Device-Architecture Co-Optimization of STT-RAM Caches for Embedded Applications

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 1 ns to 100 ns have been reported for STT-RAM. The switching current of magnetic tunnel junction (MTJ) that is the basic building block of STT-RAM is inversely proportional to the write pulse width. In this work, we provide a detailed methodology to optimize STT-RAM design for different design 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 as our optimization targets. Our study shows that the reducing write pulse width will harm read latency and energy while there might be sweet spots of write pulse width which minimize the write energy and write latency of STT-RAM caches. It's also demonstrated that the optimal value write pulse width is dependent on both MTJ spec and STT-RAM cache capacity. The simulation results indicate that by optimizing the advanced perpendicular MTJ based cache design STT-RAM can compete against SRAM for some embedded applications.**

I. Introduction

Universal memory that provides fast random access, high storage density, and non-volatility within one memory technology becomes possible thanks to the emergence of various new 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). As the ultimate goal of these NVM research is to devise a universal memory that could work across multiple layers of the memory hierarchy, each of these emerging NVM technologies has to supply a wide design space that covers a spectrum from highly latency-optimized microprocessor caches to highly density-optimized secondary storage. Among the emerging NVM technologies, STT-RAM seems to be one of the most promising candidates that has the potential to meet all the requirement of universal memory [1], [2].

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. The major drawback required current 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. Thus STT-RAM is projected to scale beyond 20nm technology node even without any material improvement [4]. To further reduce switching current and switching time, Perpendicular MTJs (PMTJ) for STT-RAM were developed [5]–[9] to achieve very low switching current while maintaining relative 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 perpendicular STT-RAM in architecture-level research and it's surprised to see the competitive results of PMTJ versus SRAM even as L1 cache replacement.

Experiments have been performed in device-level research in order to operate a MTJ (magnetic tunnel junction) at minimum energy or energy-delay-product (EDP) by applying varied write pulse width on 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 serial with a MTJ. Short write pulse induced large switching current requires large access transistor for providing enough driving current, which consequently brings more circuit design challenges of the STT-RAM prototype. Specifically, it worsens the area, latency, dynamic energy and leakage power of both memory cells 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 implement a system-level performance, energy, and area model to estimate the impact of different write pulse width on the STT-RAM macro design. We then develop a detailed device-architecture co-optimization methodology to design STT-RAM macro with different optimization goals such as area, read latency/energy, write latency, write energy by leveraging the inherent trade-off of write current and write time of MTJ. The results have potential impact on the guidelines for designing a STT-cache in different memory hierarchical levels 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-off 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 embedded system. Section VII concludes our work.

II. Related Work

Many modeling tools have been developed to enable system-level design exploration for SRAM- or DRAM-based cache and memory design. For example, CACTI [10] is a tool that has been widely used in the computer architecture community to estimate the speed, power, and area of SRAM and DRAM caches. Evans and Franzon [11] developed an energy model for SRAMs and used it to predict an optimum organization for caches. eCACTI [12] incorporated a leakage power model into CACTI. Muralimanohar *et al.* [13] modeled large capacity caches through the use of an interconnect-centric organization composed of mats and request/reply H-tree networks. In addition, CACTI has also been extended to evaluate the performance, power, and area for STT-RAM [14], PCRAM [15], NAND flash [16], and ReRAM [**?**]. 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 a tool called NVsim [15], which is a circuit-Level performance, energy, and area simulator for emerging non-volatile memories.

There have seen several work on design methodology for STT-RAM from both circuit and architecture perspective. Li *et al.* 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.* had a more through study on co-designing the sizing of the access transitor and operating voltage to achieve minimum energy dissipation. Moreover, Smullen *et al.* illustrated STT-RAM cell design for optimizing read latency and write latency separately in the presence of clock cycles. Similarly, Xu *et al.* quantitatively analyze the impact of write latency and read latency trade-off of STT-RAM on system performance. However, most of these work were using 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 by choosing the optimal write pulse width.

The contributions of this work are listed as follow,

- We provide a detailed analysis of the impact of write pulse width on area, read latency/enery, write latency/energy of STT-RAM with different capacities. Our results indicate that the write pulse width optimizing STT-RAM write latency or energy will increase as the cache capacity increase.
- To the best of our knowledge, we are the first to explore the design space of STT-RAM using advanced perpendicular MTJs. Surprisingly we find that by utilizing such MTJ spec to build L1 cache with the optimization methodology we developed, STT-RAM can have competitive write latency/energy as SRAM while maintaining equal or better read latency/energy than SRAM.

III. Preliminary

A. STT-RAM Cell Basics

STT-RAM uses MTJ as the memory storage and leverages the difference in magnetic directions to represent the memory bit. As shown in Figure 1, MTJ usually contains two ferromagnetic layers. One ferromagnetic layer is has fixed magnetization direction and it is called the reference layer, while the other layer has a free magnetization direction that can be changed by passing a write current and it is called the free layer. The relative magnetization direction of two ferromagnetic layers determines the resistance of MTJ. If two ferromagnetic layers have the parallel directions, the resistance of MTJ is low, indicating a “1” state; if two layers have anti-parallel directions, the resistance of MTJ is high, indicating a “0” state.

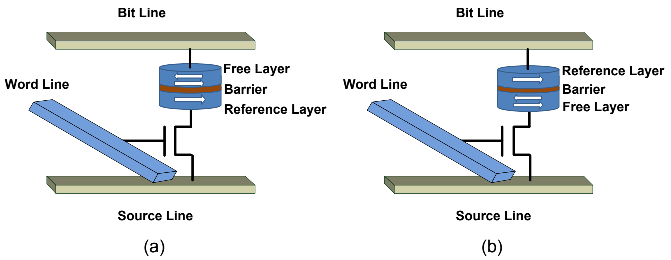

Fig. 1. Demonstration of a STT-RAM cell: (a) Conventional connection scheme; (b) Reverse connection scheme.

As shown in Figure 1, there are two possible schemes to stack MTJ atop access NMOS transistor. Conventionally, the free layer of MTJ is connected to bitline (BJ). In that scheme, when writing “1” state into STT-RAM cells, positive voltage difference is established between BL and SL and the switching current required is $I_c(AP \rightarrow P)$; when writing “0” state, negative voltage difference is established between BL and SL and the switching current required is $I_c(P \rightarrow AP)$. While the reverse connection scheme was proposed in [**?**] where the free layer of MTJ is connected to the drain of NMOS instead of BL. [**?**] argues that $I_c(P \rightarrow AP)$ is normally significantly larger than $I_c(AP \rightarrow P)$ [**?**], [**?**] due to the inherent torque asymmetry of MTJ. But the SL-to-BL current is much smaller than the BL-to-SL current under the same wordline voltage and voltage difference between BL and SL because the body effect of access transistor degrades the SL-to-BL current remarkably. Thus reserving connection scheme can relax the sizing requirement on access transistor, which results in more compact STT-RAM cell size. However, device-level efforts have been put to improve the asymmetry of switching characteristic of MTJ and $I_c(AP \rightarrow P)$ slightly larger than $I_c(P \rightarrow AP)$ was even demonstrated in [**?**]. In our work, we always choose the MJT connection scheme that is responsible for relaxed sizing requirement on access transistor.

Another important metric for an MTJ is the tunnel magnetoresistance (TMR) 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 brings faster read sensing latency or relaxes the constraint for sense amplifier design. It’s critical to introduce an equivalent metric for a STT-RAM cell which contains both the MTJ and the

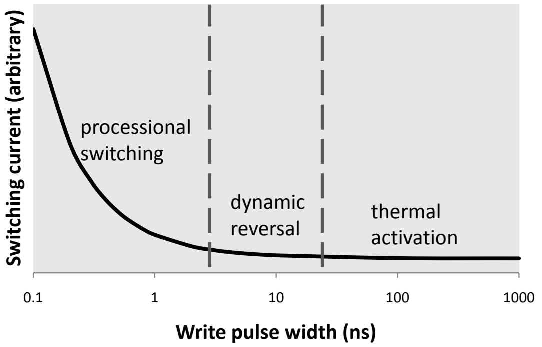

Fig. 2. Demonstration of three switching phases: thermal activation, dynamic reversal and precessional switching.

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 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 "0" state and "1" state. If we ignore the resistance difference of the access transistor for reading "0" state and "1" 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 equvalent resistance of access NMOS transistor and W is the transistor width, C is a constant related to the wordline voltage and threshold voltage of the transistor. From equation 4 we can conclude that sizing up the access transistor will make CTMR close to the inherent TMR 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 [**?**] 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 current J_c and write pulse width τ was characterized by an analytical model in [**?**]. 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 density for thermal activation, precessional switching and dynamic reversal. 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 Figure 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 decrease very slowly even we increase 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, 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 for different design goals.

C. Perpendicular MTJ

A key challenge for MTJ design is to reduce switching current while maintaining sufficient high thermal stability to not affect data retention time and write/read errors. The conventional in-plane MTJ critical switching current I_{c0} divided by the thermal barrier Δ be expressed as [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 plan anisotropy field, and H_k is the in-plane anisotropy field most of which is from 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, larger damping constant, 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

Some modeling issues will be discussed in this section.

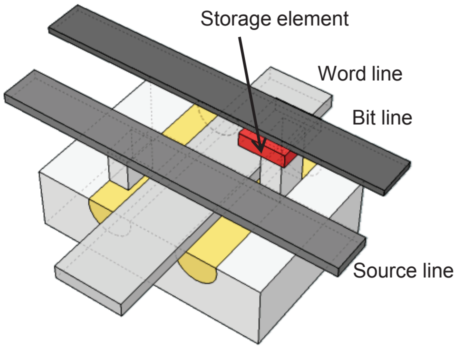

Fig. 3. Conceptual view of a MOS-accessed cell (1T1J) and its connected word line, bit line, and source line.

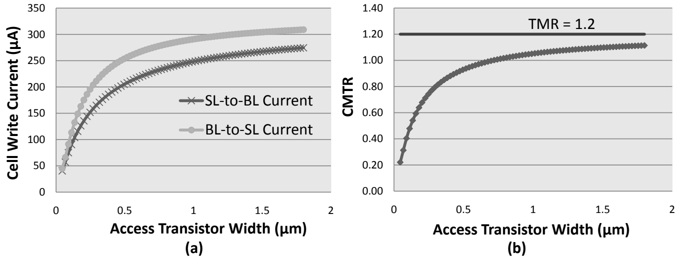

Fig. 4. (a) Driving ability, and (b) Cell TMR of access NMOS transistor.

A. Area Modeling

To simulate the performance of a single STT-RAM cell, it is important to estimate its access area first. As mentioned before, each 1T1J STT-RAM cell is composed of one NMOS and one MTJ. The NMOS access device is connected in series with the MTJ as shown in Figure 3. 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 using HSPICE with PTM 45nm HP model [**?**] is simulated. The BL-to-SL current and SL-to-BL current are obtained by assuming typical TMR (120%) and LRS ($3k\Omega$) value [**?**] and bursting wordline voltage to be 1.5V (the optimal V_{WL} value is extracted from [**?**]). As we can see in Figure 4(a), the SL-to-BL current is always smaller and saturate faster than BL-to-SL current. Such current degradation is related to the body effect of the access transistor, where the threshold voltage is boosted by a positive source-to-substrate voltage different $V_{SB} = I_c \times R$ (I_c is the current passing through MTJ and R is the resistance of the MTJ). Thus we always choose the connecting scheme which use BL-to-SL current to match the maximum of $I_c(AP \rightarrow P)$ and $I_c(P \rightarrow AP)$ and SL-to-BL to match the minimum 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 Figure fig:hspice(b), a larger access transistor increase CTRM closer to the inherent TRM value 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 satisfy 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 [17]. As a result, the cell size of a STT-RAM cell is calculated as follows,

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

B. Data sensing modeling

Three sensing modes were proposed in [**?**] 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 [**?**] to sense the state if the number of cells per bitline is larger than 64, while voltage-divider sensing the second fastest and the current-in voltage sensing is slowest. In contrast, 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-in sensing has the worst area efficiency. In this work, we will focus on current sensing scheme to reduce read latency. We adapt the current-voltage converter and sense amplifier design discussed in [14]. 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 choosing longer write pulse and smaller write current. And reduced read current have impact on the latency of 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 NVsim.

C. Cell Switching Modeling

Dynamic MTJ switching model was developed in [**?**] with consideration of the switching phenomenon involves magnetoresistive effects, which can not estimated only by RC analysis. However, this work is focusing on static analysis of STT-RAM and NVSim does not model the dynamic behavior during the switching of the cell state. Thus we use 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}).

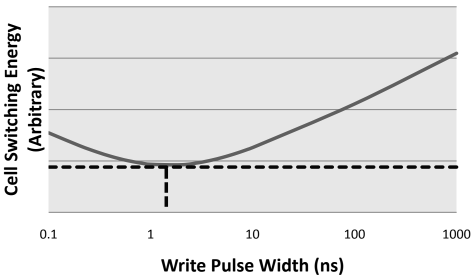

Fig. 5. Switching energy per cell versus write pulse width.

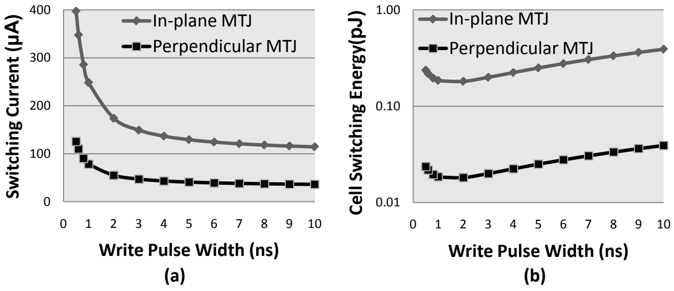

Fig. 6. Representative data curves of in-plane and perpendicular MTJs: (a)Switching current; (b) Cell switching energy.

Taking the switching current versus switching pulse width as input for Equation 15, we can easily get the relation between the energy of switching one cell and write pulse width. It can been seen in Figure 5, minimum switching energy per cell is achieved at write pulse width $\tau_{min\_cell\_energy}$ in the range of processional switching mode. However, the 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 have not been considered.

V. STT-RAM Macro Design Optimization

In this section we first choose two MTJ specs as our optimization targets. After numerous simulations we the importance of careful write pulse width selection is revealed for optimizing area, read latency/energy, write latency/enery and leakage power of STT-RAM. Then we will focus on the analysis of write energy optimization, which is both device and capacity dependent. Finally we will combine the improved architectural design together with the write pulse width optimization to illustrate a full design explore a 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 can represent the typical switching characteristics of in-plane MTJ [?] and PMTJ [8]. As seen in Figure 6, 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 III. We assume the same TMR and resistance for in-plane MTJ and PMTJ. For PMTJ, realization of matching these 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 spec is are extracted from Figure 6(b). Allowed maximum access transistor width is constrained, which has been translated to the operating range of write pulse width.

TABLE I
Simulation Parameters

Optimization target	in-plane MTJ	PMJT
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	$45ns$	
I/O Width	$32bits - 512bits$	

NVsim simulation results of $2MB$ STT-RAM macros with in-plane MTJ and PMTJ separately 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 Figure 7. In general, decreasing write pulse width will increase area, read latency/energy and leakage power. Especially reducing write pulse with in processional mode will harm these metrics badly. However, the relation between write pulse width and write latency/enery is not monotonous. Separate explanations and result analysis are given,

- Area: short write pulse induced large switching current requires large access transistor for providing enough driving current, which incur both cell and peripheral circuitry (i.e. wordline driver) area penalty. From Figure 7(a) we can see that STT-RAM generally has area advantage over SRAM. But reducing write pulse aggressively below $2ns$ will cause STT-RAM with in-plane MTJ to have 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 Figure 7(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 that of SRAM. However, read operation of STT-RAM with PMTJ can be faster than SRAM for $\tau > 0.8ns$ due to more remarkable area advantage.
- Read energy: it's affected in the similar way that read latency is affected. From Figure 7(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 thus it's 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 Figure 7(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 spec, STT-RAM capacity and I/O width. We can see from Figure 7(e) that $\tau_{min\_energy} = 5ns$ for 32-bit 2MB STT-RAM with in-plane MTJ and $\tau_{min\_energy} = 7ns$ for 32-bit 2MB STT-RAM with PMTJ. Moreover, the minimum write energy of STT-RAM with in-plane MTJ is roughly 50% more than that of SRAM while the energy of STT-RAM with PMTJ is almost half the number of SRAM.
- 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 increasesl because the leakage power for STT-RAM mainly comes from peripheral circuity. Figure 7(f) we can tell the remarkable advantage of STT-RAM over SRAM no matter what type of MTJ spec is used.

B. Write Energy Optimization

In 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 spec, STT-RAM capacity and I/O width. From the results shown in Figure 8, 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 (32bits); (3) τ_{min_energy} is a monotonic decreasing function of I/O width under fixed capacity (2MB). There are primary two reasons for (1): the baseline $\tau_{min\_cell\_energy}$ of in-plane MTJ 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, minimize the cell switching energy at $\tau_{min\_cell\_energy}$ from cell-level may not 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.

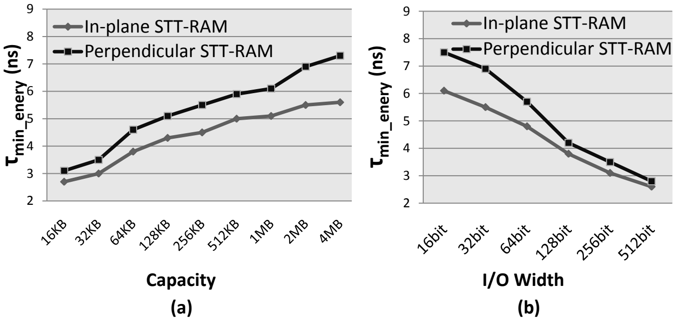

Fig. 8. Dependence of τ_{min_energy} on (a) STT-RAM Macro Capacity, and (b) I/O Width

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; (5) Differen sensing schemes for trade-off of area, latency and energy. Table II tabulates the full design spectrum this chip by listing the details of each design corner.

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 the L1 SRAM instruction cache and data cache by there different STT-RAM caches: latency-optimized STT-RAM, energy-optimized STT-RAM, or EDP-optimized STT-RAM. We are the first to explore design space of STT-RAM utilizing PMTJ as L1 cache replacement and compare the results with The optimization techniques are the same with those used in Section IV-C.

A. Experimental Setup

In our simulation, an system-level ARM simulator [?] 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 develop 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

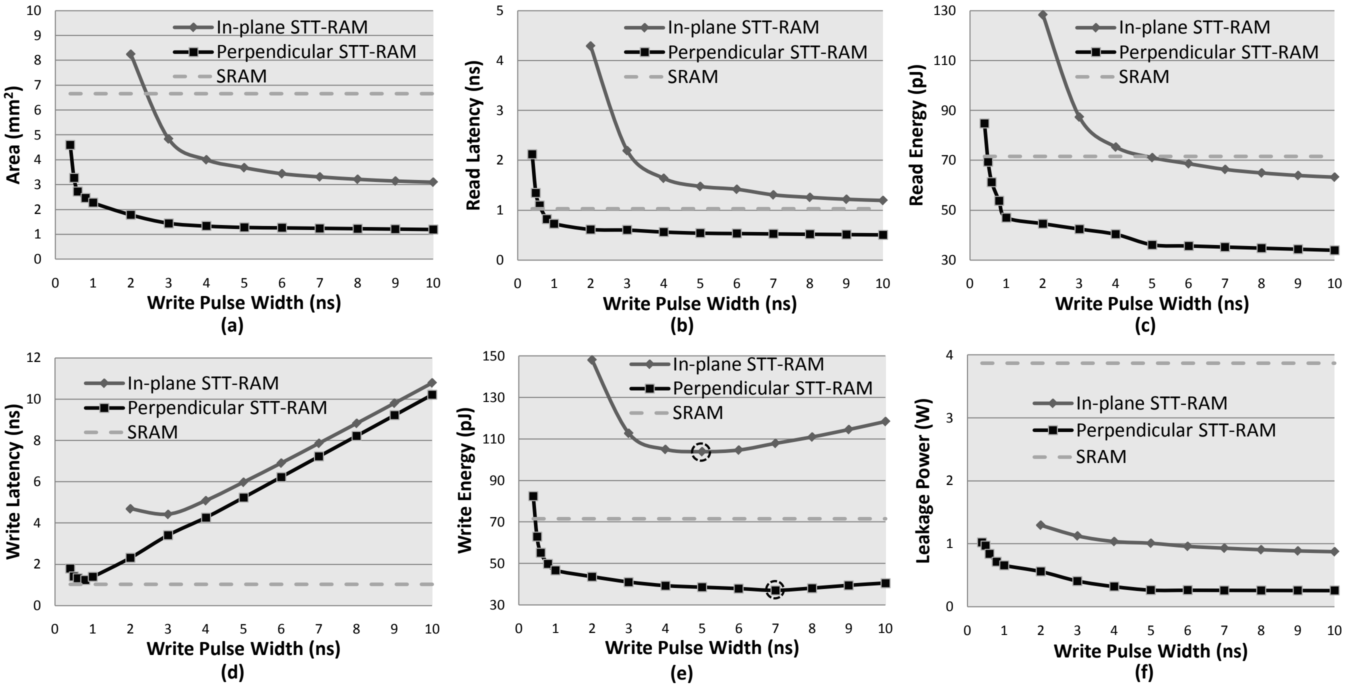

Fig. 7. Metrics of SRAM and STT-RAM built with in-plane and perpendicular MTJs: (a)Area; (b)Read latency; (c)Read energy; (d)Write latency; (e)Write energy; (f)Leakage power.

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

	Area opt.	Read latency opt.	Write latency opt.	Read energy opt.	Write energy opt.	Leakage opt.
Area (mm^2)	**3.06**	10.7	16.4	5.66	6.22	3.63
Read latency (ns)	21.8	**3.70**	4.92	9.12	9.57	9.91
Write latency (ns)	18.6	13.9	**4.01**	15.9	12.3	18.1
Read energy (nJ)	0.276	0.225	0.316	**0.105**	0.139	0.279
Write energy (nJ)	0.293	0.322	0.309	0.193	**0.131**	0.281
Leakage (W)	1.01	3.53	4.98	1.85	1.92	**0.78**
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
Sense amp placement	External	Internal	Internal	Internal	Internal	External
Sense amp type	Current-in voltage	Current	Current	Voltage-divider	Voltage-divider	Voltage-divider
Interconnect wire	Normal	Repeated	Repeated	Low-swing	Low-swing	Normal
Output buffer type	Area opt.	Latency opt.	Latency opt.	Area opt.	Area opt.	Area 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 cache 4B cache line size
Bus	32-bit address bus 32-bit data bus
Clock frequency	$1GHz$
Main memory	64MB DRAM
Benchmark	MiBench

64MB embedded DRAM. Besides, we use the MiBench [**?**]as the benchmarks 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. Figure 9 shows the simulations results in terms of normalized IPC when using the three different STT-RAM caches. Figure 10 presents the energy saving percentage of replacing SRAM by these STT-RAM caches. Figure 11 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 maintaining only 8% degraded IPC of SRAM. EDP-optimized STT-RAM has about 20% better IPC than SRAM and also nearly 40% energy saving compared to SRAM. Figure 12 shows the variation in read/write statistics for different benchmarks.

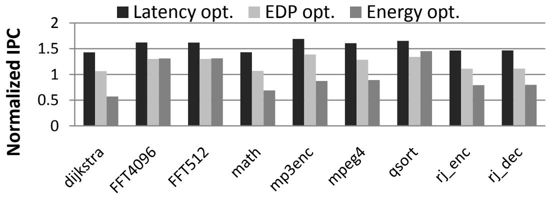

Fig. 9. Normalized IPC for STT-RAM with different optimization directions.

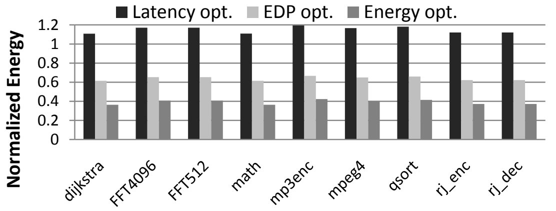

Fig. 10. Normalized energy for STT-RAM with different optimization directions.

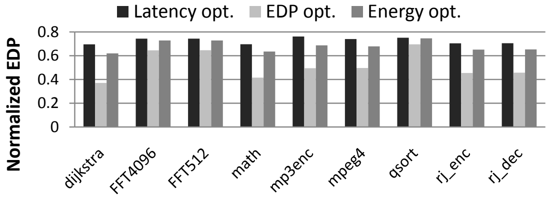

Fig. 11. Normalized EDP for STT-RAM with different optimization directions.

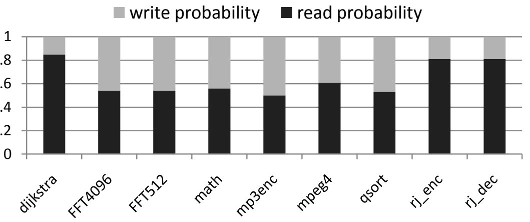

Fig. 12. Read-write ratio for different benchmarks.

VII. Conclusion

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] R. J. Evans and P. D. Franzon, "Energy consumption modeling and optimization for SRAM's," *IEEE Journal of Solid-State Circuits*, vol. 30, no. 5, pp. 571–579, 1995.

[12] M. Mamidipaka and N. Dutt, "eCACTI: An enhanced power estimation model for on-chip caches," Center for Embedded Computer Systems, Tech. Rep. TR04-28, 2004.

[13] N. Muralimanohar, R. Balasubramonian, and N. P. Jouppi, "Architecting efficient interconnects for large caches with CACTI 6.0," *IEEE Micro*, vol. 28, no. 1, pp. 69–79, 2008.

[14] 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.

[15] 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.

[16] 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.

[17] 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.