Output_Co” (~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” (see also Fig.5), 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_D) in Fig.5, STT-MRAM resistance-area product (RA) and Tunnel magnetoresistance (TMR) ratio. Fig. 9 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.
and high TMR ratio (e.g. 200%) can be achieved in PMA STT-MRAM [29-31].

As mentioned above, the critical idea of this design shown in Fig.2 is to use a programming frequency (e.g. 1 KHz) of STT-MRAM much lower than the computing frequency (see also Eq.8). 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) [23]. Moreover, the programming energy for the non-volatile data (bit B in Fig. 5) can be reduced, following the area minimization [27-28] and new material development for MTJs (e.g. to approximately 0.1 pJ/bit) [31].

**TABLE IV**

<table>
<thead>
<tr>
<th>Comparison of Proposed 1-Bit STT-MFA with CMOS-only FA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>Delay time</td>
</tr>
<tr>
<td>Dynamic power</td>
</tr>
<tr>
<td>@500MHz</td>
</tr>
<tr>
<td>Standby power</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Data transfer energy</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Die area</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

**ACKNOWLEDGMENT**

The author thanks C. Gamrat, D. Querlioz, J.M Portal, T. Devolder and G. Prenat for valuable inputs. This work was partially supported by the French research agencies through projects ANR-DIPMEM, CNRS-PePS-NVCPU and ANR-MARS.

**REFERENCES**

[9] S. Ikeda et al., “Tunnel magnetoresistance of 604% at 300 K by suppression of Ta diffusion in CoFeB/MgO/CoFeB pseudo-spin-


