Nanoscale Reconfigurable Computing Using Non-Volatile 2-D STTRAM Array

Somnath Paul
Department of EECS
Case Western Reserve U.
Cleveland, USA
sxp190@case.edu

Subho Chatterjee
Department of ECE
Georgia Institute of Tech.
Atlanta, USA
subho.chatterjee@gatech.edu

Saibal Mukhopadhyay
Department of ECE
Georgia Institute of Tech.
Atlanta, USA
saibal@gatech.edu

Swarup Bhunia
Department of EECS
Case Western Reserve U.
Cleveland, USA
skb21@case.edu

***Abstract*— In this paper, we investigate the combination of a novel computing paradigm referred to as *Memory Based Computing* (MBC) and an emerging non-volatile nanoscale memory technology, namely *Spin-Torque Transfer Random Access Memory* (STTRAM), to build a reconfigurable nanocomputing framework with high integration density, robustness and energy-delay efficiency. MBC uses a 2-D memory array as underlying computing element. Noting the read-dominant access pattern in MBC, we optimize the STTRAM cells to increase the energy-delay efficiency. Further, exploiting the asymmetric nature of the cells, we introduce the notion of preferential storage which optimizes the cell performance for '1' over '0' and skew the LUT content toward '1' for improved energy-delay product (EDP).**

Keywords-Reconfigurable Computing; Non-volatile; STTRAM; EDP Improvement

I. INTRODUCTION

STTRAM is a fast-emerging non-volatile nanoscale alternative which is being considered as the next generation universal memory [1] due to its high integration density, unlimited endurance and high speed of operation. Recent ITRS updates [14] have also recognized STTRAM as a promising candidate for non-volatile memory applications. Although recent advances in fabrication [5] and novel memory architectures [2] have been reported for STTRAM arrays, a high write current requirement remains a matter of grave concern in the operation of STTRAM based normal memory arrays. A high write current requirement is however of lesser concern in reconfigurable frameworks which are configured only infrequently. In this paper, we investigate the potential of using STTRAM in a novel memory-based reconfigurable computing platform, referred to as *MBC*, which uses a 2-D memory array as primary computing element. MBC uses a dense memory array to map logic functions into multi-input multi-output lookup tables (LUTs) and evaluate them in a time-multiplexed and topological fashion using a small CMOS-based controller [3-4]. MBC differs significantly from FPGAs in the fact that it substantially reduces the interconnect resources required to map a given application. Reduction in programmable interconnect translates into large savings in both delay and energy per vector [3]. We show that STTRAM array can be an excellent choice for implementing the memory array in a non-volatile MBC framework. This is due to the following reasons: 1) non-volatile nature of STTRAM cells that alleviates the requirement of reconfiguration at power-up; 2) high integration density; 3) high read speed and low read power; 4) near-zero standby leakage; and 5) radiation hardness. In particular, the paper makes the following contributions:

1. It proposes a novel nanoscale memory based computing framework using 2-D STTRAM array hybridized with CMOS controller logic.
2. Exploiting the read-dominant memory access pattern in MBC and the asymmetry in read energy between logic '0' and logic '1' state in STTRAM, it proposes circuit/architecture level optimization techniques to maximize the energy-delay efficiency.

The rest of the paper is organized as follows. Section II provides background on STTRAM technology and MBC framework. Section III describes the main contributions of this work focusing on circuit and architecture level optimizations to improve EDP. Section IV presents simulation results for a set of benchmark circuits. Section V concludes the paper.

II. BACKGROUND

A. STTRAM Operation

The basic building block of a STTRAM cell is the Magnetic Tunneling Junction (MTJ) (Fig. 1b-c). Each MTJ consists of two ferromagnetic layers (typically CoFe) separated by a very thin tunneling dielectric film (typically crystallized MgO). Magnetization in one of the layers (referred as pinned layer) is fixed in one direction by coupling to an anti-ferromagnetic layer (such as PtMn) [1]. The other ferromagnetic layer (referred to as free layer) is used for information storage. The direction of magnetization of free layer with respect to the pinned layer (i.e., anti-parallel or parallel) can be controlled by the injection of spin-polarized electrons. Hence the MTJ can be switched between two stable magnetic states with high (R_{AP} or R_H) or low (R_P or R_L) resistances and it retains the state without any applied power. The information stored in an MTJ is thus determined by sensing whether the resistance state is high or low. One of the quality metrics for a MTJ device is therefore the Tunneling Magneto-Resistance (referred as TMR) ratio [6], defined as $(R_H-R_L)/R_L$. An MTJ cell with high TMR ratio is desirable in order to easily distinguish between the two states ($\Delta R=R_H-R_L$). The write current for an MTJ cell is required to be larger than the switching threshold current in order to switch the magnetization of the free layer from anti-parallel to parallel spin or vice versa.

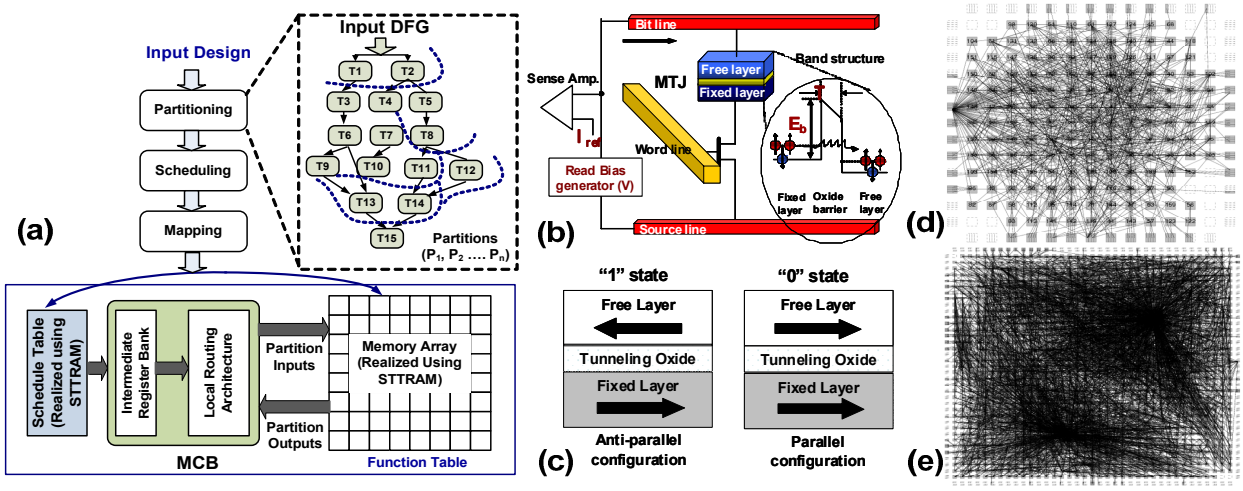

Fig. 1: (a) Functional block diagram of memory based computation; (b) STTRAM cell structure; and (c) logical states of magneto tunneling junction or MTJ (anti parallel – high resistance, and parallel – low resistance). Improvement in number of computing elements and interconnect requirement (estimated using VPR toolset [5]) for sequential benchmark s38417 when mapped to d) MBC and e) 65nm CMOS FPGA frameworks.

STTRAM has certain distinct advantages over the prevalent memories. Having essentially a 1T-1R structure it forms a dense array with high integration density and being magnetic in nature is tolerant to particle hits. In these features it is better than the SRAMs. Being non volatile, it does not require refreshing like DRAM. It has a very high endurance, which makes it a better candidate than flash memories. Another feature of STTRAM is its scalability which makes it an attractive option at scaled dimensions.

In the nanometer nodes, leakage is one of the primary forms of energy dissipation. The STTRAM structure has zero standby leakage. This is because for unselected cells in an unselected column, the bit line, source line and word line are all set to 0. Hence the leakage in all columns in standby mode and unselected columns in the active mode is negligible. There is active leakage corresponding to the unselected cells in a selected column. However, if we consider the entire memory array, unselected cells in the selected column only form a small fraction of the total cells in the array. The total leakage dissipation is therefore expected to be much less than the other memory standards.

B. Memory Based Reconfigurable Computing (MBC)

Fig. 1a illustrates the operation of the memory based reconfigurable computing framework. A target application is first partitioned into smaller multi-input, multi-output logic blocks. Next, the partitions are mapped as LUTs to the embedded memory array and finally scheduled for evaluation over multiple cycles [3]. During application mapping phase, information regarding the address, scheduling and connectivity among the partitions is stored in a register bank (denoted as *schedule table* in Fig. 1a. The embedded memory array, in which the logic partitions are mapped, is referred to as the *function table*. A bank of flip-flops stores intermediate partition outputs, to be used in the following evaluation cycles. The schedule table, the function table and the set of flip flops form the core of the computational building block, referred as *Memory-based Computational Block* or *MCB* in the proposed framework. Multiple MCBs are connected through a configurable interconnect framework similar to conventional FPGA. However, larger partition size and multilevel execution of partitions inside a single MCB greatly reduces the requirement for programmable interconnect resource (refer to Fig. 1d-e). This achieves significant improvement in performance over a conventional FPGA framework.

III. CONTRIBUTIONS

A number of non-volatile reconfigurable frameworks employing STTRAM as the primary storage element have been proposed earlier [10-11]. All these architectures employ a spatial computing model similar to FPGA where the STTRAM is used to store the configuration in small 1-D LUTs. This has two major limitations: i) it fails to exploit the high integration density of the STTRAM array; ii) the requirement for CMOS-STTRAM hybridization at the location of each configuration bit poses serious challenge in fabricating the proposed architectures.

A. Non-volatile MBC using STTRAM

In the non-volatile MBC framework, we propose to build the function table and schedule tables using conventional STTRAM arrays. This offers the following benefits: i) since the function table holds the configuration for the partitions, it occupies the maximum area inside a MCB. However a small footprint for the MTJ devices [9] ensures that the area occupied by this memory array is minimized; ii) non-volatile nature of the STTRAM array ensures that configuration bits for the logic (stored in the function table) and for the local interconnects (stored in the schedule table) is retained when power is turned down; iii) high read performance and low read power for the STTRAM array results in considerable EDP improvement for a

STTRAM based non-volatile MBC framework.

B. Circuit/Architecture Co-optimization Techniques

The design space for STTRAM is constrained by the readability and writability conditions i.e. tunnel magnetoresistance (TMR) ratio and write current requirement. Given a MTJ, the choice of the access MOSFET width (W) and its wordline voltage (V_{WL}) can be used to navigate the design space of TMR and write-current. To minimize the energy dissipation, we propose to choose the energy optimal point in the W-V_{WL} plane. The total energy is evaluated considering the write/read current through the MTJ-transistor structure and the switching energy associated with the wordline and biltine. A key aspect of the solution is its dependence on the read-write probability. Fig. 2a shows two solutions corresponding to write probabilities of 0.5 and 0.1, respectively. These two solutions are different because different write probabilities result in different read and write energies. This reflects a change in the optimization parameters. A larger write probability means a solution with larger width and smaller V_{WL} as write has a quadratic dependence on V_{WL}. From read perspective a solution with lower width is preferred due to less leakage power dissipation. For MBC with read-dominant access pattern, the W-V_{WL} configuration corresponding to equi-probable condition is not an optimal choice as it dissipates higher read energy (Fig. 2b). Hence we chose the optimal energy point corresponding to low write probabilities which provides much lower read and total energy at the expense of increased write energy (Fig. 3a-b). Fig. 2b shows 24% saving in total energy for write probability of 10^{-5} for an 8-bit 64x64 memory array with a read access time of 400ps.

From the STTRAM read operation we find that a larger current flows in the circuit corresponding to read '0' than read '1'. This is because resistance of state '0' is lower than state '1'. Thus in Fig. 3a there is a 36% difference between energy dissipated in the read '1' and read '0' operations. Write energy for '0' case dominates over the '1' case. However for the MBC application intended, the system is heavily biased towards read with a write probability in the range of 10^{-3}-10^{-5}. Hence for MBC application, we conclude that if the system is biased for more read '1's than '0's we can have considerable energy savings with STTRAMs.

Due to higher read power during a read '0' operation, it is intended that the STTRAM array contain more logic '1' than logic '0'. Considering this asymmetry, we have developed a "preferential mapping" approach to skew the LUTs to contain more logic '1' than logic '0 in order to harness the energy advantage as seen in Fig. 3a. Such a preferential application mapping scheme therefore amplifies the energy savings by storing more logic '1' than logic '0' in the schedule and function tables. The greedy heuristic for skewing the logic '0' to logic '1' ratio in the LUTs as presented in [12] is used to exploit the lower read '1' power in the STTRAM based MBC framework. The effectiveness of the preferential mapping approach was validated for a set of standard benchmark circuits chosen from MCNC benchmark suite. For the selected benchmark circuits, the preferential mapping heuristic was observed to achieve on an average about 49% increase in the logic '1' count stored in the LUTs.

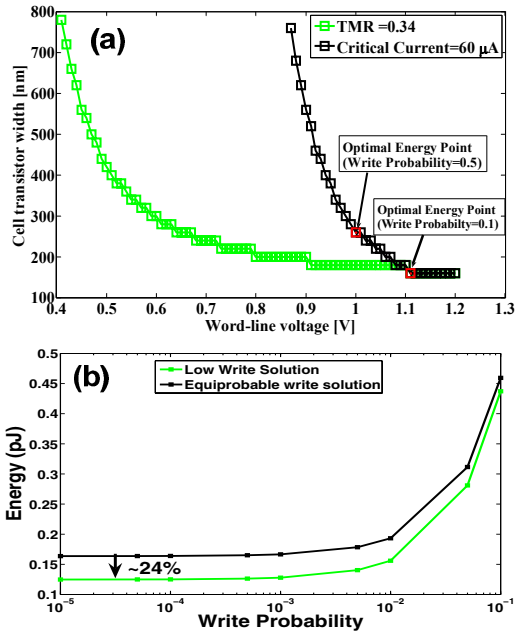

Fig. 2: a) Design of STTRAM cell for MBC framework to achieve optimal read energy; b) read energy with varying write probability.

From these discussions we conclude that as the number of '1's stored in the array increases the energy advantage will keep on increasing. A study on STTRAM array energy with varying probability of '1' storage is shown in Fig. 3c. It points to the fact that a solution with all zero storage will result in a 16% energy access overhead compared to the case when all ones are stored in the array.

IV. Simulation Results

We have performed simulations with MTJ at 65nm node with Resistance-Area product 30Ω-µm^2. The sizes of the MTJ devices have been taken as 50x90nm^2 which requires approximately 60µA of switching current assuming current density of 10^6A/cm^2. The high and low resistance states are represented by 11.1kΩ and 6.67kΩ, respectively. The transistor was designed so as to be able to drive the switching current under both write '0' and write '1' conditions.

To obtain the solution for varying write probabilities, first a host of simulations with the high and low resistance is performed for a range of V_{WL} and W. The solution space has to be extracted from the generated design space considering the constraints on minimum TMR and switching current. In this work we consider minimum TMR and switching current requirements of 0.34 and 60µA, respectively. Corresponding to this extracted feasible design space of V_{WL}-W, we evaluate the read, write, active leakage and total energy of STTRAM array. The energy evaluation considers the read and write probability ratios. The V_{WL}-W combination which gives the minimum energy is identified as the optimal energy solution. For demonstration of the results we selected write probabilities of 0.5 and 0.1 to evaluate the design points. The read and write energies for '0' and '1' are computed for these design points.

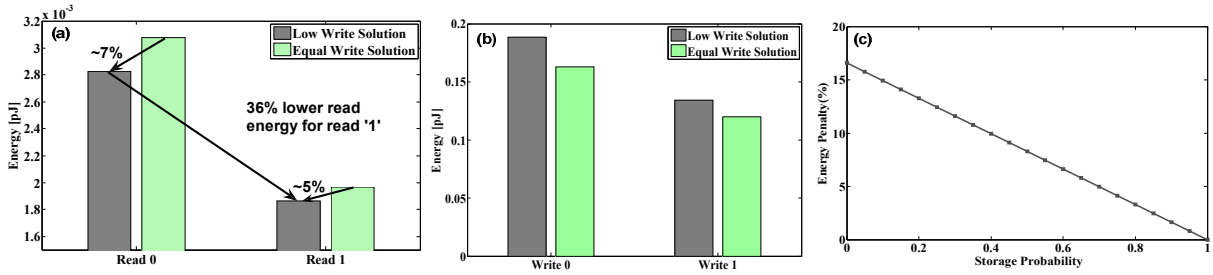

Fig. 3: a) Read and b) write energy for a cell storing logic '0' and '1'; c) Increase in read energy with increasing probability of storing '1'.

Delay and energy requirement for the CMOS elements of the MCB were obtained through SPICE simulations using BSIM4 predictive models at 65nm technology [8]. A supply voltage of 1V was used for simulation. As the cycle time for a given benchmark in the MBC framework depends on the delay through the programmable interconnects, interconnect delay for both MBC and FPGA frameworks were obtained from the VPR toolset [7]. A 65nm FPGA model [13] was used to simulate the performance of the programmable interconnects.

Fig. 4a-b shows the improvement in performance and energy-delay product for STTRAM based MBC over CMOS FPGA framework. As we note from Fig. 4a, for standard benchmark circuits on an average the MBC framework improves the performance by 45.4%. Fig. 4b compares the EDP values between the two frameworks. The non-volatile MBC framework achieves a 5% improvement in EDP over the CMOS FPGA framework. The performance and EDP computation includes the cell optimization for read operation. The EDP improvement is further enhanced through the preferential mapping step which skews the LUT contents to have more logic '1' than logic '0's. As a result of this preferential mapping, the average EDP improvement was calculated to increase from 5% to 6.64%.

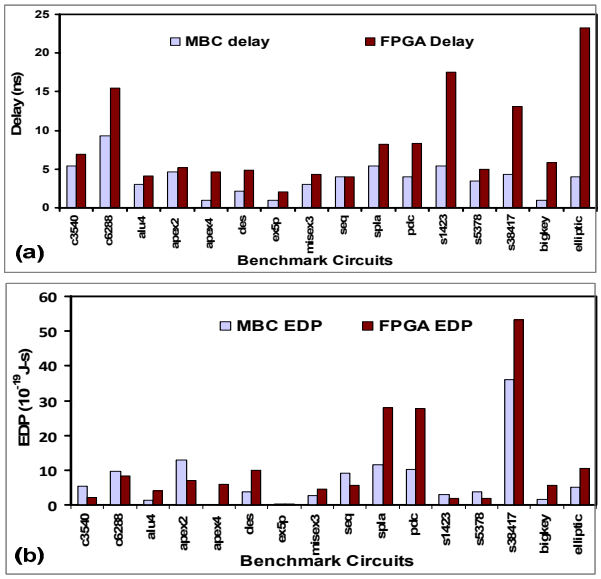

Fig. 4: Improvement in: a) delay and b) Energy Delay Product (EDP) for STTRAM MBC over conventional SRAM-based FPGA.

V. Conclusion

STTRAM holds tremendous promise as a non-volatile memory for storing the configuration bits in a reconfigurable computing framework. MBC frameworks which use conventional 2-D STTRAM array as a configuration memory are non-volatile and have high integration density compared to STTRAM based FPGA frameworks. Moreover, high read performance and low read power for the STTRAM arrays translate into considerable improvement in EDP for STTRAM based MBC over conventional CMOS based FPGA framework. This improvement was further enhanced by circuit/architecture co-optimization techniques, which involve preferential design of the MTJ cell and skewed application mapping step that increases the number of logic '1's over logic '0's in the LUT content. Simulation results for a set of benchmark circuits show that such joint circuit/architecture optimization can be very effective in improving the EDP.

References

[1] M. Hosomi et al., "A novel non-volatile nemory with spin torque transfer magnetization switching: Spin-RAM", *IEDM*, 2006.

[2] T. Kawahara et al., "2Mb Spin-Transfer Torque RAM (SPRAM) with bit-by-bit bidirectional current write and parallelizing-direction current read", *ISSCC*, 2007.

[3] S. Paul and S. Bhunia, "Reconfigurable computing using content addressable memory for improved performance and resource usage", *DAC*, 2008.

[4] S. Paul and S. Bhunia, "MBARC: A scalable memory based reconfigurable computing framework for nanoscale devices", *ASPDAC*, 2008.

[5] J. Hayakawa et al., "Current-driven magnetization switching in CoFeB/MgO/CoFeB magnetic tunnel junctions", *Japanese Journal of Applied Physics*, Vol. 44, No., 41, 2005.

[6] S. Tehrani et al., "Magnetoresistive random access memory using magnetic tunnel junctions", *Proceedings of IEEE*, Vol. 91, No. 5, 2003.

[7] Full CAD Flow for Heterogeneous FPGAs: Version 5 available online at: http://www.eecg.utoronto.ca/vpr/

[8] Predictive Technology Models: Available online at: http://www.eas.asu.edu/~ptm/latest.html

[9] ITRS 2007: Emerging Research Devices, available online at: http://www.itrs.net/Links/2007ITRS/2007_Chapters/2007_ERD.pdf

[10] N. Bruchon et al., "New non-volatile FPGA concept using Magnetic Tunneling Junction", *IEEE Computer Society Annual Symposium on Emerging VLSI Technologies and Architectures*, 2006.

[11] W. Zhao et al., "Integration of Spin-RAM Technology in FPGA circuits", *ICSICT*, 2006.

[12] S. Paul et al., "Hybrid CMOS-STTRAM non-volatile FPGA: Design challenges and optimization approaches", *ICCAD*, 2008.

[13] Intelligent FPGA Architecture Repository: Available online at: http://www.eecg.utoronto.ca/vpr/architectures/architecture_table.html

[14] ITRS 2008 update on emerging research devices. [Online]: http://www.itrs.net/Links/2008ITRS/Update/2008_Update.pdf