ODESY: A novel 3T-3MTJ cell design with Optimized area DEnsity, Scalability and latencY

Linuo Xue[1], Yuanqing Cheng[1], Jianlei Yang[2], Peiyuan Wang[3]and Yuan Xie[1]
Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106, USA[1]
BeiHang University, Beijing 100191, P.R.China[2]
Qualcomm Research, San Diego, CA 92121, USA[3]
{*linuo_xue*}*umail.ucsb.edu*[1], {*yuanqing, yuanxie*}*ece.ucsb.edu*[1], {*yuanpeiw*}*qti.qualcomm.com*[3]

ABSTRACT

The STT-RAM (Spin-Transfer Torque Magnetic RAM) technology is a promising candidate for cache memory because of its high density, low standy-power, and non-volatility. As technology scales, especially under 40nm technology node, the read disturbance becomes severe since the read current approaches closely to the switching current. In addition, the read latency and access performance degrade significantly as well. The conventional 1T-1MTJ and 2T-2MTJ cell designs cannot address these challenges efficiently. In this paper, we propose a novel 3T-3MTJ cell structure using the advanced perpendicular MTJ technology. This memory cell has higher storage density and better performance, and is particularly suitable for the deeply scaled technology node. A two-stage sensing scheme is also proposed to facilitate the read operation of the 3T-3MTJ cell design. Circuit-level and architecture-level simulations show that the proposed 3T-3MTJ cell structure can achieve a better tradeoff between storage density, access performance, energy consumption, and reliability compared to the prior 1T-1MTJ and 2T-2MTJ cell structures.

1. INTRODUCTION

The STT-RAM is a promising candidate for last-level cache design, because it has high density, fast access speed, low static power, and non-volatility. Therefore, it attracts a lot of research efforts from device-level optimizations to architecture innovations [4, 11].

The STT-RAM relies on the magnetic tunnel junction (MTJ) to store data. The MTJ has two distinct states, i.e., high and low resistance state according to the magnetization of MTJ. By sensing the MTJ resistance[1], data can be read from memory cells. Depending on the magnetic anisotropy used for data retention, the MTJ can be classified into the in-plane MTJ or the perpendicular MTJ (p-MTJ). The former takes advantage of the shape anisotropy for data retention while the latter stores data with the perpendicular magnetic anisotropy (PMA). Since the p-MTJ has better scalability as technology node shrinking [6], we focus on the p-MTJ based STT-RAM in the paper.

In order to achieve high storage density, the 1T-1MTJ cell structure is proposed and widely used in many STT-RAM based designs [8, 21]. In addition to small cell size, the simple structure makes device fabrication relatively easier. Unfortunately, it also has some drawbacks. First, the sensing procedure requires reference cell for comparison, which introduces non-negligible area overhead. Second, the sensing margin is small, which degrades read performance and threatens sensing reliability. In particularly, when MTJ size scales down to 20nm, the read disturbance will be severe because the read and write current approach to each other closely (i.e., the data are corrupted during sensing) [19]. To deal with these problems, a 2T-2MTJ cell structure is proposed to increase the read margin and approachs better scalability through the differential sensing scheme [14]. But its storage density halves, compared to that of the 1T-1MTJ cell design, which may increase miss rate and degrade access performance on the use of on-chip LLC.

In this paper, a novel 3T-3MTJ cell structure is proposed to combine the benefits of two cell structures mentioned to achieve a better tradeoff between storage density, power consumption, and scalability. In such a cell structure, 3 MTJs are used to store 2-bit data. By the appropriate mapping from data to MTJ state configurations, data pattern '00' and '11' could be sensed by the differential sensing scheme directly similar to the 2T-2MTJ case. For '01' and '10', a two-stage sensing circuit is proposed to accelerate sensing speed and increase sensing reliability taking process variations into account. Extensive circuit-level and architecture-level simulations show that the proposed 3T-3MTJ cell structure can achieve a better compromise among storage density, performance, and reliability. Our main contributions are as follows,

1. a novel 3T-3MTJ cell structure is proposed to make a better tradeoff of pertinent design metrics. This cell structure adopts differential sensing scheme to accelerate the read operation, therefore eliminating reference cells and reducing area overhead.

2. A two-stage sensing scheme is proposed to facilitate the read operation of 3T-3MTJ cell structure. What is more, the robustness of sensing process is validated considering process variation effect.

[1]Sensing resistance means sensing voltage/current difference because of resistance difference.

ICCAD '16, November 07-10, 2016, Austin, TX, USA

DOI: http://dx.doi.org/10.1145/2966986.2967060

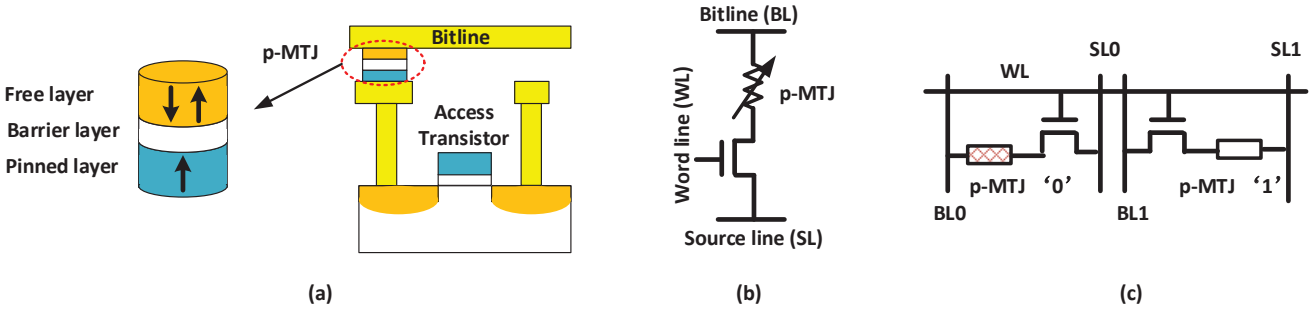

Figure 1: (a) Illustration of STT-RAM structure. MTJ is sandwitched between top metal layers and can be accessed by transistor. (b) 1T-1MTJ structure. (c) 2T-2MTJ structure.

3. Both circuit-level and architecture-level simulations are performed to compare the 3T-3MTJ to the 1T-1MTJ and 2T-2MTJ cell structures in terms of storage density, power consumption, access performance, and reliability. The simulation results indicate our proposed cell structure can achieve better scalability as technology node scales down.

The rest of the paper is organized as follows, Section 2 introduces the preliminaries of the STT-RAM and several representative cell structures. The analysis of their respective pros and cons motivate this work. Section 3 presents the 3T-3MTJ cell design and associates with the peripheral circuitry to support reliable and high speed sensing. The write operation is also discussed in this section. The validations of our cell structure and extensive comparisons with the 1T-1MTJ and 2T-2MTJ counterparts are discussed in Section 4. Section 5 makes a conclusion to the paper.

2. PRELIMINARIES OF STT-RAM TECHNOLOGY AND TYPICAL CELL STRUCTURES

In this section, we first introduce the background of the STT-RAM technology and then discuss several representative cell structures adopted in STT-RAM chip design.

2.1 Introduction of PMA STT-RAM

The STT-RAM technology uses MTJ to store data. MTJ mainly consists of the free layer, pinned layer and oxide layer as shown in Fig. 1(a). In this paper, we focus on MTJ with the perpendicular magnetic anisotropy (PMA) effect (p-MTJ) as it requires much smaller switching current and has better technology scalability compared to in-plane MTJ [6]. As shown in the figure, the magnetization of the free layer can be changed to be parallel or anti-parallel to that of the pinned layer by injecting spin polarized current into MTJ. If the magnetization of free layer is parallel to that of the pinned layer, the injected current from bitline to source line can switch the free layer into anti-parallel state, and the resistance is changed from R_p to R_{ap}[2]. As a result, '1' is written. Otherwise, '0' is written. The tunnel magnetoreistance ratio (TMR) is defined as $TMR = (R_{ap} - R_p)/R_p$. The larger TMR is, the easier we can distinguish '1' from '0'. The typical TMR value varies from 100% to 200% according to the measurements from available prototypes [10, 13]. To take the best advantages of the p-MTJ, some representative STT-RAM cell structures have been proposed and will be discussed next.

[2] R_p represents MTJ resistance in parallel state while R_{ap} denotes that in anti-parallel state.

2.2 1T-1MTJ cell structure

The 1T-1MTJ structure is one of the most widely used cell structures adopted in the STT-RAM chips [8, 12]. As shown in Fig. 1 (b), the 1T-1MTJ cell is composed of one transistor and one MTJ. To read the data out, the word line is asserted first. Then, a voltage/current is applied between bitline and source line. The current flowing through/voltage generated over MTJ is compared to that of one reference cell to identify what data is read out. The reference cell includes two MTJs in parallel and anti-parallel state respectively. Therefore, the resistance of the reference cell is about the average value of R_p and R_{ap}. Despite of the simple structure, the 1T-1MTJ cell has several drawbacks. First, since the TMR ratio is only about 100% $\sim$ 200% for the current p-MTJ technology, the sensing current/voltage difference between data and reference cell is limited by the resistance difference between these two cells. The small resistance difference makes sensing procedure prone to error. As the technology node continuously shrinks, the sensing reliability can be further limited by aggravating process variations. Second, the small sensing margin also negatively affects sensing latency and read performance. Since the STT-RAM technology is usually used for cache memory, long sensing latency may cause severe system performance degradation. Third, to enlarge the sensing margin, it is necessary to apply large read current/voltage which may introduce read disturbance [19]. Fourth, it requires reference cells for sensing signal comparisons, which incurs extra area overhead.

2.3 2T-2MTJ cell structure

To overcome drawbacks mentioned above, the 2T-2MTJ cell is proposed to improve sensing performance and reliability [14]. As shown in Fig. 1 (c), the 2T-2MTJ cell structure consists of two MTJs with complementary states. For instance, we could designate data '0' corresponding to the case where the left MTJ is in parallel state and the right one in anti-parallel state. Otherwise, data '1' is stored. This structure eliminates reference cell because data can be read out by sensing the resistance difference of the two MTJ branches. Since the resistance difference of R_{ap} and R_p is roughly 2X of that of the difference between R_{ap}/R_p and R_{ref}, the sensing margin is enlarged dramatically. Thus, sensing relia-

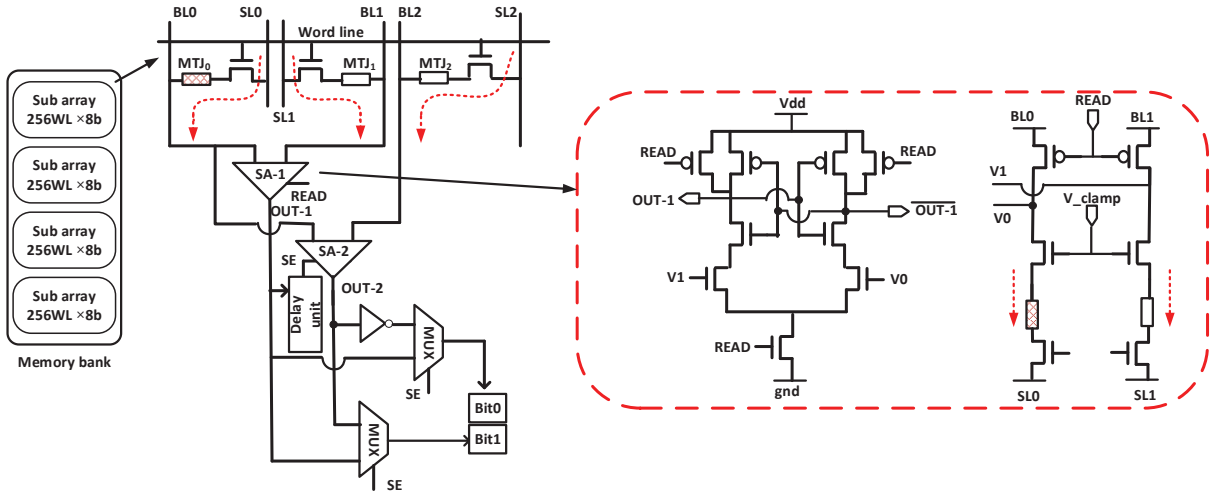

Figure 2: Illustration of 3T-3MTJ based memory bank, the schematic of 3T-3MTJ cell and peripheral circuit.

bility and sensing speed can be improved. Unfortunately, this structure also has some disadvantages. First, it uses 2 MTJs to store one bit, which reduces area efficiency by about 50%. Second, it requires to switch 2 MTJs simultaneously to change the stored data, which increases write energy significantly.

2.4 Other STT-RAM cell structures

The 4T-4MTJ is another adaptive STT-RAM cell structure based on the 2T-2MTJ [14]. It stores one bit per cell. When reading data from the cell, two word lines are asserted and each bitline sinks current flowing from two MTJs on the same side, which doubles read current and sensing speed. However, it makes storage density even worse. The 3T-2MTJ is another dual memory cell which can be used in normally-off computing [7]. Bitline pair can be driven to write data in complementary cells. The 4T-2MTJ cell structure operating like the SRAM without leakage current is proposed [15]. However, these cell structures require more transistors, higher write voltage and incur large area overhead compared to the 1T-1MTJ and 2T-2MTJ cell structures.

3. 3T-3MTJ CELL STRUCTURE AND PERIPHERAL CIRCUIT DESIGN

By considering the cons and pros of the exiting STT-RAM cell structures, a novel 3T-3MTJ cell structure is proposed, which can significantly promote memory cell density and performance of the STT-RAM. In this section, we present details of our design, including the cell structure and the peripheral circuitry to support read and write operations.

3.1 3T-3MTJ cell structure

Fig. 2 depicts the schematic of the 3T-3MTJ cell structure. As shown in the figure, three MTJs in one cell are denoted as MTJ_0, MTJ_1 and MTJ_2 respectively from the left to the right. Different from the 1T-1MTJ and 2T-2MTJ cell structures, it can store 2-bit data in one cell according to the different MTJ resistance combinations. Three MTJs in combination can present 8 different resistance states. Among them, only 4 states are used to represent 2-bit data such

Table 1: Mapping from data bits to MTJ state combinations: 4 resistance state combinations represent 2-bit data.

Data bits	MTJ_1	MTJ_2	MTJ_3
00	0	1	0
01	0	0	1
10	1	1	0
11	1	0	1

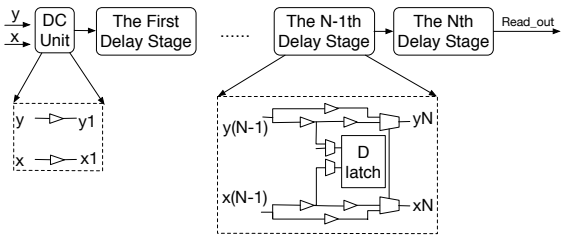

Figure 3: Illustration of the accurate on-chip path delay measurement circuit.

that fast differential sensing scheme can be used. For the ease of presentation, we classify data bits into two classes, i.e., *complementary patterns* (e.g., '01' and '10'), and *identical patterns* (e.g., '00' and '11'). Table. 1 shows the mapping of MTJ states to data bits. As shown in the table, MTJ state combinations '101' and '010' can be used to represent data '11' and '00', i.e., identical patterns. MTJ state combinations '001' and '110' are chosen from the remaining 6 MTJ state combinations to stand for complementary data patterns ('01' and '10'). The details of sensing circuit design and read operation are discussed next.

3.2 Sensing circuit and Read operation

The two-stage sensing scheme is proposed as shown in Fig. 2. Each sense amplifier (SA) contains two parts, i.e., current sensing part and amplification part, which are shown in red dashed box of Fig. 2. We adopt current sensing scheme proposed by [9] in the first part. It senses the differ-

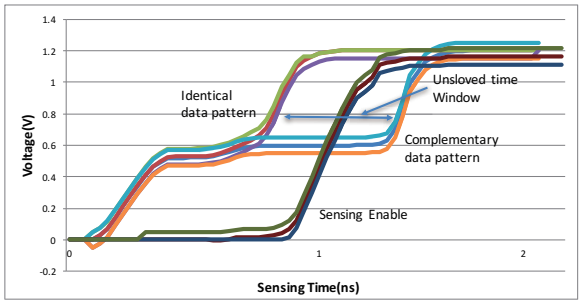

Figure 4: Simulation results of sensing delays when read identical data patterns and complementary data patterns. The timing of SE signal generation is also shown in the figure.

ence of two branches. However, due to the small signal difference generated by the amplifier, it is necessary to amplify the signal amplitude to the full swing to drive the following logic circuit. We adopt the PCSA circuit [23] in the second part for the amplification.

The read operation and working principle of the sensing circuit are described as follows. In the reading process, a read voltage is applied between bitline and source line. When word line is asserted, the current flows through each sensing branch as shown in the right part of red dashed box in Fig. 2. The current sensing circuit converts current difference to voltage difference (e.g., the difference of $V0$ and $V1$). Then, the two signals are fed into the amplification circuit in the left part of the red dashed box. It amplifies the signal difference to the full swing denoted as OUT-1 and $\overline{OUT}$-1. If bits'00' or '11' is stored in the 3T-3MTJ cell, data can be read out by sensing MTJ pair (MTJ_0 and MTJ_1). If MTJ_0 is in the anti-parallel state and MTJ_1 in the parallel state, '11' is read out. Otherwise, '00' is read out. If bits '01' or '10' are stored in the 3T-3MTJ cell, however, sense amplifier can not read the correct data in one step since the two MTJs on the left have the same resistance. It is observed that the voltages on two bitlines of MTJ_0 and MTJ_1 remain at almost the same level during early sensing stage in these two cases. After that, it finally outputs a random result depending on the resistance difference of two MTJs induced by process variations. Therefore, we could take advantage of this period to decide whether it is necessary to activate the second stage sensing. Finally, depending on data patterns, multiplexers in Fig. 2 are used to control which SA's output is used to generate the output data.

The key point in the sensing circuit design is the timing of sense enable (SE) signal generation to activate the SA-2 at an appropriate time point. As mentioned before, SA-1 gets stable output rapidly if '00' or '11' is read. Otherwise, the SA-1 can not respond in such a short time, which implies the result should be generated by SA-2. To illustrate this point, the sensing voltage waveforms corresponding to identical and complementary patterns are shown in Fig. 4. It shows that there is a 1ns time window that is the difference of sensing latencies corresponding to different data patterns. We call it "unresolved sensing window". For complementary pattern case, the SA-2 should be triggered during this window. The time point trigging SE signal is vital for sensing performance and reliability. Monte Carlo simulation results are shown in Fig. 4 to evaluate the process variation effect on unresolved sensing window and the timing of SE signal generation (refer to Section 4 for detailed simulation setup information). It indicates that the worst case latency is 1ns for identical data pattern. The SE generation should be later than this time point. On the other hand, SE generation should be as early as possible to reduce total sensing time. In this paper, we generate the SE signal after 1.25ns READ signal taking sensing reliability and performance into account. To guarantee the SE signal timing, a small delay circuit proposed in [18] is adopted. As depicted in Fig. 3, this circuit consists of a DC unit and several delay stages, which can achieve an accurate delay measurement. The read signal is delayed by the circuit to sample the output of the SA-1 to generate the SE signal. If the SE is high, the SA-2 is enabled. Otherwise, it is disabled. Fig. 4 illustrates that SE can be generated accurately even unresolved timing window varies due to process variation. So the complementary data pattern '01' and '10' can be correctly sensed as well.

3.3 Write operation

Writing to MTJ depends on two factors: the original state of MTJ and the write current direction. The Write current has two directions to switch MTJ to '0' or '1' state. This process can be achieved by applying negative or positive voltage between bitline and source line. When the word line is asserted, the write current flows through 3 MTJs simultaneously to write the corresponding bits (refer to Table. 1 for the mapping from MTJs states to digital data bits). Fig.5 shows the write operations to four different data. Taking the writing '11' case for an example, the write current from bitline to sourceline is provided to guarantee MTJ_1 and MTJ_3 being switched to the anti-parallel state. In the MTJ_2, reverse current from source line to bitline makes it switch to the parallel state.

Then, the difference of the number of MTJ flips in the 2T-2MTJ and 3T-3MTJ cell structures is analyzed as follows. Fig. 6 illustrates the number of MTJ flips in every possible write transaction for the 3T-3MTJ and 2T-2MTJ cell structure. For instance, in order to write 2-bit data to the cell of 3T-3MTJ with original data '00' stored. The probabilities of MTJ switched to '00', '01', '10' and '11' are the same, i.e., $p = 0.25$. All MTJs retain the original states if '00' is written in the next write. 2 MTJs needs to be flipped in '00' $\rightarrow$ '01' case. For '00' $\rightarrow$ '10' case, 1 MTJ will be switched. Considering the case '00' $\rightarrow$ '11', 3 MTJs should be flipped. In our example, the effective MTJ flip count, when '00' is the original data, should be $0.25\times1+0.25\times2+0.25\times3 = 1.5$ for the 3T-3MTJ structure. After that, we calculate the effective flip count considering all possible operations shown in Fig. 6 by equation:

$$O_{AveMTJFlipping} = \frac{\sum_{i=0}^{i=n} p_{state_i} \cdot N_{MTJFlipping_i}}{M_{data}} \quad (1)$$

where $O_{AveMTJFlipping}$ represents the average number of MTJ flipping per bit. P_{state_i} is the probability of original state and $N_{MtjFlipping_i}$ is the number of MTJ flips in every possible wirte transaction.

Therefore, the effective flip count per bit for the 3T-3MTJ structure is $1.5/2 = 0.75$. With the similar analysis, the effective flip count per bit of the 2T-2MTJ cell structure is 1. The 3T-3MTJ cell structure can save 25% MTJ flips compared to the 2T-2MTJ counterpart.

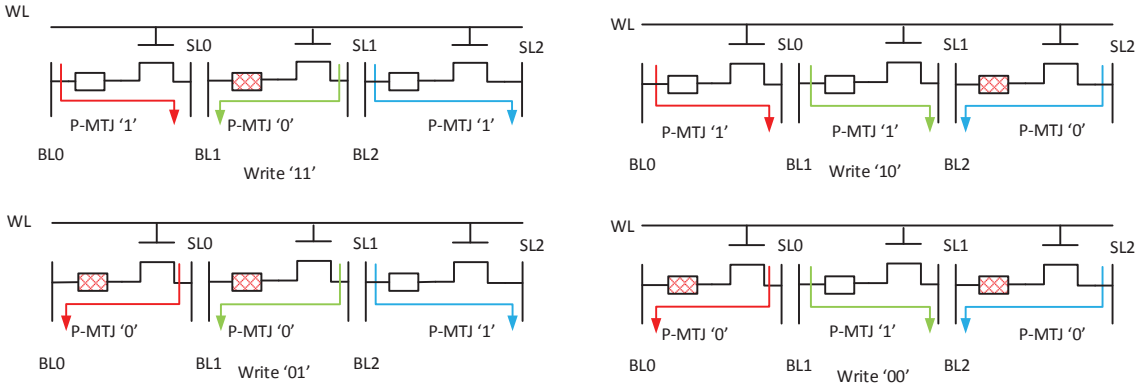

Figure 5: Illustration of write operations of 3T-3MTJ strucutre.

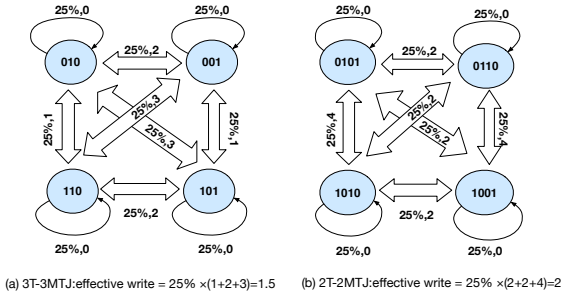

Figure 6: Illustration of the MTJ flipping situation.

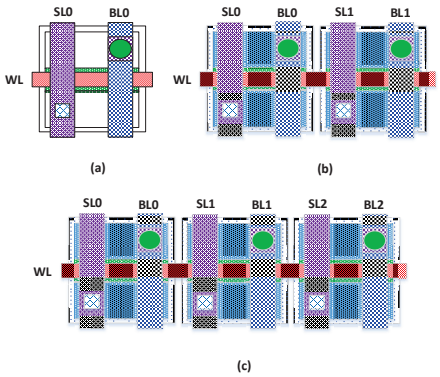

Figure 7: Layout of 1T-1MTJ, 2T-2MTJ and 3T-3MTJ cell structures. (a) 1T-1MTJ cell layout. (b) 2T-2MTJ cell layout. (c) 3T-3MTJ cell layout.

4. EXPERIMENTAL RESULTS

Table 2: Design parameters of three cell structures for experiments.

Parameters	1T-1MTJ	2T-2MTJ	3T-3MTJ
Technology node (nm)	45	45	45
R_p ($k\Omega$)	84	84	84
R_{ap} ($k\Omega$)	224	224	224
Read voltage (V)	1.2	1.2	1.2
Read current (μA)	64	36	54
Write current (μA)	100	200	150
Write latency (ns)	2	2	2
Write energy (pJ)	0.24	0.48	0.36

In the paper, p-MTJ model developed by [16] and PTM transistor model from [1] is adopted to perform hybrid CMOS /MTJ circuit simulations. Design parameters of these cell structures are derived through the cell level HSPICE simulations as shown in Table 2. Note that the read current of the 1T-1MTJ is larger than the other two cell structures because the read margin of the 1T-1MTJ is smaller, and a larger sensing current is required to guarantee sensing reliability. The read current of the 3T-3MTJ cell is slightly larger than the 2T-2MTJ since currents are flowing 3 MTJ branches for the 3T-3MTJ while 2T-2MTJ only has 2 MTJ branches. Based on these design parameters, the comparisons of three cell designs are discussed below in terms of area, performance, energy consumption and reliability.

4.1 Area analysis

Fig. 7 illustrates layouts of three cell structures. To save area, MTJ is commonly placed over the access transistor. Since the area of MTJ is much smaller than that of the access transistor, the transistor area determines the whole cell area. According to cell layout, the 1T-1MTJ cell area is $21F^2$, where F is the feature size of technology node. The cell area is coincide with measurement from [17]. For the 2T-2MTJ cell, it includes one extra transistor and MTJ area. So its cell area is $42F^2$. The 3T-3MTJ cell area is $63F^2$ because one more transistor and MTJ are introduced in the cell as shown in Fig. 7.

Then, NVSim [5] is used to perform array-level area estimation. For the 3T-3MTJ case, which uses two-stage sensing scheme, NVSim is revised to accommodate area overhead of the second sense amplifier. The area comparisons of three different structures are plotted in Fig. 8. The horizontal axis denotes STT-RAM capacity under consideration ranging from 2MB to 32MB. The vertical axis represents area overheads in mm^2. As the figure shown, 2T-2MTJ array area is on average about 30% larger than that of 3T-3MTJ with the same capacity . The obvious area improvement of the 3T-3MTJ cell structure origins from two factors. First, cell area per bit of 3T-3MTJ ($63F^2/2 = 31.5F^2$) is much smaller than 2T-2MTJ ($42F^2$). Second, since the SA can

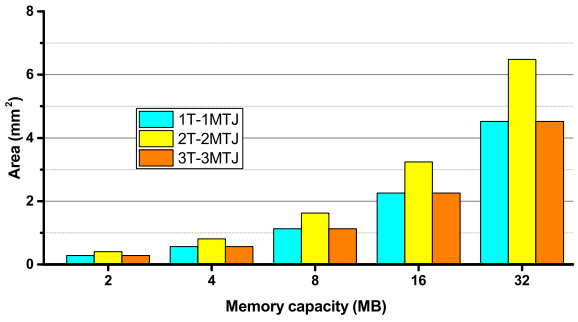

Figure 8: Area comparisons of three different cell structures with different storage capacities.

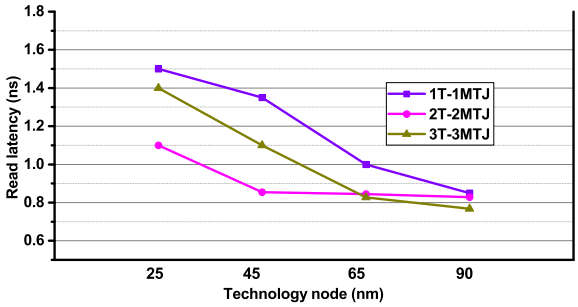

Figure 9: Read performance comparisons of three cell designs.

be shared among multiple memory cell columns, area overhead of the extra SA in the two-stage sensing circuit can be amortized.

4.2 Performance analysis

To compare read performance of different cell structures, a 2KB array is constructed as shown in the left part of Fig. 2. It is composed of 4 subarrays, and the capacity of each subarray is $256 \times 8b$. Except for 45nm technology node, other three technology nodes, i.e., 90nm, 65nm, 25nm are also considered to investigate scalabilities of different cell designs. The read latencies obtained by HSPICE simulation are plotted in Fig. 9. The access latency increases as technology node shrinks due to decreased sensing current. The 2T-2MTJ cell structure has the best access performance. Since 3T-3MTJ needs two stage sensing for complementary data patterns, the sense latency is slightly worse than that of 2T-2MTJ. 1T-1MTJ has the largest read latency because of reduced read margin. Although the read performance of 3T-3MTJ is not the best, the difference in timing is just a few hundreds of pico-seconds which can be fitted in one clock cycle in modern processor design with only negligible system performance degradation.

Considering write operation, three cell structures all use one access transistor to drive the write current into each MTJ. Therefore, MTJs in 2T-2MTJ and 3T-3MTJ can be switched simultaneously, and write latencies of all three types of cells are roughly the same, which are basically determined by the switching time of a single MTJ.

4.3 Energy analysis

Read energy comparisons are shown in Fig. 10(a). It can

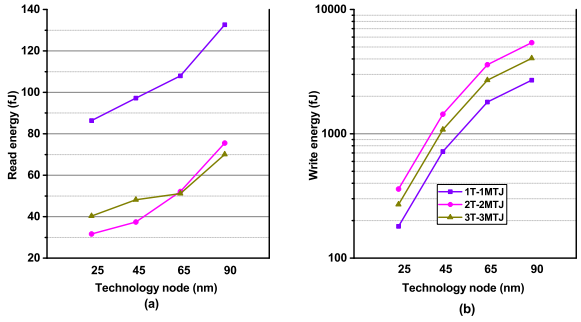

Figure 10: Energy comparisons of three cell structures at 25, 45, 65 and 90nm technology nodes. (a) Read energy. (b) Write energy.

be observed that the 3T-3MTJ cell has comparable read energy with the 2T-2MTJ, and is only 47% of that of the 1T-1MTJ on average. As technology node scales down, the 3T-3MTJ cell design can get more read energy savings because they use differential sensing scheme and read current can be reduced dramatically. Fig. 10(b) illustrates write energy comparisons. As technology node scales down, write energy reduced as well since the scaled MTJ can be switched by smaller current amplitude. Compared to 2T-2MTJ design, write energy is reduced by more than 25% when using 3T-3MTJ. One reason is that 3 MTJs are written to store 2-bit data in the 3T-3MTJ cell structure while 2 MTJs need to be written to store only 1-bit data in the 2T-2MTJ cell design. The write operations also cause different number of MTJ flips for the 2T-2MTJ and 3T-3MTJ as explained in Section 3.

4.4 Reliability analysis

We analyze the read and write reliability issues next.

4.4.1 Read disturbance and error rate

Table 3: Summary of device parameters

Parameters	Mean	Std.Dev
Channel Length	45nm	5%
Channel Width	135nm	5%
Threshold Voltage	0.6V	20mV
Low Resistance	$0.5K\Omega$	5%
High Resistance	$1.2K\Omega$	5%

The read disturbance and the read decision failure rate are two main sources of read reliability [22, 20]. Read disturbance is flipping MTJ state accidentally during the read operation. So it is a kind of destructive read operation [3]. On the other hand, read decision failure rate is caused by the process variations of memory devices and sensing margin limitation of sense amplifier. Read decision failure does not change the data stored in MTJ, however. As technology scales down, the gap between read and write current becomes narrower. In order to mitigate read disturbance,

[3]Note that read disturbance only occurs in one direction, i.e., either switching MTJ from AP to P or from P to AP, but can not make both happen.

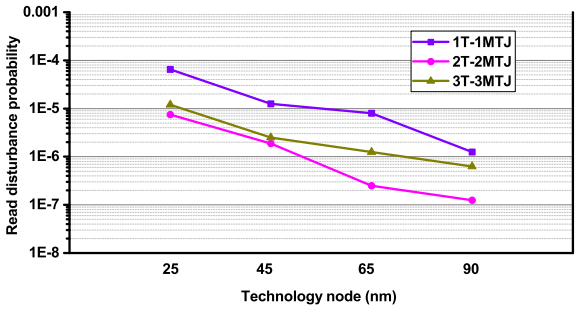

Figure 11: Read disturbance probabilities of three cell structures under 25, 45, 65, 90nm technology nodes.

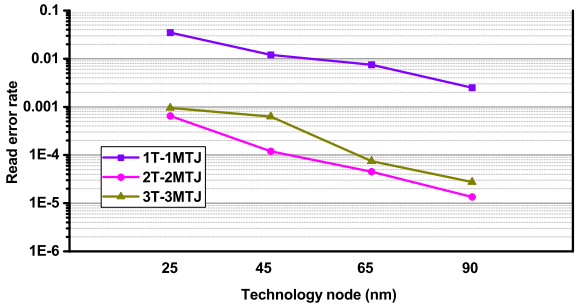

Figure 12: Read error rates of three cell structures with different technology nodes.

the read current has to be reduced, which shrinks the sensing margin and increases read decision failure rate inversely. The compromise between decision failure and read disturbance has made it even more challenging for deeply scaled STT-RAM design.

Assume that relevant device parameters follow Gaussian distributions. According to process variation settings in Table 3, we perform 1000 Monte Carlo simulations using HSPICE to obtain the read disturbance probabilities under different technology nodes. The result is shown in Fig. 11. Note that there are two bits stored in 3T-3MTJ cell. Since read disturbance can only occur in one direction, the read disturbance probability depends on data pattern stored in the cell. We assume that the read disturbance only affects data bit '1' and each data pattern occurs with the same probability. '00' would not be influenced since '0' is the safe data bit. '01' and '10' could have one bit flipped. '11' is the worst case which may have two bits flipped. Considering all above situations, we can get the read disturbance through equation 2.

$$P_{readdis} = \sum_{i=0}^{i=n} O_{state_i} \cdot Q_{mtjflip_i} \quad (2)$$

where P_{readis} represents the read disturbance probability of 2-bit data. O_{state_i} stands for the probability of current state. $Q_{mtjflip_i}$ is the probability that MTJ flips in the current state. As shown in Fig.11, it can be observed that 3T-3MTJ cell's read disturbance probabilities are similar to 2T-2MTJ case and much smaller than 1T-1MTJ design.

Considering read decision failure, it is mainly caused by the process variation and the resolution of sense amplifier.

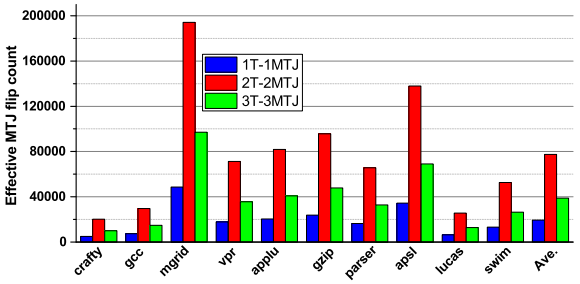

Figure 13: Effective flip count comparisons among different cell structures when 1MB STT-RAM is adopted as L2 cache.

Table 4: Parameters of write traffic simulation.

CPU	Alpha 21264 2GHz
Predictor	Bimodal predictor, BTB with 2-bit counter
IFQ Size/LSQ Size	4/8
SRAM L1 D$/I$	32KB/32KB, 64B block size, 4-way/1-way associative, LRU replacement
Unified STT-RAM L2$	1MB, 64B block size, 4-way associative, LRU replacement

To evaluate decision failure rate, 1000 Monte Carlo simulations are performed with process variation parameters listed in Table. 3. Fig. 12 shows decision failure rates of the 1T-1MTJ, 2T-2MTJ and 3T-3MTJ cells with different technology nodes. It can be observed that the 2T-2MTJ cell and 3T-3MTJ cell have the similar read error rate, which is reduced by more than one order of magnitude compared to that of the 1T-1MTJ case.

The reason of reduced read disturbance and read decision error rate of the 3T-3MTJ and 2T-2MTJ compared to 1T-1MTJ can be explained as follows. In the 1T-1MTJ cell, a reference cell is needed to generate one of the input of sense amplifier. The resistance of reference cell is typically designed as the average value of R_p and R_{ap}. Thus, the sensing signal distribution overlap between data branch and reference branch is more significant, and process variation would affect the sensing accuracy more severely for 1T-1MTJ. In contrast, the sensing margin of 3T3MTJ is doubled by sensing the difference of R_p and R_{ap}, and both read current and error rate can be reduced effectively

4.4.2 *Reliability issue due to write activities*

According to the paper [2], it is known that the STT-RAM endurance can be affected by the write current amplitude and write frequencies. Since the write current injected to each MTJ is the same for three types of cell designs, we will investigate STT-RAM lifetime from the write traffic perspective. To evaluate the write activities when different cell structure are employed, we simulate cache access patterns on an alpha processor using SimpleScalar [3] with 64KB SRAM L1 cache (instruction cache + data cache) and 1MB STT-

RAM L2 cache. The architecture configurations are listed in Table. 4. The write traffics to L2 cache are obtained through simulations on selective SPEC 2000 benchmarks as shown in Fig. 13. For each benchmark, we fast forward 100 million instructions for cache warm-up, and execute another 100 million instructions for write traffic extraction. Additionally, the write traffic is transformed to effective MTJ flip count according to the definition mentioned in Section 3. As shown in the figure, it can be observed that the 3T-3MTJ cell suffers from much less effective MTJ flips compared to 2T-2MTJ cell. Therefore, the 3T-3MTJ cell is more friendly to the memory's lifetime.

5. CONCLUSIONS

In this paper, we propose a novel 3T-3MTJ cell structure with advanced perpendicular MTJ achieving better tradeoff between area, read/write performance and energy consumption. One 3T-3MTJ cell can store 2-bit data via the different resistance state combinations of three MTJs. The peripheral circuit and the adaptive memory access control scheme are also proposed to support the 3T-3MTJ read and write operations considering process variation effect. The simulation results performed on circuit-level and architecture-level indicate that the 3T-3MTJ cell has better read performance and reliability than 1T-1MTJ cell since differential sensing scheme accelerates the read operation and reduces the read error rate. Experiment results from cell layout and bank level simulations show that the storage density of the 3T-3MTJ improves 30% compared to the 2T-2MTJ counterpart. The energy consumption of the 3T-3MTJ cell approaches that of the 2T-2MTJ, and much smaller than the 1T-1MTJ cell structure. The reliability evaluation also validates that the 3T-3MTJ cell structure is suitable for the deeply scaled STT-RAM design.

6. ACKNOWLEDGMENT

Linuo Xue and Yuan Xie were supported in part by Qualcomm gift, NSF 1213052,1533933,1461698 and 1500848. Yuanqing Cheng and Jianlei Yang were supported in part by China NSFC No.61401008 and Beijing NSF No.4154076.

7. REFERENCES

[1] Predictive technology model, http://ptm.asu.edu.

[2] S. Amara-Dababi et al. Modelling of time-dependent dielectric barrier breakdown mechanisms in MgO-based magnetic tunnel junctions. *Journal of Physics D: Applied Physics*, 45(29):295002, 2012.

[3] T. Austin et al. Simplescalar: an infrastructure for computer system modeling. *Computer*, 35(2):59–67, 2002.

[4] Y. Chen et al. Design margin exploration of spin-transfer torque RAM(STT-RAM) in scaled technologies. *IEEE Transactions on VLSI Systems*, 18(12):1724–1734, 2010.

[5] X. Dong et al. Nvsim: A circuit-level performance, energy, and area model for emerging nonvolatile memory. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 31(7):994–1007, 2012.

[6] S. Ikeda et al. A perpendicular-anisotropy CoFeB-MgO magnetic tunnel junction. *Nature Materials*, 9(9):721–724, 2010.

[7] A. Kawasumi et al. Circuit techniques in realizing voltage-generator-less STT MRAM suitable for normally-off-type non-volatile l2 cache memory. In *IMW*, 2013.

[8] C. Kim et al. A covalent-bonded cross-coupled current-mode sense amplifier for STT-MRAM with 1T1MTJ common source-line structure array. In *ISSCC*, 2015.

[9] J. Kim et al. A novel sensing circuit for deep submicron spin transfer torque MRAM (STT-MRAM). *IEEE Transactions on VLSI Systems*, 20(1):181–186, 2012.

[10] W. Kim et al. Extended scalability of perpendicular STT-MRAM towards sub-20nm MTJ node. In *IEDM*, 2011.

[11] E. Kultursay et al. Evaluating STT-RAM as an energy-efficient main memory alternative. In *ISPASS*, 2013.

[12] Lin et al. 45nm low power CMOS logic compatible embedded STT MRAM utilizing a reverse-connection 1T/1MTJ cell. In *IEDM*, 2009.

[13] Noguchi et al. A 250-MHz 256b-I/O 1-Mb STT-MRAM with advanced perpendicular MTJ based dual cell for nonvolatile magnetic caches to reduce active power of processors. In *VLSIC*, 2013.

[14] H. Noguchi et al. Variable nonvolatile memory arrays for adaptive computing systems. In *IEDM*, 2013.

[15] T. Ohsawa et al. 1Mb 4T-2MTJ nonvolatile STT-RAM for embedded memories using 32b fine-grained power gating technique with 1.0ns/200ps wake-up/power-off times. In *VLSIC*, 2012.

[16] G. Panagopoulos et al. A framework for simulating hybrid MTJ/CMOS circuits: Atoms to system approach. In *DATE*, 2012.

[17] C. Park et al. Systematic optimization of 1 Gbit perpendicular magnetic tunnel junction arrays for 28 nm embedded STT-MRAM and beyond. In *IEDM*, 2015.

[18] S. Pei et al. A high-precision on-chip path delay measurement architecture. *Very Large Scale Integration Systems, IEEE Transactions on*, 20(9):1565–1577, 2012.

[19] A. Raychowdhury et al. Design space and scalability exploration of 1T-1STT MTJ memory arrays in the presence of variability and disturbances. In *IEDM*, 2009.

[20] R. Wang et al. Selective restore: an energy efficient read disturbance mitigation scheme for future STT-MRAM. In *DAC*, 2015.

[21] H.-C. Yu et al. Cycling endurance optimization scheme for 1Mb STT-MRAM in 40nm technology. In *ISSCC*, 2013.

[22] Y. Zhang et al. Adams: Asymmetric differential STT-RAM cell structure for reliable and high-performance applications. In *ICCAD*, 2013.

[23] W. Zhao et al. High speed, high stability and low power sensing amplifier for MTJ/CMOS hybrid logic circuits. *Magnetics, IEEE Transactions on*, 45(10):3784–3787, 2009.