All Spin Artificial Neural Networks Based on Compound Spintronic Synapse and Neuron

Deming Zhang, *Student Member, IEEE*, Lang Zeng, *Member, IEEE*, Kaihua Cao, *Student Member, IEEE*, Mengxing Wang, *Student Member, IEEE*, Shouzhong Peng, *Student Member, IEEE*, Yue Zhang, *Member, IEEE*, Youguang Zhang, *Member, IEEE*, Jacques-Olivier Klein, *Member, IEEE*, Yu Wang, *Senior Member, IEEE*, and Weisheng Zhao, *Senior Member, IEEE*

***Abstract*—Artificial synaptic devices implemented by emerging post-CMOS non-volatile memory technologies such as Resistive RAM (RRAM) have made great progress recently. However, it is still a big challenge to fabricate stable and controllable multilevel RRAM. Benefitting from the control of electron spin instead of electron charge, spintronic devices, e.g., magnetic tunnel junction (MTJ) as a binary device, have been explored for neuromorphic computing with low power dissipation. In this paper, a compound spintronic device consisting of multiple vertically stacked MTJs is proposed to jointly behave as a synaptic device, termed as compound spintronic synapse (CSS). Based on our theoretical and experimental work, it has been demonstrated that the proposed compound spintronic device can achieve designable and stable multiple resistance states by interfacial and materials engineering of its components. Additionally, a compound spintronic neuron (CSN) circuit based on the proposed compound spintronic device is presented, enabling a multi-step transfer function. Then, an All Spin Artificial Neural Network (ASANN) is constructed with the CSS and CSN circuit. By conducting system-level simulations on the MNIST database for handwritten digital recognition, the performance of such ASANN has been investigated. Moreover, the impact of the resolution of both the CSS and CSN and device variation on the system performance are discussed in this work.**

***Index Terms*—All Spin Artificial Neural Network (ASANN), artificial synaptic device, compound spintronic device, magnetic tunnel junction (MTJ), post-CMOS non-volatile memory (NVM) technologies, resistive RAM (RRAM).**

I. Introduction

NEUROMORPHIC computing based on Artificial Neural Networks (ANNs) has advantages over Von-Neumann architecture based digital computing paradigm in terms of its massive parallelism, adaptivity to the varying and complex input information, energy-efficiency and inherent tolerance to fault and variation [1], [2]. In the ANNs, neurons and synapses are the two basic computing elements. The neural computation is performed by weighted summation of input neuron signals subjecting to transfer function of output neurons. For example, if the input neuron signals are represented by $\{I_i\}$ and the corresponding synaptic weights are represented by $\{w_i\}$, then the output neuron signals can be represented by $y = f(\sum w_i \cdot I_i)$ [3], where f represents the transfer function of the output neuron, which could be a step, linear or sigmoid function. The synapse contributes to the computation by modulating its weight according to the output signals of its pre- and post-neuron [4]. However, custom analog/digital CMOS implementations of the neurons [5] and synapses [6] have been proved to be extremely area and power inefficiency, since they cannot provide a direct mapping to the thresholding operation and tunable weights involved in neuromorphic computing [7].

In human brain, there are 10^{14} synapses and 10^{10} neurons. It is clear that the synapses outnumber the neurons by several orders of magnitude [4]. Thus, research work has been focused on how to mimic the biological synapse functionality using the emerging post-CMOS non-volatile memory (NVM) technologies such as resistive RAM (RRAM) and spin transfer torque magnetic RAM (STT-MRAM) with low power dissipation and high density in the form of crossbar [8]–[26]. Although it has made great progress recently, there still exists great difficulties to fabricate controllable and stable multilevel RRAM. While in the case of using STT-MRAM as synapses, the situation and difficulty are different. In order to acquire learning ability, its intrinsic STT switching stochasticity is introduced since each magnetic tunnel junction (MTJ), the basic storage cell of STT-MRAM, is a binary device [8]. However, this leads to a worse learning accuracy and redundancy is required to achieve learning accuracy as high as that achieved by the analog-like synapse with RRAM.

In our previous study [27], we have shown that a compound spintronic device (CSD) that consists of multiple vertically stacked MTJs can achieve multiple resistance states by manipulating the critical switching current and resistance area

Manuscript received November 25, 2015; revised February 5, 2016; accepted February 17, 2016. Date of publication May 17, 2016; date of current version July 26, 2016. This work was supported by the Chinese National Natural Science Foundation projects (No. 61306104, 61471015, 61571023, 61504006, 61373026), by the International Collaboration Project 2015DFE12880 from the Ministry of Science and Technology in China and by the Beijing Municipal of Science and Technology (Grant No. D15110300320000). Deming Zhang and Lang Zeng contributed equally to this work. This paper was recommended by Associate Editor A. Basu. *(Corresponding author: Weisheng Zhao.)*

D. Zhang, L. Zeng, K. Cao, M. Wang, S. Peng, Y. Zhang, Y. Zhang and W. Zhao are with Spintronics Interdisciplinary Center (SIC), Beihang University, 100191, Beijing, China, and also with the School of Electrical and Information Engineering, Beihang University, Beijing, 100191, China (e-mail: deming.zhang@buaa.edu.cn; zenglang@buaa.edu.cn; kaihua.cao@buaa.edu.cn; wangmxbj@163.com; shouzhong.peng@buaa.edu.cn; yz@buaa.edu.cn; zhangyouguang@vip.sina.com; weisheng.zhao@buaa.edu.cn).

J.-O. Klein is with IEF, University of Paris-Sud, Orsay, 91405, France (e-mail: Jacques-olivier.klein@u-psud.fr).

Y. Wang is with the Department of Electronic Engineering, Tsinghua University, Beijing, 100084, China (e-mail: yu-wang@mail.tsinghua.edu.cn).

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/TBCAS.2016.2533798

product of its components. Different from the RRAM, the multilevel CSD is highly controllable. Even the separation of any two adjacent resistance levels can be finely adjusted to fulfill the requirement of the whole neuromorphic computing system. In our novel approach, the critical switching current and resistance area product of the MTJ can be adjusted by the interfacial perpendicular magnetic anisotropy (PMA) with interfacial and material engineering. In this paper, a compound spintronic synapse (CSS) is proposed by using the multilevel CSD. Crossbar structure of the proposed CSS is employed for high density. On the other hand, a limitation in the existing work [28] is that the spintronic neuron could only emulate the step transfer function in the ANNs (corresponding to the switching between its two state states by a resultant synaptic current), whereas the non-step like (e.g., linear and sigmoid) transfer function can be more attractive for complex pattern recognition tasks since more information can be encoded in the neuron activity. In order to address this problem, a compound spintronic neuron (CSN) circuit based on the multilevel CSD is also proposed, which can implement a multi-step transfer function. Additionally, we construct an all spin artificial neural network (ASANN) by combining the proposed CSS and CSN circuit together. Furthermore, a study case is made on MNIST database for handwritten digital recognition.

This article is organized as follows. Section II presents the design methodology and fabrication process for multilevel compound spintronic device (CSD). In Section III, the compound spintronic synapse (CSS) and neuron (CSN) circuit are proposed. Circuit-level simulations of the proposed CSS and CSN circuit are performed with considerations of device variation. Then, an all spin artificial neural network (ASANN) consisting of the proposed CSS and CSN circuit is illustrated and the performance of such ASANN with off-line training is investigated by performing system-level simulations in Section IV. Section V concludes this paper and then discusses future research directions.

II. Design Methodology for Proposed Multilevel Compound Spintronic Device

A. *Fundamentals of STT-MTJ*

By exploiting the electron spin property instead of electron charge, spintronic devices can offer numerous novel features, such non-volatility, low power, etc. Most spintronic devices are based on magnetic tunnel junction (MTJ) structure as shown in Fig. 1(a). A MTJ is mainly composed of one one ultra-thin tunneling barrier (TB) layer (e.g., MgO) sandwiched by two ultra-thin ferromagnetic (FM) layers (e.g., CoFeB). In order to store one bit binary information, these two FM layers are set to be asymmetric. It means that the magnetization orientation of one FM layer is fixed in one direction, called pinned layer (PL), while that of the other FM layer is reversible, named free layer (FL). For the TB layer, it functions as a spin filter. If the magnetization orientation of these two FM layers is parallel, the current flow depends on the tunneling electron exchange between majority spin polarization of both the PL and the FL. Due to the large Density of State (DOS) of majority spin, the resistance of the MTJ in this configuration is small. On the other hand, the current flow depends on the tunneling electron exchange between the majority and minority spin polarization of both the PL and the FL. Due to the small DOS of minority spin, the resistance of the MTJ in this configuration is large. Generally, the resistance ratio of the MTJ in the parallel and anti-parallel configuration is characterized by the tunneling magnetoresistance (TMR), which is defined as

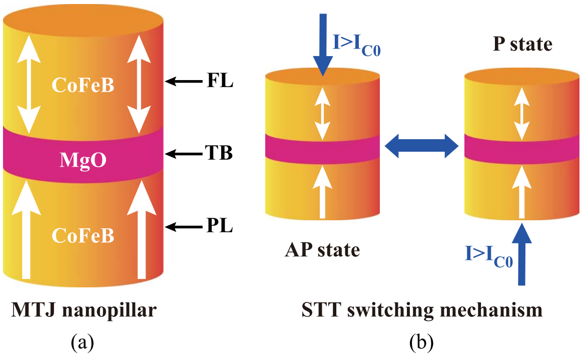


Fig. 1. (a) Vertical structure schematic of a MTJ nanopillar with perpendicular magnetization anisotropy composed of CoFeB/MgO/CoFeB thin films. (b) Spin transfer torque (STT) switching mechanism: the resistance state of the MTJ nanopillar can be reversed by a bidirectional current. Here, FL, TB and PL are short for free layer, tunnel barrier and pinned layer, respectively.

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

where R_{AP} is the electrical resistance in the anti-parallel (AP) state, whereas R_P is the resistance in the parallel (P) state.

The spin transfer torque (STT) effect is an effect in which the magnetization orientation of the FL in one MTJ can be switched by applying a spin-polarized current, and it is an inverse effect of spin polarization. As shown in Fig. 1(b), an unpolarized charge current (consisting of 50% spin-up and 50% spin-down electrons) passes through the PL to the FL and becomes a spin-polarized current. When the spin-polarized current arrives into the FL, angular momentum carried by the spin-polarized current can be transferred to this layer, changing its orientation to be parallel with the PL. If an unpolarized electric current passes through the FL to the PL. The electrons with spin direction parallel to majority spin in the PL can tunnel into the PL, and those electrons with the other spin direction is left in the FL. Angular momentum carried by these electrons left in the FL is transferred to this layer, changing its orientation to be anti-parallel with the PL. Since the STT effect is purely an electrical effect without magnetic field, it is critical for STT-MRAM to achieve high speed, low power and scalability.

The explosive growth of computation power needed by the modern information society demands the memory chips with high capacity and low power. In order to meet this requirement, the dimension of the MTJ should be scaled down to below 40 nm. However, the critical switching current of the MTJ with in-plane magnetic anisotropy (i-MTJ) cannot be continuously reduced with shrinking device size. Recent material progress showed that the MTJ with perpendicular magnetic anisotropy (p-MTJ) as seen in Fig. 1(a) can offer lower critical switching current and higher switching speed in comparison with the

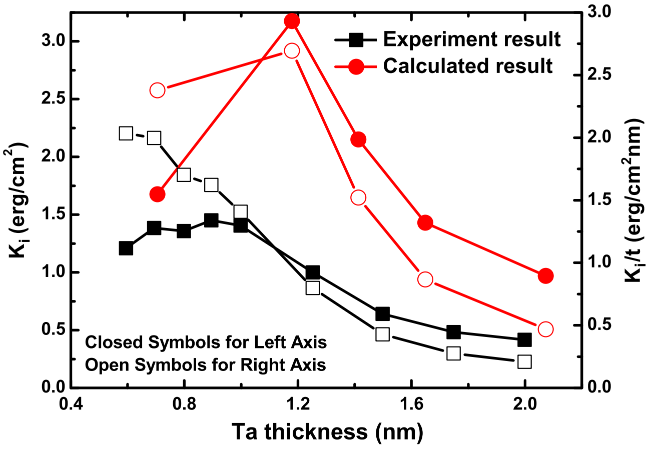


Fig. 2. The calculated K_i and K_i/t from first-principle calculations as well as experimental results are shown. It can be seen that the theoretical calculations give the same trends as experimental measurements.

i-MTJ. The critical switching current can be expressed as for the p-MTJ [29]

$$I_{c0} = \alpha \frac{\gamma e}{\mu_B g} (\mu_0 M_s) H_k V \tag{2}$$

where α is the Gilbert damping factor, γ is the gyromagnetic ratio, e is the electron charge, μ_B is the Bohr magneton, g is a parameter corresponding to the spin polarization and the magnetization orientation angel between the FL and the PL, μ_0 is the permeability in the free space, M_s is the saturation magnetization, H_k is the perpendicular magnetic anisotropy (PMA) field and V is the volume of the FL. According to this equation, it is implied that the critical switching current can be manipulated if the PMA field H_k can be tuned.

B. Interface Engineering for PMA Field

The relationship between the PMA field H_k and the effective magnetic anisotropy constant K_{eff} is written as [27]

$$K_{\text{eff}} = \frac{M_s H_k}{2}. \tag{3}$$

For ultra-thin film structure, both bulk and interface effect contribute to the effective magnetic anisotropy constant K_{eff}. Thus K_{eff} can be described by [27]

$$K_{\text{eff}} = K_b + K_d + K_i/t \tag{4}$$

where K_b is bulk anisotropy, K_d is demagnetization field and can be termed as $K_d = -2\pi M_s^2$, and K_i is interfacial PMA and t is the thickness of magnetic layer.

According to our recent theoretical work [30], it has been revealed that the interfacial PMA in the CoFeB/MgO/CoFeB ultra-thin film structure comes from both the MgO/CoFe and CoFe/capping layer interfaces. Furthermore, we have shown that the contribution of interfacial PMA from the MgO/CoFe and CoFe/capping layer interfaces are additive, which can be tuned separately. In Fig. 2, the calculated K_i and K_i/t from first-principle calculations as well as experimental results are shown. As seen, the theoretical calculations exhibit the same trends with experimental measurements. The interfacial PMA changes significantly as the thickness of the Ta film varies. The maximum value of interfacial PMA is at the Ta thickness of ~1–1.2 nm.

Further, the interfacial PMA originated from other capping layer materials have been calculated as shown in Table I [30]. The second last column shows the sum from the magnetic anisotropy energy (MAE) at the interface of the MgO/CoFe and CoFe/X, where the X includes Ru, Ta and Hf. It can be observed that the interfacial PMA can be tuned by different capping layer materials as well as different capping layer thickness. By carefully designing the capping layer materials and their thickness, we can achieve critical switching current with several times of changing range.

C. Proposed Compound Spintronic Device

To achieve multiple resistance states, we proposed a compound spintronic device, which consists of multiple vertically stacked MTJs as shown in Fig. 3(a). The critical switching current and resistance area product (RA) of each MTJ can be modulated separately. The former can be tuned by choosing different capping layers materials and their thickness, while the latter by varying the MgO thickness and other fabrication process parameters. It is apparent that by using our proposed design methodology, the proposed compound spintronic device can exhibit multiple more designable and controllable resistance states in comparison with the RRAM device.

Furthermore, we are also trying to fabricate the proposed device, and Fig. 3(b) shows the TEM image of the fabricated compound spintronic device with two vertically stacked MTJs. The 14 layer stack is Ta(5)/Ru(10)/Ta(5)/CoFeB(1)/MgO(0.85)/CoFeB(1.4)/Ta(5)/Ru(5)/Ta(5)/CoFeB(1.2)/MgO(1)/CoFeB(1.6)/Ta(5)/Ru(5), where the numbers indicate the layer thickness in nanometer. These two vertically stacked MTJs with different MgO and capping layer thickness are implemented. The critical switching current and RA of these two MTJs are different for the capping layer and MgO tunneling barrier are intendedly deposited with different thickness. Since we are working on the optimization of the multi-step etching process of these complex layer stack to avoid deteriorating of the MTJ profile and performance, we have not had a chance to perform experimental measurements on the proposed compound spintronic device.

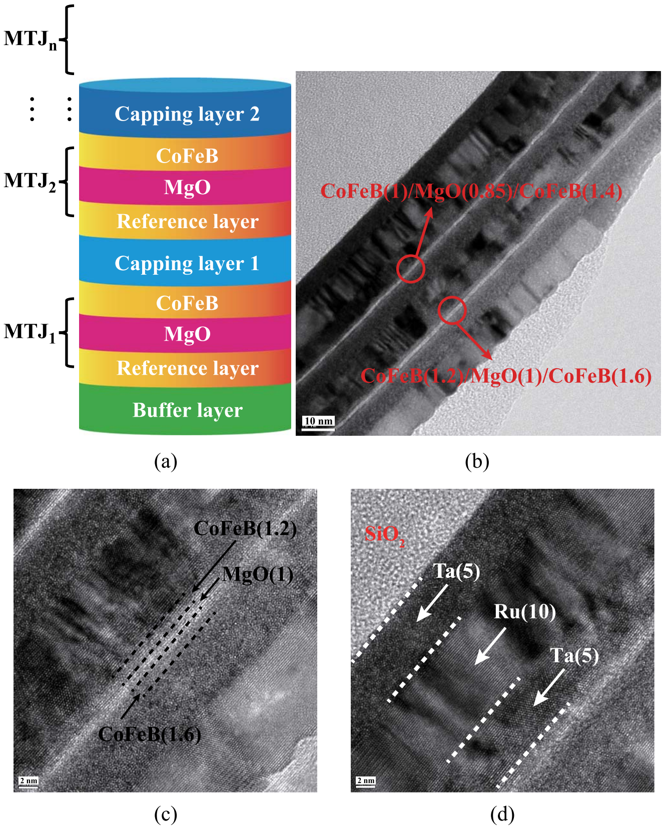
However, the performance of the proposed compound spintronic device with two vertically stacked MTJs was evaluated with a physics-based STT-MTJ compact model [31], which was programmed with the Verilog-A language and performed on the Cadence Platform. In the model, M_s was set to be 1150 emu/cm^3 according to the experimental results published in [32] and [33]. H_k of each MTJ is calculated from simulation results and chosen to be 3057 and 1520 Oe for the Ta film with thickness of ~1.35 and 1.65 nm, respectively. The TMR is set to be 200% for both the MTJs. In order to achieve maximum separations between each resistance state for better system level performance, the relation between RA of these two MTJs is written as

$$\text{RA}_2 = 2 \times \text{RA}_1 \tag{5}$$

where $\text{RA}_1 = 5\ \Omega \cdot \mu\text{m}^2$.

TABLE I
CALCULATED MAE VALUES (erg/cm^2) FOR DIFFERENT STRUCTURES

X	CoFe surface	MgO/CoFe interface	CoFe/X interface	Sum of MAE in two interfaces	MgO/CoFe/X structure
Ru	0.41	0.57	0.52	1.09	0.98
Ta	0.41	0.57	1.13	1.70	1.77
Hf	0.41	0.57	1.65	2.22	2.28



(a) (b)

(c) (d)

Fig. 3. (a) Schematic of the proposed compound spintronic device. The capping layer materials/thickness and MgO tunneling barrier thickness can be manipulated separately to obtain desired multilevel states. In ideal case, N vertically stacked MTJs can exhibit 2^N states. (b). TEM picture of the fabricated two vertically stacked MTJs with capping layers. (c) Zoom in of the CoFeB/MgO/CoFeB structure (d) Zoom in of the Ta/Ru/Ta capping layers.

By performing the DC simulation with the above parameters, Fig. 4 shows the $R - I$ loop of the proposed compound spintronic device (with two MTJs) corresponding to fabricated devices in Fig. 3(b). In the simulation, the size of both the MTJs is set to be 40 nm. It can be observed that by properly adjusting the capping layer and MgO tunneling barrier thickness, four distinct states can be obtained with well separated critical switching current and relevant resistance. The critical switching currents of these two MTJs are $\sim$50 μA and $\sim$100 μA, respectively. The simulation clearly verifies that the proposed compound spintronic device is designable and stable in comparison with almost uncontrollable multilevel RRAM. Moreover, it can be inferred that by vertically stacking more MTJs, more resistance states can be obtained. In the ideal case, the 2^N resistance states can be achieved by stacking N MTJs.

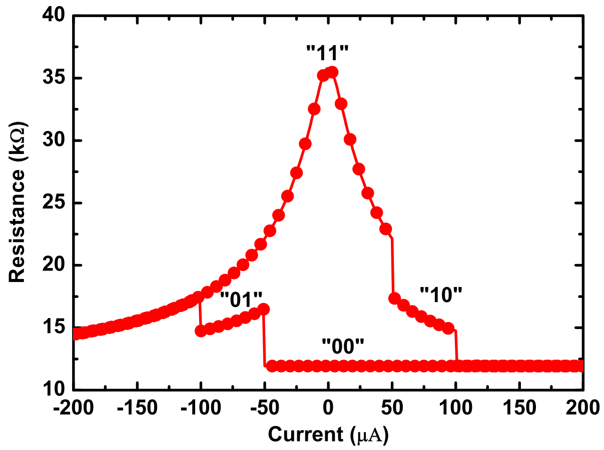


Fig. 4. Electrical characterazation of $R - I$ loop of 2 bit multilevel STT MRAM corresponding to fabricated devices in Fig. 3(b). The MTJ size is set to be 40 nm in the simulation.

III. PROPOSED COMPOUND SPINTRONIC SYNAPSE AND NEURON

A. Compound Spintronic Synapse

Generally, each synapse in the ANNs is characterized by a weight, transmitting the weighted signal from its pre- to post-neuron. And an essential property of the synapse is its analog tunable synaptic weight. In the emerging field of the synaptic electronics, the synaptic weights can be characterized by the conductances of synaptic devices. Unfortunately, a MTJ device could only exhibit two stable conductance states. In order to exactly mimic the functionality of the synapse, the proposed compound spintronic device is employed to behave as a single synapse, termed as compound spintronic synapse (CSS). As a result, such CSS can achieve multiple discrete conductance states with well design and manipulation of materials/thickness of the capping layer and thickness of MgO tunneling barrier. With the number of MTJs in one CSS increasing, its conductance stats increase more rapidly, thus achieving an analog-like synaptic weight spectrum. It is worth noting that the generated current by the read voltage V_R of input neuron is required to be much smaller than the critical switching current of the MTJ nano-pillar to ensure that the synaptic weights cannot be changed during the transmission stage of the weighted signals.

In order to validate its feasibility, a CSS with three vertically stacked MTJs was simulated. The main parameters of these three MTJs in this simulation are shown in Table II. And the simulation results are shown in Fig. 5. As seen, such CSS can achieve 8 discrete resistance states. By generating different input current pulses, the proposed CSS can be set into a specific weight state as needed by the whole neuromorphic computation system. Additionally, the impact of device variations originated in the fabrication process is also taken into consideration. With

TABLE II
PARAMETERS AND VARIABLES PRESENT IN THE FITTING FUNCTIONS

Parameter	Description	Default value
Area	MTJ surface	40nm$*$40nm$*\pi/4$
TMR(0)	TMR ratio with zero V_{bias}	200%
t_f	Free layer height	1.3nm
V	Volume of free layer	Area*t_f
RA_1	Resistance*Area product of MTJ_1	$5\Omega \cdot \mu m^2$
RA_2	Resistance*Area product of MTJ_2	$10\Omega \cdot \mu m^2$
RA_3	Resistance*Area product of MTJ_3	$20\Omega \cdot \mu m^2$
I_{c1}	Critical switching current of MTJ_1	$50\mu A$
I_{c2}	Critical switching current of MTJ_2	$75\mu A$
I_{c3}	Critical switching current of MTJ_3	$100\mu A$

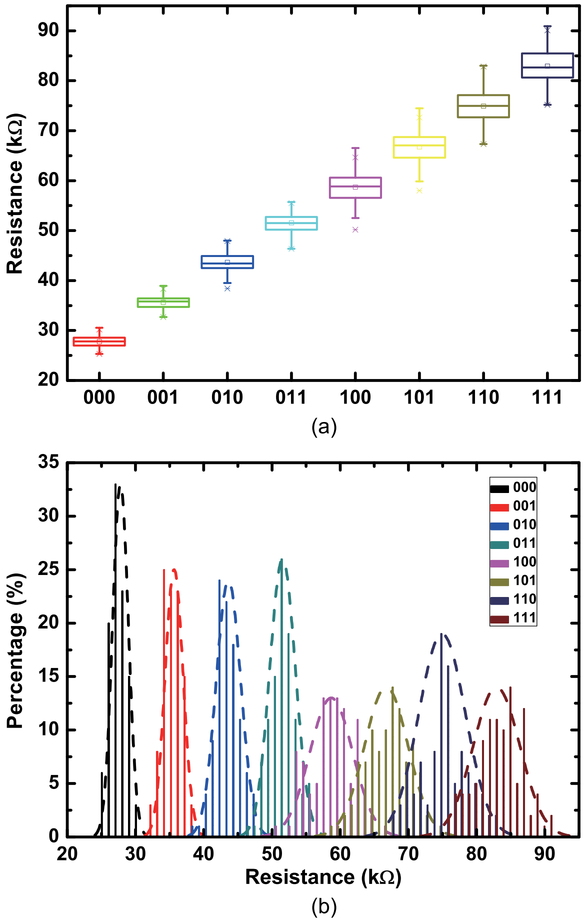


Fig. 5. Compound spintronic synapse (CSS) with 8 discrete synaptic weights. For the multilevel CSS, the impact of device variation on the resistance of separated synaptic states are also shown. The variation for temperature, MgO tunneling barrier thickness and TMR are set to be 3%. It is clear that even with such device variation, the CSS can achieve 8 distinct resistance levels.

5% variations for perpendicular magnetic anisotropy field H_k and resistance area product RA, well separated critical switching current and resistance states still can be achieved. It is worth noting that our proposed all spin artificial neural networks performs off-line learning paradigm (The training is performed by system-level simulation, then the simulated synaptic weights are mapped into CSSs discrete resistance states.). As a consequence, only variations in resistance area product RA which affects the separation of resistance states would have effect on the performance of neuromorphic computation system.

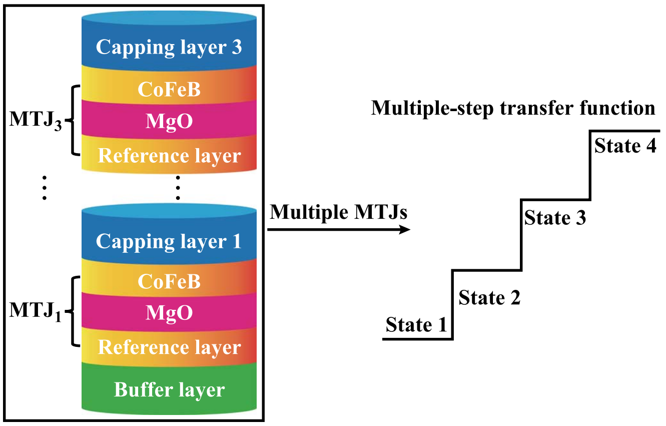


Fig. 6. Proposed compound spintronic neuron (CSN), which can implement a multiple-step transfer function.

B. Compound Spintronic Neuron

As the basic computational element in ANNs, each neuron is generally characterized by a transfer function (activation function). The neuron generates an output signal depending on the magnitude of the summation of its synaptic weighted input signals. And then the neuron's output signal is transmitted via the axon as the input of its fan-out neurons. In [28], a single MTJ device has been demonstrated to emulate the step transfer function (corresponding to the switching between its two stable states by a resultant synaptic current) in ANNs. Non-step (linear and sigmoid) transfer function could be more attractive for complex pattern recognition tasks since more information can be encoded in the neuron's output. For this, we proposed a compound spintronic neuron (CSN) that also employs multiple MTJs to function as a single neuron as shown in Fig. 6. Since vertically stacked MTJs can exhibit multiple discrete resistance states due to designed perpendicular magnetic anisotropy field H_k and resistance area product RA, analog-like information can be encoded in the CSN output to implement a multiple-step (linear-like) transfer function. It is worth noting that in case of CSN, $N+1$ steps for the transfer function can be implemented with N stacked MTJs, different from the 2^N discrete states for CSS.

To implement a multiple-step (linear-like) transfer function, a CSN circuit was proposed as shown in Fig. 7(a), which involves three operation phases, i.e., encode, read and reset phase. During the encode phase in Fig. 7(b), the NMOS transistor N1 turns on and the N2, N3 and the transmission gate (TG) are in off state. As a result, the resultant synaptic current, I_{synapse}, can be encoded into the CSN resistance state. Higher the magnitude of the resultant synaptic current I_{synapse}, higher resistance state the CSN can be switched into. Then, the CSN circuit enters the read phase as shown in Fig. 7(c). During it, the N2 and the TG turn on and the N1 and N3 are in off state. As a result, the CSN can generate a voltage signal V_{out} depending on its resistance state to the next layer neurons. Also, the output voltage signal V_{out} can be tuned by varying I_{read}. After the information is transferred to the next layer, the CSN circuit is reset for next operation as shown in Fig. 7(d). That is, all the MTJs are set to be in parallel configuration in reset stage. As a result, the

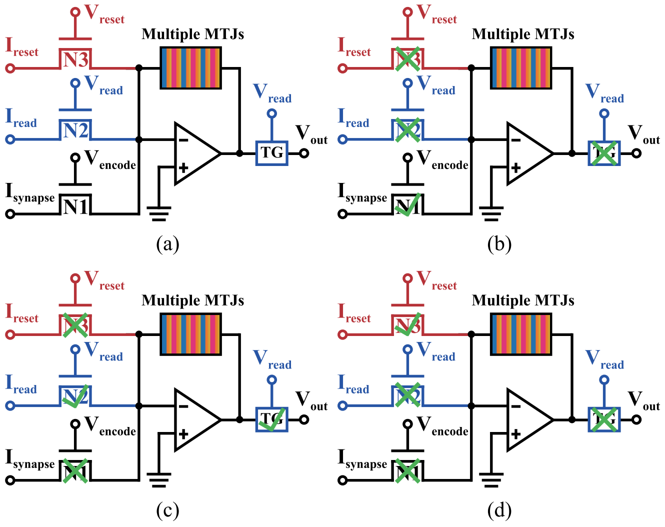


Fig. 7. Operation of the proposed compound spintronic neuron. (a) CSN circuit. (b) Encode phase. (c) Read phase. (d) Reset phase.

transfer function realized by the CSN can be characterized by the relationship of the I_{synapse} and the output voltage signal V_{out}.

To validate the functionality of the proposed CSN circuit, hybrid CMOS/MTJ circuit simulation was performed with the commercial CMOS 40 nm design kit. In the simulation, the CSN is composed of three vertically stacked MTJs as shown in Table II. Fig. 8(a) shows the relationship of the resultant synaptic current I_{synapse} of 10 ns and the resistance state of the CSN during the encode phase. As seen, the I_{synapse} can be encoded into four discrete resistance states of the CSN at 69/97/123 μA. For example, if the I_{synapse} is larger than 69 μA but smaller than 97 μA, the CSN is encoded in the "001" state. Fig. 8(b) shows the relationship of the output voltage signal V_{out} and the resistance state of the CSN, considering the variations for temperature, MgO tunneling barrier thickness and TMR of 3%. In the simulation, the I_{read} is set to be -1 μA to avoid the voltage bias effect on the resistance of the CSN.

IV. Proposed All Spin Artificial Neural Networks

A. *Structure of Proposed All Spin Artificial Neural Network*

Based on the proposed CSS and CSN, an all spin artificial neural network (ASANN) is proposed. It is designed as a fully connected feed-forward three-layer network. The network between the input and hidden layer as well as that between the hidden and output layer can be implemented by a crossbar architecture. Fig. 9 shows the network between the input and hidden layer, where each crosspoint consists of a CSS and its resistance state encodes the synaptic weight. Each input neuron generates a corresponding voltage depending on the training data. These input voltages are applied to the corresponding horizontal metal lines of the CSS crossbar. Then, these input voltages are modulated by the corresponding CSS in each vertical mental line, generating the weighted input signals (i.e., the synaptic currents). These weighted input signals are summed into the CSN circuit located at the end of each vertical mental

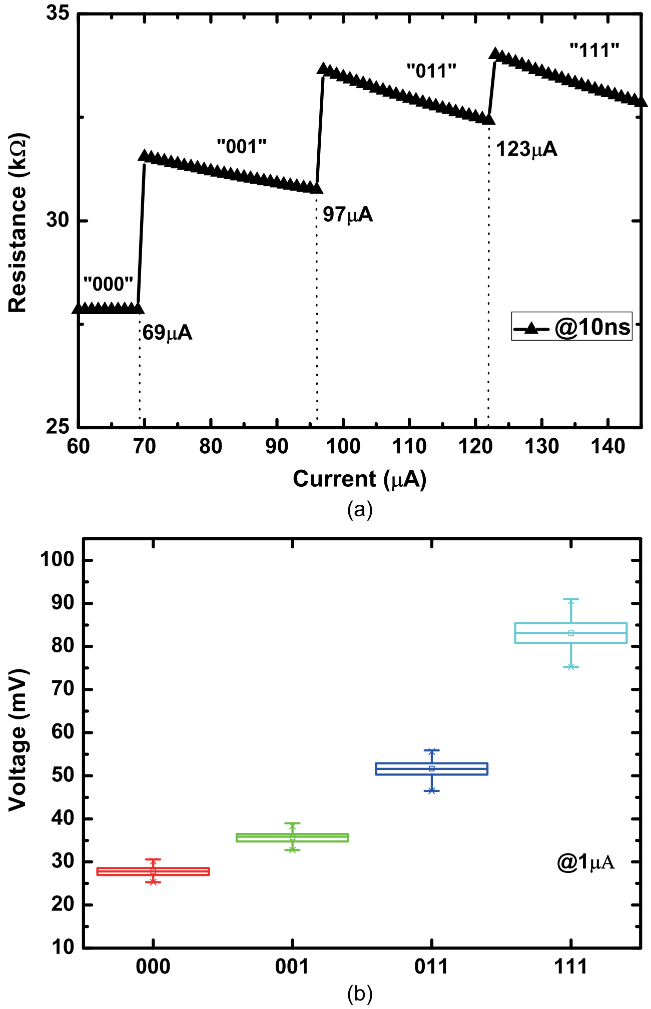


Fig. 8. (a) The simulated threshold switching current for different CSN resistance states and its resistance. 10 ns pulse is used for the simulation. (b) The output voltage of the CSN for different states. Device variations are considered in the simulation. Taken (a) and (b) together, the transfer function of the proposed CSN can be obtained.

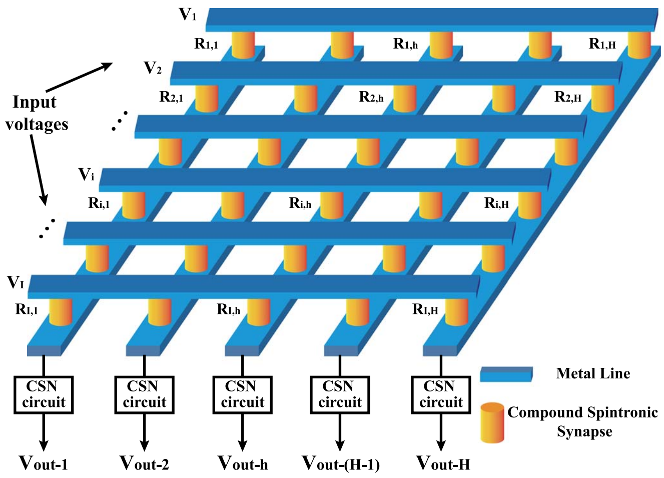
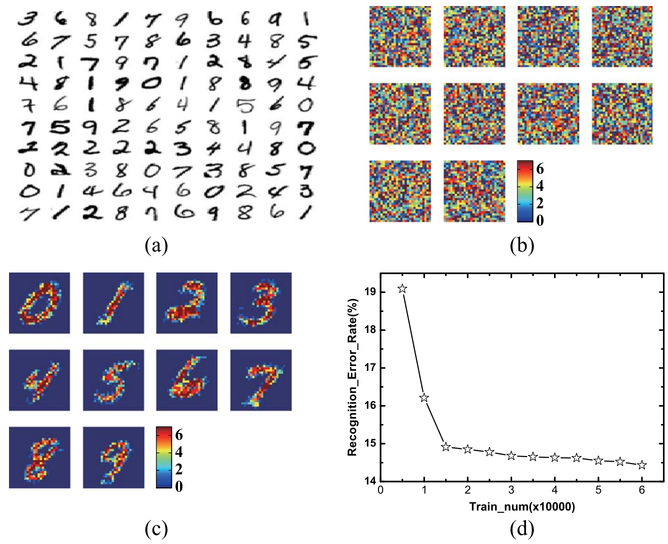


Fig. 9. The crossbar structure between the input and hidden layer with the proposed CSS and CSN circuit.



(a) (b) (c) (d)

Fig. 10. System-level simulation results. (a) Examples of digital patterns in the MNIST database. (b) Initialization of the synaptic weight matrix of the network between the input and hidden layer. (c) The learned handwritten digital patterns. (d) The recognition error rate during the training process.

line. In the encode phase of the CSN in hidden layer, the CSN is encoded into a specific resistance state according to the magnitude of the summed synaptic currents. Then, in the read phase of the CSN, the CSN outputs a voltage signal, whose magnitude depends on the resistance state of the corresponding CSN. And the output voltage signal is transferred to the network between hidden and output layer as a input signal. The network between hidden and output layer is the same as that in Fig. 9, and the output signals of the neurons in the output layer give the recognition result.

To investigate the performance of the proposed ASANN, a study case was made on the MNIST which is a widely used database for the handwritten digital recognition including 60 000 and 10 000 handwritten digital patterns for training and testing respectively [34]. Since there are 10 kinds of different handwritten digits in the database and each digit is composed of $28 * 28 = 784$ pixels with 256 level gray levels, the number of input neurons is set to be 784 and the number of output neurons is set to be 10. In order to more accurately learn all the handwritten digits, the number of the hidden neurons is set to 100. The network between the input and hidden layer was used to learn the handwritten digits and that between the hidden and output layer to classify the learned handwritten digits. The standard back-propagation algorithm was employed for the training process [34]. After the training process is completed, the trained weight of each synapse is mapped into our CSS crossbar network with discrete weight. Then recognition test will be performed on the proposed ASANN consisting of the CSS and CSN with real physical parameters listed in Table II. It is worth noting that the off-line learning is used since the back-propagation is easier for software implementation but more difficulty for hardware implementation.

Taking each CSS with 3 MTJs and each CSN with 1 MTJ for instance, Fig. 10 shows the simulation results. Fig. 10(a) shows the examples of the handwritten digits from the database. Initially, each CSS weight was set randomly as shown in Fig. 10(b). Fig. 10(c) shows the learned handwritten digital patterns. To quantitatively evaluate the learning accuracy of the proposed ASANN, Fig. 10(d) shows the recognition error rate during the training process. As can be seen, the recognition error rate could reach about 9.15%.

TABLE III
RECOGNITION ERROR RATE (RER) UNDER DIFFERENT RESOLUTION OF BOTH THE CSS WEIGHT AND CSN OUTPUT SIGNAL

RER CSS CSN	N=1	N=2	N=3
N=1	33.28%	20.51%	14.49%
N=2	32.19%	19.40%	13.26%
N=3	31.59%	18.62%	12.55%
Linear	30.01%	16.30%	9.73%

TABLE IV
THE INFLUENCE OF THE DEVICE VARIATION ON RECOGNITION ERROR RATE (RER) IN THE CASE OF THE CSN WITH 3 MTJs

RER CSS Var.	N=1	N=2	N=3
0	33.28%	20.51%	14.49%
10%	33.43%	20.71%	14.63%
20%	33.63%	21.77%	15.28%
30%	35.77%	24.02%	16.69%

B. Impact of the Resolution of Both the CSS Weight and CSN Output Signal on System Performance

It is worth noting that the number of the MTJs (N) in each compound spintronic device determines the resolution of both the CSS weight and CSN output signal. Apparently, the larger N, the higher is the learning accuracy of the proposed ASANN. Unfortunately, higher N would increase the difficulty of device fabrication and energy dissipation. To investigate the impact of the low resolution of the CSS weight and CSN output signal on the learning accuracy of the proposed ASANN, we repeated the simulation in the above subsection. The simulation results for varying the number of the MTJs in both the CSS and CSN are shown in Table III. As we can see, obviously higher resolution of the CSS weight and CSN output signal resolution leads to lower recognition error rate. Detailed analysis shows that the weight resolution of the CSS has a more significant influence on the learning accuracy of the proposed ASANN than that of the CSN output signal. This implies different roles of the CSS weight and CSN output signal resolution in the proposed ASANN.

C. Impact of the Device Variation on System Performance

For large scale implementation of the proposed compound spintronic device, there exist massive device variation on its resistance in different states, which will directly influence the CSS weight. In the following, the impact of device variation on the recognition error rate was investigated. By performing the circuit-level Monte Carlo simulation, the distributions of the resistance of the CSS in different states were obtained. Then, the simulation in the above subsection was repeated with different CSS weight resolutions, where the CSN was with 3 MTJs. Simulation results are shown in Table IV. We can observe that the larger device variation results in the larger recognition error rate. That is, it can degrade the accuracy of neuromorphic computation. Fortunately, the higher the weight resolution, the greater tolerance the neural network system has.

V. Conclusion

In this paper, an all spin artificial neural network (ASANN) is constructed with the proposed compound spintronic synapse (CSS) and neuron (CSN). Both the CSS and CSN are based on a compound spintronic device, which consists of multiple vertically stacked MTJs. From the views of the device physics, fabrication and circuit-level simulation, we have demonstrated that the proposed CSS and CSN can achieve multiple discrete weight levels and implement a multi-step transfer function, respectively. Additionally, the performance of the built ASANN has been investigated by performing system-level simulations with off-line training on the MNIST database for handwritten digital recognition. Moreover, the impact of both the resolution of the proposed CSS and CSN and device variation on system performance has also been evaluated. It has been proved that the proposed CSS and CSN can be employed to build the next generation neural computation platform. However, more physical evaluation and fabrication work will be taken in the future research work.

References

[1] G. Snider, "Instar and outstar learning with memristive nanodevices," *Nanotechnology*, vol. 22, no. 1, 2011, Art. no. 015201.

[2] D. S. Modha, R. Ananthanarayanan, S. K. Esser, A. Ndirango, A. J. Sherbondy, and R. Singh, "Cognitive computing," *Commun. ACM*, vol. 54, no. 8, pp. 62–71, 2011.

[3] Z. Wang, W. Zhao, W. Kang, Y. Zhang, J.-O. Klein, and C. Chappert, "Ferroelectric tunnel memristor-based neuromorphic network with 1t1r crossbar architecture," in *Proc. IEEE Int. Joint. Conf. Neural Networks*, 2014, pp. 29–34.

[4] D. Kuzum, S. Yu, and H. P. Wong, "Synaptic electronics: Materials, devices and applications," *Nanotechnology*, vol. 24, no. 38, 2013, Art. no. 382001.

[5] A. Basu, S. Ramakrishnan, C. Petre, S. Koziol, S. Brink, and P. E. Hasler, "Neural dynamics in reconfigurable silicon," *IEEE Trans. Biomed. Circuits Syst.*, vol. 4, no. 5, pp. 311–319, Oct. 2010.

[6] S. Ramakrishnan, P. E. Hasler, and C. Gordon, "Floating gate synapses with spike-time-dependent plasticity," *IEEE Trans. Biomed. Circuits Syst.*, vol. 5, no. 3, pp. 244–252, Jun. 2011.

[7] A. Sengupta, Y. Shim, and K. Roy, "Simulation studies of an all-spin artificial neural network: Emulating neural and synaptic functionalities through domain wall motion in ferromagnets," *arXiv preprint arXiv:1510.00459*, 2015.

[8] A. Vincent, J. Larroque, N. Locatelli, N. Ben Romdhane, O. Bichler, C. Gamrat, W. S. Zhao, J.-O. Klein, S. Galdin-Retailleau, and D. Querlioz, "Spin-transfer torque magnetic memory as a stochastic memristive synapse for neuromorphic systems," *IEEE Trans. Biomed. Circuits Syst.*, vol. 9, no. 2, pp. 166–174, Apr. 2015.

[9] D. Zhang, L. Zeng, Y. Qu, Z. M. Wang, W. Zhao, T. Tang, and Y. Wang *et al.*, "Energy-efficient neuromorphic computation based on compound spin synapse with stochastic learning," in *Proc. IEEE Int. Symp. Circuits and Systems*, 2015, pp. 1538–1541.

[10] D. Chabi, Z. Wang, C. Bennett, J.-O. Klein, and W. Zhao, "Ultrahigh density memristor neural crossbar for on-chip supervised learning," *IEEE Trans. Nanotechnol.*, vol. 14, no. 6, pp. 954–962, 2015.

[11] N. Locatelli, A. F. Vincent, A. Mizrahi, J. S. Friedman, D. Vodenicarevic, J.-V. Kim, J.-O. Klein, W. Zhao, J. Grollier, and D. Querlioz, "Spintronic devices as key elements for energy-efficient neuroinspired architectures," in *Proc. Design, Automation & Test in Europe Conf. and Exhib.*, 2015, pp. 994–999, EDA Consortium.

[12] W. Zhao and G. Prenat, Spintronics-Based Computing, 2015.

[13] M. Prezioso, F. Merrikh-Bayat, B. Hoskins, G. Adam, K. K. Likharev, and D. B. Strukov, "Training and operation of an integrated neuromorphic network based on metal-oxide memristors," *Nature*, vol. 521, no. 7550, pp. 61–64, 2015.

[14] S. Park, H. Kim, M. Choo, J. Noh, A. Sheri, S. Jung, K. Seo, J. Park, S. Kim, and W. Lee *et al.*, "Rram-based synapse for neuromorphic system with pattern recognition function," in *Proc. Electron Devices Meeting*, 2012, pp. 10–12.

[15] W. Zhao, D. Querlioz, J.-O. Klein, D. Chabi, and C. Chappert, "Nanodevice-based novel computing paradigms and the neuromorphic approach," in *Proc. IEEE Int. Symp. Circuits and Systems*, 2012, pp. 2509–2512.

[16] S. Yu, Y. Wu, R. Jeyasingh, D. Kuzum, and H.-S. P. Wong, "An electronic synapse device based on metal oxide resistive switching memory for neuromorphic computation," *IEEE Trans. Electron Devices*, vol. 58, no. 8, pp. 2729–2737, 2011.

[17] O. Bichler, W. Zhao, F. Alibart, S. Pleutin, S. Lenfant, D. Vuillaume, and C. Gamrat, "Pavlov's dog associative learning demonstrated on synaptic-like organic transistors," *Neural Comput.*, vol. 25, no. 2, pp. 549–566, 2013.

[18] W. Zhao, G. Agnus, V. Derycke, A. Filoramo, J. Bourgoin, and C. Gamrat, "Nanotube devices based crossbar architecture: Toward neuromorphic computing," *Nanotechnology*, vol. 21, no. 17, 2010, Art. no. 175202.

[19] T. Chang, Y. Yang, and W. Lu, "Building neuromorphic circuits with memristive devices," *IEEE Circuits Syst. Mag.*, vol. 13, no. 2, pp. 56–73, 2013.

[20] M. Hu, H. Li, Q. Wu, and G. S. Rose, "Hardware realization of bsb recall function using memristor crossbar arrays," in *Proc 49th Annu. Design Automation Conf.*, 2012, pp. 498–503.

[21] D. Querlioz, O. Bichler, and C. Gamrat, "Simulation of a memristor-based spiking neural network immune to device variations," in *Proc. IEEE Int. Joint Conf. Neural Networks*, 2011, pp. 1775–1781.

[22] D. Querlioz, O. Bichler, P. Dollfus, and C. Gamrat, "Immunity to device variations in a spiking neural network with memristive nanodevices," *IEEE Trans. Nanotechnol.*, vol. 12, no. 3, pp. 288–295, 2013.

[23] S. Yu, B. Gao, Z. Fang, H. Yu, J. Kang, and H.-S. P. Wong, "Stochastic learning in oxide binary synaptic device for neuromorphic computing," *Fronties Neurosci.*, vol. 7, 2013, Art. no. 186.

[24] W. Zhao, S. Chaudhuri, C. Accoto, J.-O. Klein, C. Chappert, and P. Mazoyer, "Cross-point architecture for spin-transfer torque magnetic random access memory," *IEEE Trans. Nanotechnol.*, vol. 11, no. 5, pp. 907–917, 2012.

[25] V. Erokhin, T. Berzina, P. Camorani, A. Smerieri, D. Vavoulis, J. Feng, and M. P. Fontana, "Material memristive device circuits with synaptic plasticity: Learning and memory," *BioNanoSci.*, vol. 1, no. 1–2, pp. 24–30, 2011.

[26] S. Gaba, P. Sheridan, J. Zhou, S. Choi, and W. Lu, "Stochastic memristive devices for computing and neuromorphic applications," *Nanoscale*, vol. 5, no. 13, pp. 5872–5878, 2013.

[27] M. Wang, S. Peng, Y. Zhang, Y. Zhang, Y. Zhang, Q. Zhang, D. Ravelosona, and W. Zhao, "A multilevel cell for stt-mram realized by capping layer adjustment," *IEEE Trans. Magn.*, vol. 51, no. 11, pp. 1–4, 2015.

[28] A. Sengupta and K. Roy, "Spin-transfer torque magnetic neuron for low power neuromorphic computing," in *Proc. IEEE Int. Joint Conf. Neural Networks*, 2015, pp. 1–7.

[29] S. Ikeda, K. Miura, H. Yamamoto, K. Mizunuma, H. Gan, M. Endo, S. Kanai, J. Hayakawa, F. Matsukura, and H. Ohno, "A perpendicular-anisotropy cofeb-mgo magnetic tunnel junction," *Nature Mater.*, vol. 9, no. 9, pp. 721–724, 2010.

[30] S. Peng, M. Wang, H. Yang, L. Zeng, J. Nan, J. Zhou, Y. Zhang, A. Hallal, M. Chshiev, and K. L. Wang *et al.*, "Origin of interfacial perpendicular magnetic anisotropy in mgo/cofe/metallic capping layer structures," *Scientific Rep.*, vol. 5, 2015, Art. no. 18173.

[31] Y. Zhang, W. Zhao, G. Prenat, T. Devolder, J.-O. Klein, C. Chappert, B. Dieny, and D. Ravelosona, "Electrical modeling of stochastic spin transfer torque writing in magnetic tunnel junctions for memory and logic applications," *IEEE Trans. Magn.*, vol. 49, no. 7, pp. 4375–4378, 2013.

[32] T.-I. Cheng, C.-W. Cheng, and G. Chern, "Perpendicular magnetic anisotropy induced by a cap layer in ultrathin mgo/cofeb/nb," *J. Appl. Phys.*, vol. 112, no. 3, 2012, Art. no. 033910.

[33] D.-S. Lee, H.-T. Chang, C.-W. Cheng, and G. Chern, "Perpendicular magnetic anisotropy in mgo/cofeb/nb and a comparison of the cap layer effect," *IEEE Trans. Magn.*, vol. 50, no. 7, pp. 1–4, 2014.

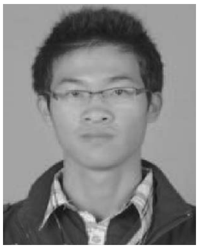
[34] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

Deming Zhang (S'15) received the B.S. degree in electronic and information engineering from Beihang University, Beijing, China, in 2011.

Currently, he is working toward the Ph.D. degree in microelectronics and solid state electronics at the Spintronics Interdisciplinary Center, Beihang University. His interests include the emerging NVM-based IC design and neuromorphic computation.

Lang Zeng (S'07–M'12) received the B.S. and Ph.D degrees in microelectronics from Peking University, Beijing, China, in 2007 and 2012, respectively.

From 2009–2011, he was a Visiting Scholar at Purdue University, West Lafayette, IN, USA. From 2012–2014, he was a Postdoctoral Associate at the Institute of Microelectronics, Peking University. In 2014, he joined Beihang University, Beijing, China, as an Associate Professor. His research interest includes carrier transport in nanoscale devices and 2D material, All Spin Logic devices, neuromorphic computation based on spintronics, and reliability issues of PMA STT-MRAM.

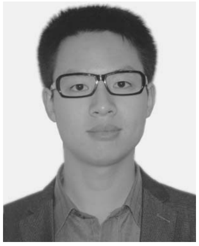


Kaihua Cao (S'16) was born in China in 1992. He received the B.S. degree in microelectronics from the Qingdao University of Physics, Shandong, China, in 2014.

Currently, he is working toward the Ph.D degree in microelectronics and solid electronics at the Fert Beijing Institute, Beihang University, Beijing, China. His main research interests are the nanofabrication and measurement of spin devices.

Mengxing Wang (S'15) received the B.S. degree in electronic and information engineering from Beihang University, Beijing, China, in 2012.

Currently, she is working toward the Ph.D. degree in microelectronics and solid-state electronics at Spintronics Interdisciplinary Research Center, Beihang University. Her research interest is the fabrication of perpendicular magnetic tunnel junction with high performance.



Shouzhong Peng (S'15) received the B.S. degree in electronic and information engineering from Beihang University, Beijing, China, in 2013.

Currently, he is working toward the Ph.D. degree in microelectronics and solid-state electronics at Spintronics Interdisciplinary Research Center, Beihang University. His research interests include first-principles study of electronic and magnetic properties of magnetic tunnel junctions.

Yue Zhang (S'11–M'14) received the B.S. degree in optoelectronics from the Huazhong University of Science and Technology, Wuhan, China, in 2009, and the M.S. and Ph.D. degrees in microelectronics from University of Paris-Sud, Orsay, France, in 2011 and 2014, respectively.

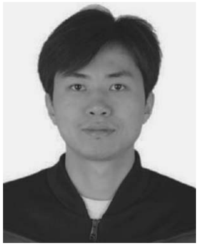
Currently, he is an Associate Professor at Beihang University, Beijing, China. His research focuses on emerging non-volatile memory technologies and hybrid low-power circuit designs. He has authored more than 50 scientific papers in referred journals and conferences and received several best paper awards from international conferences, including NANOARCH, NEWCAS, and ESREF.

Youguang Zhang (M'13) received the M.S. degree in mathematics from Peking University, Beijing, China, and the Ph.D. degree in communication and electronic systems from Beihang University, Beijing, China, in 1987 and 1990, respectively.

Currently, he is a Professor in the School of Electronic and Information Engineering, Beihang University. His research interests include microelectronics and wireless communication. He has participated in several NSF projects and and has authored numerous papers.

Jacques-Olivier Klein (M'90) received the Ph.D. degree and the Habilitation in electronic engineering from the University of Paris-Sud, Orsay, France, in 1995 and 2009, respectively.

Currently, he is a Professor in the Institut d'Electronique Fondamentale, University of Paris-Sud, where he leads the nano-computing group that focuses on architecture of circuits and systems based on emerging devices in the field of nano-magnetism and bio-inspired nano-electronics.



Yu Wang (S'05–M'07–SM'14) received the B.S. degree and Ph.D. degree (with honors) from Tsinghua University, Beijing, China, in 2002 and 2007, respectively.

Currently, he is an Associate Professor in the Department of Electronic Engineering, Tsinghua University. His research interests include parallel circuit analysis, application specific hardware computing (especially on the Brain related problems), and power/reliability aware system design methodology. He has authored or coauthored more than 140 papers in refereed journals and conferences.

Dr. Wang was the recipient of the IBM X10 Faculty Award in 2010, Best Paper Award at ISVLSI 2012, Best Poster Award at HEART 2012, and six Best Paper Nominations at ASPDAC/CODES/ISLPED. He serves as the associate editor for IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems and *Journal of Circuits, Systems, and Computers*. He was the TPC Co-Chair of ICFPT 2011, Finance Chair of ISLPED 2012–2015, and serves as a TPC member at many conferences, including DAC, FPGA, DATE, ASPDAC, ISLPED, ISQED, ICFPT, and ISVLSI.

Weisheng Zhao (S'04–M'07–SM'14) received the Ph.D. degree in physics from the University of Paris-Sud, Orsay, France, in 2007.

In 2009, he joined the CNRS as a tenured Research Scientist and his interests include the hybrid integration of spintronics nanodevices with CMOS circuits. In 2014, he became a Distinguished Professor at Beihang University, Beijing, China. He has authored more than 150 scientific papers for leading journals and conferences, including *Nature Communications*, *Advanced Materials*, IEEE TED, IEEE TCAS-I, IEEE TBIOCAS, IEEE TNANO, ISCA, and DAC.

Dr. Zhao serves as associate editor of two SCI journals, IEEE Transactions on Nanotechnology and *IET Electronics Letters*, guest editor for IEEE Transactions on Multi-Scale Computing, and is the general chair for ACM/IEEE Nanoarch 2016.