INVITED: A Pathway to Enable Exponential Scaling for the Beyond-CMOS Era

Jian-Ping Wang, Sachin S. Sapatnekar, Chris H. Kim, Paul Crowell, Steve Koester
University of Minnesota
{jpwang, sachin, chriskim, crowell, koester}@umn.edu

Supriyo Datta, Kaushik Roy, Anand Raghunathan
Purdue University
{kaushik, Raghunathan, datta}@purdue.edu

X. Sharon Hu, Michael Niemier
Notre Dame University
{shu, mniemier}@nd.edu

Azad Naeemi
Georgia Tech
azad@gatech.edu

Chia-Ling Chien
Johns Hopkins University
clchien@jhu.edu

Caroline Ross
MIT
caross@mit.edu

Roland Kawakami
Ohio State University
kawakami.15@osu.edu

ABSTRACT

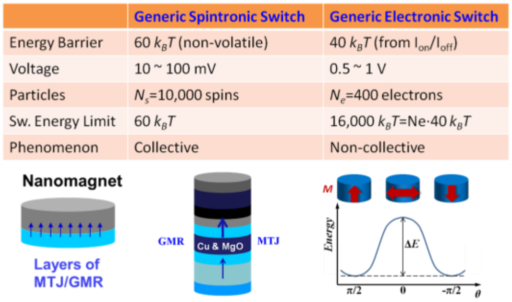
Many key technologies of our society, including so-called artificial intelligence (AI) and big data, have been enabled by the invention of transistor and its ever-decreasing size and ever-increasing integration at a large scale. However, conventional technologies are confronted with a clear scaling limit. Many recently proposed advanced transistor concepts are also facing an uphill battle in the lab because of necessary performance tradeoffs and limited scaling potential. We argue for a new pathway that could enable exponential scaling for multiple generations. This pathway involves layering multiple technologies that enable new functions beyond those available from conventional and newly proposed transistors. The key principles for this new pathway have been demonstrated through an interdisciplinary team effort at C-SPIN (a STARnet center), where systems designers, device builders, materials scientists and physicists have all worked under one umbrella to overcome key technology barriers. This paper reviews several successful outcomes from this effort on topics such as the spin memory, logic-in-memory, cognitive computing, stochastic and probabilistic computing and reconfigurable information processing.

KEYWORDS: Spintronics, spin logic, spin memory, beyond-CMOS, post-CMOS, neuromorphic computing, stochastic computing, logic-in-memory, probabilistic computing, nonvolatile computing.

1 INTRODUCTION

The inherent properties of ferromagnetic materials operating at room temperature and at the nanoscale couple with various aspects of spin physics (transport, switching, etc.), to offer abundant possibilities for developing novel memory and information processing devices . This has been a new and fruitful research direction[1,2,3,4,5] that diverges significantly from prior spintronics research. The fundamental advantage of this approach over the semiconductor-based switch concept is its projected low operation energy. Fig. 1 compares a generic spintronic switch and a generic electronic switch.

There are many unique features that arise from nanomagnet-based spintronic devices. The most apparent is nonvolatility and a superior endurance behavior, where spin-based devices outperform other nonvolatile devices for designing embedded nonvolatile memory, nonvolatile processors, and logic-in-memory arrays, as shown in Fig. 2[6,7,8,9,10,11,12,13].



	Generic Spintronic Switch	Generic Electronic Switch
Energy Barrier	60 k_BT (non-volatile)	40 k_BT (from I_{on}/I_{off})
Voltage	10 ~ 100 mV	0.5 ~ 1 V
Particles	N_s=10,000 spins	N_e=400 electrons
Sw. Energy Limit	60 k_BT	16,000 k_BT=Ne·40 k_BT
Phenomenon	Collective	Non-collective

Fig. 1 Operation energy comparison for a thermally stable nanomagnet (e.g., a free layer of an MTJ) and a generic electronic switch; Collective behavior of spin-polarized electrons coupled through different quantum mechanisms for the nanomagnet leads to unique operation energy advantages for spintronics.

In recent decades, there has been exciting progress in implementing spintronic devices with low switching energy. Fig. 3 summarizes the experimental demonstration and theoretical predictions of the switching energy based on various switching mechanisms and materials[4,14,15,16,17,18,19,20,21,22,23,24,]. Several device concepts have been predicted with high energy efficiency to approach the ideal case of 60 k_BT.

Spintronics can enable the efficient implementation of important primitive functionalities. For example: controllable interactions between spin-polarized currents and/or electrical field and nanomagnetic states open the door for the efficient implementation of functions such as the dot product; Magnetic Tunnel Junctions (MTJs) provide low-cost solutions for nonlinear activation functions; and random number generators can be built simply by using the intrinsic stochastic behavior of nanomagnets with low energy barriers.

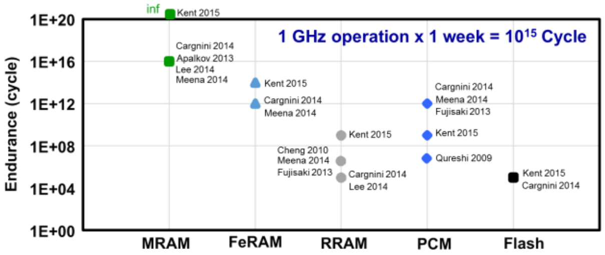


Fig. 2 Endurance of nonvolatile memory and computation devices; MRAM: magnetic random access memory cell (MTJ); FeRAM: Ferroelectric random access memory cell; RRAM: Resistive random access memory cell; PCM: Phase change memory cell;

DAC '17, June 18-22, 2017, Austin, TX, USA

ACM 978-1-4503-4927-7/17/06...$15.00
http://dx.doi.org/10.1145/3061639.3072951

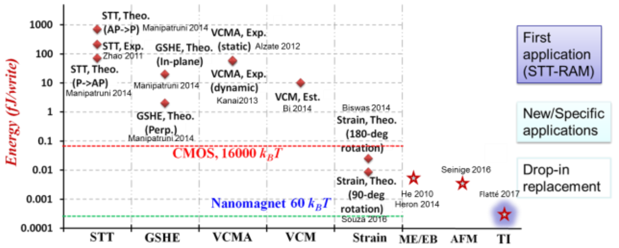


Fig. 3 Experimental demonstration and theoretical prediction of switching energy for various mechanisms. STT: Spin transfer torque; GSHE: giant spin-Hall effect; VCMA: voltage controlled magnetic anisotropy; VCM: voltage controlled magnetism; ME/EB: magnetoelectric/Exchange Bias; AFM: antiferromagnetic; TI: topological insulator;

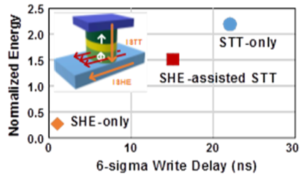
We believe that significant opportunities exit to realize compounded energy reductions beyond the improvements achieved via device and material optimizations alone when the intrinsic spin device behavior is appropriately blended with circuit, architectural, and algorithmic innovations, providing a pathway to exponential scaling. The rest of this paper focuses on key components of such a pathway.

2 NOVEL COMPUTING PARADIGMS ENABLED BY SPINTRONIC DEVICES

2.1 Spin-based Random Access Memory

For memory, spintronic devices are among the most promising candidates due to the smaller area per bit and zero leakage current[25,26]. Spin transfer torque magnetoresistive random access memory (STT-MRAM) using magnetic tunnel junctions (MTJs) has already been commercialized for specific applications, such as data centers, cloud storage, energy, industrial, automotive, consumer, and transportation markets. Toshiba and SK Hynix[27], and Samsung [28] demonstrated their prototypes for STT-MRAM in 2016. STT-MRAM offers 3x-5x higher memory cell density compared to a 6-transistor static random access memory cell, and its nonvolatility ensures that its state is maintained, without consuming leakage power, when the memory is powered down. However, the write energy of this device remains high as a large current is required for fast switching. Device scaling poses another set of challenges for STT-MRAM. Another potential candidate for spin-based memory is the spin-Hall effect MRAM (SHE-MRAM). SHE-MRAM has a decoupled read and write path with competitive memory density, and its performance advantages are shown in Fig. 4.

There are two types of magnetic anisotropy that are used for memory, in-plane magnetic anisotropy (IMA) and perpendicular magnetic anisotropy (PMA)[29]. Of these, PMA magnets are considered more suitable for scaling devices. There are challenges in working with PMA-based SHE-MRAM as SHE requires an external field to deterministically switch the magnetization. Several solutions have been proposed to address the PMA switching with SHE. Nevertheless, many of these solutions require specific fabrication processes or face challenges with scaling. One of the recently proposed solutions has used a simple composite structure that can switch a PMA without any external field, with no scaling or fabrication challenges [30].



		SHE-only		SHE-assisted STT		STT-only
Spin Hall Angle		0.4	18.8	0.4	18.8	N/A
Write Current	STT	0	0	100μA	100μA	100μA
	SHE	320μA	20μA	50μA	50μA	0
Delay Improv.		24x	24x	1.8x	28x	1X

Metrics	PMA SHE-MRAM
MTJ type	Crystalline perpendicular
Thermal stability factor	70
Free layer material	FePdX
Free layer dimension	45nmx45nmx0.7nm
Saturation Magnetization	1185 emu/cc
Damping constant, α	0.02
Crystal anisotropy, Ku (10^6 erg/cc)	9.0
TMR	130%
RA ($\Omega.\mu m^2$)	5
SHM dimension	60nmx45nmx5nm
SHM spin diffusion length	3.5nm
SHM resistivity ($\mu\Omega\cdot cm^2$)	200
Spin hall angle, θ_{sh}	0.4, 18.8
Effective external field	30 mT

Fig. 4. Write energy and write delay comparison between STT, SHE-assisted STT, and SHE. Material parameters and simulation results are provided in the two tables above.

2.2 Spin-based Logic-in-memory

In-memory processing is widely recognized as an effective approach to overcome the energy and latency bottleneck associated with fetching data from memory to a processor. In one approach[31], suitable modifications are made to peripheral circuits that enable standard STT-MRAM arrays to perform bitwise, arithmetic, and complex vector operations, providing system performance improvements of 3.93X on average (up to 12.4X), and memory system energy reductions of 3.83X on average (up to 12.4X).

An alternative solution proposes spin-based computational RAM (CRAM)[32] structures, which offer a means for true in-memory computation and can provide over 18-28X better energy-efficiency with 2.8X speed gains. The state of an MTJ-based memory cell is characterized by its resistance, and this can be leveraged to implement logic functions entirely within the array. A subarray of three 2T1MTJ CRAM bit-cells is shown in Fig. 5(a). In normal operation, the dotted transistor acts as the access transistor and the solid transistor is off. In logic mode, BL0 and BL1 are connected to Vdd and BL2 to ground, creating the resistor configuration in Fig. 5(b) and the current through the rightmost MTJ depends on the states (resistances) of the two bit-cells at left; depending on the current, this MTJ may be switched. The scheme can be used to implement functionalities such as NAND, NOR, MAJ, and others. Parallel operations can be performed simultaneously in the array, as shown in Fig. 5(c), which shows a snapshot of a dot product computation in four rows of the CRAM (the encircled cells are active).

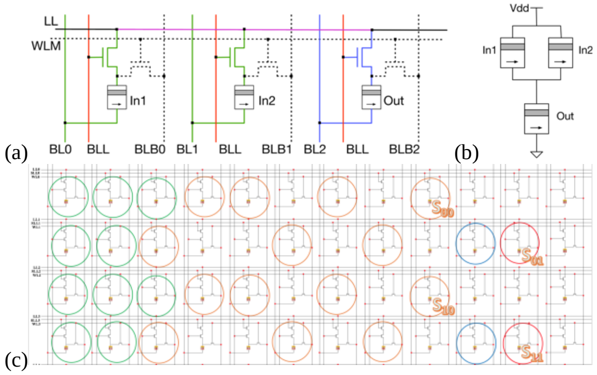


Fig. 5: (a) CRAM structure and (b) a NAND (c) a CRAM computation.

2.3 Spin-based Nonvolatile Processor

The inherent nonvolatility of spin devices is not only promising in the context of memory, but also inspires novel pathways towards extremely energy-efficient information processing. A nonvolatile processor (NVP), where the intermediate state of the processor can be saved with near-zero time/energy overhead, allows ultra-fine-grained power management, and could tolerate arbitrary power supply interruption during information processing. Such an NVP can either be based on a traditional von Neumann architecture or consist of reconfigurable computing fabrics. Depending on how the state is saved, we can classify NVPs into three categories (Fig. 6)[33]: (i) NVP with explicit backup (EB-NVP), (ii) NVP with implicit backup (IB-NVP), and (iii) NVP with hybrid backup (HB-NVP). In EB-NVPs, processor states must be explicitly backed up to and restored from NV memory. In IB-NVPs, NV devices are used to realize all state-storing elements, and there is no need for a separate backup NV memory. For HB-NVPs, the retention time of the storage elements cannot be treated as “infinitely” long, and NV memory is still needed. If the time of a power outage is shorter than the retention time, no backup/recovery is needed. Thus, HB-NVP is an effective way to trade off operating energy with backup/ recovery overhead. Most existing NVPs belong to the EB-NVP category.

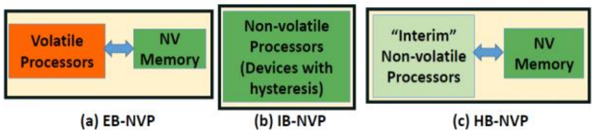

Volatile Processors
NV Memory
Non-volatile Processors (Devices with hysteresis)
"Interim" Non-volatile Processors
NV Memory
(a) EB-NVP
(b) IB-NVP
(c) HB-NVP

Fig. 6 Types of NVPs: (a) NVP with explicit backup, (b) NVP with implicit backup, and (c) NVP with hybrid backup.

Spin devices, such as ASL[3] and CoMET[5], can be used to construct IB-NVPs and HB-NVPs. We have examined two intermittent processing scenarios where such NVPs can help save significant amounts of energy. The first considers applications powered by harvested energy sources, which are frequently unreliable. Using an IB-NVP, we can eliminate the need for backup/recovery to/from NV memory, as well as the energy and delays associated with the backup and recovery operations. The second is from applications with idle intervals due to stall cycles. In both scenarios, there are benefits from the near-zero backup/restore overheads of IB-NVPs and HB-NVPs, as well as extremely low sleep state overheads. Fig. 7 illustrates energy/instruction results from a case study comparing an ASL-based IB-NVP (1st bar from left), two ASL based HB-NVPs (2nd and 3rd bars), CoMET-based IB-NVP (4th bar), CMOS+STT-RAM-based EB-NVPs (5th, 7th, and 9th bars), and CMOS+SHE-RAM-based EB-NVPs (6th, 8th, and 10th bars). Major savings are possible as the amount of backup/restore overhead can be avoided by using IB-NVPs or HB-NVPs.

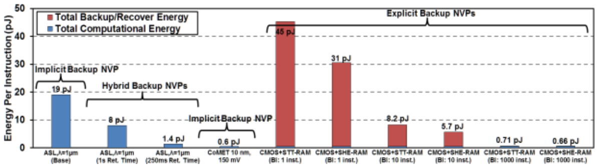


Fig. 7. Energy/instruction comparing implicit, explicit, and hybrid backup/recovery strategies with different technologies.

2.4 Spin-based Neuromorphic Computing

Recent experiments on spin-orbit torque driven domain wall motion in ferromagnet-heavy metal bilayers have opened the possibility of emulating neural and synaptic operations by single device structures. As shown in Fig. 8(a), input current flowing through an underlying heavy metal (between terminals WRITE and GND) results in spin-orbit torque induced domain wall motion in a ferromagnet lying on top[34]. The magnet is also part of a tunneling junction whose conductance is modulated by the domain wall position. The domain wall displacement, being a function of the input current magnitude, determines the final resistance state of the MTJ. Such a device structure can be used to mimic the synaptic functionality since the read current through it (due to a constant voltage between terminals READ and GND) is weighted by the device conductance[35]. Similar device structures mimicking neural operations (non-spiking[36] and spiking[37] functionalities) have been proposed. Device-circuit-algorithm co-design suggests that such all-spin neuromorphic architectures potentially yield two orders of magnitude lower energy over corresponding CMOS implementations[38].

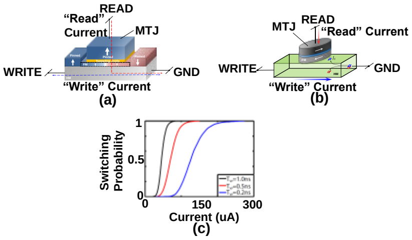


Figure 8. (a) Domain wall motion based spintronic device that acts as a building block for All-Spin Neural Networks. "Write" current through the heavy metal (HM) programs the domain wall position in the ferromagnet (FM), which modulates the conductance of the MTJ, (b) A mono-domain FM lying on top of an HM switches probabilistically due to flow of charge current through the HM.

As device dimensions start scaling, such domain wall motion based devices may not continue to exhibit such multi-bit precision. Thermal noise prevalent in such devices becomes increasingly dominant at scaled device dimensions, thereby leading to stochastic behavior. Fig. 8(b) shows a three-terminal device structure where current through an underlying heavy metal probabilistically switches a mono-domain magnet lying on top. A corresponding stochastic switching characteristic is also shown in Fig. 8(c), where the probabilistic switching characteristics can be modulated by the pulse width duration. Neuromorphic computing with such stochastic single-bit neurons[39] and synapses[40] have been recently demonstrated where the multi-bit precision requirements are replaced by probabilistic synaptic and neural updates over time. Such stochastic devices can thereby lead to highly compact neuromorphic hardware where the computing methodology leverages the underlying device stochasticity[41, 42].

2.5 Spin-based Error-Resilient and Stochastic Computing

Several major applications (e.g., image or video processing, or neural network tasks) show inherent resilience to errors. Two versions of an JPEG-compressed image, with and without approximations, are essentially indistinguishable (Fig. 9). Spin-based approximate computing leverages tradeoffs between error and circuit performance in spin-based computing structures such as all-spin logic (ASL) to reduce circuit power and increase speed with a controlled amount of injected error. Approximate logic[43,44] can reduce the number of magnets in the ASL gates that implement a functionality (e.g., in a full adder (FA)), or by providing an early clock to a computation. In each case, errors may be introduced within the truth table, but with performance benefits. ASL gates can be optimized to deliver trade-offs between power, delay, and error. For example, at quantified error levels, a four-magnet FA can be configured to reduce the delay of an accurate five-magnet FA by 46%, or its area by 42%. Realistically, to limit the maximum error, these errors are introduced to lower significant bits of a computation (e.g., when FAs are configured as *n*-bit adders). Executed appropriately, this approach can significantly improve power and delay over a conventional implementation, with about 40% power savings at iso-delay.

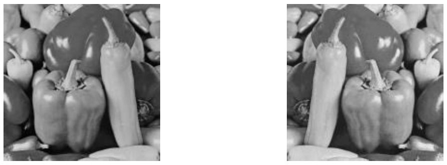
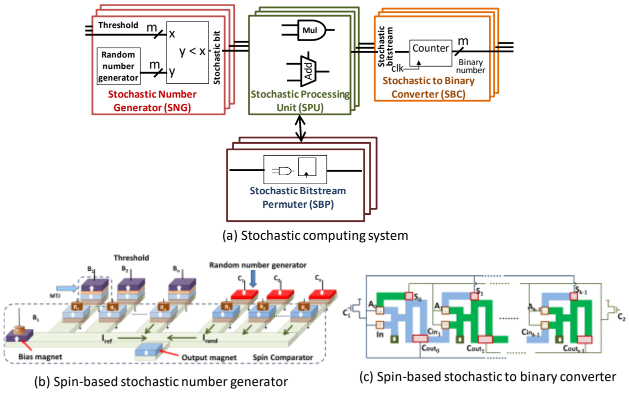


Fig. 9: The result of exact (left, PSNR=35.29) and approximate (right, PSNR=30.76) JPEG computations.

Another approach uses the principles of Shannon-inspired computing[45] to overcome high error rates within a single device to deliver reliable system-level computing. The idea is to use a high-complexity main block with low-energy gates that could have high error levels. The errors are compensated by a low-complexity estimator using low-error blocks and a fusion block that determines the best estimate of the output using the outputs of the main block and the estimator. The concept is applied to a support vector machine classifier for EEG seizure detection. A 10^{13}-fold increase in tolerable device rates while maintaining system performance has been reported.

Stochastic computing[46], which represents and processes information in the form of stochastic bit-streams, can exploit the unique characteristics of spintronic devices to realize computations in an energy-efficient manner[47]. In a stochastic computing system [Fig. 10(a)], binary numbers are converted to stochastic bit-streams using stochastic number generators (SNGs), processed using low-complexity stochastic processing units (SPUs), and converted back to binary using stochastic-to-binary converters

(SBCs). Additionally, stochastic bit-stream perimeters (SBPs) are used to ensure that the inputs to SPUs are uncorrelated. A key advantage of spintronics in realizing stochastic computing systems is that the SNGs, STBs and SBPs can be realized in a highly compact and power-efficient manner, as illustrated in Fig. 10(b) and 10(c). A second key advantage is that the low complexity and logic depth of the processing units masks the inefficiencies of stochastic logic such as static power and slow switching time. Third, the processing units can be operated with lower switching currents and/or switching times, resulting in improved energy efficiency at the cost of errors, which can be tolerated by the intrinsically fault-tolerant nature of stochastic computing. Although stochastic computing does suffer from higher processing latency due to the serial nature of stochastic bit-streams, the fine-grained parallelism across bits in a bitstream can be leveraged for vectorization or pipelining. Evaluations on a suite of signal processing, CMOS image processing and machine learning benchmarks suggest that spin-based stochastic logic implementations were ~9X more energy efficient than CMOS.



(a) Stochastic computing system

(b) Spin-based stochastic number generator

(c) Spin-based stochastic to binary converter

Fig. 10: Spin-based stochastic computing system

2.6 Spin-based Probabilistic Computing

Conventional logic and memory devices use stable deterministic units such as standard MOS transistors, or nanomagnets, with energy barriers in excess of 40−60 kT, which represent a bit (0 or 1). A very different paradigm is based on a "p-bit" that continuously fluctuates between 0 and 1, a behavior that arises naturally from the physics of low barrier nanomagnets.

Such stochastic nanomagnets can be driven by the spin current from a spin-Hall material to construct a three-terminal unit (Fig.10a) whose output $m_i(t)$ fluctuates between 0 and 1 with a mean value that can be tuned with an analog signal $I_i(t)$ applied to the input terminal (Fig.10b). We call this tunable random bit generator a p-transistor. If these can be interconnected to build p-circuits, a new class of circuits could provide novel functionality. This not only includes non-Boolean functions like optimization and inference[48, 49], but also precise Boolean logic that is invertible unlike standard digital circuits[50, 51].

The compact model[50] describing such p-circuits is essentially the same as the equations for Boltzmann machines[52, 53], which are key to machine learning, but are usually implemented in software. The physics of low barrier nanomagnets driven by the spin-Hall effect provides a natural hardware for p-transistors that could be built out of state-of-the-art materials and phenomena. Other p-transistor realizations are also possible.

Large numbers of p-transistors (Fig. 11(a)) can be interconnected into networks (Fig. 11(c)) of correlated p-bits that can perform many novel functions. Fig. 11(d) shows an example of a 32-bit adder implemented using an interconnected network of nearly 500 p-bits. Initially, when the connections are weak relative to the noise, the sum bits (S) fluctuate in an uncorrelated manner. But once the connections are turned on, they overcome the noise, and the magnets get precisely correlated to converge on the one correct answer out of 2^{33} (~ 8 billion) possibilities. When we quench a molten liquid we expect a solid full of uncontrolled defects. Instead our design yields a perfect crystal every time[50]. Remarkably, the adder is invertible as well. For example, when the output (S) is clamped to a fixed number, the inputs (A) and (B) fluctuate in a correlated manner to make A+B=S. This ability of a system to implement the inverse function has far-reaching possibilities. For example, we have shown that a 4-bit multiplier acting in the inverse mode performs integer factorization, suggesting that probabilistic computers based on robust room temperature p-bits could provide practically useful solutions to many challenging problems by rapidly sampling the phase space in hardware.

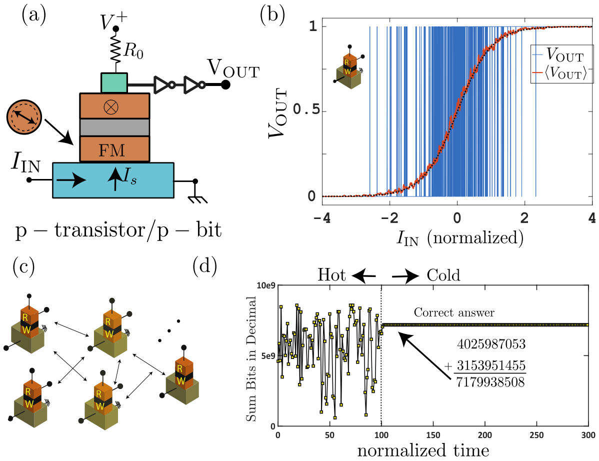


Fig. 11 (a) A possible implementation of a p-transistor to generate the required input-output characteristics for a p-bit combining a spin-Hall metal and an MTJ with a low-barrier free layer. (b) The input-output characteristics of the idealized p-bit based on the generic model. The blue line shows the real-time response of the p-bit, and the red line is the RC averaged p-bit value that follows a sigmoidal behavior. (c) A network of correlated p-bits operating as a p-circuit. (d) An example illustrating a p-circuit used to implement precise Boolean logic: a 32-bit adder implemented using a network of p-bits. Remarkably, the operation is invertible as discussed in the text.

We are currently using SPICE simulations to evaluate the energy and delay for different realizations of p-transistors which compare well with standard CMOS implementations[54] since the randomness and the summation of multiple inputs come naturally from the underlying physics[55]. More importantly, p-transistors can enable functionalities such as invertible logic that are truly novel compared to existing digital logic.

3 SPINTRONIC DEVICE BENCHMARKING

The recent benchmarking research for Boolean circuits, such as 32-bit adders, has projected a limited performance gain for only a few beyond-CMOS device candidates[56]. Research in beyond-CMOS devices is progressing fast, and the proposed devices are being continuously revised and reinvented. While such innovations are hard to predict, there is little doubt that they will make emerging devices more competitive. However, one needs to recognize that conventional CMOS devices and their corresponding circuits and architectures have evolved together over many years. Some of the emerging beyond-CMOS devices offer fundamentally different (and in some cases unique) characteristics requiring novel and nontraditional circuit concepts to realize their full potential.

To better utilize emerging spin-based technologies, alternative non-Boolean platforms based on neuromorphic circuits are quite attractive[57,58,59]. Biologically-inspired computing platforms are highly efficient for solving many problems, particularly in voice, image, and video processing, by taking advantages of massive parallel low-power computing blocks[60, 61]. Fig. 12 shows results from a uniform non-Boolean benchmarking performed for a variety of beyond-CMOS devices based on the Cellular Neural Network (CeNN) architecture. The CeNN is a suitable platform for benchmarking because a variety of charge- and spin-based devices can be used to implement CNNs efficiently[62,63,64]. Moreover, the mathematical framework for CNN circuits is well-defined and understood, facilitating benchmarking of various implementations for a given task and desired accuracy.

For the charge-based CNN implementation, CMOS HP and LV devices are employed to quantify the performance of the digital CNN and to compare against their analog counterparts. Comparing the benchmarking results for Boolean and non-Boolean circuits[65], shown in reference [56] and Fig. 12, respectively, spintronic devices shift much closer to the preferred corner and are competitive compared to charge-based devices. This is because a single magnet can mimic the functionality of a neuron and these spintronic devices operate at a low supply voltage. The domain wall device provides the best performance in terms of the EDP thanks to its low critical current requirement.

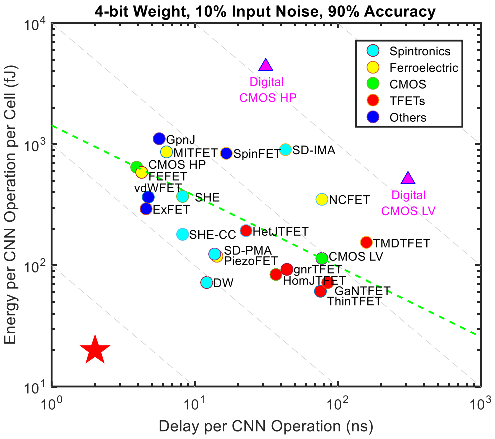


Fig. 12. Comparison of energy and delay per operation among various beyond-CMOS technologies based on analog, digital, and spintronic implementations. Triangle and circular points of charge-based devices represent the digital and analog CNN implementation, respectively. For the text labels of spintronic CNN implementation, SD, SHE, and DW stand for spin diffusion, spin-Hall effect, and domain wall motion, respectively, and CC represents for the copper collector.

4 CHALLENGES AND OPPORTUNITIES

Several challenges remain for future spintronic devices and materials. If the switching speed can be improved through experimental demonstrations, more impactful applications will be expected. Fig. 13 summarizes the experimental demonstration and theoretical prediction of nanomagnet switching speeds for various switching mechanisms[14, 15, 16,17, 66,67,68,19,69,70,71]. A demonstration of switching at 10ps could happen in the near future.

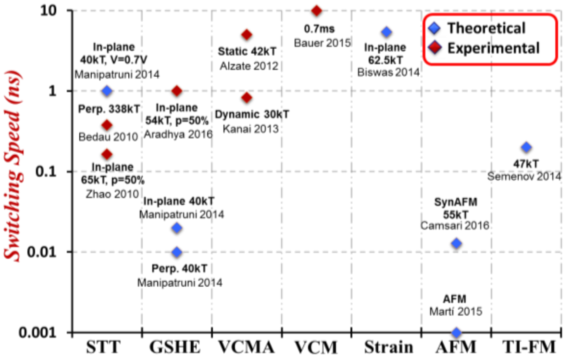


Fig. 13 Experimental demonstration and theoretical prediction of nanomagnet switching speed based on different switching mechanisms.

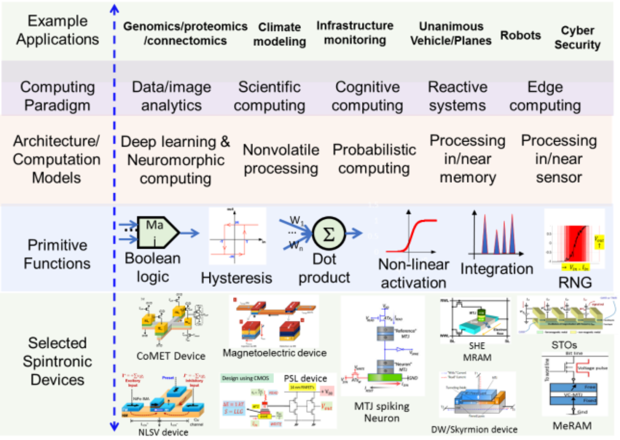


Fig. 14. Outlook for spintronic devices, from devices to applications.

There are many remaining challenges to develop practical materials that could further reduce the operation voltage of spintronic devices down to several tens of mV. Future research should emphasize heterostructured, hybrid, and composite materials that could meet a package of strict device requirements and be implemented for future spintronic devices and systems.

5 OUTLOOK

In summary, we have reviewed the opportunities and challenges of spintronic devices and several selected enabled circuits and architectures.

We believed that MRAM would become the mainstream embedded NV memory nearly a decade ago. The recent experimental and theoretical progress on spintronic materials and devices further confirm its potential to go beyond memory applications, e.g. cognitive computing and memory chips. This is not only based on the fundamental potential of the projected operation energy (60 k_BT) of nanomagnets, but also on the unique package of primitive functions of spintronic devices such as superior endurance performance and easily implemented dot product position as shown in Fig. 14. Fig. 14 also shows both the bottom-up and top-down views between spintronic devices and important applications.

ACKNOWLEDGMENTS

This work was supported by C-SPIN, one of six centers of STARnet, a Semiconductor Research Corporation program, sponsored by MARCO and DARPA. Authors thank all C-SPIN professors, students and postdoctoral fellows (www.cspin.umn.edu). Authors also gratefully thank the fruitful and encouraging discussion with Drs. I. A. Young, Manipatruni and D. E. Nikonov (Intel), Dr. C. T. Chen (IBM) and Drs. G. Sandhu, S. Kramer, Wiltod Kula (Micron). JP Wang thanks ZY Zhao and Dr. DL Zhang for preparing some figures.

REFERENCES

[1] J. Wang, et al, 2005. Programmable Spintronics Logic Device Based on a Magnetic Tunnel Junction Element. *J. Appl. Phys.* 97, 10D509. DOI: https://doi.org/10.1063/1.1857655

[2] A. Lyle, et al, 2010. Direct Communication Between Magnetic Tunnel Junctions for Nonvolatile Logic Fan-out Architecture. *Appl. Phys. Lett.* 97, 152504. DOI: https://doi.org/10.1063/1.3499427

[3] B. Behin-Aein, et al, 2010. Proposal for an All-spin Logic Device with Built-in Memory. *Nat. Nanotechnol.* 5, 4 (April 2010), 266–270. DOI: https://doi.org/10.1038/nnano.2010.31

[4] S. Manipatruni, D. E. Nikonov, and I. A. Young. 2015. Spin-Orbit Logic with Magnetoelectric Nodes: A Scalable Charge Mediated Nonvolatile Spintronic Logic. *arXiv: 1512.05428* (2015).

[5] M. G. Mankalale, et al, 2016. CoMET: Composite-Input Magnetoelectric-based Logic Technology. *arXiv: 1611.09714v2, IEEE J. Explor. Solid-State Comput. Devices Circuits*

[6] A. D. Kent and D. C. Worledge. 2015. A New Spin on Magnetic Memories. *Nat. Nanotechnol*. 10, 3 (2015), 187–191. DOI: https://doi.org/10.1038/nnano.2015

[7] L. V. Cargnini, et al, 2014. Embedded Memory Hierarchy Exploration Based on Magnetic Random Access Memory. *J. Low Power Electron. Appl.* 4, 214–230. DOI: https://doi.org/10.3390/jlpea4030214

[8] M. K. Qureshi, V. Srinivasan, and J. Rivers. 2009. Scalable High Performance Main Memory System Using Phase-change Memory Technology. *Proc. 36th Annu. Int. Symp. Comput. Archit. - ISCA '09* (2009), 24–33. DOI: https://doi.org/10.1145/1555815.1555760

[9] C.H. Cheng, A. Chin, and F.S. Yeh. 2010. Novel Ultra-low Power RRAM with Good Endurance and Retention. *Dig. Tech. Pap. - Symp. VLSI Technol.*, 5, 85–86. DOI: https://doi.org/10.1109/VLSIT.2010.5556180

[10] D. Apalkov et al. 2013. Spin-Transfer Torque Magnetic Random Access Memory (STT-MRAM). *ACM J. Emerg. Technol. Comput. Syst*. 9, 2, 1–35. DOI: https://doi.org/10.1145/2463585.2463589

[11] K. Lee, J. J. Kan, and S. H. Kang. 2014. Unified Embedded Non-volatile Memory for Emerging Mobile Markets. *Proc. 2014 Int. Symp. Low power Electron. Des. - ISLPED '14*, 131–136. DOI: https://doi.org/10.1145/2627369.2631641

[12] J. Meena, S. Sze, U. Chand, and T.-Y. Tseng. 2014. Overview of Emerging Nonvolatile Memory Technologies. *Nanoscale Res. Lett.* 9, 1, 526. DOI: https://doi.org/10.1186/1556-276X-9-526

[13] Y. Fujisaki. 2013. Review of Emerging New Solid-State Non-Volatile Memories. *Jpn. J. Appl. Phys.* 52, 40001. DOI: https://doi.org/10.7567/JJAP.52.040001

[14] H. Zhao et al. 2011. Sub-200 ps Spin Transfer Torque Switching in In-plane Magnetic Tunnel Junctions with Interface Perpendicular Anisotropy. *J. Phys. D. Appl. Phys*. 45, 2, 25001. DOI: https://doi.org/10.1088/0022-3727/45/2/025001

[15] S. Manipatruni, D. E. Nikonov, and I. A. Young. 2014. Energy-delay Performance of Giant Spin Hall Effect Switching for Dense Magnetic Memory. *Appl. Phys. Express* 7, 10, 103001. DOI: https://doi.org/10.7567/APEX.7.103001

[16] J.G. Alzate et al. 2012. Voltage-induced Switching of Nanoscale Magnetic Tunnel Junctions. *Tech. Dig. - Int. Electron Devices Meet. IEDM* (2012), 681–684. DOI: https://doi.org/10.1109/IEDM.2012.6479130

[17] S. Kanai, et al, 2013. In-plane Magnetic Field Dependence of Electric Field-induced Magnetization Switching. *Appl. Phys. Lett.* 103, 7. DOI: https://doi.org/10.1063/1.4818676

[18] C. Bi et al. 2014. Reversible Control of Co Magnetism by Voltage-Induced Oxidation. *Phys. Rev. Lett.* 113, 267202. DOI: https://doi.org/10.1103/PhysRevLett.113.267202

[19] A. K. Biswas, S. Bandyopadhyay, and J. Atulasimha. 2014. Complete Magnetization Reversal in a Magnetostrictive Nanomagnet with Voltage-generated Stress: A Reliable Energy-efficient Non-volatile Magneto-elastic Memory. *Appl. Phys. Lett*. 105, 7, 72408. DOI: https://doi.org/10.1063/1.4893617

[20] N. D'Souza, et al, 2016. Experimental Clocking of Nanomagnets with Strain for Ultralow Power Boolean Logic. *Nano Lett*. 16, 1069, DOI: https://doi.org/10.1021/acs.nanolett.5b04205

[21] X. He et al. 2010. Robust Isothermal Electric Control of Exchange Bias at Room Temperature. *Nat. Mater*. 9, 7 (July 2010), 579–85. DOI: https://doi.org/10.1038/nmat2785

[22] J. T. Heron et al. 2014. Deterministic Switching of Ferromagnetism at Room Temperature Using an Electric Field. *Nature* 516, 7531, 370–3. DOI: https://doi.org/10.1038/nature14004

[23] H. Seinige et al. 2016. Electrically Tunable Transport and High-frequency Dynamics in Sr3Ir2O7. *Phys. Rev. B* 94, 21, 214434. DOI: https://doi.org/10.1103/PhysRevB.94.214434

[24] M. E. Flatté. 2017. Voltage-driven Magnetization Control in Topological Insulator/Magnetic Insulator Heterostructures. *AIP Adv.* 7, 55923. DOI: https://doi.org/10.1063/1.4975692

[25] J.-G. Zhu. 2008. Magnetoresistive Random Access Memory: the Path to Competitiveness and Scalability. *Proc. IEEE* 96, 1786. DOI: https://doi.org/10.1109/JPROC.2008.2004313

[26] S. A. Wolf, J. Lu, M. R. Stan, E. Chen, and D. M. Treger. 2010. The Promise of Nanomagnetics and Spintronics for Future Logic and Universal Memory. *Proc. IEEE* 98, 12, 2155. DOI: https://doi.org/10.1109/JPROC.2010.2064150

[27] S.-W. Chung et al. 2016. 4Gbit Density STT-MRAM Using Perpendicular MTJ Realized with Compact Cell Structure. *2016 IEEE Int. Electron Devices Meet.* (IEDM). 659–62. DOI: http://dx.doi.org/10.1109/IEDM.2016.7838490

[28] Y.J. Song et al. 2016. Highly Functional and Reliable 8Mb STT-MRAM Embedded in 28nm Logic. *2016 IEEE Int. Electron Devices Meet.* (IEDM). 663–666. DOI: https://doi.org/10.1109/IEDM.2016.7838491

[29] L. Liu, et al, 2012. Spin-torque Switching with the Giant Spin Hall Effect of Tantalum. *Science* 336, 6081, 555–8. DOI: https://doi.org/10.1126/science.1218197

[30] A. K. Smith, et al, 2016. External Field Free Spin Hall Effect Device for Perpendicular Magnetization Reversal Using a Composite Structure with Biasing Layer. *arXiv:1630.09624*

[31] S. Jain, A. Ranjan, K. Roy, and A. Raghunathan. 2017. Computing in Memory with Spin-Transfer Torque Magnetic RAM. *arXiv: 1703.02118v3* (2017).

[32] J.-P. Wang and J. Harms. 2015. General Structure for Computational Random Access Memory (CRAM). US Patent 9224447 B2

[33] R. Perricone, et al, 2017. Advanced Spintronic Memory and Logic for Non-volatile Processors. In *Proceedings of Design, Automation Test in Europe Conference* (DATE)

[34] S. Emori, et al, 2013. Current-driven Dynamics of Chiral Ferromagnetic Domain Walls. *Nat. Mater.* 12, 7, 611–6. DOI: https://doi.org/10.1038/nmat3675

[35] A. Sengupta, A. Banerjee, and K. Roy. 2015. Hybrid Spintronic-CMOS Spiking Neural Network with On-chip Learning: Devices, Circuits and Systems. *Phys. Rev. Appl.* 6, 6, 64003. DOI: https://doi.org/10.1103/PhysRevApplied.6.064003

[36] A. Sengupta, Y. Shim, and K. Roy. 2016. Proposal for an All-Spin Artificial Neural Network: Emulating Neural and Synaptic Functionalities through Domain Wall Motion in Ferromagnets. *IEEE Trans. Biomed. Circuits Syst.* 10, 6, 1152–1160. DOI: https://doi.org/10.1109/TBCAS.2016.2525823

[37] A. Sengupta and K. Roy. 2016. A Vision for All-Spin Neural Networks: A Device to System Perspective. *IEEE Trans. Circuits Syst. I Regul. Pap.* 63, 12 (2016), 2267–2277. DOI: https://doi.org/10.1109/TCSI.2016.2615312

[38] A. Sengupta, and K. Roy. 2017. Performance Analysis and Benchmarking of All-Spin Spiking Neural Networks. *2017 International Joint Conference on Neural Networks*

[39] A. Sengupta, et al, 2016. Magnetic Tunnel Junction Mimics Stochastic Cortical Spiking Neurons. *Sci. Rep.* 6, (2016), 30039. DOI: https://doi.org/10.1038/srep30039

[40] G. Srinivasan, A. Sengupta, and K. Roy. 2016. Magnetic Tunnel Junction Based Long-Term Short-Term Stochastic Synapse for a Spiking Neural Network with On-Chip STDP Learning. *Sci. Rep.* 6, 29545. DOI: https://doi.org/10.1038/srep29545

[41] A. Sengupta, et al, 2016. Probabilistic Deep Spiking Neural Systems Enabled by Magnetic Tunnel Junction. *IEEE Trans. Electron Devices* 63, 2963. DOI: https://doi.org/10.1109/TED.2016.2568762

[42] G. Srinivasan, A. Sengupta, and K. Roy. 2017. Magnetic Tunnel Junction Enabled All-Spin Stochastic Spiking Neural Network. In *Proceedings of Design, Automation and Test in Europe Conference* (DATE). IEEE, 2017.

[43] V. Gupta, et al, 2013. Low-Power Digital Signal Processing Using Approximate Adders. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 32(1) 124.

[44] F. Sharmin Snigdha, et al, 2016. Optimal Design of JPEG Hardware Under the Approximate Computing Paradigm. In *ACM/EDAC/IEEE Design Automation 2016.*

[45] A. D. Patil, S. Manipatruni, D. Nikonov, I. A. Young, and N. R. Shanbhag, 2017. Shannon-inspired Computing to Enable Spintronics. arxiv:1702:06119, 2017.

[46] B.R. Gaines. 1969. Stochastic Computing Systems. *Advances in Information Systems Science*, 37–172.

[47] R. Venkatesan, et al, 2015. Spintastic: Spin-based Stochastic Logic for Energy-efficient Computing. In *Proceedings of the 2015 Design, Automation & Test in Europe Conference & Exhibition* (DATE '15). IEEE, 1575-1578.

[48] B. Behin-Aein, V. Diep, and S. Datta. 2016. A Building Block for Hardware Belief Networks. *Sci. Rep.* 6, 29893. DOI: https://doi.org/10.1038/srep29893

[49] B. Sutton, et al, 2017. Intrinsic Optimization Using Stochastic Nanomagnets. *Sci. Rep.* 7, 44370. DOI: https://doi.org/10.1038/srep44370

[50] K. Y. Camsari, et al, 2016. Stochastic p-bits for Invertible Spin Logic. *arXiv:1610.00377*.

[51] R. Faria, K. Y. Camsari, and S. Datta. 2017. Low Barrier Nanomagnets as p-bits for Spin Logic. *IEEE Magn. Lett.* 1-1, 99, DOI: https://doi.org/10.1109/LMAG.2017.2685358

[52] D. H. Ackley, G.E Hinton and T.J Sejnowski. 1985. A Learning Algorithm for Boltzmann Machines, *Cogn. Sci.* 9, 1, 147. DOI: https://doi.org/10.1016/S0364-0213(85)80012-4

[53] Y. LeCun, Y. Bengio, and G. Hinton. 2015. Deep Learning. *Nature* 521, 7553 (2015), 436–444. DOI: https://doi.org/10.1038/nature14539

[54] M. Yamaoka, et al, 2015. 20k-spin Ising Chip for Combinational Optimization Problem with CMOS Annealing. *Dig. Tech. Pap. - IEEE Int. Solid-State Circuits Conf.* (ISSCC) 58, 432–433. DOI: https://doi.org/10.1109/ISSCC.2015.7063111

[55] K. Y. Camsari, O. Hassan, R. Faria, S. Ganguly, and S. Datta, unpublished.

[56] D. E. Nikonov and I. A. Young. 2015. Benchmarking of Beyond-CMOS Exploratory Devices for Logic Integrated Circuits. *IEEE J. Explor. Solid-State Comput. Devices Circuits* 1, 3–11. DOI: https://doi.org/10.1109/JXCDC.2015.2418033

[57] H. Markram. 2006. The Blue Brain Project. *Nat. Rev. Neurosci.* 7, 2, 153–60. DOI: https://doi.org/10.1038/nrn1848

[58] P. A. Merolla et al. 2014. A Million Spiking-neuron Integrated Circuit with a Scalable Communication Network and Interface. *Science* 345, 6197, 668–673. DOI: https://doi.org/10.1126/science.1254642

[59] H. Aghasi, et al, 2016. Smart Detector Cell: A Scalable All-Spin Circuit for Low Power Non-Boolean Pattern Recognition. *IEEE Trans. Nanotechnol.* 15, 3, 356–366. DOI: https://doi.org/10.1109/TNANO.2016.2530779

[60] D. Monroe. 2014. Neuromorphic Computing Gets Ready for the (Really) Big Time. *Commun. ACM* 57, 6, 13–15. DOI: https://doi.org/10.1145/2601069

[61] S. Venkataramani, A. Ranjan, K. Roy, and A. Raghunathan. 2014. AxNN: Energy-Efficient Neuromorphic Systems using Approximate Computing. In *Proceedings of the 2014 International Symposium on Low Power Electronics and Design* (ISLPED '14). New York, New York, USA: ACM Press, 27–32. DOI: https://doi.org/10.1145/2627369.2627613

[62] A. R. Trivedi and S. Mukhopadhyay. 2014. Potential of Ultralow-Power Cellular Neural Image Processing With Si/Ge Tunnel FET. *IEEE Trans. Nanotechnol.* 13, 4, 627–629. DOI: https://doi.org/10.1109/TNANO.2014.2318046

[63] I. Palit, X. S. Hu, J. Nahas, and M. Niemier. 2013. TFET-based Cellular Neural Network Architectures. In *International Symposium on Low Power Electronics and Design* (ISLPED). IEEE, 236–241. DOI: https://doi.org/10.1109/ISLPED.2013.6629301

[64] C. Pan and A. Naeemi. 2016. A Proposal for Energy-Efficient Cellular Neural Network Based on Spintronic Devices. *IEEE Trans. Nanotechnol.* 15, 5, 820–827. DOI: https://doi.org/10.1109/TNANO.2016.2598147

[65] C. Pan and A. Naeemi. 2016. Non-Boolean Computing Benchmarking for Beyond-CMOS Devices Based on Cellular Neural Network. *IEEE J. Explor. Solid-State Comput. Devices Circuits* 2, 36–43. DOI: https://doi.org/10.1109/JXCDC.2016.2633251

[66] D. Bedau et al. 2010. Ultrafast Spin-transfer Switching in Spin Valve Nanopillars with Perpendicular Anisotropy. *Appl. Phys. Lett*. 96, 2. DOI: https://doi.org/10.1063/1.3284515

[67] S. V. Aradhya, et al, 2016. Nanosecond-Timescale Low Energy Switching of In-Plane Magnetic Tunnel Junctions through Dynamic Oersted-Field-Assisted Spin Hall Effect. *Nano Lett*. 16, 10, 5987–5992. DOI: https://doi.org/10.1021/acs.nanolett.6b01443

[68] U. Bauer et al. 2015. Magneto-ionic Control of Interfacial Magnetism. *Nat. Mater*. 14, 2 (2015), 174–81. DOI: https://doi.org/10.1038/nmat4134

[69] X. Marti, I. Fina, and T. Jungwirth. 2015. Prospect for Antiferromagnetic Spintronics. *IEEE Trans. Magn*. 51, 4 (2015), 5–8. DOI: https://doi.org/10.1109/TMAG.2014.2358939

[70] K. Y. Camsari, et al, 2016. Ultrafast Spin-Transfer-Torque Switching of Synthetic Ferrimagnets. *IEEE Magn. Lett.* 7, 3107205. DOI: https://doi.org/10.1109/LMAG.2016.2610942

[71] Y. G. Semenov, X. Duan, and K. W. Kim. 2014. Voltage-driven Magnetic Bifurcations in Nanomagnet-topological Insulator Heterostructures. *Phys. Rev. B* 89, 20, 1–5. DOI: https://doi.org/10.1103/PhysRevB.89.201405