Device-Architecture Co-Optimization of STT-RAM Based Memory for Low Power Embedded Systems

Cong Xu†, Dimin Niu†, Xiaochun Zhu†‡, Seung H. Kang†‡, Matt Nowak†‡, Yuan Xie†
†Department of Computer Science and Engineering, Pennsylvania State University
‡Emerging Memory Technology Group, Qualcomm Incorpotation

***Abstract*—Spin-transfer torque random access memory (STT-RAM) is a fast, scalable, durable non-volatile memory which can be embedded into standard CMOS process. A wide range of write speeds from $1ns$ to $100ns$ have been reported for STT-RAM. The switching current of magnetic tunnel junction (MTJ) (which is the storage element of STT-RAM) is inversely proportional to the write pulse width. In this work, we propose a methodology to design STT-RAM for different optimization goals such as read performance, write performance and write energy by leveraging the trade-off between write current and write time of MTJ. We take the typical in-plane MTJ and advanced perpendicular MTJ (PMTJ) as our optimization targets. Our study shows that reducing write pulse width will harm read latency and energy. It is observed that "sweet spots" of write pulse width which minimize the write energy or write latency of STT-RAM caches may exist. The optimal write pulse width depends on MTJ specifications, STT-RAM capacity and I/O width. The simulation results indicate that by utilizing PMTJ, the optimized STT-RAM can compete against SRAM and DRAM as universal memory replacement in low power embedded systems.**[1]

I. Introduction

The goal of universal memory is to combine the best attributes such as fast random access, high storage density and non-volatility into one memory technology. Industry and academia are seeking solutions in some of the emerging non-volatile memory (NVM) technologies, such as spin-torque-transfer random-access memory (STT-RAM, or MRAM), phase-change random-access memory (PCRAM), and resistive random-access memory (ReRAM). Universal memory, as indicated by its name, should work across multiple layers of the memory hierarchy. More importantly, it is expected to provide a large design space which has the potential to replace both performance-critical caches and cost-optimized secondary storage. Among the emerging NVM technologies, STT-RAM is one of the most promising candidates that has the potential to meet all the requirement of universal memory [1], [2]. Realization of STT-RAM based universal memory seems to be more practical in low power embedded systems rather than high performance computer thanks to the less demanding performance requirement and more constrains on power budget and battery life.

STT-RAM was invented as the second generation of Magnetic RAM (MRAM) [3] to conquer the two major problems for conventional MRAM: high write energy and poor scalability. Conventional MRAM uses the magnetic fields produced by electrical currents to change the resistance of the MTJ and the required current increases as technology scales down. However, in STT-RAM, by applying the spin polarized current through the MTJ element to switch the memory states, the required switching current decreases as technology scales down. STT-RAM is projected to scale beyond 20nm technology node [4]. To further reduce switching current and switching time, perpendicular MTJs for STT-RAM were developed [5]–[9] to achieve low switching current while maintaining high thermal stability for non-volatility of STT-RAM. To the best of our knowledge, we are the first to explore the design space of such PMTJ-based STT-RAM in architecture-level research.

Experiments have been performed in device-level research in order to operate an MTJ at minimum energy or energy-delay-product (EDP) by applying varied write pulse width on the MTJ. However, the optimal operating write pulse width from cell-level point of view is not necessarily the best operating point from system-level point of view. Normally, an STT-RAM memory cell consists of an access transistor in series with an MTJ. Short write pulses induce large switching currents requiring large access transistors for providing enough driving current, which consequently brings more circuit design challenges for STT-RAM macros. Specifically, reducing write pulse width aggressively will incur penalty on area, latency, dynamic energy and leakage power of both access transistors and peripheral circuitry. Thus it's imperative to offer a methodology for system-level analysis of the memory macro to quantitatively address the trade-off of all the metrics of STT-RAM.

In this work, we implemented a system-level performance, energy and area model to estimate the impact of different write pulse widths on the STT-RAM macro design. We then develop a detailed device-architecture co-optimization methodology to design STT-RAM macros with different optimization goals such as area, read latency/energy, write latency/energy by leveraging the inherent trade-off of write current and write time of MTJ. The goal of this work is to provide design implications of STT-RAM based cache or main memory with different capacities and different optimization goals.

The rest of the paper is organized as follows: Section II presents related work. Section III discusses the basics of STT-RAM and inherent trade-offs of write current and write time of MTJ. Section IV provides a system-level modeling of STT-RAM macro. Section V analyzes the optimization methodology of STT-RAM with different optimization goals. Section VI shows a case study of replacing L1 cache with STT-RAM in an embedded system. Section VII concludes our work.

[1]This work is supported in part by Qualcomm, SRC grant, NSF 1147388, 0903432 and by DoE under Award Number DE-SC0005026.

978-1-4577-1400-9/11/$26.00 ©2011 IEEE

II. Related Work

Several modeling tools have been developed to perform system-level simulation for different memory technologies. The most authoritative tool called CACTI [10] has been widely used to estimate the speed, power, and area of SRAM and DRAM. In addition, CACTI has also been extended to evaluate the performance, power, and area for STT-RAM [11], PCRAM [12], NAND flash [13], and ReRAM [14]. However, fixed write pulse width were assumed in all the the mentioned work. In our work we developed a system-level modeling of STT-RAM with varied write pulse width coupling corresponding write current and integrated the model in an in-house estimation tool, which is a circuit-level performance, energy, and area simulator based on CACTI for emerging non-volatile memories.

There have been several work on design methodology for STT-RAM from both circuit and architecture perspectives. Li *et al.* [15] developed a physics-based MTJ model and their analysis results showed that the sizing of access NMOS transistor has critical impact on the stability and the density of STT-RAM. Chatterjee *et al.* [16] had a more thorough study on co-designing the sizing of the access transistor and operating voltage to achieve minimum energy dissipation. Moreover, Smullen *et al.* [17] illustrated STT-RAM cell design for optimizing read latency and write latency separately in the presence of clock cycles. Similarly, Xu *et al.* [18] quantitatively analyzed the impact of write latency and read latency trade-off of STT-RAM on overall system performance. However, most of these work were used STT-RAM as a last-level cache replacement and few of them gave a comprehensive study on how to design a STT-RAM macro with minimum write latency or write energy. This paper is targeting at evaluating low-energy STT-RAM as universal memory replacement in low power embedded systems.

The contributions of this work are listed as follows,

- We propose a detailed analysis of the impact of write pulse width on area, read latency/energy, write latency/energy of STT-RAM with architectural configurations. Our results indicate that the write pulse width optimizing STT-RAM write energy is a function of capacity and I/O width.
- To the best of our knowledge, we are the first to explore the design space of STT-RAM using advanced perpendicular MTJs. Simulation results show that by device-architecture co-optimized PMTJ-based STT-RAM is a very promising candidate as both L1 cache and main memory replacement in low power embedded system.

III. Preliminary

A. STT-RAM Cell Basics

MTJ, the key component in a STT-RAM cell, is used to store bit information by its two different resistance states. As shown in Figure 1, an conventional MTJ has two ferromagnetic layers with fixed magnetization direction in one layer (reference layer) and free rotate magnetization direction in the other layer (free layer). The magnetization direction of the free layer can be changed by passing a large enough current through the MTJ. If the magnetization directions of the reference layer and free layer are parallel, MTJ is in low resistance state (LRS); if the magnetization directions of the two layers are anti-parallel, MTJ is in high resistance state (HRS).

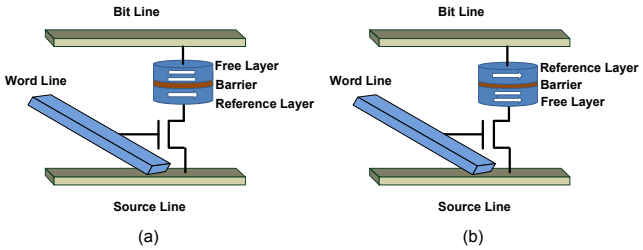

Fig. 1. Demonstration of a STT-RAM cell: (a) Conventional connection scheme; (b) Reverse connection scheme.

As shown in Fig. 1, there are two possible schemes to connect an MTJ to access NMOS transistor. Conventionally, the free layer of MTJ is connected to bitline (BL). In this scheme, when writing LRS into STT-RAM cells, positive voltage difference is established between BL and SL and the anti-parallel to parallel switching current ($I_c(AP \rightarrow P)$) is required; when writing HRS state, negative voltage difference is established between BL and SL and the parallel to anti-parallel switching current ($I_c(P \rightarrow AP)$) is required. Another scheme was proposed in [19] by connecting the free layer of MTJ to the drain of NMOS instead of BL. It's mentioned in [19]–[21]that $I_c(P \rightarrow AP)$ is normally 20% to 50% larger than $I_c(AP \rightarrow P)$ due to the inherent torque asymmetry of MTJ. But the SL-to-BL driving current is much smaller than the BL-to-SL driving current under the same voltage bias. Thus the reserve connection scheme can relax the sizing requirement on the access transistor, which results in more compact STT-RAM cell size. However, device-level efforts have been put to improve the asymmetry of switching characteristics of an MTJ. And $I_c(AP \rightarrow P)$ slightly larger than $I_c(P \rightarrow AP)$ was even demonstrated in [22]. In our work, we always choose the MTJ connection scheme which requires the least demanding access transistor sizing.

Another important metric for an MTJ is the tunnel magnetoresistance (TMR) ratio which is defined as,

$$TMR = \frac{R_{AP} - R_P}{R_P} \tag{1}$$

where R_{AP} is the electrical resistance in the anti-parallel state, whereas R_P is the resistance in the parallel state. A large TMR means big gap between low resistance state (LRS) and high resistance state (HRS), which could essentially bring faster read sensing latency or relax constraints for sense amplifier design. It's also critical to introduce an equivalent metric for an STT-RAM cell which contains both MTJ and access transistor. Similarly the cell TMR (CTMR) is defined as,

$$CTMR = \frac{R_{cell,AP} - R_{cell,P}}{R_{cell,P}} \tag{2}$$

where $R_{cell,AP}$ is the total cell resistance when the MTJ is in the anti-parallel state, whereas $R_{cell,P}$ is the total cell resistance when the MTJ is in the parallel state. CTMR can

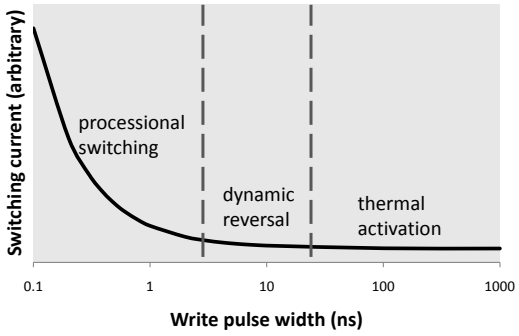

Fig. 2. Demonstration of three switching phases: thermal activation, dynamic reversal and precessional switching.

be expressed by another equation,

$$CTMR = \frac{I_P - I_{AP}}{I_{AP}} \tag{3}$$

where I_{AP} and I_P are the currents for reading "1" state and "0" state. If we ignore the resistance difference of the access transistor for reading "1" state and "0" state. CTMR can be interpreted as,

$$CTMR = \frac{R_{AP} - R_P}{R_P + R_{NMOS}} = \frac{R_{AP} - R_P}{R_P + \frac{C}{W}} \tag{4}$$

where R_{NMOS} is the equivalent resistance of access NMOS transistor and W is the transistor width, C is a constant dependent on the threshold voltage of the NMOS and the wordline voltage applied on it. From equation 4 we can conclude that sizing up the access transistor will make CTMR closer to the inherent TMR ratio of MTJ.

B. Write current versus write pulse width trade-off

The current amplitude required to reverse the direction of the free ferromagnetic layer is determined by a lot of factors such as material property, device geometry and importantly the write pulse duration. Generally, the longer the write pulse is applied, the less the switching current is needed to switch the MTJ state. Three distinct switching modes were identified [23] according to the operating range of switching pulse width τ: thermal activation ($\tau > 20ns$), processional switching ($\tau < 3ns$) and dynamic reversal ($3ns < \tau < 20ns$).

The relationship between switching current density J_c and write pulse width τ was characterized by an analytical model in [24]. The equations are listed as follows,

$$J_{c,TA}(\tau) = J_{c0}\{1 - (\frac{k_B T}{E_b}) ln(\frac{\tau}{\tau_0})\} \tag{5}$$

$$J_{c,PS}(\tau) = J_{c0} + \frac{C}{\tau^{\gamma}} \tag{6}$$

$$J_{c,DR}(\tau) = \frac{J_{c,TA}(\tau) + J_{c,PS}(\tau) e^{-k(\tau - \tau_c)}}{1 + e^{-k(\tau - \tau_c)}} \tag{7}$$

where $J_{c,TA}$, $J_{c,PS}$, $J_{c,DR}$ are the switching current densities for thermal activation, processional switching and dynamic reversal respectively. J_{c0} is the critical switching current density, k_B is the Boltzmann constant, T is the temperature, E_b is the thermal barrier, and τ_0 is inverse of the attempt frequency. C, γ, k, and τ_c are fitting constants. Based on the observation from Fig. 2 and analysis of the analytical model, we found very different switching characteristics in the three switching modes. For example, in thermal activation mode, the required switching current increases very slowly even we decrease the write pulse width by orders of magnitude, thus short write pulse width is more favorable in this regime because reducing write pulse can reduce both write latency and energy without much penalty on read latency and energy. While in processional switching regime, write current goes up rapidly if we further reduce write pulse width, therefore minimum write energy of the MTJ is achieved at some particular write pulse width in this regime. Consequently, this paper will focus on the exploration of write pulse width in processional switching and dynamic reversal to optimize MTJ for different design goals.

C. Perpendicular MTJ

A key challenge for MTJ design is to reduce switching current while maintaining sufficiently high thermal stability in order not to affect data retention time and write/read errors. The conventional in-plane MTJ critical switching current I_{c0} divided by the thermal barrier Δ can be expressed as in [7],

$$\frac{I_{c0}}{\Delta} = \frac{\alpha}{\eta} \times (1 + \frac{H_d}{2H_k}) \tag{8}$$

where α is the damping constant, η is the STT efficiency, H_d is the out of plane demagnetization field, and H_k is the in-plane anisotropy field dominated by the shape anisotropy. The typical value of $H_d/2H_k$ is about 20-150 [2]. From Equation 8 we can see that the magnetization has to overcome a very large out-of-plane demagnetizing field before it can switch to the opposite direction. However, only H_k not H_d will contribute to thermal stability [7]. Perpendicular MTJ were investigated as a promising solution [5], [7]–[9] as the critical switching current of PMTJ can be described as,

$$\frac{I_{c0}}{\Delta} = \frac{\alpha}{\eta} \tag{9}$$

Therefore, PMTJ can have much smaller switching current than in-plane devices if the same ratio of α/η can be maintained. Indeed, very low switching current density of MTJ was demonstrated while maintaining high enough thermal stability factor [8]. There are some issues to be solved for PMTJ such as degraded compatibility with CMOS process, relative large damping constant, and potential lattice mismatch for high TMR ratio and STT efficiency. In this paper, we take both the near-commercialized in-plane MTJ and advanced PMTJ as our optimization targets.

IV. STT-RAM Macro Modeling

A. Area Modeling

To simulate the performance of STT-RAM macro, it is important to estimate its cell area first. As mentioned before, each STT-RAM cell is composed of one NMOS and one MTJ (1T1J). The NMOS access device is connected in series with the MTJ as shown in Fig. 1. The size of NMOS is constrained by both $I_c(AP \rightarrow P)$ and $I_c(P \rightarrow AP)$, which are inversely proportional to the writing pulse width. In order to estimate the current driving ability of MOSFET devices, a small test circuit

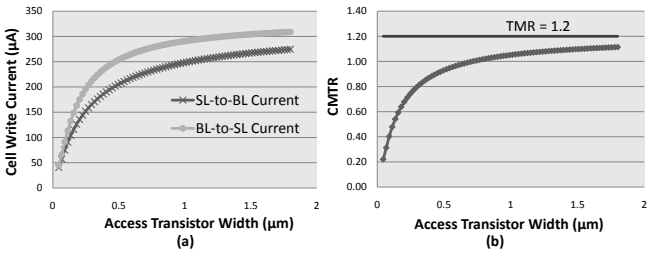

Fig. 3. (a) Driving ability, and (b) Cell TMR of access NMOS transistor.

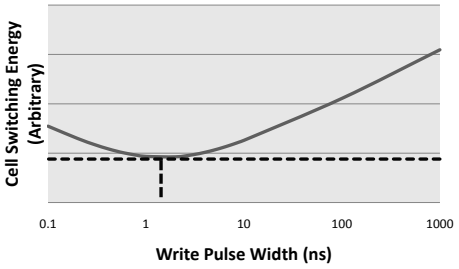

Fig. 4. Switching energy per cell versus write pulse width.

using HSPICE with PTM 45nm HP model [25] is simulated. The BL-to-SL current and SL-to-BL current are obtained by assuming typical TMR (120%) and LRS ($3k\Omega$) value [19] and bursting wordline voltage to be 1.5V (the optimal value is extracted from [16]). As we can see in Fig. 3(a), the SL-to-BL current is always smaller and saturates faster than BL-to-SL current. Such current degradation is related to the voltage drop on MTJ. There is a positive source-to-substrate voltage difference $V_{SB} = I_c \times R$ for the SL-to-BL current (I_c is the current passing through MTJ and R is the resistance of the MTJ). First, it causes the gate-to-source voltage V_{GS} for the SL-to-BL current to be smaller than the gate bias voltage. Second, the body effect of the access transistor degrades the SL-to-BL current further because the threshold voltage is boosted by V_{SB}. In our model, we always choose the connecting scheme which uses BL-to-SL current to match the larger value of $I_c(AP \rightarrow P)$ and $I_c(P \rightarrow AP)$ and SL-to-BL current to match the smaller value of them. The corresponding access transistor width must satisfy the following conditions,

$$I_{BS}(W_{BS}) \geq max(I_c(AP \rightarrow P), I_c(P \rightarrow AP)) \quad (10)$$

$$I_{SB}(W_{SB}) \geq min(I_c(AP \rightarrow P), I_c(P \rightarrow AP)) \quad (11)$$

where $I_{BS}(W_{BS})$ is the current from BL to SL with transistor width W_{BS} and $I_{SB}(W_{SB})$ is the current from SL to BL with transistor width W_{SB}.

The relationship between access transistor width and CTMR defined in Section III-A was also simulated. As can be seen in Fig. 3(b), a larger access transistor can improve CTMR closer to the inherent TMR ratio of MTJ. It's necessary to set a lower bound $CTMR_{min}$ for CTMR to guarantee the correctness of read operation. Thus the transistor width must be large enough to satisfy the minimum CTMR requirement,

$$CTMR(W_{CTMR}) \geq CTMR_{min} \quad (12)$$

Finally we will choose a transistor width W which satisfies all the above requirements,

$$W = max(W_{BS}, W_{SB}, W_{CTMR}) \quad (13)$$

To achieve high cell density, we model the STT-RAM cell area by referring to DRAM design rules [26]. As a result, the cell size of a STT-RAM cell is calculated as below,

$$\text{Area}_{\text{cell}} = 3\,(W/L + 1)(F^2) \quad (14)$$

B. Data sensing modeling

Three sensing modes were proposed in [14] to sense resistance-based NVMs including STT-RAM, PCRAM and ReRAM: current sensing, current-in voltage sensing, and voltage-divider sensing. There are trade-offs between the area, latency and energy of the three sensing modes. For example, current sensing is the fastest approach [27] to sense the memory state if the number of cells per bitline is large enough, while voltage-divider sensing is the second fastest and the current-in voltage sensing is the slowest. On the other hand, current-in voltage sensing has the best area efficiency which is defined as the ratio of NVM cell area to the prototype area while current sensing has the worst area efficiency. For current sensing modeling, we adapt the current-voltage converter and sense amplifier design discussed in [11]. The current-voltage converter in our current sensing scheme is actually the first-level sense amplifier, and the conventional voltage sense amplifier is still kept as the final stage of the sensing scheme. In order to maintain low rate of read disturbance, it's necessary to reduce read current when smaller switching current is used. The reduced read current will have an impact on the latency of the current-voltage converter and sense amplifier. Therefore we use HSPICE to simulate the latency, energy and leakage of the two-stage sense amplifier with different read current and build a look-up table in our model.

C. Cell Switching Modeling

A dynamic MTJ switching model was developed in [15] with consideration of the switching phenomenon involving magnetoresistive effects, which can not be estimated only by RC analysis. However, this work is focusing on static analysis of STT-RAM and our architecture-level tool does not model the dynamic behavior during the switching of the cell state. Thus we simply calculate switching energy (i.e. cell write energy) by using Joule's first law that is,

$$\text{Energy}_{\text{cell\_switching}} = I_c^2 R\tau \quad (15)$$

in which the resistance value R can be the equivalent resistance of the corresponding LRS or HRS (i.e. R_P or R_{AP}). Taking the coupled I_c and τ as input for Equation 15, we can easily get the relation between the energy of switching one cell and write pulse width. From Fig. 4 we can see that minimum switching energy per cell is achieved at write pulse width $\tau_{min\_cell\_energy}$ in the range of processional switching mode. However, the optimal operating write pulse width from cell-level point of view is not necessarily the best operating point from system-level point of view because the effect of access transistor and peripheral circuitry has not been considered.

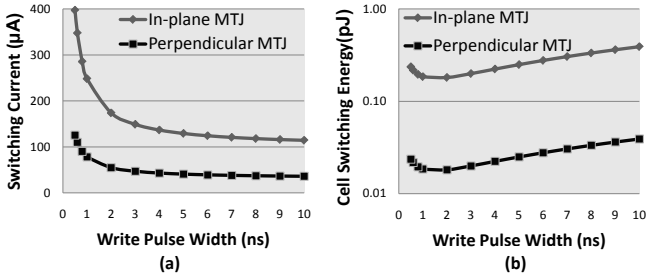

Fig. 5. Representative data curves of in-plane and perpendicular MTJs: (a)Switching current; (b) Cell switching energy.

TABLE I
SIMULATION PARAMETERS

Optimization target	in-plane MTJ	PMTJ
Write pulse width operating range	$2ns - 10ns$	$0.4ns - 10ns$
$\tau_{min\_cell\_energy}$	$1.3ns$	$1.8ns$
TMR	120%	
LRS Resistance	$3k\Omega$	
$CTMR_{min}$	60%	
Macro Capacity	$16KB - 4MB$	
Technology Node	$45nm$	
I/O Width	$32bits - 512bits$	

V. STT-RAM MACRO DESIGN OPTIMIZATION

In this section we first choose two MTJ species as our optimization targets. After numerous simulations the importance of careful write pulse width selection is revealed for optimizing area, read latency/energy, write latency/energy and leakage power of STT-RAM. Then we will focus on the analysis of write energy optimization, which is both device and architecture dependent. Finally we will combine the improved architectural design together with the write pulse width optimization to explore a full design space for a $64MB$ STT-RAM chip.

A. Impact of Write Pulse Width

As discussed in Section III-B, a wide range of coupled switching current and switching time can be operated for STT-RAM cell. In this work, we will focus on processional switching mode and dynamic reversal mode, particularly, for $0.4ns < \tau < 10ns$. We choose two curves which represent the typical switching characteristics of in-plane MTJ [19] and PMTJ [8]. As seen in Fig. 5, PMTJ has remarkable advantages over in-plane MTJ in both switching current and switching energy for any given write pulse width assuming the same ratio of damping constant to STT efficiency. The other simulation parameters are listed in Table III. We assume the same TMR and resistance for in-plane MTJ and PMTJ. For PMTJ, to achieve those parameters of in-plane MTJ needs some device-level efforts as mentioned in Section III-C. The minimum write pulse width $\tau_{min\_cell\_energy}$ for each MTJ species is extracted from Fig. 5(b).

The simulation results of $2MB$ STT-RAM macros with in-plane MTJ and PMTJ are compared with SRAM of the same capacity. Different impacts of write pulse width on area, read latency/energy, write latency/energy and leakage power are demonstrated in Fig. 6. In general, decreasing write pulse width will increase area, read latency/energy and leakage power. Especially reducing write pulse width in processional mode will harm these metrics badly. However, the relation between write pulse width and write latency/enery is non-monotonous. Separate explanations and result analysis are given as follows,

- Area: short write pulse induced large switching current requires large access transistors for providing enough driving current, which leads to both cell and peripheral circuitry (i.e. wordline driver) area penalty. From Fig. 6(a) we can see that STT-RAM generally has area advantage over SRAM. But reducing write pulse aggressively below $2ns$ will result in STT-RAM with in-plane MTJ having larger area than SRAM.
- Read latency: the read timing can be approximately divided into four components: (1) H-tree input/output delay; (2) Decoder + wordline delay; (3) Bitline delay; (4) Sense amplifier delay. (1) is affected because larger area essentially means longer routing distance and interconnection RC delay. Moreover, the increased gate and drain capacitance of larger access transistor will contribute to wordline capacitance and bitline load capacitance, which increase (2) and (3). From Fig. 6(b) we can see that STT-RAM with in-plane MTJ is slightly slower than SRAM mainly because sensing the state of STT-RAM cell takes longer than SRAM. However, read operation of STT-RAM with PMTJ can be faster than SRAM for $\tau > 0.8ns$ due to remarkable area advantage.
- Read energy: it's affected in the similar way as read latency is affected. From Fig. 6(c) we can see that the read energy of STT-RAM with in-plane MTJ is comparable to that of SRAM while PMTJ improves STT-RAM read energy significantly and makes it better than SRAM.
- Write latency: the write timing can be approximately divided into four components: (1) H-tree input latency; (2) Decoder + wordline delay; (3) Write pulse width. (1) and (2) both increase as (3) decreases therefore "sweet spots" may exist, which is approved by Fig. 6(d). The write latency of STT-RAM with in-plane MTJ can no longer be improved when $\tau < 3ns$. While the minimum write latency STT-RAM with PMTJ is achieved at $\tau = 0.8ns$ and the latency value is comparable to SRAM.
- Write energy: it consists of two parts: the energy of cell switching (cell energy) and the energy of the circuitry (peripheral energy) most of which is shared with read operation. When reducing write pulse width from $10ns$ to $\tau_{min\_cell\_energy}$, cell energy decreases while peripheral energy goes up in the same manner with read energy. Later we'll show that the optimal write pulse width τ_{min_energy} for minimum write energy is dependent on MTJ species, STT-RAM capacity and I/O width. We can see from Fig. 6(e) that $\tau_{min\_energy} = 5ns$ for STT-RAM with in-plane MTJ and $\tau_{min\_energy} = 7ns$ for STT-RAM with PMTJ, both of which are $2MB$ STT-RAM with 32-bit I/O width. Moreover, the minimum write energy of STT-RAM with in-plane MTJ is roughly 50% larger than that of SRAM while the energy of STT-RAM with PMTJ is almost half the number of SRAM.

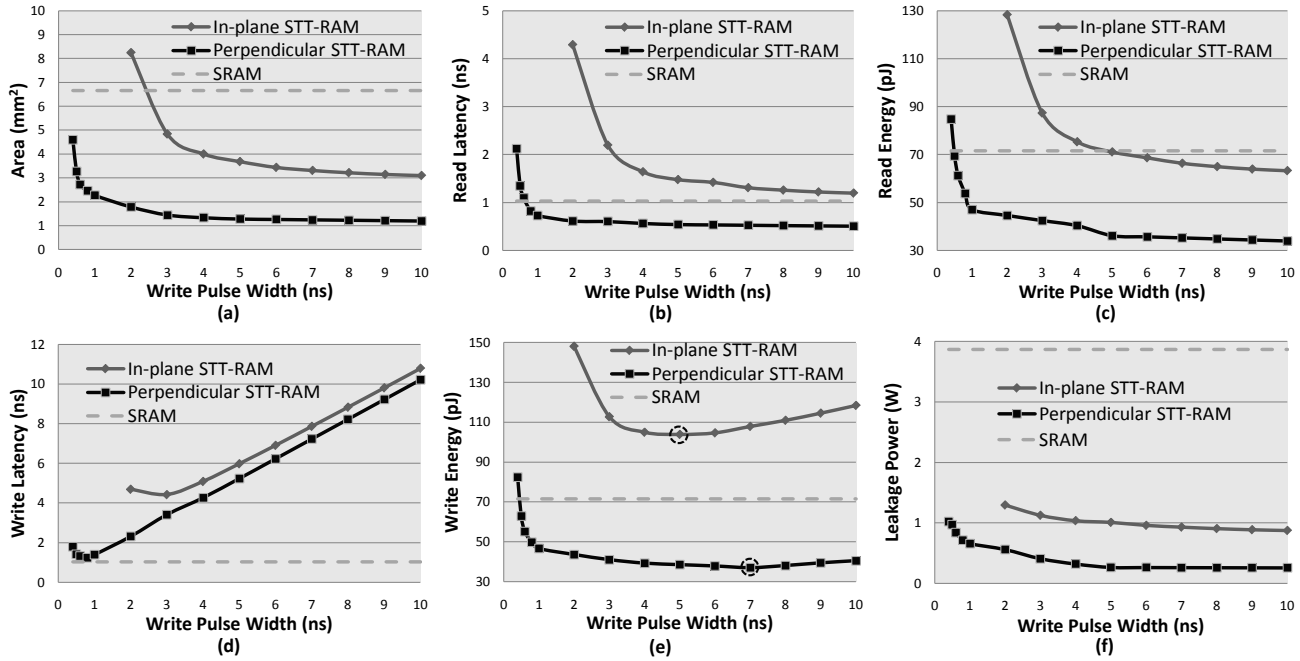

Fig. 6. Metrics of SRAM and STT-RAM built with in-plane and perpendicular MTJs: (a)Area; (b)Read latency; (c)Read energy; (d)Write latency; (e)Write energy; (f)Leakage power.

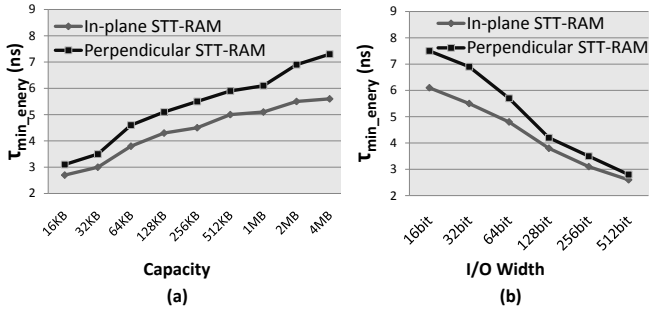

Fig. 7. Dependence of τ_{min_energy} on (a) STT-RAM Macro Capacity, and (b) I/O Width

- Leakage power: leakage power is basically proportional to the sizing of transistor contributing the leakage current. Thus, it increases as the area of peripheral circuity increases because the leakage power for STT-RAM mainly comes from peripheral circuity. From Fig. 6(f), we can see the remarkable advantage of STT-RAM over SRAM no matter what type of MTJ species is used.

B. Write Energy Optimization

In the previous section we note that τ_{min_energy} is different for STT-RAM with in-plane MTJ and PMTJ. Thus, we did more study on the determining factors and found out τ_{min_energy} depends on MTJ species, STT-RAM capacity and I/O width. From the results shown in Fig. 7, we observed that: (1) The optimal write pulse width τ_{min_energy} for STT-RAM with in-plane MTJ is smaller than that of STT-RAM with PMTJ under the same capacity and I/O Width; (2) τ_{min_energy} is a monotonic increasing function of STT-RAM capacity under fixed I/O width; (3) τ_{min_energy} is a monotonic decreasing function of I/O width under fixed capacity. There are primary two reasons for (1): the baseline $\tau_{min\_cell\_energy}$ of in-plane MTJ is smaller than that of PMTJ; proportion of cell energy in PMTJ is much smaller and peripheral energy has more weights on minimizing total energy. (2) is because the proportion of cell energy decreases as capacity increases and peripheral energy begins to decimate at large memory capacity. (3) is due to the similar reason. The conclusions can be viewed as design implications of MTJ device, that is, minimizing the cell switching energy at $\tau_{min\_cell\_energy}$ from cell-level may not be enough, while it's more important to reduce to cell switching energy at τ_{min_energy} for particular capacity and I/O width from system-level point of view.

C. Device-Architecture Co-Optimization

Finally we combine the write pulse width optimization together with other circuit- and architectural- level techniques to design a $64MB$ STT-RAM prototype with 64-bit I/O width under $45nm$ technology node using PMTJ. These optimization choices include: (1) Insert repeaters in interconnection wire to reduce routing delay at the penalty of area and energy; (2) Use partial swing signal for data transfer to reduce energy at the penalty of latency; (3) External sensing scheme with Non-H-Tree routing to reduce chip area; (4) Different buffer design styles for area optimization or latency optimization; and (5) Different sensing schemes for trade-off of area, latency and energy. Table II tabulates the full design spectrum of this chip by listing the details of each design corner. Compared the metrics of STT-RAM chip to those of DRAM in the rightmost column, we can see that replacing DRAM by STT-RAM can greatly reduce energy while maintaining performance competitive to DRAM. This is primarily because of elimination of refresh operation due to inherent non-volatility of STT-RAM.

TABLE II
DEVICE-ARCHITECTURE CO-OPTIMIZATION OF A 45NM 64MB STT-RAM CHIP AND COMPARISON WITH DRAM

	Area opt.	Read latency opt.	Write latency opt.	Read energy opt.	Write energy opt.	Leakage opt.	DRAM
Area (mm^2)	**3.06**	10.7	16.4	5.66	6.22	3.63	8.55
Read latency (ns)	21.8	**3.70**	4.92	9.12	9.57	9.91	6.24
Write latency (ns)	18.6	13.9	**4.01**	15.9	12.3	18.1	6.24
Read energy (nJ)	0.276	0.225	0.316	**0.105**	0.139	0.279	5.21
Write energy (nJ)	0.293	0.322	0.309	0.193	**0.131**	0.281	4.82
Leakage (W)	1.01	3.53	4.98	1.85	1.92	**0.78**	1.83
Write pulse width (ns)	10	10	2	10	6	10	-
Inter-array routing	Non-H-tree	H-tree	H-tree	H-tree	H-tree	Non-H-tree	H-tree
Sense amp placement	External	Internal	Internal	Internal	Internal	External	Internal
Sense amp type	Current	Current	Current	Voltage	Voltage	Voltage	Voltage
Interconnect wire	Normal	Repeated	Repeated	Low-swing	Low-swing	Normal	Repeated
Output buffer type	Area opt.	Latency opt.	Latency opt.	Area opt.	Area opt.	Area opt.	Latency opt.

TABLE III
SIMULATION PARAMETERS

Components	Features
Simulator kernel	SystemC 2.2.0
CPU core	ARM9TDMI and XScale compatible
Cache configurations	4-way associative 16K L1 data cache 4-way associative 16K L1 instruction cache 4B cache line size
Bus	32-bit address bus 32-bit data bus
Clock frequency	$1GHz$
Main memory	64MB DRAM
Benchmark	MiBench

VI. CASE STUDY

In this section we will conduct one case study to demonstrate how the device-architecture co-optimization methodology can help design STT-RAM cache with different optimization directions. We will replace L1 SRAM instruction cache and data cache by different STT-RAM caches: latency-optimized STT-RAM, energy-optimized STT-RAM, or EDP-optimized STT-RAM. Most the optimization techniques are the same as those used to optimize the $64MB$ STT-RAM chip in Section IV-C. We are the first to explore design space of STT-RAM utilizing PMTJ as L1 cache replacement and compare the results with SRAM.

A. Experimental Setup

In our simulation, an system-level ARM simulator [28] is modified to conduct the evaluation of the latency and energy consumption of the system. As shown in Table.III, the simulator underlying kernel is SystemC 2.2.0 and it is compatible to ARM9TDMI and XScale architecture. Based on this simulator, we developed a STT-RAM cache module with precise timing and energy model. The 4-way associated L1 cache in our simulation has the size of 16KB with the cache line size of 32bits. The main memory is implemented with a 64MB embedded DRAM. Besides this, we also use MiBench [29] as the benchmark for our simulation.

B. Experimental Results

We normalize instruction per cycle (IPC), energy and EDP to SRAM-based cache. Note that energy here includes read dynamic energy, write dynamic energy and leakage energy. Fig. 8 shows the simulations results in terms of normalized IPC when using the three different STT-RAM caches. Fig. 9 presents the energy saving percentage of replacing SRAM by these STT-RAM caches. Fig. 10 illustrates the EDP value of STT-RAM caches for different benchmarks. We can see that the latency-optimized STT-RAM has almost 50% better IPC than SRAM, while it has negative energy saving compared to SRAM. Energy-optimized STT-RAM has approximately 60% average energy saving compared to SRAM while IPC is degraded 8% from SRAM. EDP-optimized STT-RAM has

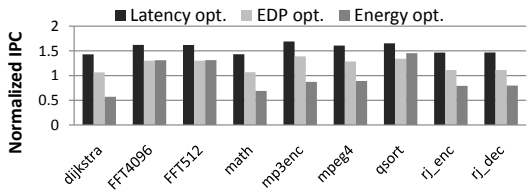

Fig. 8. Normalized IPC for STT-RAM with different optimization directions.

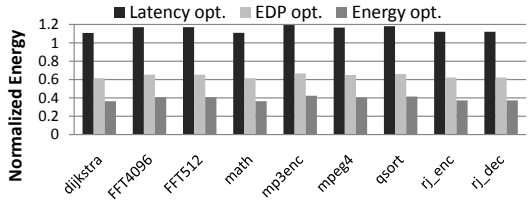

Fig. 9. Normalized energy for STT-RAM with different optimization directions.

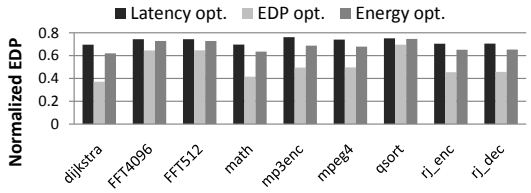

Fig. 10. Normalized EDP for STT-RAM with different optimization directions.

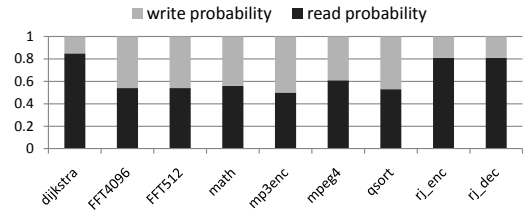

Fig. 11. Read-write ratio for different benchmarks.

about 20% better IPC than SRAM and also nearly 40% energy saving compared to SRAM. Fig. 11 shows the variation in read/write statistics for different benchmarks.

VII. Conclusion

In this paper, we analyze the impact of write pulse width on area, performance, and energy of STT-RAM array and develop a methodology for device-architecture co-design of STT-RAM macros with different optimization goals. We take both near-commercialized in-plane MTJ and advanced PMTJ as optimization targets. Our study shows that for a given MTJ species the quality of STT-RAM macro strongly depends on the write pulse width. In general, reducing write pulse width will harm area, read operation and leakage. These metrics become worse when the write pulse width is below some point in processional mode. Since write latency/energy is not a non-monotonic function of write pulse width. Therefore it's important to find the optimal write pulse width for minimum write latency or energy. We combine the write pulse width optimization with other architectural techniques to design STT-RAM macro as DRAM or SRAM replacement. A $64MB$ STT-RAM chip design spectrum was demonstrated as potential main memory replacement in low power embedded system. Three STT-RAM caches are verified as L1 cache replacement in an embedded system and the simulation results show that by utilizing advanced PMTJ STT-RAM based L1 cache can outperform SRAM in both system performance and energy.

References

[1] S. Wolf, J. Lu, M. Stan, E. Chen, and D. Treger, "The promise of nanomagnetics and spintronics for future logic and universal memory," *Proceedings of the IEEE*, vol. 98, no. 12, pp. 2155 –2168, 2010.

[2] E. Chen, D. Apalkov, Z. Diao, A. Driskill-Smith, D. Druist, D. Lottis, V. Nikitin, X. Tang, S. Watts, S. Wang, S. Wolf, A. Ghosh, J. Lu, S. Poon, M. Stan, W. Butler, S. Gupta, C. Mewes, T. Mewes, and P. Visscher, "Advances and future prospects of spin-transfer torque random access memory," *IEEE Transactions on Magnetics*, vol. 46, no. 6, pp. 1873 –1878, 2010.

[3] M. Hosomi, H. Y. Yamagishi, T. *et al.*, "A novel nonvolatile memory with spin torque transfer magnetization switching: Spin-RAM," in *Proceedings of International Electron Devices Meeting*, 2005, pp. 459–462.

[4] A. Driskill-Smith, "Latest Advances in STT-RAM," in *2nd Annual Non-Volatile Memories Workshop*, 2011.

[5] H. Meng and J.-P. Wang, "Spin transfer in nanomagnetic devices with perpendicular anisotropy," *Applied Physics Letters*, vol. 88, no. 17, pp. 172 506 –172 506–3, Apr. 2006.

[6] P. Khalili Amiri, Z. M. Zeng, J. Langer, H. Zhao, G. Rowlands, Y.-J. Chen, I. N. Krivorotov, J.-P. Wang, H. W. Jiang, J. A. Katine, Y. Huai, K. Galatsis, and K. L. Wang, "Switching current reduction using perpendicular anisotropy in CoFeB-MgO magnetic tunnel junctions," *Applied Physics Letters*, vol. 98, no. 11, pp. 112 507 –112 507–3, Mar. 2011.

[7] Z. R. Tadisina, A. Natarajarathinam, B. D. Clark, A. L. Highsmith, T. Mewes, S. Gupta, E. Chen, and S. Wang, "Perpendicular magnetic tunnel junctions using co-based multilayers," *Journal of Applied Physics*, vol. 107, no. 9, pp. 09C703 –09C703–3, May 2010.

[8] T. Kishi, H. Yoda, T. Kai, T. Nagase, E. Kitagawa, M. Yoshikawa, K. Nishiyama, T. Daibou, M. Nagamine, M. Amano, S. Takahashi, M. Nakayama, N. Shimomura, H. Aikawa, S. Ikegawa, S. Yuasa, K. Yakushiji, H. Kubota, A. Fukushima, M. Oogane, T. Miyazaki, and K. Ando, "Lower-current and fast switching of a perpendicular TMR for high speed and high density spin-transfer-torque MRAM," in *Proceedings of International Electron Devices Meeting*, 2008, pp. 1 –4.

[9] X. Zhu and J.-G. Zhu, "Spin torque and field-driven perpendicular MRAM designs Scalable to multi-Gb/chip capacity," *IEEE Transactions on Magnetics*, vol. 42, no. 10, pp. 2739 –2741, 2006.

[10] S. Thoziyoor, N. Muralimanohar, J.-H. Ahn, and N. P. Jouppi, "CACTI 5.1 technical report," HP Labs, Tech. Rep. HPL-2008-20, 2008.

[11] X. Dong, X. Wu, G. Sun, Y. Xie, H. Li *et al.*, "Circuit and Microarchitecture Evaluation of 3D Stacking Magnetic RAM (MRAM) as a Universal Memory Replacement," in *Proceedings of the Design Automation Conference*, 2008, pp. 554–559.

[12] X. Dong, N. P. Jouppi, and Y. Xie, "PCRAMsim: System-level performance, energy, and area modeling for phase-change RAM," in *Proceedings of the International Conference on Computer-Aided Design*, 2009, pp. 269–275.

[13] V. Mohan, S. Gurumurthi, and M. R. Stan, "FlashPower: A detailed power model for NAND flash memory," in *Proceedings of Design, Automation and Test in Europe*, 2010, pp. 502–507.

[14] C. Xu, X. Dong, N. P. Jouppi, and Y. Xie, "Design implications of memristor-based RRAM cross-point structures," in *Proceedings of Design, Automation & Test in Europe*, 2011, pp. 1–6.

[15] J. Li, P. Ndai, A. Goel, S. Salahuddin, and K. Roy, "Design paradigm for robust spin-torque transfer magnetic RAM (STT MRAM) from circuit/architecture perspective," *IEEE Transactions on Very Large Scale Integration*, vol. 18, no. 12, pp. 1710 –1723, 2010.

[16] S. Chatterjee, M. Rasquinha, S. Yalamanchili, and S. Mukhopadhyay, "A scalable design methodology for energy minimization of STTRAM: a circuit and architecture perspective," *IEEE Transactions on Very Large Scale Integration*, 2010.

[17] C. W. Smullen, V. Mohan, A. Nigam, S. Gurumurthi, and M. R. Stan, "Relaxing non-volatility for fast and energy-efficient STT-RAM caches," in *Proceedings of the International Symposium on High Performance Computer Architecture*, 2011, pp. 50–61.

[18] W. Xu, H. Sun, X. Wang, Y. Chen, and T. Zhang, "Design of last-level on-chip cache using spin-torque transfer RAM (STT RAM)," *IEEE Transactions on Very Large Scale Integration*, vol. 19, no. 3, pp. 483 –493, 2011.

[19] C. Lin, S. Kang, Y. Wang, K. Lee, X. Zhu, W. Chen, X. Li, W. Hsu, Y. Kao, M. Liu, Y. Lin, M. Nowak, N. Yu, and L. Tran, "45nm low power CMOS logic compatible embedded STT MRAM utilizing a reverse-connection 1T/1MTJ cell," in *Proceedings of International Electron Devices Meeting*, 2009, pp. 57 –59.

[20] Z. Diao, D. Apalkov, M. Pakala, Y. Ding, A. Panchula, and Y. Huai, "Spin transfer switching and spin polarization in magnetic tunnel junctions with MgO and AlOx barriers," *Applied Physics Letters*, vol. 87, no. 23, pp. 232 502 –232 502–3, Dec. 2005.

[21] J. C. Slonczewski, "Currents, torques, and polarization factors in magnetic tunnel junctions," *Physical Review B*, vol. 71, no. 2, p. 024411, Jan 2005.

[22] Z. Diao, A. Panchula, Y. Ding, M. Pakala, S. Wang, Z. Li, D. Apalkov, H. Nagai, A. Driskill-Smith, L.-C. Wang, E. Chen, and Y. Huai, "Spin transfer switching in dual MgO magnetic tunnel junctions," *Applied Physics Letters*, vol. 90, no. 13, pp. 132 508 –132 508–3, Mar. 2007.

[23] Z. Diao, Z. Li, S. Wang, Y. Ding, A. Panchula, E. Chen, L.-C. Wang, and Y. Huai, "Spin-transfer torque switching in magnetic tunnel junctions and spin-transfer torque random access memory," *Journal of Physics: Condensed Matter*, vol. 19, no. 16, p. 165209, 2007.

[24] A. Raychowdhury, D. Somasekhar, T. Karnik, and V. De, "Design space and scalability exploration of 1T-1STT MTJ memory arrays in the presence of variability and disturbances," in *Proceedings of International Electron Devices Meeting*, 2009, pp. 707–710.

[25] W. Zhao and Y. Cao, "New generation of predictive technology model for sub-45 nm early design exploration," *IEEE Transactions on Electron Devices*, vol. 53, no. 11, pp. 2816 –2823, nov. 2006.

[26] F. Fishburn, B. Busch, J. Dale, D. Hwang *et al.*, "A 78nm $6F^2$ DRAM technology for multigigabit densities," in *Proceedings of the Symposium on VLSI Technology*, 2004, pp. 28–29.

[27] S.-S. S. et al., "A 4Mb embedded SLC resistive-RAM macro with 7.2ns read-write random-access time and 160ns MLC-access capability," in *Proceedings of the IEEE International Solid-State Circuits Conference*, 2011, pp. 200 –202.

[28] J. Lee, J. Kim, C. Jang, S. Kim, B. Egger, K. Kim, and S. Han, "Facsim: a fast and cycle-accurate architecture simulator for embedded systems," in *Proceedings of the 2008 ACM SIGPLAN-SIGBED conference on Languages, compilers, and tools for embedded systems*, ser. LCTES '08, 2008, pp. 89–100.

[29] M. R. Guthaus, J. S. Ringenberg, D. Ernst, T. M. Austin, T. Mudge, and R. B. Brown, "MiBench: A free, commercially representative embedded benchmark suite," in *IEEE 4th Annual Workshop on Workload Characterization*, 2001, pp. 3–14.