Overview of Circuits, Systems, and Applications of Spintronics

Swaroop Ghosh, Anirudh Iyengar, Seyedhamidreza Motaman, Rekha Govindaraj, Jae-Won Jang, Jinil Chung, Jongsun Park, Xin Li, Rajiv Joshi, and Dinesh Somasekhar

***Abstract*— Spintronic technologies have demonstrated significant promise due to multitude of features that can find applications in storage, cache, non-volatile combinational logic, sequential logic, search engines, security primitives, and, neuro-inspired computing to name a few. This paper reviews well-known spintronic devices, and circuits and systems that exploit their special properties for novel applications. Analysis indicates substantial benefits in area, energy-efficiency and performance compared to conventional complementary metal oxide semiconductor (CMOS) technology.**

***Index Terms*— Associative computing, convolutional neural network, digital signal processor, domain wall memory, embedded memory, hardware security, spin-transfer torque RAM (STTRAM), spintronics.**

I. Introduction

SPINTRONIC technology operates on the principle of manipulation of electron spins to perform computation and storage. Contrary to charge-based computing spintronic computing requires less energy to switch the output making them energy-efficient. The non-volatility is also desirable for many applications especially energy-constrained Internet-of-Things (IoT) [1] that mostly stay OFF and occasionally perform computing. Persistence of information during inactive cycles saves re-initialization energy.

The structure of spintronic devices is an active area of research. Magnetic RAM (MRAM), spin-transfer torque RAM (STTRAM), domain wall memory (DWM), and spin memristors are some of the most investigated spintronic devices [2], [3]. Interestingly, a variety of new structures have been proposed to suit particular applications. Examples include spintronic devices for interconnects, full adders, neurons, synapses, analog-to-digital converters (ADC) and digital-to-analog converters (DAC) [38]. The feasibility of such structures have been validated using experimental demonstration as well as through micro-magnetic simulations.

Manuscript received July 31, 2016; accepted August 8, 2016. Date of publication September 7, 2016; date of current version September 9, 2016. This work was supported by the NSF under Grant CNS-1441757 and SRC Grant 2442.001.

S. Ghosh, A. Iyengar, S. Motaman, and J.-W. Jang are with Pennsylvania State University, Old Main, State College, PA-16801 (e-mail: szg212@psu.edu; asi7@psu.edu; sxm884@psu.edu; jxj328@psu.edu).

J. Chung, and J. Park are with Korea University, (e-mail: jinil_chung@korea.ac.kr; jongsun@korea.ac.kr).

R. Govindaraj is with the University of South Florida, (e-mail: rekhag@mail.usf.edu).

X. Li is with Carnegie Mellon University, (e-mail: xinli@ece.cmu.edu).

R. Joshi is with IBM T.J. Watson Research Center, (e-mail: rvjoshi@us.ibm.com).

D. Somasekhar is with Intel Labs, (e-mail: dinesh.somasekhar@intel.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JETCAS.2016.2601310

Spintronic devices possess properties such as serial access, polarization dependent resistance, asymmetric read/write latencies, current-based switching and dynamics of magnetization, non-linearity, chaotic dynamics, noise sensitivity, and, non-volatility [9]–[18]. These properties can be tied with appropriate applications for area, energy-efficiency and quality. In this paper, we illustrate application of STTRAM and DWM for associative search, state retentive sequential elements, digital signal processing, neuro-inspired computing, hardware security and last level cache (LLC).

The rest of the paper is organized as follows. Section II reviews few flavors of spintronic devices such as STTRAM, MRAM, and DWM. The circuit design to realize CAMs and sequential elements and are presented in Section III. The application of spintronics for hardware security primitives and cache is described in Section IV and V respectively. DWM-based DSP modules and neural network are presented in Section VI. The outstanding issues are described in Section VII. Conclusions are drawn in Section VIII.

II. Basics of Spintronic Elements, Opportunities, and Challenges

A. Technology

1) STTRAM and MRAM: STTRAM cell contains magnetic tunnel junction (MTJ) as the storage element. MTJ contains a free layer and a pinned magnetic layer (a schematic is shown in Fig. 1(a)). The resistance of the MTJ stack is high (low) if the free layer magnetic orientation is anti-parallel (parallel) compared to the fixed layer. The configuration of the MTJ can be changed from parallel (P) to anti-parallel (AP) or vice versa. The switching of free layer is achieved by field-driven or current-driven techniques. The field-driven MTJ is the basis for MRAM technology [3] which is promising due to high-density, low standby power, and high-speed operation. STTRAM [4] is energy-efficient variant of MRAM where the switching of magnetization is based on spin-transfer-torque using current. Fig. 1(a) and (b) shows the schematic of MRAM and STTRAM bitcell respectively. In MRAM a torque in appropriate polarity is induced on the free layer of the MTJ during write by generating fields through digitline and bitline (the isolation transistor is kept off). In STTRAM, the write is done by injecting current from the source-line to the bitline or vice versa.

2) DWM: DWM consists of three components: (a) write head, (b) read head, and (c) magnetic Nanowire (NW).

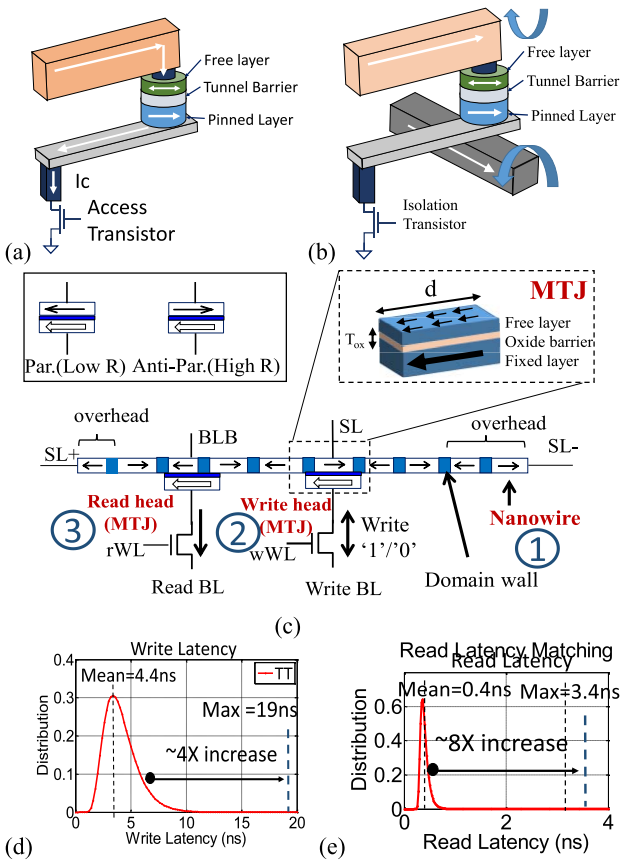

Fig. 1. Schematic of (a) STTRAM, (b) MRAM, (c) DWM bitcell, (d) write latency distribution, and (e) read latency distribution.

As shown in Fig. 1(c), the read and write heads are similar to the conventional magnetic tunnel junction (MTJ) whereas NW holds the bits in the form of magnetic polarity. Since a single bit-cell can hold multiple bits in the NW, this memory technology provides high-density. The NW is analogous to a shift register and typically contains physical notches to move the DW in a lockstep fashion [15], [21], [22]. It also ensures that the DW does not land in between two notches. The shift pulse is enough to dislodge the DW and shift along the NW. Note that the MTJ forms naturally between the NW and the fixed magnetic layer that are separated by the tunnel oxide barrier. The parallel(anti-parallel) magnetic orientation in the NW can be regarded as '0' ('1').

The domain walls (DWs) form between domains of opposite polarities where the local magnetization changes its polarity. The DWs can be shifted forward and backward by injecting charge current from left-shift (SL+) and right-shift (SL−) contacts. The new bits are written by first pushing current through shift contacts to move the bits in lockstep fashion to bring the desired bit under write head. Next, spin polarized current is injected through write MTJ (using Write BL and SL) in positive or negative direction to write a '1' or '0' in the NW. The writing involves current induced spin-transfer torque to flip the magnetization of the free layer (NW in this case). The bits are shifted back to initial state after the write operation. Read is performed by bringing the desired bit under read head using shift and sensing the resistance of MTJ formed by DW under the read head [using Read BL and BLB in Fig. 1(c)]. The bits are shifted back to initial state after the read operation. For random access, the worst case latency is the summation of number of shifts and read/write latency. However, for serial access the latency is a summation of single shift and read/write latency.

B. Opportunities

1) *Non-volatility:* Non-volatility is desirable in many applications to eliminate leakage power. It can provide memory bandwidth in high-performance computing, instant-ON feature and energy-efficiency in mobiles, and, power-efficiency in IoT.

2) *Resistive storage:* The ability of the memory to store the bits in terms of resistance is desirable for associative computing applications. It is also useful in non-Boolean computations such as sorting, max/min and floor/ceiling operations.

3) *Shift-based access:* Ability to access bit using shift especially in DWM is desirable in signal processing applications where access in serial in nature.

4) *Relationship between write current and latency:* The write latency of spintronic memories is a function of write current. The latency decreases with higher current which can be exploited to split the cache in faster and slower segments for energy-efficiency.

5) *Relationship between shift current and latency:* The shift latency of spintronic memories is a function of shift current. The latency decreases with higher current which can be exploited to split the cache for energy-efficiency.

6) *Small footprint:* The footprints are 6–20 F^2 for STTRAM and 2.56 F^2 for DWM [23]. The small footprint is important due to two reasons: 1) it increases the size of the memory while staying within the same footprint as SRAM; and 2) it reduces the interconnect length if the memory size is kept same.

C. Challenges

The challenges in spintronic devices include long and asymmetric write latency, write current and poor sense margin. Process variation acerbate these issues further.

1) Long write and read latency: One of the primary challenge of STTRAM, MRAM, and DWM is long write and read latency. Additionally, process variations in the STTRAM bitcell increases write latency further for large cache. Similarly the read latency is also degraded due to process variations, thus resulting in a long tail in write and read latencies which leads to significant performance degradation and area overhead [Fig. 1(d) and (e)]. It must also be noted that the process variations in the read head can reduce the TMR and read current, which in turn can increase the sense time. For DWM, the read/write access latency is the summation of read/write latency and shift latency.

2) Higher write current: The write current of conventional STTRAM, DWM, and MRAM is high and asymmetric. Reducing write current is important to minimize the dynamic power. New device structures such as perpendicular magnetic anisotropy (PMA) STTRAM has been investigated to

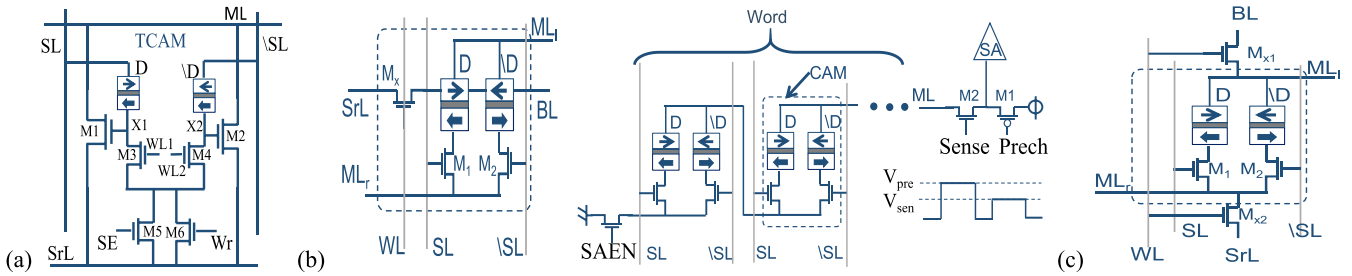

Fig. 2. Circuits of proposed TCAM cell (a) 6T-2MTJ TCAM [31], (b) 3T-2DW BCAM and the array [32], and (c) 4T-2MTJ TCAM [32].

minimize the write current. Additionally, the write operation in MTJs is heterogeneous due to its inherent asymmetry in writing the two logic states. This arises due to the variation in current polarization, which is higher in one state than the other, and the variation of charge current injection.

3) Poor sense margin (SM): Sense margin of STTRAM depends on TMR (tunnel magnetic ratio). Due to poor TMR, the voltage/current differential between high and low resistance decreases which degrades the SM. Besides, SM is asymmetric in nature [26]. In [5], a sizing methodology is proposed to improve the SM of MRAM arrays. A new self-referencing and ratio matching method is introduced in [6] to alleviate read disturb and improve the SM. Negative resistance read and write technique has been described in [7] to eliminate read disturb and reduce the write power. Reference voltage (V_{ref}) biasing has been explored in [8] to shift margins between polarities to improve the robustness. A simultaneous transistor sizing and clamp and reference voltage biasing technique is proposed [53] to maximize the sense margin.

III. Application in Associative Search And State Retention

In this section, we present the application of spintronics in associative computation such as content addressable memory (CAM) and state retentive sequential elements.

A. Basics of CAM

CAM [30] finds numerous applications in pattern matching, internet data processing, packet forwarding, and storage of tag bits in processor cache. Conventional CAM suffers from area and power overhead both due to storage bit and comparison circuitry. The need to store and match "*don't care*" requires two storage bits which further worsens the area overhead. CMOS CAM is power hungry due to power consumed in match line (ML), search line and leakage of the bit cell. In nanometer technologies leakage power constitutes a major fraction of the total power consumed in CAM memory. Spintronic technologies are area-efficient and can also provide zero leakage. In the following paragraphs we have presented examples of spintronic binary and ternary CAMs (BCAMs and TCAMs) [33].

1) 6T-2MTJ TCAM [31]: i) CAM Design: The proposed TCAM is shown in Fig. 2 (a). Two MTJs store data (D) and complementary data (\D) respectively. Transistors M1 and M2 form ML discharge network depending on the result of data comparison with the search lines SL and \SL. During search, transistors M3/M5 and M4/M5 along with MTJs make a voltage divider network in which the drain voltages of M3/M4 drive the gates of discharge transistors M1/M2. The cell is designed in such a way that during *match* the voltage of node X1 and X2 are below the threshold voltage of M1 and M2, and the ML stays precharged. However, during a *mismatch* the voltage of X1 (or X2) rises above the threshold voltage of M1 (or M2) discharging the ML.

Transistor M3/M4 are the word line (WL1/WL2) selection transistors and M6 is the write access transistor that turns ON only during write (Wr) operation. Transistor M6 is sized larger to allow sufficient write current. Transistor M5 is driven by Search Enable (SE) signal and, sized to limit the MTJ current for read disturb free search operation. *Don't care* bit can be stored in the cell by storing '1' in both D and \D bits. The search bit can be masked by driving SL= \SL= 0 on the search lines. The source line (SrL) is used for two purposes namely (a) write operation when the SrL is connected to 0 or V_{dd} depending on the write data to the MTJs and (b) search operation when SrL is driven to 0 to allow voltage division.

ii) *Operation:* In the proposed TCAM [Fig. 2(a)] the search lines SL and \SL are used to write data to the MTJs. Writing '1' and '0' consume two cycles to write to the two MTJs while 'X' can be written in a single cycle. During write the ML precharge is disabled to avoid power consumption from the ML. This is achieved by pulling the ' *precharge* ' signal high. NMOS transistor M6 is turned ON during write by WR signal. The SE signal is pulled to ground which disables transistor M5. The WL is turned ON only for the selected word so that the unselected cells are unaffected. The source line SrL is controlled appropriately to write a '1' or '0'. The respective word line (WL1/WL2) is activated by pulling high in the two cycles of a write operation. The 'X' state can be stored by writing logic 1 to both D and \D. The SrL is pulled to $V_{dd}$and the search lines SL and \SL are pulled low.

Search is a single cycle operation in CAM. The ML is precharged to V_{dd} and Wr is pulled to ground. The SrL is pulled to ground throughout the search operation. Next SE and WL is pulled high to enable the conducting path through M5 and M3/M4. The mismatch or match voltage is developed depending on the match or mismatch respectively at the gate of M1 or M2. The search lines SL is pulled to V_{dd} and \SL are pulled low to search a logic '1'. Similarly, SL is pulled low and, \SL is pulled to V_{dd} to search for logic '0'. Both SL

and \SL are pulled low to search '*X*'. The proposed TCAM consumes 4.7 fJ/bit search, can support up to 256 bits word search.

2) 3T-2DW BCAM [32]: i) CAM Design: The CAM bitcell shown in Fig. 2(b) contains two DW-based MTJs (D and \D) and three transistors (M_1, M_2 and M_x). The bit cell requires search lines (SL and \SL), wordline (WL), bit line (BL), source line (SrL), and match lines (ML_r and ML_l). The 3T-2DW CAM allows storage of complementary bits in MTJ. The high resistance corresponds to mismatch whereas low resistance corresponds to match. Therefore, it can only allow BCAM functionality.

ii) *Operation:* The write operation is performed by turning ON transistor M_x and shifting the DWs in the MTJs using SrL and BL. The MTJs are connected to write complementary bits. The write polarity on MTJs is controlled by modulating the direction of current. A '0' is written by making (SrL, BL)=(1,0) whereas a '1' is written by (SrL, BL)=(0,1). The search line transistors M_1 and M_2 are kept OFF which in turn isolate the CAM bitcell by disconnecting ML_l and ML_r. The write speed is the time needed to shift the DW under the read MTJ which is $\sim$0.5 ns for a 10 n$\times$10n $\times$ 10n nanowire. The search operation is performed by turning OFF write access transistor M_x and putting the search value on SL and \SL. During match (mismatch) a high (low) resistance is connected between ML_l and ML_r.

The BCAM array [Fig. 2(b)] contains a precharge transistor (Mp) on one end of match line and a search enable transistor (Ms) on the other end. A sense transistor (Msen) is connected between ML and output node. The ML is precharged to Vdd before search operation. The search begins by enabling the search transistor Ms and pulsing the sense transistor using voltage Vsen. During full match the resistance of the ML stack is high resulting in a discharge rate. The mismatch lowers the effective resistance of the ML increasing the discharge rate. A very fine voltage difference between match and mismatch develops when Vsen is OFF. Once the ML discharges depending on the match or mismatch the sense transistor Msen is pulsed with second voltage Vsen. The threshold voltage of the Msen and gate voltage Vsen is crucial to distinguish full match and one-bit mismatch for robust sensing operation. During match the Msen turns ON and output discharges quickly however during mismatch Msen turns OFF (or conducts weakly) discharging the output slowly. Consequently a sense margin develops between match and mismatch cases that can be sensed by a sense amplifier by employing a reference voltage. The value of reference voltage should be between the match and mismatch voltages. The proposed TCAM can support 16 bits and consumes 0.09 fJ/bit search in 16-bit word.

3) 4T-2MTJ TCAM [32]: (i) CAM Design: The writing in proposed CAM is based on DW shift. Furthermore, the CAM lacks ternary search capability. In absence of DW nanowire the design can be modified to incorporate MTJ and additional write circuitry. The circuitry can also be adjusted to allow TCAM functionality. Fig. 2(c) shows a 4T-2MTJ CAM bit cell. The core remains same except DW nanowires are replaced by MTJs. The access transistor M_x is replaced by two new transistors M_{x1} and M_{x2}. The peripheral contacts (SL, \SL, WL, BL, SrL, ML_r, and ML_l) remain same as before. The CAM structure allows storage of all possible combinations of values in MTJs. Therefore TCAM functionality can be performed using the proposed CAM.

(ii) Operation: The write operation is performed by turning ON access transistors M_{x1} and M_{x2}, and passing current in the required direction by controlling the voltage of SrL and BL appropriately. The write polarity on MTJs is controlled by modulating the direction of current. A '0' is written by making (SrL, BL)=(1,0) whereas a '1' is written by (SrL, BL)=(0,1). The search line transistors M_1 and M_2 are turned ON one-by-one to write MTJ1 and MTJ2, respectively. The match lines ML_l and ML_r are disconnected from unselected neighboring cells by keeping their SL and \SL at '0'. Writing of 'X' is accomplished by turning both M_1 and M_2 ON and writing '1' in parallel. In the proposed architecture writing complementary values in MTJ1 and MTJ2 is done serially whereas writing 'X' is done in parallel. Alternatively, the MTJs can also be flipped with respect to each other to enable parallel writing of complementary bits and serial writing of 'X' value. The search operation is performed by turning OFF write access transistors M_{x1} and M_{x2} (same as before) and putting the search value on SL and \SL. During match (mismatch) a high (low) resistance is connected between ML_l and ML_r.

B. Enhanced Scan Enabled Non-Volatile Flip-Flop (ES-NVFF)

Enhanced scan flip-flops are widely accepted form of sequential design-for-test technique to enable two-pattern delay testing [56]. We incorporate the store and restore functionality in the hold latch of the enhanced scan circuitry. We propose following flavors of ES-NVFF:

1) Base ES-NVFF [57]: It consists of two parallel latches to allow normal, enhanced scan and store-restore operations [Fig. 3(a)]. The output of the master latch is provided to the slave as well as the NV latch. The HOLD and REST signals control the operating mode of the FF. The primary property of this design is that the writing of the MTJ is only during the negative phase of the clock cycle. Furthermore, the write of MTJs take place serially. In the following paragraph we describe the operation of ES-NVFF in detail.

Normal Mode: During this mode, both the HOLD and REST signals are low, which sets the 'ST' signal, enabling transmission gate T1 and disabling T2 that is controlled by the 'HOLD' signal. The data from the master stage is fed to both the parallel latches. The output Q is driven by the slave latch. While the slave is pushing data out, it is also stored into the NVFF in parallel using control signals (SEN and CTRL).

Store Mode: Signal 'SEN' is enabled (activates the access transistor (T_A)) during the negative CLK phase. The CTRL signal is pulsed high for half of the 'SEN' signal to enable the writing of the MTJ with '0' node voltage (SN1/SN2). A voltage difference is created between the nodes (SN1/SN2) and CTRL, which provides the current to switch the state of the MTJ magnetization. In the next half phase, the CTRL is made low to enable the MTJ with '1' node voltage (SN1/SN2)

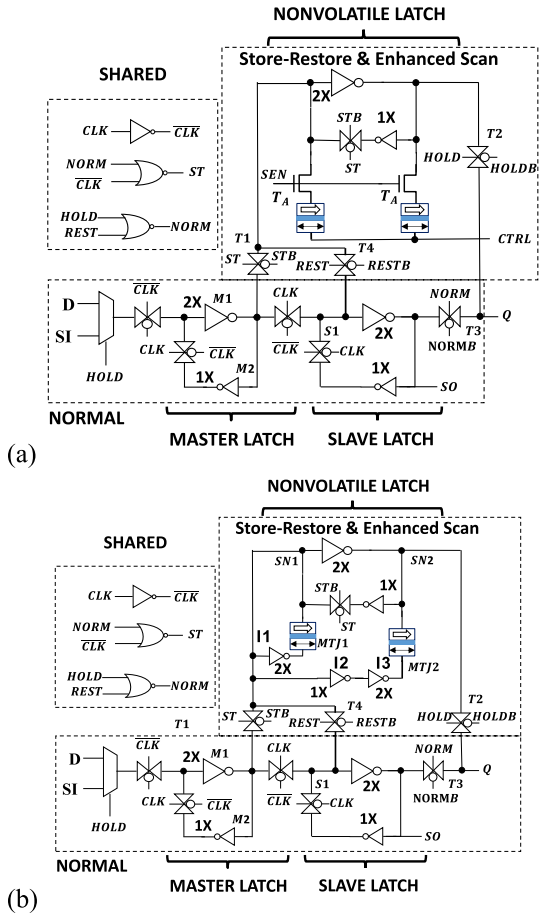

Fig. 3. Schematic of the proposed (a) base ES-NVFF circuit and (b) HPES-NVFF circuit.

to generate a current that switches its magnetization state to '1'. Note that all the MTJ stores are done sequentially which makes the operational frequency of the flip-flop dependent on the MTJ write latency.

Enhanced Scan Mode: The test pattern V_1 is first shifted into the NV latch of all the flip-flops by the scan-out chain (SO). The HOLD signal is then asserted which makes NORM= 0 (disabling transmission gate T3) and enables transmission gate T2. The output is driven by NV latch while the second test pattern V_2 is scanned through scan-in (SI) port and shifted in the scan chain. Next HOLD is made '0' (re-enabling T3) and the two-pattern transition is injected into the combinational logic. Note that the test clock is typically slower than functional clock. Therefore, MTJ write latency is not critical for performance.

Restore Mode: Initially, the Vdd and SEN are ramped up while maintaining a low CLK signal. Due to a difference between the resistances of the two MTJs (opposite magnetization states), their respective current drivability equally vary. As the node voltage (SN1/SN2) rises, the voltage is being drained through the MTJ resistance. Due to a difference in the path resistance, a mismatch in the current is generated, which in turn leads to a voltage difference between SN1 and SN2. This enables the back-to-back inverters to latch to the corresponding logic state [as observed in Fig. 3(a)]. Once latched, T4 is activated (with a high REST signal) to set the slave latch node voltage S1 with minimal contention, thus completing the restore operation.

2) High Performance ES-NVFF (HPES-NVFF) [57]: The base ES-NVFF stores the data serially during the negative phase of CLK cycle. Since MTJ write is delay intensive, it limits the frequency of the flip-flop. The primary feature of HPES-NVFF design [Fig. 3(b)] is that it allows the MTJs to be written in parallel, thereby increasing the frequency of operation. We parallelize the write of MTJs by removing the access transistor (T_A) and CTRL and using output of the master stage to drive the MTJ. Therefore the CTRL is replaced by data input of the nonvolatile latch (NVL), through the inverters (I_1, I_2 and $I_{3)}$. Inverters I_1, I_2 and I_3 form the necessary complementary inputs to the MTJ. By providing separate write drivers to the MTJs, the write operation can be performed during the entire CLK cycle.

We observe a similar functionality as that of the base ES-NVFF, however, both the MTJs are written in parallel upon receiving a new input (during high CLK). Additionally, by removing the access transistor, the resistance of the path 1 is reduced. This allows for a 5-8X larger current (by correspondingly sizing the INVs) to flow into the MTJs, thus reducing their effective write time which in turn leads to an increase in operation frequency. As much as 5X frequency benefit can be obtained compared to the base ES-NVFF at the cost of extra area overhead due to drivers. By controlling the transmission gates (T1, T2, T3, and T4) both ES-NVFF and HPES-NVFFs can store the information every cycle or backup data prior to power gating.

IV. Application in Hardware Security

In this section, we present the application of spintronics in designing physically unclonable function (PUF).

A. DWM-Based PUF [27], [48], [52]

Traditionally, unique keys were generated by the ICs for important applications such as IP security, counter-plagiarism etc. These keys are then stored on the on-chip non-volatile memory that was thought to be impervious to illegal access and duplication. However, adversaries can decode the secret key through Reverse Engineering. In order to address such issues, an auxiliary circuit i.e., PUF [28] is incorporated in the authentic chips.

PUFs are grouped under two general categories: "strong" and, "weak" PUF. A strong PUF is one where the challenge-response pairs grows exponentially with physical size and challenge parameters, whereas a weak PUF is one where the challenge-response pairs grows linearly with the size of implementation. In this work, we describe two flavors of DW based PUFs; (a) DW-relay PUF, which falls under the category of strong PUF since the number of challenge-response pairs can be exponentially increased by varying the shift pulse width, pulse magnitude and pulse frequency (as seen in Fig. 4) and, (b) memory-PUF which falls under the category of a weak PUF.

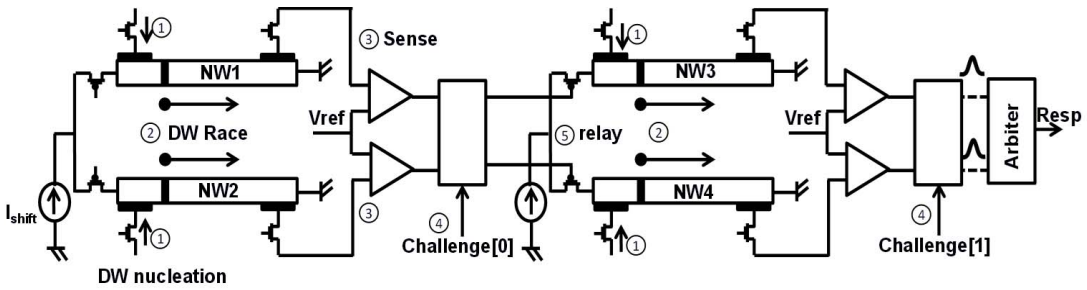

Fig. 4. Overview of the relay-PUF comprising of multiple stages of parallel NWs, write and read heads, sensing circuitry, switching block and an arbiter.

1) Relay-PUF [27], [48], [52]: In this PUF design, we exploit the concept of DW velocity variation due to inherent process variation by combining multiple NWs in multiple stages in sets of 2 [illustrated in Fig. 4]. A switching circuitry is used between each stage to switch between paths in accordance to a challenge pattern. By increasing the number of stages, the randomness of the relay-race is improved. An arbiter block is placed at the end to compare the arrival times of the respective DWs.

Challenge: In contrast to conventional delay-PUF where only the select signal to the switching circuitry (muxing) are used as challenges, the relay-PUF provides three additional sets of challenges namely shift pulse magnitude (PM), shift pulse width (PW), and the shift pulse frequency (PF). These new challenges can be employed to increase the size of challenge-response pairs.

DW nucleation and relay race: In order to incorporate adequate amount of randomness, a long chain of such NWs is used. The first step of operation is to nucleate the DWs in all the NWs. The DW race is triggered next. The read head is activated by pulsing the read word line. As soon as the resistance sensed by the read head changes (by sensing the magnetization change), the shifting of the stage is stopped. Once the read head detects the arrival of the DW (the DW reaches the end of NW), the shift signal of the following stage is fired, thus relaying the DW information to the next stage. The mux select determines whether the upper or lower DW will be fired in the following stage.

Response: The response of the relay-PUF is determined by an arbiter in accordance to the arrival of the DWs in the parallel NWs. If the top (bottom) DW reaches first the response of the PUF is 0 (1). Depending on the path the DW takes, the outcome of the race can be varied. For example, a fast DW in one NW can travel through a NW with higher surface roughness that causes it to slow down and vice versa. The response is also dependent on shift pulse challenges. The die-to-die average HD is found to be 45%. An average of 25% separation between the inter-die and the intra-die variation is observed for relay-PUF which shows good randomness and stability.

2) Memory-PUF: This PUF is similar to SRAM based PUF where the entire memory bank is used to obtain the response. The DWs in all NWs in the memory banks are fired simultaneously. The race concludes when the read signal is asserted. The DWs winning the race are set to 1 whereas the others are set to 0.

Challenge: In contrast to SRAM-PUF where the memory pattern is solely dependent on power up and variations, the DWM memory-PUF depends on both variations and shift pulse characteristics (magnitude and width). The challenges are the address of the array and shift pulse.

DW nucleation and race: Similar to relay-PUF, first a single DW is nucleated in all NWs present in the array. Next, the DWs are shifted/raced by a shift pulse challenge. The read wordline is fired after a conservative time to screen the pinned DWs at the end of the race for determining the outcome.

Response: The response of this PUF is the output of the array when a certain address is accessed for a particular pulse setting. The value of the bitcell is '1' ('0') if a high (low) resistance is read from the read head as discussed before.

The average inter-HD is found to be 50%. An average of 45% separation between the inter-die and the intra-HD is observed for memory-PUF.

B. STTRAM-Based PUF [51]

SRAM is susceptible to process variations which makes it randomly converge into one of stable states, 1 or 0. Since the strength of a cross-coupled inverter is different for each SRAM bitcell, they will have a preferred unique initializations. This will cause each bitcell to power-up to a random state and generate a response. However, the response can flip over time due to temperature and voltage fluctuations. We propose a non-volatile (NV) 7 T SRAM with embedded STTRAMs to enhance the robustness 2.3X to 20X compared to SRAM PUF) while lowering the leakage power and area overhead [Fig. 5(a)]. After the power-ON sequence the SRAM is initialized to random values. The values are programmed in the STTRAMs by turning ON the enable signal (EN) for the gating multiplexer and access transistor N4 and following two steps. 1) When PL= 0, the node storing a '1' writes a high resistance to the STTRAM. The node storing a '0' is idle since the potential difference is zero. 2) When PL= 1, the node storing '0' writes a low resistance to the STTRAM. The node storing a '1' stays idle. Once the STTRAMs are programmed it stays there and reinforces the values. When the power is turned-OFF, the SRAM forgets the initialized value however, the STTRAMs remember them. During re-initialization, the STTRAM with low resistance reinforces the '0' side ensuring that the SRAM is brought to the same state as before. This is achieved by enabling the gating multiplexer and

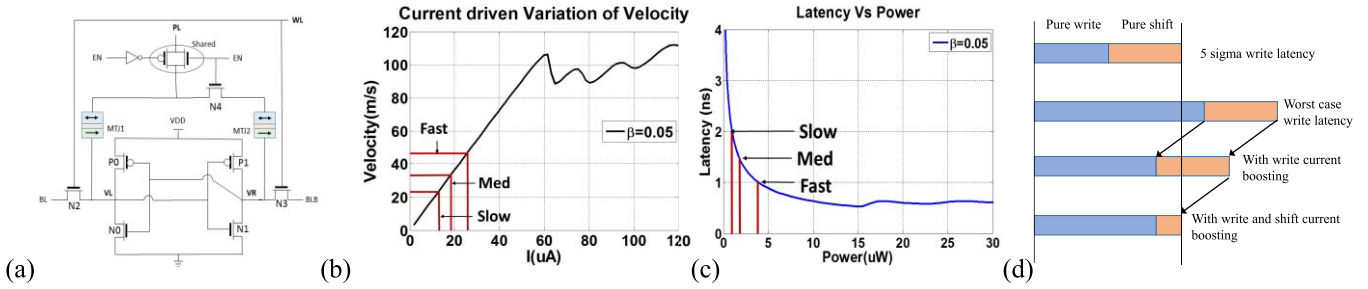

Fig. 5. (a) Schematic of NV 7T-SRAM PUFs. The gating multiplexer is distributed in column area and it is shared by all columns. (b) DW velocity versus input current using 1D model. (c) Shift latency versus power. Power for fast, medium and slow shift are shown. (d) Mitigation of process variation on write latency by write and shift current boosting.

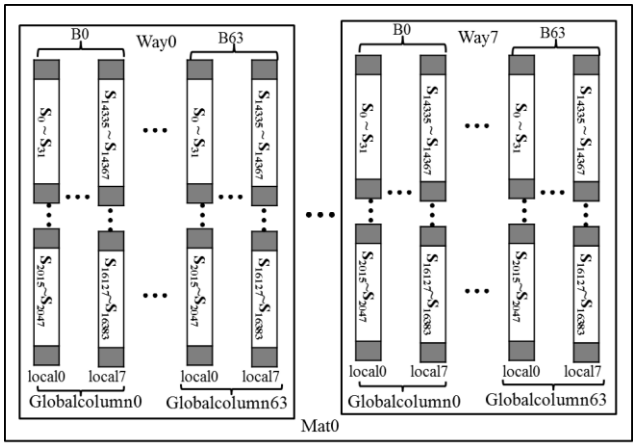

Fig. 6. Logical to physical mapping of a Mat.

N4 ON with PL= 0 before the next initialization. We compare the CMOS based arbiter and SRAM PUF with respect to DWM-based relay-PUF and memory-PUF. We use 75% of the CRPs towards training the machine learning algorithm and the rest 25% towards test. The probability of correct prediction for arbiter-PUF (relay-PUF) is 50.8% (48.4%) whereas for memory-PUF, it is 65.6% (69.1%). Therefore, the proposed spintronic PUFs perform at par with CMOS PUFs.

V. Application in Cache Architecture

In this section, we propose adaptive write and shift current modulation to exploit the relationship between shift current and DW velocity, and, write current, and write latency.

A. Cache Segregation in DWM-based Memory [23], [54]

The DW motion depends on the shift current. Higher current increases the DW velocity but increases the power consumption as well. Fig. 5(b) shows the DW velocity versus shift current by using the 1D NW model described in [25]. The corresponding DW shift latency with shift power is plotted in Fig. 5(c). We leverage this property to trade-off between shift power and latency. The fast, medium and slow caches are shifted with high, medium and low currents respectively. We assume the shift latency for the fast, medium and slow cache to be 1 ns, 1.5 ns, and 2 ns, respectively. The shift circuit of the fast, medium and slow cache is sized accordingly to enable variable shift latency.

The shift latency depends on the offset of the bit from the head. The worst case read/write latency is experienced by the bit which needs most number of shifts to reach the slowest heads. Therefore, boosting shift speed and write current together can accelerate the worst case bits. Modulation of shift speed can also be employed to fix read latency degradation. Since read latency variation is relatively less severe compared to write latency, shift boosting is sufficient to mitigate the delay degradation. Note that the current boosting for write and shift is associated with power consumption. Therefore, these knobs should be used only for the tail bits to improve the performance with minimal impact of dynamic power. The proposed approach is summarized in Fig. 5(d).

1) Cache Organization and Replacement Policy [54]: The L2 cache is divided into: i) sub-array; ii) mat that consists of a group of sub-arrays which share a common pre-decoder. Each mat contains multiple ways. A group of mats provides output cache-line (e.g., eight mats provide 64 bits each totaling 512 bits); and iii) bank that operates independently. Each way in L2 is implemented in a different subarray in mat for parallelism. For fast tag comparison, the Tag array is implemented using SRAM. The detailed logical to physical mapping is shown in Fig. 6. The sets are labeled in the NW. Each mat provides 64-bits of data by accessing a subarray. For example, way0 is accessed by enabling SA [0] of Mat [7:0] providing 512 bits of cache line. Each subarray contains 64 rows and 512 columns of 32-bit NWs. This amounts to 1 Mb data. Each mat is composed of eight subarrays (SA [7:0]). The write and shift drivers of each subarray receives global column based boost signal. Each bank contains eight mats (mat [7:0]) of total size 8MB. There are four independent banks (bank [3:0]) in the cache.

2) Cache replacement policy: If an access to L2 cache is a hit, we check whether this access is to fast way or not. If so, the access is granted and the way is marked as most recently used (MRU). For the medium way access the block is moved to fast way and marked as MRU after granting the access. LRU block from the fast way is replaced. The block replacement policy in fast way can be explained as follows: During cache access both the tag and data array are accessed simultaneously. The data is temporarily buffered in each mat. In case of hit, the content of buffer is routed to I/O ports. The latency from edge of mat to the CPU is longest and the block can be replaced

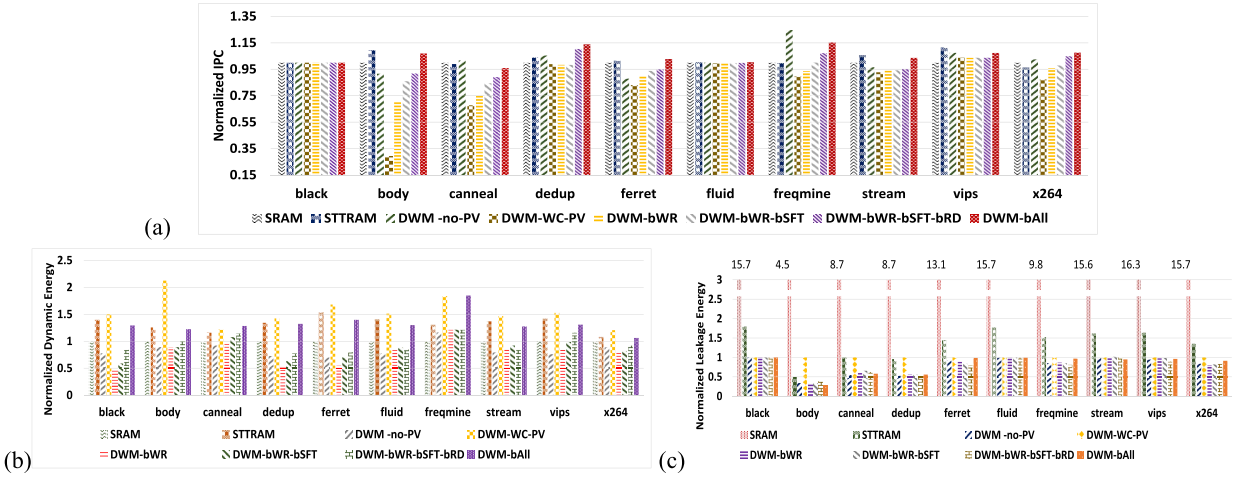

Fig. 7. (a) IPC. (b) Dynamic energy. (c) Leakage energy.

during that interval by embedding swap-enable in each way. A hit signal to a slow and medium way will trigger the swap-enable. For example, if the desired data is present in way5 and way0 is LRU way in fast ways, the accessed set from way0 is copied to way5 and the corresponding set of way5 from buffer will be placed into way0. Hence, the latency due to block swapping could be hidden.

3) Simulation Results: We performed the evaluation on a 4-core Alpha processor in Gem5 [44]. Gem5 is modified accordingly to implement cache segmentation and replacement policy. The simulations are performed over a wide range of Parsec benchmarks [45]. The cache latency and energy is achieved using CACTI [47] and hspice model of DWM [25]. For power simulation we used McPAT [46] multi-core power simulator with modified CACTI. We evaluate SRAM, STTRAM and several cases of DWM in terms of power and performance. The 32 MB cache contains 32 million MTJs. We simulate process variation for 5000 runs of Monte Carlo and find a model to fit the distribution in Matlab. Next the model is used to estimate the write and read latency distributions for 32 million MTJs. We have simulated following cases to evaluate DWM under process variations: (a) DWM-no-PV: DWM without any process variation. (b) DWM-WC-PV: DWM with worst-case write and read latency due to process variation. (c) DWM-bWR: DWM with write boosting of slow columns. (d) DWM-bWR-bSFT: DWM with write and shift boosting. (e) DWM-bWR-bSFT-bRD: DWM with write and shift boost for slow write and shift boost for slow read. (f) DWM-bAll: DWM with write and shift boosting of all columns.

The parameters from [54] are used for simulations. Mean write latency is considered for DWM-no-PV whereas worst case write latency is considered for DWM-WC-PV. We use write current of 70uA for DWM-no-PV and DWM-WC-PV and 85 μA for boosted cases. Fig. 7(a) shows the performance result represented by the normalized instruction per cycle (IPC). DWM-no-PV provides 2% performance improvement over SRAM. However DWM-WC-PV indicates that process variation can degrade the IPC by 17% on average compared to DWM-no-PV. Boosting the write current (DWM-bWR) can improve the IPC. The maximum benefit is observed for write intensive benchmarks such as dedup, body and freqmine. Boosting both write and shift current (DWM-bWR-bSFT) improves the IPC by 13% compared to DWM-WC-PV. Finally, when slow reads are fixed by boosting the shift current 18% IPC gain is observed. We also plot the IPC improvement when all global columns are boosted. This is a power intensive operation which improves the IPC by 24%.

The DWM architecture shows $\sim$ 12X saving compared to SRAM. This is owing to elimination of bitcell leakage and reduction in peripheral leakage. Fig. 7(b) and (c) shows the breakdown of total energy into leakage and dynamic energy. The proposed DWM-bWR-bSFT-bRD reduces the dynamic energy consumption by 40% compared to DWM-WC-PV due to shorter write pulse width. Furthermore, it reduces the dynamic energy by 30% relative to DWM-bAll. Therefore, the proposed read and write boosting shows 30% dynamic energy improvement compared to boosting all bit-cells and 18% performance improvement compared to worst case latency due to process variation.

B. Cache Segregation in STTRAM-Based Memory [49], [50]

We evaluate SRAM and several cases of STTRAM in terms of power and performance. The evaluations are performed on a 4-core Alpha processor in Gem5 over a wide range of parsec benchmarks. We have simulated following cases to evaluate STTRAM under process variations: (a) STTRAM-no-PV: STTRAM without any process variation. (b) STTRAM-WC-PV: STTRAM with worst-case write and read latency due to process variation. (c) STTRAM-bWR: STTRAM with write boosting of slow columns. (d) STTRAM-bAll: STTRAM with write boosting of all columns.

The parameters from [49] are used for simulations. Mean write latency is considered for STTRAM-no-PV whereas worst case write latency is considered for STTRAM-WC-PV. We use write current of 70 μA for STTRAM-no-PV and

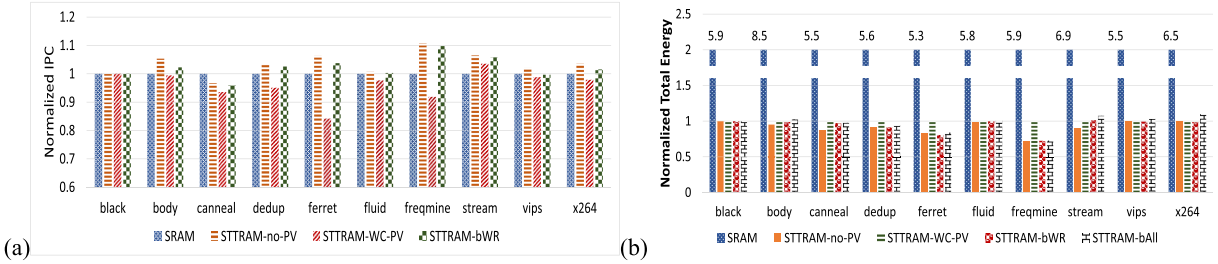

Fig. 8. (a) IPC and (b) L2 total energy comparison.

STTRAM-WC-PV and 85μ A for boosted cases. For boosted cases, we assume four sigma write latencies for normal columns and boosted columns. Fig. 8(a) shows the performance result represented by the normalized (normalized to SRAM) instruction per cycle (IPC). STTRAM-no-PV provides 4% performance improvement over SRAM. However, STTRAM-WC-PV indicates that process variation can degrade the IPC by 10% on average compared to STTRAM-no-PV. Boosting the write current (STTRAM-bWR) can improve the IPC by 13% compared to STTRAM-WC-PV. The maximum benefit is observed for write intensive benchmarks such as dedup. Fig. 8(b) shows the normalized total energy dissipation with respect to STTRAM-WC-PV. The STTRAM architecture shows ~6.4X saving compared to SRAM. This is owing to elimination of bitcell leakage and reduction in peripheral leakage. STTRAM-bAll increases the power for benchmarks dedup and freqmine because they are write intensive. The other benchmarks observe power reduction due to lower peripheral leakage.

VI. Application in Digital Signal Processing and Neuro-Inspired Computing

A. Digital Signal Processing [34], [35]

Conventionally, SRAM and flip-flop based shift registers are used as the memories in digital signal processors (DSPs). These embedded memories take a significant portion of area and power in DSP chips [4]. One of the interesting observations in DSP is that the memory access patterns are sequential and/or predictable. To save area and power of the embedded memories in DSP, the unique serial access mechanism, non-volatility and small footprint of spintronic DWM can be efficiently exploited. The area and power efficient DWM based DSP architectures have been presented in [34], [35]. Considering the shift directions, DWM based memories are classified as last-in first-out (LIFO) and first-in first-out (FIFO) as shown in Fig. 9. The number of inputs and outputs on a single nanowire (NW) are also used for the classification—i.e., single-input single-output (SISO) or single-input parallel-output (SIPO).

Fig. 10 (a) shows the architecture of the survivor memory of Viterbi decoder with constraint length $K = 7$. The survivor memory consists of three banks of single-port SRAM, where the memory write address direction is forward (address index increases) during write (WR) operation, and the memory read address direction is backward (address index decreases)

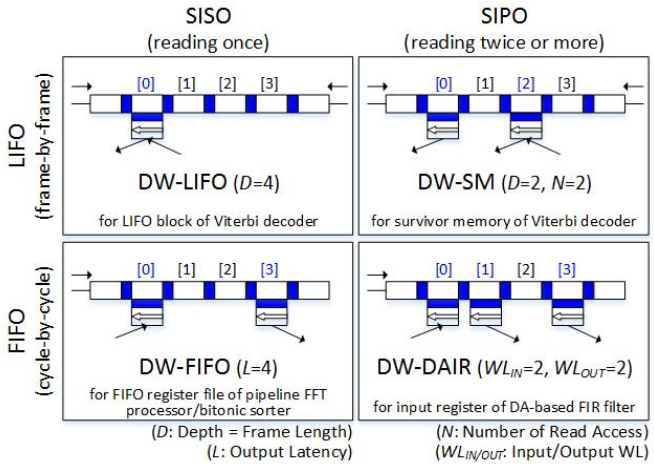

Fig. 9. Classifications of the DWM-based memories.

during traceback (TB) read and decode (DC) read operations. Those predictable survivor memory operations can be easily implemented with DWM based memory, using the LIFO with SIPO structures shown in Fig. 9. Fig. 10 (b) shows the configuration of domain wall survivor memory (DW-SM). Each nanowire of DW-SM consists of one merged read/write head, one read head, one write bit line, two read bit lines, common left/right shift inputs with 64 cells, and 64 redundant cells. Since survivor memory has a 128 bit interface, 128 nanowires construct a single DW-SM bank. To support TB read and DC read operations of survivor memory, where a written data has to be read two times, the redundant cells with read heads are added in the DW-SM. With a small number of read/write heads, this redundant DWM cells are not a large area overhead considering the conventional design approach.

Another DWM application is the Distributed Arithmetic (DA) based Finite Impulse Response (FIR) filter with 16-bit input data and constant 64 coefficients. In the conventional DA based FIR shown in Fig. 11 (a), ROM (look-up table) occupies the significant portion of the area, followed by the input regiser with DFF-based shift registers. Since the input register receives single input (bit-serial data) and generates parallel outputs (the 64-bit address for ROM), it can be replaced with Domain Wall Distributed Arithmetic Input Registers (DW-DAIR) which is composed of FIFO with SIPO input/outputs. Fig. 11 (b) shows the schematic of the DW-DAIR bank. The write head corresponds to the input of

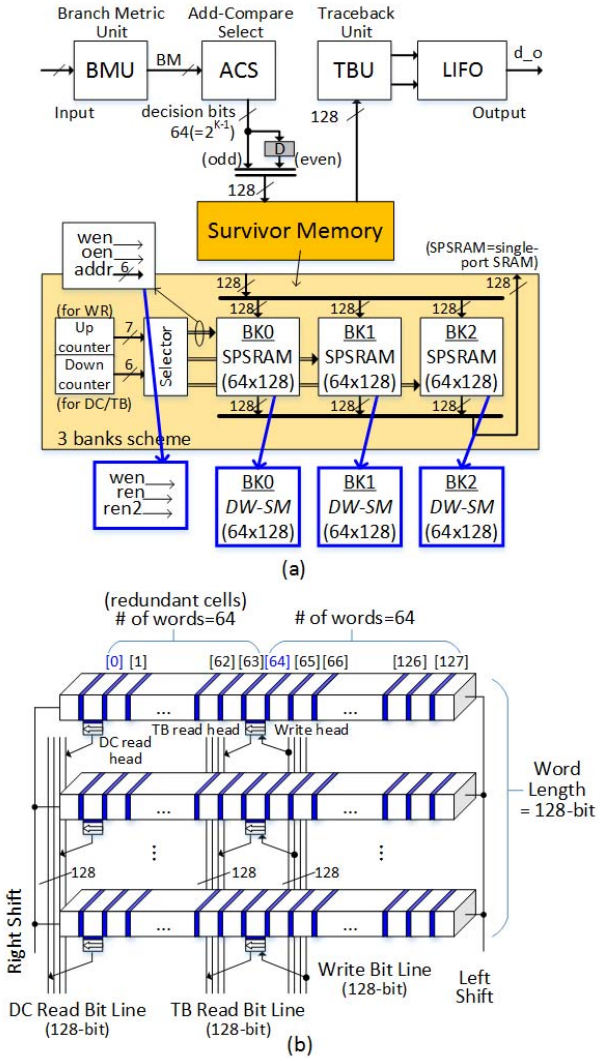

Fig. 10. Design example 1)–(a) block diagram of Viterbi decoder ($K = 7$), and (b) schematic of a single DW-SM bank (64 × 128-bit).

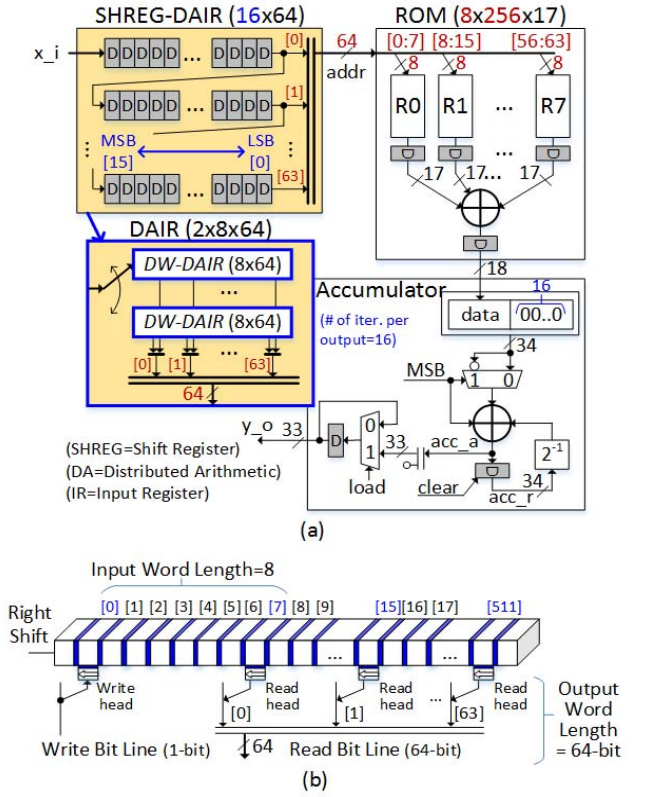

Fig. 11. Design example 2)—(a) block diagram of DA-based FIR filter ($L = 64$) and (b) schematic of a single DW-DAIR bank (8 × 64-bit).

TABLE I

LOGIC SYNTHESIS RESULTS OF VITERBI DECODER ($K = 7$) AND DA-BASED 64 TAP FIR FILTER (AT TYPICAL 1.2 V 25 ° C AND 200 MHz USING 65 nm CMOS PROCESS)

Design example with sequential addressing	Area [gate count]			Total Power [mW]		
	SRAM/ SHREG	STTRAM	DWM	SRAM/ SHREG	STTRAM	DWM
Viterbi Decoder (*K*=7)	109,721 (100%)	36,265 (33.1%)	36,873 (33.6%)	17.6 (100%)	11.3 (64.2%)	7.03 (40.0%)
DA-FIR Filter (*L*=64)	72,600 (100%)	-	61,197 (84.3%)	15.5 (100%)	-	13.1 (84.5%)

SHREG-DAIR shown in Fig. 11 (b), and the read heads is same with the 64 outputs of SHREG-DAIR that are used as the ROM address. Table I presents the numerical results of Viterbi decoder (K = 7) and DA-based 64 tap FIR filter using several embedded memories and/or shift registers. Compared to the conventional SRAM-based designs, the DWM-based designs achieve significant area and power savings. In the input register of DA-based FIR filter design, replacing the DFF-based shift registers with the DWM-based design presents 15% of area and power savings.

In [36], [37], DW-based in-memory computing approaches at memory cell and block level are proposed to reduce the communication traffic between logic and memory. Big-data and security applications [e.g., extreme learning machine (ELM) and advanced encryption standard (AES)] are presented in [36] by employing the block-level in-memory architectures.

B. Neuro-Inspired Computing [58]

Conventionally, Artificial Neural Networks (ANN) for various machine learning algorithms have been implemented using CPU/GPU based design platform. As the ANN applications are running on the mobile devices in the near future, dedicated hardware implementation of ANN for low power consumption has been of particular interests. The basic operations of an artificial neuron in ANN involve summing the N weighted inputs (or performing the dot product between N weights and N inputs) and passing the result through a thresholding (or activation) function. Spin-transfer torque (STT) device based neuromorphic computing approaches has been proposed in [38]. The STT devices are used to implement an energy-efficient analog dot product and activation function with CMOS circuits. In [39], an ANN implementation using spin-based neuron combines memristor-based crossbar (MBC) array (used as synapse) with domain wall neuron (DWN). In the approach, the MBC arrays are widely used for low cost

implementation of multiple dot product. In [40], in order to reduce the ADC/DAC overheads in MBC-based ANN, analog interconnection between MBC arrays is employed without ADC/DACs. In the deep ANN designs such as convolutional neural networks (CNNs) and deep neural networks (DNNs), one of the most challenging design issues is to have good enough computational accuracy for achieving similar accuracy with CPU/GPU based implementations [41], [42]. In order to have the accuracy of GPUs that supports floating point operations [41], [42], the required bit-width of the parameters has to be larger than 7 $\sim$ 8-bit, which is a critical programming limitation in a memristor device [43]. Without innovative memristor device improvements and analog circuit techniques, MBC-based design is hard to be extended to high accuracy applications. Hence it cannot replace the GPU-based systems as the training accelerator. In order to implement high accuracy training accelerator for image classification applications, bit-width extendability should be supported in deep ANN implementations with reasonable energy-efficiency.

To address those bit-width extendability issue, the partial dot product implementation using DWM-based cell string can be applied to the CNN convolutional layer which is the core building block of CNNs. In the DWM-based architecture (Fig. 12), partial dot products are merged together in the cell arrays. The DWM-based cell array consists of the DWM-based cell sub-arrays and ADC sub-arrays, and the current reference circuit and/or several adders are located outside the cell array. The DWM-based cell array architecture shown in Fig. 12 implement the dot product operation, where the number of weights is 25 with the output depth of 20. The operation can be considered as 20 parallel filtering operations with each filter having 25 filter taps with shared inputs (bit width of each weight= 16 and input bit width = 8). The operation can be expressed as

$$y_l[m][n] = \sum_{k=0}^{24} x_k[m] \cdot w_{k,l}[n],$$

where l = output depth index, k = weight index, m = the bit index of input, n = the bit index of weight, and w = weight. In Fig. 12, the blue box shows an example of $\sum_{k=0}^{6} x_k[m] \cdot w_k[n]$ with bit-serial operations. The operation in blue box is $x_0[m] \cdot w_0[n] + x_1[m] \cdot w_1[n] + x_2[m] \cdot w_2[n] + x_3[m] \cdot w_3[n] + x_4[m] \cdot w_4[n] + x_5[m] \cdot w_5[n] + x_6[m] \cdot w_6[n]$, where m changes from 0 to 7 and n changes from 0 to 15 in bit-serial manner. As shown in the figure, each of $w_k[n]$ is stored in MTJ and $x_k[m]$ is applied to the gate input of the selective transistor. The output of $\sum_{k=0}^{6} x_k[m] \cdot w_k[n]$ operation can be modeled as series of resistors (referred as the DWM based cell string in the figure), where the resistance values are dependent on the input $x_k[m]$ and weight $w_k[n]$.

As shown in Fig. 12, the DWM-based cell string is composed of serial connection of selective transistor and MTJ pairs. Since MTJ cell can be modeled as R_P ('0' in logic) or R_{AP} ('1' in logic) resistor, and the selective transistor has R_{on} or open state depending on the gate input $x_k[m]$, each pair can be a resistor with four resistance value– R_P, R_{AP}, $(R_{on}||R_P)$, and $(R_{on}||R_{AP})$. By assuming that $R_{on} < R_P <$

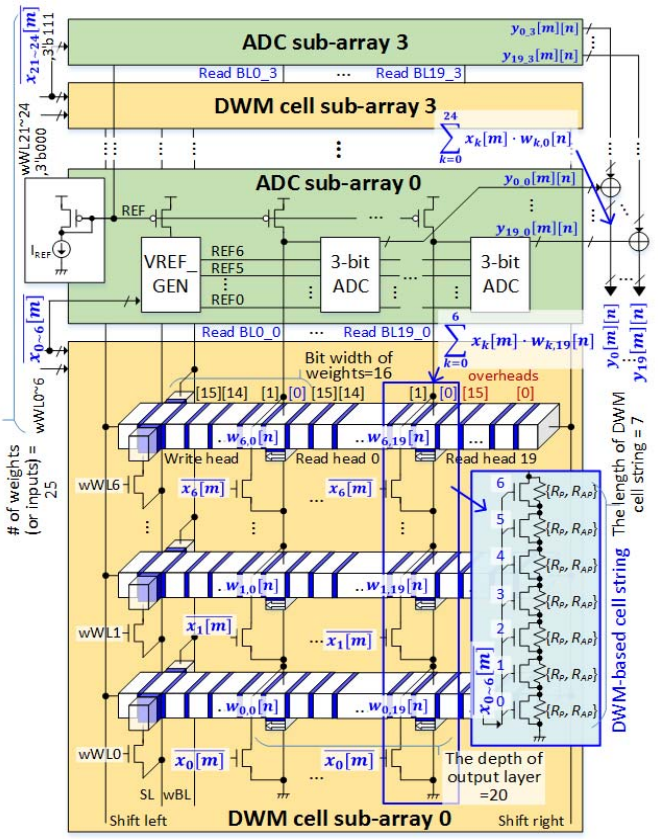

Fig. 12. Schematic of the DWM-based cell array (the number of weights = 25, the bit width of weights = 16, the depth of output = 20, and the length of DWM cell string = 7).

R_{AP} and $(R_{on}||R_P) \approx (R_{on}||R_{AP})$ in this work, each pair can be modeled as a resistor with three variable resistances $R_{AP} > R_P > R_S$ where $R_S = R_{on}||R_P$ (or R_{AP}). Here, the selective transistor performs a masking operation for the MTJ value based on the gate input. The logic value of the pair is same as AND gate output between the inverted input and the bit stored in MTJ (=the bit information of weights). The DWM-based cell string accumulates the resistance values of these pairs, and the accumulated resistance value is same as the number of '1 s' in terms of logical value. As a result, DWM-based cell string for 25 weights performs (1). The output of DWM-based CNN convolutional layer with input depth of $p = 16$, can be expressed as the following equation:

$$A_l[m][n] = \sum_{p=0}^{15} y_{l,p}[m][n].$$

$$z_l = \sum_{m=0}^{7} \left(-A_l[m][0] + \sum_{n=1}^{15} A_l[m][n] \cdot 2^{-n} \right) \cdot 2^m + b_l$$

where A = the output of adder module, b = bias, and z = the output of DWM-based CNN convolutional layer.

Since DWM-based cell string in the voltage reference circuit has only R_P of MTJ cells (however, the DWM-based cell sub-array has both R_P and R_{AP}), the reference voltage cannot completely track the read BL of MTJ cells. This effect is

amplified with long length of DWM-based cell string. Here, the length of DWM-based cell string is limited to 7. When the number of weights is more than 8, additional adders between the outputs of ADCs are required, however, this overhead is minor considering the overall design cost (in local-connected CNNs).

VII. Future Outlook and Challenges

Spintronic technologies hold significant promise due to variety of features offered by them. This paper investigates a subset of possible applications, however, we believe that other applications are also possible in non-memory areas such as non-volatile logic, data analytics and neuromorphic computing. Although spintronic technology is promising, it brings new design challenges. One particular issue pertains to non-volatility due to which spintronic memories retain data after power is turned OFF. This provides instant ON experience as the operating system and application software are retained in an initialized and executable state. But, at the same time sensitive data like passwords, cryptographic keys, credit card details, etc. are also retained after power OFF making it susceptible to scavenging, stealing and other types of attack. In volatile cache memories like SRAM and eDRAM the sensitive user data is automatically lost at power OFF. Thus, they are inherently more secure and also the embedded tamper-sensing unit turns OFF the power in the event of tampering. In case of NVMs the attacks are launched by the adversary when the user has logged off or turned OFF the machine and then the adversary logs in. The adversary can issue a read signal and in case of read hit, the data from the LLC moves to the CPU [60]. Additionally, the adversary can deliberately alter the cache content through non-invasive tampering using magnetic field and temperature to set as many valid (or dirty) bits as possible to increase the chances of retrieving information [59]. Further research in this direction can expose other security and privacy issues and appropriate countermeasures.

The other issues include the susceptibility of spintronic devices to both intrinsic variations such as saturation magnetization, anisotropy, polarization, and, physical process variations. The variations manifest themselves in long access latency and unpredictable retention time. Ensuring resilience to variations is important especially for memory applications such as LLC, main memory and storage. Finally, this paper only covers selected properties of spintronic devices. Research effort on new device structures and materials are required to identify new properties and their targeted applications.

VIII. Conclusion

We presented the applications of spintronics beyond storage such as in associative search, state retentive sequantial hardware security and digital signal processing. These applications exploit various aspects of spintronic technology for orders of magnitude area and energy reduction and substantial performance improvement. We also described future outlook and outstanding challenges to motivate further research on circuits, systems and applications of spintronics.

References

[1] F.-L. Luo, W. Williams, R. M. Rao, R. Narasimha, and M.-J. Montpetit, "Trends in signal processing applications and industry technology [in the spotlight]," *IEEE Signal Process. Mag.*, vol. 29, no. 1, pp. 174–184, Jan. 2012.

[2] M. Hosomi *et al.*, "A novel nonvolatile memory with spin torque transfer magnetization switching: Spin-RAM," in *IEDM Tech. Dig.*, 2005, pp. 459–462.

[3] M. H. Kryder and C. S. Kim, "After hard drives—What comes next?" *IEEE Trans. Magn.*, vol. 45, no. 10, pp. 3406–3413, Oct. 2009.

[4] N. D. Rizzo *et al.*, "A fully functional 64 Mb DDR3 ST-MRAM built on 90 nm CMOS technology," *IEEE Trans. Magn.*, vol. 49, no. 7, pp. 4441–4446, Jul. 2013.

[5] J.-H. Song, J. Kim, S. H. Kang, S.-S. Yoon, and S.-O. Jung, "Sensing margin trend with technology scaling in MRAM," *Int. J. Circuit Theory Appl.*, vol. 39, no. 3, pp. 313–325, 2011.

[6] Y. Chen, H. Li, X. Wang, W. Zhu, W. Xu, and T. Zhang, "A nondestructive self-reference scheme for spin-transfer torque random access memory (STT-RAM)," in *Proc. DATE*, 2010, pp. 148–153.

[7] D. Halupka *et al.*, "Negative-resistance read and write schemes for STT-MRAM in 0.13μm CMOS," in *Proc. ISSCC*, 2010, pp. 256–257.

[8] F. Ren, H. Park, R. Dorrance, Y. Toriyama, C.-K. K. Yang, and D. Marković, "A body-voltage-sensing-based short pulse reading circuit for spin-torque transfer RAMs (STT-RAMs)," in *Proc. ISQED*, 2012, pp. 275–282.

[9] A. Thiaville and Y. Nakatani, "Domain-wall dynamics in nanowiresand nanostrips," in *Spin Dynamics in Confined Magnetic Structures III*, B. Hillebrands and A. Thiaville, Eds. Berlin, Germany: Springer, 2006, pp. 161–205.

[10] J. C. Slonczewski, "Theory of domain-wall motion in magnetic films and platelets," *J. Appl. Phys.*, vol. 44, no. 4, pp. 1759–1770, 1973.

[11] S. S. P. Parkin, M. Hayashi, and L. Thomas, "Magnetic domain-wall racetrack memory," *Science*, vol. 320, no. 5873, pp. 190–194, 2008.

[12] L. Thomas, M. Hayashi, X. Jiang, R. Moriya, C. Rettner, and S. S. Parkin, "Oscillatory dependence of current-driven magnetic domain wall motion on current pulse length," *Nature*, vol. 443, no. 7108, pp. 197–200, Sep. 2006.

[13] Y. Zhang, W. S. Zhao, D. Ravelosona, J.-O. Klein, J. V. Kim, and C. Chappert, "Perpendicular-magnetic-anisotropy CoFeB racetrack memory," *J. Appl. Phys.*, vol. 111, no. 9, p. 093925, May 2012.

[14] S. Ghosh, "Path to a TeraByte of on-chip memory for petabit per second bandwidth with $<$ 5Watts of power," in *Proc. Design Autom. Conf. (DAC)*, 2013, pp. 1–2.

[15] R. Venkatesan, V. Kozhikkottu, C. Augustine, A. Raychowdhury, K. Roy, and A. Raghunathan, "TapeCache: A high density, energy efficient cache based on domain wall memory," in *Proc. Int. Symp. Low Power Electron. Design (ISLPED)*, 2012, pp. 185–190.

[16] A. J. Annunziata *et al.*, "Racetrack memory cell array with integrated magnetic tunnel junction readout," in *Proc. Int. Electron Device Meet.*, 2011, pp. 24.3.1–24.3.4.

[17] M. Mao, W. Wen, Y. Zhang, Y. Chen, and H. Li, "Exploration of GPGPU register file architecture using domain-wall-shift-write based racetrack memory," in *Proc. Design Autom. Conf. (DAC)*, 2014, pp. 1–6.

[18] R. Venkatesan, S. G. Ramasubramanian, S. Venkataramani, K. Roy, and A. Raghunathan, "STAG: Spintronic-tape architecture for GPGPU cache hierarchies," in *Proc. Int. Symp. Comput. Arch. (ISCA)*, 2014, pp. 253–264.

[19] H. Yu *et al.*, "Energy efficient in-memory machine learning for data intensive image-processing by non-volatile domain-wall memory," in *Proc. Asia South Pacific Design Autom. Conf. (ASP-DAC)*, 2014, pp. 191–196.

[20] Y. Wang, H. Yu, D. Sylvester, and P. Kong, "Energy efficient in-memory AES encryption based on nonvolatile domain-wall nanowire," in *Proc. Conf. Design, Autom. Test Eur. (DATE)*, 2014, pp. 1–4.

[21] C. Augustine, "Spintronic memory and logic: From atoms to systems," Ph.D. dissertation, Dept. Elect. Eng., Purdue Univ., West Lafayette, IN, USA, 2011.

[22] R. Venkatesan, M. Sharad, K. Roy, and A. Raghunathan, "DWM-TAPESTRI—An energy efficient all-spin cache using domain wall shift based writes," in *Proc. Conf. Design, Autom. Test Eur. (DATE)*, 2013, pp. 1825–1830.

[23] S. Motaman, A. Iyengar, and S. Ghosh, "Synergistic circuit and system design for energy-efficient and robust domain wall caches," in *Proc. Int. Symp. Low Power Electron. Design (ISLPED)*, 2014, pp. 195–200.

[24] M. Hayashi, "Current driven dynamics of magnetic domain walls in permalloy nanowires," Ph.D. dissertation, Mater. Sci. Eng., Stanford Univ., Stanford, CA, USA, 2006.
[25] A. Iyengar and S. Ghosh, "Modeling and analysis of domain wall dynamics for robust and low-power embedded memory," in *Proc. Design Autom. Conf. (DAC)*, 2014, pp. 1–6.
[26] Y. Emre, C. Yang, K. Sutaria, Y. Cao, and C. Chakrabarti, "Enhancing the reliability of STT-RAM through circuit and system level techniques," in *Proc. SiPS*, 2012, pp. 125–130.
[27] A. Iyengar, K. Ramclam, and S. Ghosh, "DWM-PUF: A low-overhead, memory-based security primitive," in *Proc. IEEE Int. Symp. Hardw.-Oriented Secur. Trust (HOST)*, May 2014, pp. 154–159.
[28] G. E. Suh and S. Devadas, "Physical unclonable functions for device authentication and secret key generation," in *Proc. 44th Annu. Design Autom. Conf.*, Jun. 2007, pp. 9–14.
[29] S. Srinivasan, "All spin logic: Modeling multi-magnet networks interacting via spin currents," Ph.D. dissertation, Purdue Univ., West Lafayette, IN, USA, 2012.
[30] K. Pagiamtzis and A. Sheikholeslami, "Content-addressable memory (CAM) circuits and architectures: A tutorial and survey," *IEEE J. Solid-State Circuits*, vol. 41, no. 3, pp. 712–727, Mar. 2006.
[31] R. Govindaraj and S. Ghosh, "Design and analysis of 6-T 2-MTJ ternary content addressable memory," in *Proc. IEEE/ACM Int. Symp. Low Power Electron. Design (ISLPED)*, Jul. 2015, pp. 309–314.
[32] S. Ghosh and R. Govindaraj, "Spintronics for associative computation and hardware security," in *Proc. IEEE 58th Int. Midwest Symp. Circuits Syst. (MWSCAS)*, Aug. 2015, pp. 1–4.
[33] R. Karam, R. Puri, S. Ghosh, and S. Bhunia, "Emerging trends in design and applications of memory-based computing and content-addressable memories," *Proc. IEEE*, vol. 103, no. 8, pp. 1311–1330, Aug. 2015.
[34] J. Chung, K. Ramclam, J. Park, and S. Ghosh, "Domain wall memory based digital signal processors for area and energy-efficiency," in *Proc. Design Autom. Conf. (DAC)*, 2015, pp. 1–6.
[35] J. Chung, K. Ramclam, J. Park, and S. Ghosh, "Exploiting serial access and asymmetric read/write of domain wall memory for area and energy-efficient digital signal processor design," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 63, no. 1, pp. 91–102, Jan. 2016.
[36] H. Yu and Y. Wang, *Design Exploration of Emerging Nano-Scale Non-Volatile Memory*. New York, NY, USA: Springer, 2014.
[37] W. Zhao and G. Prenat, *Spintronics-Based Computing*. Cham, Switzerland: Springer, 2015.
[38] K. Roy *et al.*, "Exploring spin transfer torque devices for unconventional computing," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 5, no. 1, pp. 5–16, Mar. 2015.
[39] S. G. Ramasubramanian, R. Venkatesan, M. Sharad, K. Roy, and A. Raghunathan, "SPINDLE: Spintronic deep learning engine for large-scale neuromorphic computing," in *Proc. ISLPED*, 2014, pp. 15–20.
[40] X. Liu *et al.*, "RENO: A high-efficient reconfigurable neuromorphic computing accelerator design," in *Proc. DAC*, 2015, pp. 1–6.
[41] M. Courbariaux, Y. Bengio, and J.-P. David. (2014). "*Training deep neural networks with low precision multiplications*." [Online]. Available: https://arxiv.org/abs/1412.7024
[42] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan. (2015). "*Deep learning with limited numerical precision*." [Online]. Available: https://arxiv.org/abs/1502.02551
[43] F. Alibart *et al.*, "High precision tuning of state for memristive devices by adaptable variation-tolerant algorithm," *Nanotechnology*, vol. 23, no. 7, p. 075201, 2012.
[44] Gem5. [Online]. Available: http://www.gem5.org
[45] C. Bienia, S. Kumar, J. P. Singh, and K. Li, "The PARSEC benchmark suite: Characterization and architectural implications," in *Proc. 17th Int. Conf. Parallel Archit. Compilation Techn.*, 2008, pp. 72–81.
[46] McPAT. [Online]. Available: http://www.hpl.hp.com/research/mcpat
[47] CACTI. [Online]. Available: http://www.hpl.hp.com/research/cacti/
[48] Anirudh *et al.*, "Spintronic PUFs for security, trust, and authentication," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 13, no. 1, 2016, Art. no. 4.
[49] S. Motaman and S. Ghosh, "Adaptive write and shift current modulation for process variation tolerance in domain wall caches," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 24, no. 3, pp. 944–953, Mar. 2016.
[50] S. Motaman, S. Ghosh, and N. Rathi, "Impact of process-variations in STTRAM and adaptive boosting for robustness," in *Proc. Design, Autom. Test Eur. Conf. Exhibit.*, 2015, pp. 1431–1436.
[51] J.-W. Jang and S. Ghosh, "Design and analysis of novel SRAM PUFs with embedded latch for robustness," in *Proc. 16th Int. Symp. Quality Electron. Design*, 2015, pp. 298–302.
[52] A. S. Iyengar, S. Ghosh, and K. Ramclam, "Domain wall magnets for embedded memory and hardware security," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 5, no. 1, pp. 40–50, Mar. 2015.
[53] S. Motaman and S. Ghosh, "Simultaneous sizing, reference voltage and clamp voltage biasing for robustness, self-calibration and testability of STTRAM arrays," in *Proc. 51st Annu. Design Autom. Conf.*, 2014, pp. 1–6.
[54] S. Motaman, A. S. Iyengar, and S. Ghosh, "Domain wall memory-layout, circuit and synergistic systems," *IEEE Trans. Nanotechnol.*, vol. 14, no. 2, pp. 282–291, Mar. 2015.
[55] J. Chung, K. Ramclam, J. Park, and S. Ghosh, "Exploiting serial access and asymmetric read/write of domain wall memory for area and energy-efficient digital signal processor design," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 63, no. 1, pp. 91–102, Jan. 2016.
[56] M. Bushnell and V. Agrawal, *Essentials of Electronic Testing for Digital, Memory and Mixed-Signal VLSI Circuits*, vol. 17. Springer, 2000.
[57] A. S. Iyengar, S. Ghosh, and J.-W. Jang, "MTJ-based state retentive flip-flop with enhanced-scan capability to sustain sudden power failure," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 62, no. 8, pp. 2062–2068, Aug. 2015.
[58] J. Chung, J. Park, and S. Ghosh, "Domain wall memory based convolutional neural networks for bit-width extendability and energy-efficiency," in *Proc. IEEE/ACM Int. Symp. Low Power Electron. Design (ISLPED)*, Aug. 2016, pp. 332–337.
[59] J.-W. Jang, J. Park, S. Ghosh, and S. Bhunia, "Self-correcting STTRAM under magnetic field attacks," in *Proc. 52nd Annu. Design Autom. Conf.*, 2015, pp. 1–6.
[60] N. Rathi, S. Ghosh, A. Iyengar, and H. Naeimi, "Data privacy in non-volatile cache: Challenges, attack models and solutions," in *Proc. 21st Asia South Pacific Design Autom. Conf. (ASP-DAC)*, 2016, pp. 348–353.

Swaroop Ghosh (S'04–SM'13) received the B.E. (Hons.) degree from IIT, Roorkee, India, in 2000, the M.S. degree from University of Cincinnati, Cincinnati, OH, USA, in 2004, and the Ph.D. degree from Purdue University, West Lafayette, IN, USA, in 2008.

He is an Assistant Professor in School of Electrical Engineering and Computer Science at Pennsylvania State University, State College, PA, USA. He was Senior Research and Development Engineer in Advanced Design, Intel Corp. from 2008 to 2012 and an Assistant Professor at University of South Florida from 2012 to 2016. His research interests include low-power circuit design, hardware security and digital testing.

Dr. Ghosh served as Associate Editor of IEEE Transactions on Circuits and Systems—I:Regular Papers, lead guest editor of IEEE Journal on Emerging and Selected Topics in Circuits and Systems and Journal of Low Power Electronics and Applications. He served as the Chair of DAC Ph.D. Forum (2016) and served on the organizing committees of DAC Ph.D. Forum (2015), ISQED (2016) and System Level InterconnectPrediction (2016). He has served in the technical program committees of DAC, DATE, ICCAD, CICC, ISLPED, HOST,Nanoarch, VLSI Design, ISQED, ASQED, ISVLSI, FAC and VLSI-SOC. He is a recipient of DARPA Young FacultyAward (2015), ACM SIGDA Outstanding New Faculty Award (2016), USF Outstanding Research Achievement Award (2015) and USF College of Engineering Outstanding Research Achievement Award (2015).

Anirudh Iyengar received the B.E. degree from Manipal Institute of Technology, Manipal, India, in 2010 and the M.Sc. degree in electrical engineering from the University of South Florida, Tampa, FL, USA, in 2013. He is currently pursuing the Ph.D. degree in computer science and engineering at Pennsylvania State University, State College, PA, USA.

His research is low-power and secure circuits and systems.

Seyedhamidreza Motaman received the M.Sc. degree in electronic engineering from Tehran Polytechnic, Tehran, Iran, in 2013. He is currently pursuing the Ph.D. degree in computer science and engineering at Pennsylvania State University, State College, PA, USA.

His research interests include low power circuit and system design.

Rekha Govindaraj received the Bachelor of Engineering degree (with honors) from Visweswaraya Technological University, Karnataka, India, in 2009, the Master of Technology from Indian Institute of Technology, Kharagpur, India, in 2012. She is a doctorate student at University of South Florida, Tampa, FL, USA.

She worked at Qualcomm Inc. for two years before starting her doctorate degree.Her research interests include low power VLSI circuits and systems design.

Jae-Won Jang received the B.Sc. degree in computer engineering and the Masters degree from the University of South Florida, Tampa, FL, USA. He is currently pursuing the Ph.D. degree in computer science and engineering at Pennsylvania State University, State College, PA, USA.

His research interest is hardware security.

Jinil Chung (S'13) received the B.S. and M.S. degrees from Konkuk University, Seoul, Korea, in 2002 and 2004, respectively. He is currently working toward the Ph.D. degree in the School of Electrical Engineering from Korea University, Seoul, Korea.

Since 2004, he has been with DRAM design team, SK Hynix Inc., Korea. His research interests include error correction coding design and low power reconfigurable system design.

Jongsun Park (M'05–SM'13) received the B.S. degree in electronics engineering from Korea University, Seoul, Korea, in 1998, and the M.S. and Ph.D. degrees in electrical and computer engineering from Purdue University, West Lafayette, IN, USA, in 2000 and 2005, respectively.

He joined the Electrical Engineering faculty of the Korea University, Seoul, Korea, in 2008. From 2005 to 2008, he was with the Signal Processing Technology Group, Marvell Semiconductor Inc., Santa Clara, CA, USA. He was also with the Digital Radio Processor System Design Group, Texas Instruments, Dallas, TX, USA, in Summer of 2002. His research interests focus on variation-tolerant, low-power and high-performance VLSI architectures and circuit designs for digital signal processing and digital communications.

Xin Li (S'01–M'06–SM'10) received the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, USA, in 2005.

From 2009 to 2012, he was the Assistant Director at the FCRP Focus Research Center for Circuit and System Solutions. He is currently an Associate Professor at the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA. His research interests include integrated circuit and signal processing.

Rajiv Joshi (F'02) received the B.Tech. degree from I.I.T, Bombay, India, the M.S. degree from the Massachusetts Institute of Technology, Cambridge, MA, USA, and the Dr. Eng. Sc. from Columbia University, New York, NY, USA.

He is a research staff member at T. J. Watson research center, IBM. His novel interconnects processes and structures for aluminum, tungsten and copper technologies which are widely used in IBM for various technologies from sub-0.5μ m to 14 nm. He has led successfully pervasive statistical methodology for yield prediction and also the technology-driven SRAM at IBM Server Group. He commercialized these techniques. He received three Outstanding Technical Achievement l (OTAs), three highest Corporate Patent Portfolio awards for licensing contributions, holds 55 invention plateaus and has over 200 U.S. patents and over 350 including international patents. He has authored and co-authored over 180 papers.

Dr. Joshi is recipient of 2015 BMM award. He is inducted into New Jersey Inventor Hall of Fame in August 2014 along with pioneer Nikola Tesla. He is a recipient of 2013 IEEE CAS Industrial Pioneer award and 2013 Mehboob Khan Award from Semiconductor Research Corporation. He is a member of IBM Academy of technology and master inventor. He is Distinguished Lecturer for IEEE CAS and EDS society. He is ISQED fellow and distinguished alumnus of IIT Bombay. He is on the Board of Governors of IEEE CAS society. He serves as an Associate Editor of the IEEE Transactions on Very Large Scale Integration Systems. He serves on executive committee of ISLPED and served on committees of IEEE VLSI design, IEEE CICC, IEEE International SOI conference, ISQED and IITC/AMC.

Dinesh Somasekhar received the B.E. degree in electronics engineering from the Maharaja Sayajirao University Baroda, India, in 1989, the M.E. degree in electrical communications engineering from Indian Institute of Science, Bangalore, India, in 1990, and the Ph.D. degree from Purdue University, West Lafayette, IN, USA, in 1999.

He is a Principal Engineer at Intel. His current role is Server Technology Definition for future nodes as part of Server Technology and Pathfinding, Intel, Hillsboro OR, USA.

Dr. Somasekhar served as Mentor at the Semiconductor Research Consortium, and has participated in the Technical Program Committee of ISLPED, ISQED, DATE, GLVLSI, and CICC.