Roadmap

Roadmap for Unconventional Computing with Nanotechnology

Giovanni Finocchio[1,37,38], Supriyo Bandyopadhyay[2,37,38], Peng Lin[3], Gang Pan[3], J. Joshua Yang[4], Riccardo Tomasello[5], Christos Panagopoulos[6], Mario Carpentieri[5], Vito Puliafito[5], Johan Åkerman[7], Hiroki Takesue[8], Amit Ranjan Trivedi[9], Saibal Mukhopadhyay[10], Kaushik Roy[11], Vinod K. Sangwan[12], Mark C. Hersam[12], Anna Giordano[1], Huynsoo Yang[13], Julie Grollier[14], Kerem Camsari[15], Peter Mcmahon[16], Supriyo Datta[11], Jean Anne Incorvia[17], Joseph Friedman[18], Sorin Cotofana[19], Florin Ciubotaru[20], Andrii Chumak[21], Azad J. Naeemi[10], Brajesh Kumar Kaushik[22], Yao Zhu[23], Kang Wang[24], Belita Koiller[25], Gabriel Aguilar[25], Guilherme Temporão[26], Kremena Makasheva[27], Aida Tordi- Sanial[28], Jennifer Hasler[10], William Levy[29], Vwani Roychowdhury[30], Samiran Ganguly[31], Avik Ghosh[31], Davi Rodriquez[5], Satoshi Sunada[32], Karin Evershor-Sitte[33], Amit Lal[16], Shubham Jadhav[16], Massimiliano Di Ventra[34], Yuriy Pershin[35], Kosuke Tatsumura[36], Hayato Goto[36]

[1] Department of Electrical Engineering, University of Messina, Italy
[2] Department of Electrical and Computer Engineering, Virginia Commonwealth University, USA
[3] College of Computer Science and Technology, Zhejiang University, China
[4] Department of Electrical and Computer Engineering, University of Southern California, USA
[5] Dipartimento di Ingegneria Elettrica e dell'Informazione, Politecnico di Bari
[6] School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore
[7] Department of Physics, University of Gothenburg, Sweden
[8] Graduate School of Engineering Science, Osaka University, Japan
[9] Department of Electrical and Computer Engineering, University of Illinois at Chicago, USA
[10] School of Electrical Engineering, Georgia Institute of Technology, USA
[11] School of Electrical Engineering, Purdue University, USA
[12] Department of Materials Science and Engineering, Northwestern University, USA
[13] Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[14] Unité Mixte de Physique, CNRS, Thales, France
[15] Department of Electrical and Computer Engineering, University of California at Santa Barbara, USA
[16] School of Applied and Engineering Physics, Cornell University, USA
[17] Department of Electrical and Computer Engineering, University of Texas at Austin, USA
[18] Department of Electrical and Computer Engineering, University of Texas at Dallas, USA
[19] Computer Engineering Laboratory, Technical University – Delft, Netherlands
[20] Interuniversity Microelectronics Center (IMEC), Belgium
[21] Department of Physics, University of Vienna, Austria
[22] Department of Electronics and Communications Engineering, Indian Institute of Technology – Roorkee, India
[23] A-STAR, Singapore
[24] School of Integrated Circuit Science and Engineering, Beihang University, China
[25] Institute of Physics, Federal University of Rio de Janeiro, Brazil
[26] Center for Telecommunication Studies, Pontifical Catholic University of Rio de Janeiro, Brazil
[27] CNRS Laboratory on Plasma and Conversion of Energy, France
[28] Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, France
[29] Department of Neurological Surgery, University of Virginia, USA
[30] Department of Electrical and Computer Engineering, University of California at Los Angeles, USA
[31] Department of Electrical and Computer Engineering, University of Virginia, USA
[32] Department of Mechanical Engineering, Kanazawa University, Japan
[33] Department of Physics, University of Duisberg-Essen, Germany
[34] Department of Physics, University of California at San Diego, USA

[35] Department of Physics, University of South Carolina, USA
[36] Toshiba Corporation, Japan

[37] Guest editors of the Roadmap.
[38]Author to whom any correspondence should be addressed.

E-mails: giovanni.finocchio@unime.it (Giovanni Finocchio) sbandy@vcu.edu (Supriyo Bandyopadhyay)

currently under peer review with Nano

Abstract

In the "Beyond Moore's Law" era, with increasing edge intelligence, domain-specific computing embracing unconventional approaches will become increasingly prevalent. At the same time, the adoption of a wide variety of nanotechnologies will offer benefits in energy cost, computational speed, reduced footprint, cyber-resilience and processing prowess. The time is ripe to lay out a roadmap for unconventional computing with nanotechnologies to guide future research and this collection aims to fulfill that need. The authors provide a comprehensive roadmap for neuromorphic computing with electron spins, memristive devices, two-dimensional nanomaterials, nanomagnets and assorted dynamical systems. They also address other paradigms such as Ising machines, Bayesian inference engines, probabilistic computing with p-bits, processing in memory, quantum memories and algorithms, computing with skyrmions and spin waves, and brain inspired computing for incremental learning and solving problems in severely resource constrained environments. All of these approaches have advantages over conventional Boolean computing predicated on the von-Neumann architecture. With the computational need for artificial intelligence growing at a rate 50 faster than Moore's law for electronics, more unconventional approaches to computing and signal processing will appear on the horizon and this roadmap will aid in identifying future needs and challenges.

Contents

7. Compute in memory

7.1 Compute-in-Memory with Nanoscale CMOS Technologies, Amit Ranjan Trivedi, Saibal Mukhopadhyay, Kaushik Roy

7.2 In-memory computing using non-volatile memories, I-Ting Wang, Wang Kang, Yao Zhu, Brajesh Kumar Kaushik

8. **Computing with dynamical systems**

8.1 Computing with Dynamical Systems, Davi Rohe Rodrigues, Satoshi Sunada, Karin Everschor-Sitte

8.2 Computing with Ising Machines realized through coupled nano-oscillators, Vito Puliafito, Johan Akerman, Hiroki Takesue

9. **Brain inspired computing**

9.1 – Brain-Inspired Unconventional Computing, Jennifer Hasler, Samiran Ganguly, Avik Ghosh, William Levy, Vwani Roychowdhury, Supriyo Bandyopadhyay

10. **Quantum memories**

10.1 Quantum computers with spin-based qubits in silicon, Belita Koiller, Gabriel H. Aguilar, Guilherme P. Temporao

1.1 – Magnetic Architectures for Unconventional Computing

Jean Anne Incorvia, University of Texas at Austin, Austin, TX 78712
incorvia@austin.utexas.edu

Joseph Friedman, University of Texas at Dallas, Richardson, TX 75080
Joseph.Friedman@utdallas.edu

Supriyo Bandyopadhyay, Virginia Commonwealth University, Richmond, VA 23284
sbandy@vcu.edu

Status

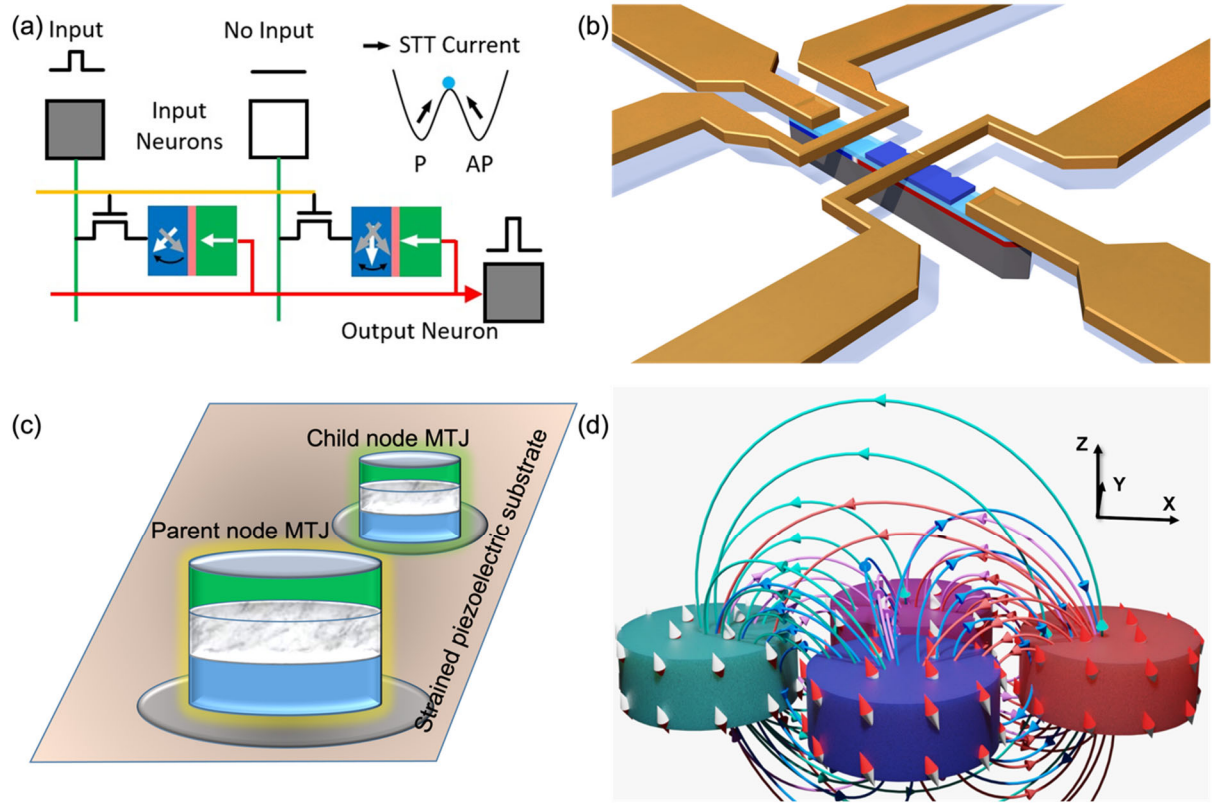

Figure 1.1 (a) Stochastic write MTJ. (b) Example domain wall MTJ, here patterned as a synapse, with patterned blue/white/red domain wall track and blue output tunnel junction, including notches to control the domain wall position. (c) A two node Bayesian network implemented with two MTJs. (d) Depiction of a reservoir using MTJs.

Next generation unconventional computing will address key needs and problems for processing the increasingly large and unstructured data workloads, as well as the increase in edge computing devices and corresponding energy constraints. Some problems that unconventional computing addresses include the bottleneck between compute and memory; the large energy and delay penalty of analog-to-digital conversion; computing with small energy budgets; and application specific computing with balanced energy, time, and precision needs, since precise computing is not always needed.

Magnetic thin films, both continuous and patterned into nanomagnets, have a long history in computing, starting with hard disk drives and including today's spin transfer torque and spin orbit torque based magnetic random access memory (STT-MRAM, SOT-MRAM). Unconventional computing hardware (neuromorphic, Bayesian, Boltzmann machines) implemented with magnetic devices, e.g., magnetic tunnel junctions (MTJs), are attractive since the constituent elements are non-

volatile and could be extremely energy efficient [SB1]. The device characteristics and inter device interactions, which depend on the energy barrier within the free layer of the MTJ, can be easily tailored, as well as controlled through multiple simultaneous knobs, such as current, voltage, strain, magnetic fields, and by both DC and AC inputs. This offers immense flexibility in designing hardware accelerators for machine learning such as binary stochastic neurons (BSNs) [SB2], neuromorphic components like synapses, Ising machines, etc.

The MTJ and corresponding devices also benefit from high endurance in switching the magnetic state, and from the fact that, under normal operation, the resistance states can be set in a controllable way, without drift over time or over cycles. This stability and robustness of the bit state control (not necessarily the states themselves, which can be tuned between stable and stochastic) can compensate for some of the challenges MTJs face.

Whereas MTJs with low energy barriers exhibit constant stochastic switching between resistance states, useful for BSNs, MTJ memory devices with high energy barriers exhibit an alternative stochastic phenomenon: the switching between the two stable states is intrinsically stochastic. This stochastic writing process provides analog behavior to these binary memory devices, enabling their use in neuromorphic systems of the type described by [F1]. Furthermore, the binary MTJ states are inherently robust against the variations and stochastic behavior that plagues memristors and phase - change memory, thereby making non ∀olatile MTJ synapses a promising technology for neuromorphic computing.

Neural network crossbar arrays can be implemented using the nanomagnet as both the artificial synapse and the artificial neuron. By using a top þinned MTJ stack and extending the bottom magnetic layer into a longer track, the MTJ can be configured as a domain wall ₥agnetic tunnel junction (DW - MTJ). Subsequent choice of patterning can then have the device show analog resistance states as a synapse [JA1, JA2, JA3] or leaky, integrate, and fire as a neuron [JA4]. While a domain wall, or similarly a magnetic skyrmion, can be harder to control than a single ₫omain nanomagnet, it provides additional bio ₥imetic functions for unconventional computing such as time delays, stochastic pinning and depinning, and frequency ƀased switching [JA3]. It can also benefit from magnetic field interactions between the domain walls of the devices [JA5].

Belief networks (Bayesian inference engines) are another genre of unconventional computers for computing in the presence of uncertainty. They are difficult to implement with most technologies since they require *non ₹eciprocal* synapses. Simple 2 ₦ode networks consist of a parent and a child node where the child node's state is correlated with that of the parent, but not the other way around. Two dipole €oupled MTJs of different shapes built on a piezoelectric substrate can implement this paradigm easily. The degree of correlation or anti €orrelation between the nodes can be varied with global strain applied to both MTJs via the piezoelectric [SB3] and this can enable Bayesian inference [SB4]. The synaptic connection between the nodes is dipole coupling, which consumes no area on the chip and dissipates no energy since it does not involve current flow.

While trained neuromorphic computing systems promise exceptional capabilities, the training process incurs significant hardware costs in terms of energy, area, and speed. Reservoir computing therefore provides an opportunity to avoid those costs by using a system that requires minimal training. In

particular, the bulk of the system is untrained, while only a single output layer must be trained. Nanomagnetism naturally provides such reservoirs, as irregular arrays of closely packed nanomagnets exhibit frustration that produces complex physical dynamics and hysteresis. All these extra ordinary capabilities make magnetic architectures for unconventional computing unique and attractive.

Current and Future Challenges

The year 2021 heralded the first three experimental demonstrations of neural networks with synapse weights encoded in binary MTJ states; all three performed some type of recognition task. In the simplest of these experiments, a 4x2 single layer neuromorphic network was directly implemented with MTJ synapses [F2] to perform vector matrix multiplication in the manner described by [F1]. More complex MTJ based synapse structures were used in a 64x64 single layer network [F3] and a two layer (13×6 + 6×3) network [F4]. The key future challenges for this neuromorphic computing approach are scaling to large network dimensions and the experimental demonstration of learning through stochastic switching.

DW MTJs also have been demonstrated recently [JA6, JA1, JA7]. Clear needs are better understanding and control of the domain wall behavior over many cycles, especially without needing to refresh the devices; all electrical control without the need of external magnetic fields to aid domain wall movement; scaling down to modern feature sizes; and scaling up for larger circuit demonstrations, including better understanding of device to device variations and their impact on the unconventional computing applications.

As a first step towards the development of reservoir computers based on frustrated nanomagnetism, micromagnetic simulation studies have demonstrated their memory capacity and expressivity [F5]. These systems have been shown to successfully perform complex classification tasks, including waveform identification, Boolean operations based on previous inputs, and observation and prediction of dynamical discrete time series [F6]. Furthermore, comparative simulation studies indicate a 60x improvement in energy efficiency relative to conventional CMOS systems. However, experimental demonstration and proof of concept remain a significant challenge [F6].

Neuromorphic computing is generally much more forgiving of switching errors than Boolean logic, but it is not necessarily very tolerant of large device to device variations. The response time of BSNs, for example, can change dramatically in the presence of fabrication defects [SB5] or slight shape variations [SB6], which results in significant device to device variations that is a challenge for large scale networks. One way to counter this is to adopt hardware aware in situ learning [SB7]. Another is to replace common ferromagnets used in MTJs with dilute magnetic semiconductors which have several orders of magnitude lower saturation magnetization. That makes the energy barriers in the nanomagnets much less sensitive to shape and size variations and suppresses device to device variations [SB8].

A challenge with ferromagnetic devices is the relatively slow switching speed of ~1 ns which creates a bottleneck in training and inference in both recurrent and deep neural networks. There has been some recent interest in harnessing anti ferromagnetic materials for synapses [SB9] and they are capable of much higher speed. This is a nascent field, but important discoveries may be around the corner. All

these challenges make magnetic architectures for unconventional computing a fertile field of research.

Advances in Science and Technology to Meet Challenges

One of the critical challenges for neural networks based on both conventional MTJs and DW-MTJs is efficient network training. As conventional supervised learning algorithms (e.g., backpropagation) become increasingly complex when scaled to large and deep networks, the hardware costs for implementing this mathematical circuitry hinder the development of neural networks with online learning. Preliminary explorations of unsupervised learning algorithms with MTJs [F7, F8] and DW-MTJs [F9] indicate that local Hebbian learning rules can be used with feedback circuits to efficiently train neural networks with minimal energy, speed, and area costs.

Realization of reservoir computing systems based on frustrated nanomagnets will require advances in experimental techniques for providing the input signals while contemporaneously measuring the magnetization of the various nanomagnets that make up the reservoir. The inputs can be provided via STT switching, and the output magnetizations can be read through an MTJ (preliminary experimental efforts may focus on imaging the output). These output signals must be fed to a single trained layer of the type described in [F1].

Domain wall creep and the stochasticity of domain wall motion at room temperature pose significant obstacles to DW-MTJ technologies. Recent progress with notched structures [JAC et al] have ameliorated some of the difficulties, but further material research is needed to find possibly simpler solutions where the intrinsic material properties may be able to suppress or control the stochastic behaviour of domain wall motion.

A well known challenge with magnetic devices such as MTJs is the low on/off ratio which typically results in low training accuracies in neuromorphic architectures [SB10]. Research is needed to find proper material combinations to increase the tunnelling magneto-resistance or on/off ratios of MTJs to alleviate this problem. This field has a long history and unfortunately has been slow in making progress. However, its importance cannot be overstated since a high on/off ratio provides a wider range of synaptic weights and improves error tolerance.

Challenges in patterning MTJ-based structures leads to device-to-device variation that can increase with reducing feature size [F3, JA8]. This challenge compounds with the low on/off ratio to blur the difference between the 0 and 1 state, and is even more of an issue if more resistance levels are desired between the 0 and 1. While the good control of the device resistance states can help with this issue, better patterning methods are needed, as well as more effort in circuit design to design around these challenges.

Concluding Remarks

The human brain consumes 1-100 fJ of energy per synaptic event [SB11]. Magnetic devices can rival (or even eclipse) this energy efficiency [SB1]. Their non-volatility offers additional architectural advantages, e.g., in reservoir computing [F6].

The low energy consumption has other benefits; it provides hardware security as well, which is very important for artificial intelligence. Because of the low power requirement, these architectures can be embedded in edge devices that have minimal contact with the cloud and are therefore somewhat insulated from cloud borne attacks. Additionally, they are inherently resilient against malign hardware. A hardware Trojan, no matter how surreptitious, will consume some energy and that can become comparable to (or exceed) the energy consumed by magnetic hardware. Therefore, Trojans can be easily detected with side channel monitoring.

Finally, neuromorphic computing with anti ferromagnetic devices is a burgeoning area of research laden with promise and it can spawn new devices and architectures that will speed up training and inference tasks immensely. This is an exciting area of research that is about to bear fruit.

Acknowledgements

The work of S. B. in this field is currently supported by the National Science Foundation under grants CCF 2001255 and CCF 2006843. The work of J. A. I. in this field is currently supported by the National Science Foundation under grants EPMD 2225744, CCF 1910997, and CCF 2006753. Joe add yours.

References

[(Separate from the two page limit) Maximum 20 References. Please provide the full author list, and article title, for each reference to maintain style consistency in the combined roadmap article. Style should be consistent with all other contributions - use IEEE style]

[F1] J. J. Yang *et al.*….can I reference the first paper in this roadmap?

[F2] P. Zhou, A. J. Edwards, F. B. Mancoff, D. Houssameddine, S. Aggarwal, J. S. Friedman, "Experimental Demonstration of Neuromorphic Network with STT MTJ Synapses," arXiv, 2112.04749, 2021.

[F3] S. Jung, H. Lee, S. Myung, H. Kim, S. K. Yoon, S. W. Kwon, Y. Ju, M. Kim, W. Yi, S. Han, B. Kwon, B. Seo, K. Lee, G. H. Koh, K. Lee, Y. Song, C. Choi, D. Ham, S. J. Kim, "A crossbar array of magnetoresistive memory devices for in memory computing," Nature, vol. 601, pp. 211 216, 2022.

[F4] J. Goodwill, N. Prasad, B. Hoskins, M. Daniels, A. Madhavan, L. Wan, T. Santos, M. Tran, J. Katine, P. Braganca, M. Stiles, J. McClelland, "Implementation of a Binary Neural Network on a Passive Array of Magnetic Tunnel Junctions," arXiv, 2112.09159 2021.

[F5] P. Zhou, N. R. McDonald, A. J. Edwards, L. Loomis, C. D. Thiem, J. S. Friedman, "Reservoir Computing with Planar Nanomagnet Arrays," arXiv, 2003.10948, 2020.

[F6] A. J. Edwards, D. Bhattacharya, P. Zhou, N. R. McDonald, L. Loomis, C. D. Thiem, J. Atulasimha, J. S. Friedman, "Frustrated Arrays of Nanomagnets for Efficient Reservoir Computing," arxiv, 2103.09353, 2021.

[F7] A. F. Vincent, J. Larroque, N. Locatelli, N. B. 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 Transactions on Biomedical Circuits and Systems, vol. 9, no. 2, pp. 166–174, 2015.

[F8] P. Zhou, J. A. Smith, L. Deremo, S. K. Heinrich Barna, J. S. Friedman, Synchronous Unsupervised STDP Learning with Stochastic STT MRAM Switching, arXiv, 2112.05707, 2021.

[F9] A. Velasquez, C. H. Bennett, N. Hassan, W. H. Brigner, O. G. Akinola, J. A. C. Incorvia, M. J. Marinella, J. S. Friedman, "Unsupervised Competitive Hardware Learning Rule for Spintronic Clustering Architecture," arXiv, 2003.11120, 2020.

[SB1] S. Bandyopadhyay, J. Atulasimha and A. Barman, "Magnetic Straintronics: Manipulating the Magnetization of Magnetostrictive Nanomagnets with Strain for Energy Efficient Applications", Appl. Phys. Rev., vol. 8, no. 4, 041323 (2022).

[SB2] O. Hassan, R. Faria, K. Y. Camsari, J. Z. Sun and S. Datta, "Low Barrier Magnet Design for Efficient Hardware Binary Stochastic Neurons", IEEE Magn, Lett., vol. 10, 4502805 (2019).

[SB3] M. T. McCray, M. A. Abeed and S. Bandyopadhyay, "Electrically Programmable Probabilistic Bit Anti Correlator on a Nanomagnetic Platform", Sci. Rep., vol. 10, 12361 (2020).

[SB4] S. Nasrin, J. L. Drobitch, P. Shukla, T. Tulabandhula, S. Bandyopadhyay and A. R. Trivedi, "Bayesian Reasoning Machine on a Magneto Tunneling Junction Network", Nanotechnology, vol. 31, 484001 (2020).

[SB5] M. A. Abeed and S. Bandyopadhyay, "Low Energy Barrier Nanomagnet Design for Binary Stochastic Neurons: Design Challenges for Real Nanomagnets with Fabrication Defects", IEEE Magn. Lett., vol. 10, 4504405 (2019).

[SB6] R. Rahman and S. Bandyopadhyay, "Variability of Binary Stochastic Neurons Employing Low Energy Barrier Nanomagnets with In Plane Anisotropy", arXiv:2108.04319.

[SB7] J. Kaiser, W. A. Borders, K. Y. Camsari, S. Fukami, H. Ohno and S. Datta, "Hardware Aware In Situ Learning Based on Stochastic Magnetic Tunnel Junctions", Phys. Rev. Appl., vol. 17, 014016 (2022).

[SB8] R. Rahman and S. Bandyopadhyay, "Robustness of Binary Stochastic Neurons Implemented with Low Barrier Nanomagnets Made of Dilute Magnetic Semiconductors", arXiv:2205.01793.

[SB9] A. Kurenkov, S. Fukami and H. Ohno, "Neuromorphic Computing with Anti Ferromagnetic Spintronics", J. Appl. Phys., vol. 128, 010902 (2020).

[SB10] A. Gutsche, S. Siegel, J. Zhang, S. Hambsch and R. Dittman, "Exploring Area Dependent $Pr_{0.7}Ca_{0.3}MnO_3$ Based Memristive Devices as Synapses in Spiking and Artificial Neural Networks", Front. Neurosci, vol. 15, 661261 (2021).

[SB11] R. Merkle, "Energy Limits to the Computational Power of the Human Brain", Foresight Update 6, (Foresight Institute, 1989).

[JA1] T. Leonard, S. Liu, M. Alamdar, C. Cui, O. G. Akinola, L. Xue, T. P. Xiao, J. S. Friedman, M. J. Marinella, C. H. Bennett, and J. A. C. Incorvia. "Shape Dependent Multi Weight Magnetic Artificial Synapses for Neuromorphic Computing." ArXiv, 2111.11516, 2021.

[JA2] S. A. Siddiqui, S. Dutta, A. Tang, L. Liu, C. A. Ross & M. A. Baldo. "Magnetic Domain Wall Based Synaptic and Activation Function Generator for Neuromorphic Accelerators." *Nano Lett.* vol. 20, 1033–1040 (2020).

[JA3] S. Liu, T. P. Xiao, C. Cui, J. A. C. Incorvia, C. H. Bennett, and M. J. Marinella. "A domain wall - magnetic tunnel junction artificial synapse with notched geometry for accurate and efficient training of deep neural networks." *Applied Physics Letters* vol. 118, 202405 (2021).

[JA4] N. Hassan, X. Hu, L. Jiang Wei, W. H. Brigner, C. H. Bennett, O. G. Akinola, J. A. C. Incorvia, and J. S. Friedman. "Magnetic domain wall neuron with lateral inhibition." *Journal of Applied Physics* 124:15, 152127 (2018).

[JA5] C. Cui, O. G. Akinola, N. Hassan, C. H. Bennett, M. J. Marinella, J. S. Friedman, and J. A. C. Incorvia. "Maximized lateral inhibition in paired magnetic domain wall racetracks for neuromorphic computing." *IOP Nanotechnology* vol. 31, 29 (2020).

[JA6] M. Alamdar, T. Leonard, C. Cui, B. P. Rimal, L. Xue, O. G. Akinola, P. Xiao, J. S. Friedman, C. H. Bennett, M. J. Marinella, and J. A. C. Incorvia. "Domain wall magnetic tunnel junction spin orbit torque device and circuit prototypes for in memory computing". *Applied Physics Letters* vol. 118, 112401 (2021).

[JA7] E. Raymenants, O. Bultynck, D. Wan, T. Devolder, K. Garello, L. Souriau, A. Thiam, D. Tsvetanova, Y. Canvel, D. E. Nikonov, I. A. Young, M. Heyns, B. Soree, I. Asselberghs, I. Radu, S. Couet & V. D. Nguyen. "Nanoscale domain wall devices with magnetic tunnel junction read and write." *Nat. Electron.* vol. 4, 392–398 (2021).

[JA8] L. Xue, C. Ching, A. Kontos, J. Ahn, X. Wang, R. Whig, H. W. Tseng, J. Howarth, S. Hassan, H. Chen, M. Bangar, S. Liang, R. Wang & M. Pakala. Process optimization of perpendicular magnetic tunnel junction arrays for last level cache beyond 7 nm node. *Dig. Tech. Pap. Symp. VLSI Technol.* 117–118 (2018).

1.2- The impact of spintronics in neuromorphic computing

Qu Yang[1] (eleyaq@nus.edu.sg), Anna Giordano (agiordano@unime.it)[2], Julie Grollier (julie.grollier@cnrs-thales.fr),[3] Giovanni Finocchio (gfinocchio@unime.it),[4] Hyunsoo Yang[1] (eleyang@nus.edu.sg)

[1]Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117576

[2]Department of Engineering, University of Messina, Italy

[3]Unité Mixte de Physique CNRS/Thales, CNRS, Université Paris Saclay, 91767, Palaiseau, France

[4]Department of Mathematical and Computer Sciences, Physical Sciences and Earth Sciences, university of Messina, Italy

Status

Neuromorphic spintronics aims to develop spintronic hardware devices and circuits with brain-inspired principles [1]. The conventional complementary metal–oxide–semiconductor (CMOS) neuron and synapse designs require numerous transistors and feedback mechanisms and would be unsuitable for developing modern artificial intelligence systems. Spintronics is a promising approach to neuromorphic computing as it potentially enables energy and area-efficient embedded applications by mimicking key features of biological synapses and neurons with a single device instead of using multiple electronic components [2, 3].

Currently, the main building block for neuromorphic spintronics is the magnetic tunnel junction (MTJ), which exhibits several unique characteristics over other technologies, including CMOS compatibility, low power consumption, outstanding read/write endurance, non-volatility, and fast speed [4]. Krysteczko et al. carried out the first work on the spintronic implementation of memristive functionalities by voltage-induced switching in MTJs [5]. Later on, different MTJ-based spintronic structures have been proposed to potentially offer solutions to neuronal computations with bio-fidelity [6]. For example, MTJs have been used for the realization of memristors for storing synaptic weights [6], activation functions of a neuron (such as ReLU-like and sigmoidal) [6, 7], reservoir computing [8] and very recently as crossbar array for in-memory computing [9].

The versatile behaviour of MTJs is the winning point of this technology, as we stress that the functionality described below can be realized at device level. Fig. 1 summarizes some examples of building blocks such as superparamagnetic MTJs (S-MTJ), three terminal MTJ where both spin-orbit torque (SOT) and spin-transfer torque can be combined together, spintronic diodes (DIODE) for microwave operations, spin-torque nano-oscillators (STNO) useful for computing tasks, and MTJs with non-uniform ground state driven by the stabilization of a domain wall (DWM) which can be used for emulating synaptic behaviour. Concerning the latter, Fig. 2 illustrates an MTJ-based integrate-and-fire spiking neuron [6]. In this case, the current-induced SOT integrates the domain wall motion (DWM). When the domain wall reaches the critical position (threshold), the neuron device spikes a “fire” signal. With similar principles, more sophisticated MTJ-based neuron models have been further developed [10], and similar structures with magnetic skyrmions instead of DWM as information carrier have also been constructed [11]. Moreover, various SOT neuromorphic solutions also have been demonstrated experimentally for the realization of ultrafast neuromorphic spintronics [12], field-free artificial neuron [3], and auto-reset stochastic neuron [13].

Spintronics could be powerful in the development of neuromorphic computing because it enables the data processing and storage at a very local level. To this end, there have been a rich variety of spintronic materials and device designs for proof-of-concept neuromorphic computing

implementations. Regarding the neural networks, the paradigm is now shifting from frame-based to event-based exploiting the idea of spiking neurons in spiking neural networks, an approach that is closer to the brain working principle. These novel research progresses have further aroused a research enthusiasm towards developing large-scale brain-inspired spintronic systems.

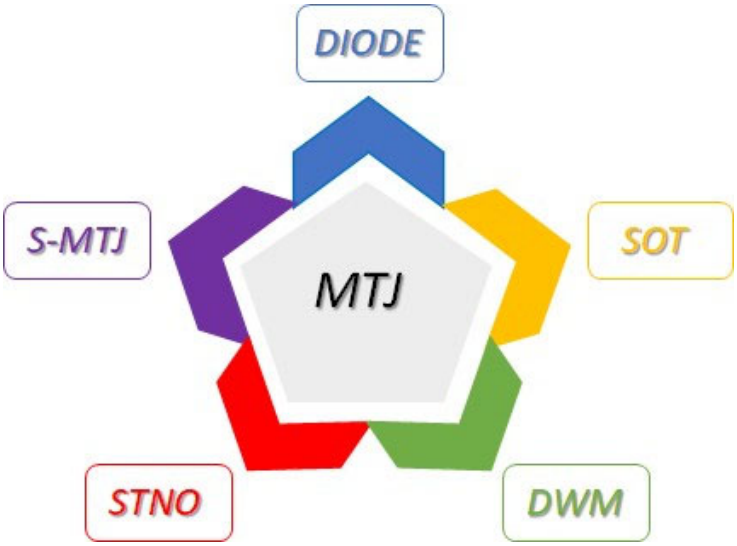

Figure 1. Illustration of the potential use of a magnetic tunnel junction (MTJ). Superparamagnetic MTJ (s-MTJ) can be used for example to generate random numbers or for inference. SOT is a three terminal MTJ where it is possible to take advantage from both spin-orbit-torque and spin-transfer-torque. STNO and DIODE refer to spintronic oscillators and diodes which can used both for computing and the development of a microwave-based neuromorphic computing technology. Finally, MTJ with non-uniform ground state (DWM) can be used for the implementation of synaptic behaviour.

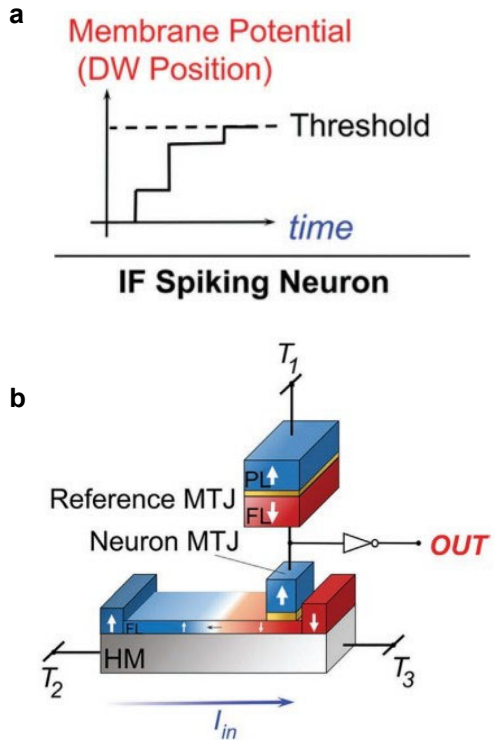

Figure 2. Integrate-fire spiking artificial neuron with an MTJ located at the edge of the free layer. Reprinted with permission from [6]. Copyright 2017, American Institute of Physics (AIP).

Current and Future Challenges

There are important challenges to be overcome for further development of neuromorphic spintronics. One of the biggest challenges is that the read-out signal of spintronic approaches is quite

small, making it difficult to read quickly. Integration of spintronic synapse/neuron devices in the MTJ-based the magnetic random-access memory (MRAM) architecture will increase the read-out signal via the tunnelling magnetoresistance (TMR). However, the resistance changes of MTJs (OFF/ON ratios are typically one to three) are still small compared to other memory technologies [14]. To further address this issue requires complementing junctions with CMOS. For example, every MTJ can be accompanied by a field-effect transistor switch [9]. As a result, one can pursue a higher ON/OFF ratio and lower leakage current. However, the integration of transistors also introduces a drawback as it limits the achievable density of the synaptic arrays.

Regarding spintronic neurons, typically a reset pulse with a sufficiently high magnitude (equal to or several times larger than that required for writing) and of opposite polarity is necessary [7, 12]. This not only increases energy consumption and complexity of the chip, but also lowers the areal density for peripheral circuits required. Besides, an extra resetting step will decrease the operational speed of the neural circuit. The neural device will not be usable till it has been reset by a reset-pulse. Therefore, a bio-realistic neural device with the auto-reset functionality is desirable for energy-efficient and densely packed artificial neural networks.

Moreover, implementing spintronic hardware in neural networks has challenges in coupling control of each neuron. Synchronization of device properties instead of changing their individual response would be one promising way to extend spintronic approaches to multilayer neural networks [15]. There is also an inevitable device variability and energy consumption issue to be overcome for each neuron to be connected up to several thousand synapses in a neural network algorithm. For further large scaling, it is necessary to explore novel materials to maintain the analog behaviours when reducing dimensions. In addition, it is important to integrate spintronics and CMOS-based technology at the system level as well as to minimize the usage of peripheral circuits for compactness.

Advances in Science and Technology to Meet Challenges

Research efforts have been put to address the challenges encountered in spintronic neuromorphic computing. The spintronic memristor is one of the promising technologies that has been developed to mimic analog, non-volatile and plastic storage elements like synapses that allow stored information to be modified. To date, spintronic memristors have been experimentally demonstrated in various ways such as the current-driven domain wall motion in MTJs [16], accumulation and dissipation of magnetic skyrmions [17], and electric field manipulation of the magnetization based on oxygen migrations [2].

Another spintronic technology is the S-MTJ. The intrinsic stochasticity allows an S-MTJ to perform extremely low-energy probabilistic computing such as stochastic number generation and probabilistic spin logic operation. Besides, the switching probability of S-MTJ is adjustable by an applied electric bias, and thus the junction can be utilized for the emulation of the Poisson neuron which generates a spiking train with a tunable firing rate [18].

STNOs are specific types of MTJ, which can excite spontaneous microwave oscillations when applying an injected direct current. The STNO is appealing for neuromorphic computing for its unique features. The finite magnetization relaxation of oscillation amplitudes can imitate the leaky integration of neurons. The nonlinear precession dynamics allows STNO to be modelled as nonlinear activation functions. Using time multiplexing, the single oscillator can function as a reservoir computer, which is a special type of neural network for time series analysis [8]. In addition, the high tunability of STNOs facilitates the coupling with other oscillator devices and could emulate the synchronization of neurons [19]. This is important for information sharing and processing. The classification of vowels at microwave frequencies has been experimentally demonstrated through the synchronization of STNOs [15].

Spin-diodes are MTJs exhibiting the reverse effect compared to STNOs: when they receive a microwave current at their input, a direct voltage is generated across the junction. The response can be used to mimic neurons in the non-linear regime [20], and synapses in the linear regime [21]. Frequency multiplexing appears as a possible solution to build multilayer neural networks with STNO neurons and spin diodes synapses [21].

Future research efforts should focus on the integration of spintronic devices (e.g. SOT devices) in the MTJ-based MRAM architecture to increase the read-out signal via TMR. The shared write channel-based SOT architecture can be explored to reduce the transistor count for large scaling [22]. Adding a gate to these devices to enable volatile or non-volatile voltage-controlled anisotropy will certainly be critical to enhance the computational capabilities of SOT-based architectures. To further enlarge the resistance change of MRAM and improve the scaling, researchers have put the effort into investigating new crossbar array architecture [9]. As shown in Fig. 3, instead of using the standard current sum in the analog multiply-accumulate (MAC) operation, this 64 × 64 MRAM array uses resistance summation for MAC operations. Much less power is required when this technique is used for image classification and face-detection tasks.

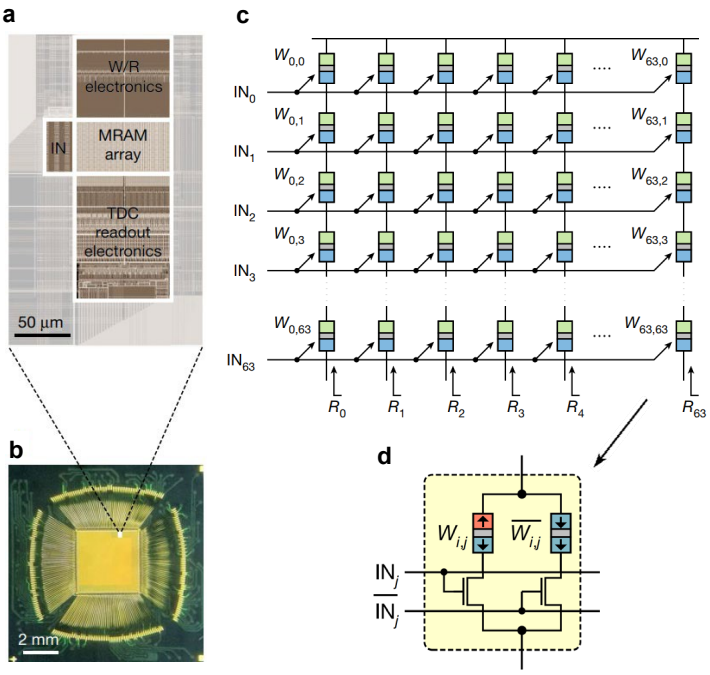

Figure 3. Layout (a) and micrograph (b) and of the 64 × 64 MRAM crossbar array. (c) MRAM crossbar array architecture and (d) configuration of each bit cell. Reprinted with permission from [9]. Copyright 2022, Springer Nature.

Concluding Remarks

In summary, spintronics offers compelling opportunities for the development of neuromorphic computing as it provides diverse bio-plausible hardware solutions. Various spintronic artificial synapses and neurons under different physical mechanisms (such as SOT, DWM, and magnetic skyrmions) have been demonstrated and can potentially be further implemented in large-scale brain-inspired spintronic systems. Nevertheless, several challenges remain to be addressed such as increasing the read-out signal, further investigation of bio-realistic neural devices, coupling control of neurons, and large scaling of compact and energy-efficient artificial neural networks. The state-of-the-art spintronic technologies have been discussed to meet these challenges including spintronic memristors, S-MTJ, STNO, spin-diodes, and new design of MTJ-based MRAM architecture. Spintronic neuromorphic computing is

currently a technologically fast evolving field. The experimental demonstration of spintronics-based network-level neuromorphic computing remains to be further explored and to be implemented into large-scale hardware neural networks.

Acknowledgements

The work is supported partially by the SpOT-LITE Programme (A*STAR grant, A18A6b0057) through RIE2020 funds, the National University of Singapore Advanced Research and Technology Innovation Centre (A-0005947-19-00), National Research Foundation (NRF) Singapore (NRF-000214-00), Samsung Electronics' University R&D Programme, the project PRIN 2020LWPKH7 funded by the Italian Ministry of University and Research and by the European Union's Horizon 2020 research and innovation program under grant RadioSpin No 101017098 and under grant SWAN-on-chip No. 101070287 HORIZON-CL4-2021-DIGITAL-EMERGING-01.

References

[1] J. Grollier, D. Querlioz, and M. D. Stiles, "Spintronic nanodevices for bioinspired computing," *Proceedings of the IEEE,* vol. 104, pp. 2024-2039, 2016.

[2] R. Mishra, D. Kumar, and H. Yang, "Oxygen-migration-based spintronic device emulating a biological synapse," *Physical Review Applied,* vol. 11, p. 054065, 2019.

[3] J. Zhou, T. Zhao, X. Shu, L. Liu, W. Lin, S. Chen*, et al.*, "Spin–Orbit Torque Induced Domain Nucleation for Neuromorphic Computing," *Advanced Materials,* vol. 33, p. 2103672, 2021.

[4] J. Grollier, D. Querlioz, K. Camsari, K. Everschor-Sitte, S. Fukami, and M. D. Stiles, "Neuromorphic spintronics," *Nature electronics,* vol. 3, pp. 360-370, 2020.

[5] P. Krzysteczko, J. Münchenberger, M. Schäfers, G. Reiss, and A. Thomas, "The Memristive Magnetic Tunnel Junction as a Nanoscopic Synapse-Neuron System," *Advanced Materials,* vol. 24, pp. 762-766, 2012.

[6] A. Sengupta and K. Roy, "Encoding neural and synaptic functionalities in electron spin: A pathway to efficient neuromorphic computing," *Applied Physics Reviews,* vol. 4, p. 041105, 2017.

[7] A. Sengupta, P. Panda, P. Wijesinghe, Y. Kim, and K. Roy, "Magnetic tunnel junction mimics stochastic cortical spiking neurons," *Scientific reports,* vol. 6, pp. 1-8, 2016.

[8] J. Torrejon, M. Riou, F. A. Araujo, S. Tsunegi, G. Khalsa, D. Querlioz*, et al.*, "Neuromorphic computing with nanoscale spintronic oscillators," *Nature,* vol. 547, pp. 428-431, 2017.

[9] S. Jung, H. Lee, S. Myung, H. Kim, S. K. Yoon, S.-W. Kwon*, et al.*, "A crossbar array of magnetoresistive memory devices for in-memory computing," *Nature,* vol. 601, pp. 211-216, 2022.

[10] W. H. Brigner, N. Hassan, L. Jiang-Wei, X. Hu, D. Saha, C. H. Bennett*, et al.*, "Shape-based magnetic domain wall drift for an artificial spintronic leaky integrate-and-fire neuron," *IEEE Transactions on Electron Devices,* vol. 66, pp. 4970-4975, 2019.

[11] X. Chen, W. Kang, D. Zhu, X. Zhang, N. Lei, Y. Zhang*, et al.*, "A compact skyrmionic leaky–integrate–fire spiking neuron device," *Nanoscale,* vol. 10, pp. 6139-6146, 2018.

[12] J. Liu, T. Xu, H. Feng, L. Zhao, J. Tang, L. Fang*, et al.*, "Compensated ferrimagnet based artificial synapse and neuron for ultrafast neuromorphic computing," *Advanced Functional Materials,* vol. 32, p. 2107870, 2022.

[13] Q. Yang, R. Mishra, Y. Cen, G. Shi, R. Sharma, X. Fong*, et al.*, "Spintronic Integrate-Fire-Reset Neuron with Stochasticity for Neuromorphic Computing," *Nano Letters,* 2022.

[14] J. J. Yang, D. B. Strukov, and D. R. Stewart, "Memristive devices for computing," *Nature Nanotechnology,* vol. 8, pp. 13-24, 2013.

[15] M. Romera, P. Talatchian, S. Tsunegi, F. Abreu Araujo, V. Cros, P. Bortolotti*, et al.*, "Vowel recognition with four coupled spin-torque nano-oscillators," *Nature,* vol. 563, pp. 230-234, 2018.

[16] S. Lequeux, J. Sampaio, V. Cros, K. Yakushiji, A. Fukushima, R. Matsumoto*, et al.*, "A magnetic synapse: multilevel spin-torque memristor with perpendicular anisotropy," *Scientific reports,* vol. 6, pp. 1-7, 2016.

[17] K. M. Song, J.-S. Jeong, B. Pan, X. Zhang, J. Xia, S. Cha*, et al.*, "Skyrmion-based artificial synapses for neuromorphic computing," *Nature Electronics,* vol. 3, pp. 148-155, 2020.

[18] N. Locatelli, A. Mizrahi, A. Accioly, R. Matsumoto, A. Fukushima, H. Kubota*, et al.*, "Noise-Enhanced Synchronization of Stochastic Magnetic Oscillators," *Physical Review Applied,* vol. 2, p. 034009, 2014.

[19] A. Slavin and V. Tiberkevich, "Nonlinear auto-oscillator theory of microwave generation by spin-polarized current," *IEEE Transactions on Magnetics,* vol. 45, pp. 1875-1918, 2009.

[20] L. Mazza, V. Puliafito, E. Raimondo, A. Giordano, Z. Zeng, M. Carpentieri*, et al.*, "Computing with injection-locked spintronic diodes," *Physical Review Applied,* vol. 17, p. 014045, 2022.

[21] N. Leroux, A. De Riz, D. Sanz-Hernández, D. Markovi ć A. Mizrahi, and J. Grollier, "Convolutional neural networks with radio-frequency spintronic nano-devices," *Neuromorphic Computing and Engineering,* vol. 2, p. 034002, 2022.

[22] R. Mishra, T. Kim, J. Park, and H. Yang, "Shared-Write-Channel-Based Device for High-Density Spin-Orbit-Torque Magnetic Random-Access Memory," *Physical Review Applied,* vol. 15, p. 024063, 2021.

2.1 – Neuromorphic computing with memristive devices

Peng Lin[1], Zhejiang University (penglin@zju.edu.cn)
Gang Pan[1], Zhejiang University (gpan@zju.edu.cn)
J. Joshua Yang[2], University of Southern California (jjoshuay@usc.edu)

[1]Address: College of Computer Science and Technology, Zhejiang University, Hangzhou, 310013, China

[2]Address: Electrical and Computer Engineering Department, University of Southern California, Los Angeles, CA 90089, USA

Status

Neuromorphic computing is a promising paradigm of artificial intelligence (AI) systems that aims at developing an efficient computing architecture with great physical and functional resemblance to the biological brains. Redox memristors, which represent a group of two terminal devices, can reversibly change their conductance as a result of ion migration inside the switching layer [1]. The ionic switching nature of memristors shares some similarities at the physics level with ion transport in nerve cells (e.g., Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}), which equips memristors with some desirable ion dynamics that can potentially enable a variety of neuronal and synaptic functions more efficiently for neuromorphic computing [2], [3].

Although not fully ready for large-scale standalone memory applications yet, memristors with decent array sizes have been used as static synapses for physical implementations of artificial neural networks (ANNs), which are low hanging fruits for memristors because neural network applications take advantages of the strengths and avoid the weakness of typical memristive devices as revealed in Ref [4]. This trend is reflected by a significant increase in application-oriented publications since 2018 (fig. 1). Each memristor is not only a weight storage unit, but can also directly apply weight function to upstream voltage inputs in the form of voltage-conductance multiplications [5], and thus co-locates the memory and processing functions within the same cell. A memristor array is essentially a physical neural network in between two neuron layers with many possible array arrangements, and can be extended to 3D layout to implement complex neural networks [6]. Owing to its low power, highly parallel computing paradigm, memristor based systems have found numerous successes in ANN related applications and demonstrated excellent computing efficiency exceeding 50 TOPS/W [7].

In the meantime, dynamical neuronal and synaptic properties of memristors are also under extensively study to provide more capable and compact building blocks for neuromorphic computing. In ANN applications, linear conductance modulation of memristors using a burst of identical pulses have been reported [8], which shows promises to implement accurate weight updates for fast on-chip network training of ANNs. Meanwhile, a lot of efforts have been put to develop functional memristive devices to implement spiking neural network (SNNs) – a potentially more efficient neural network model with a high bio-plausibility [9]. Important neuron models such as leaky-integrate-and-fire (LIF), Hodgkin-Huxley (HH) and plastic synaptic behaviors such as paired-pulse-facilitation (PPF) and spike-timing-dependent-plasticity (STDP) have been achieved by harvesting more dynamical behaviors of the memristors [2]. These novel functions were achieved in a very compact form normally with a couple of memristive devices instead of bulky circuits with many transistors in pure CMOS implementations. However, the overall sizes of these demonstrations were still limited to small-scale arrays, which is significantly hampering the overall capability of neuromorphic computing systems in practical applications.

Current and Future Challenges

Even though building a neuromorphic system requires a synergistic effort from both hardware and software, the device performance of memristors still plays a decisive role in determining the ultimate capability and functionality of the neuromorphic system. Currently, there are a few prominent challenges for memristive devices.

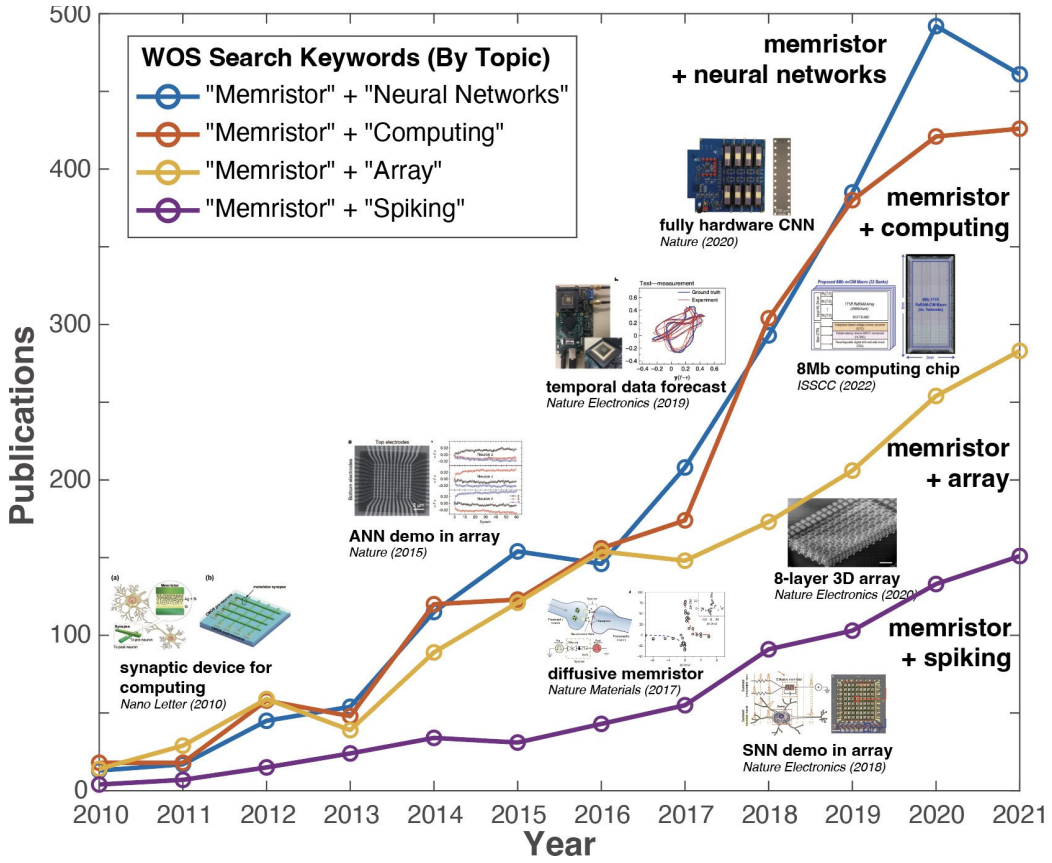

Figure 1. Number of publications related to memristor based neuromorphic computing system, retrieved from Web of Science (WOS) database, data source: Clarivate. Inset: synaptic device for computing [10]; ANN demo in array [11]; diffusive memristor [2]; temporal data forecast [12]; fully hardware CNN [13]; SNN demo in array; [14]; 8-layer 3D computing array [6]; 8Mb computing chip [[15].

First, it is still challenging to build a large-scale memristor arrays without an access transistor. At present, the majority of the memristor chips are based on the so-called one-transistor-one-memristor (1T1R) cell design, for which a transistor is integrated with a memristor and served as a current regulator for the memristor cell. The access transistor can (1) suppress the leakage current from the unselected cell, (2) use current compliance to achieve accurate conductance tuning and mitigate switching variations. However, the downside of having an access transistor is that it essentially limits the use of any array-wise parallel programming strategies, and raises challenges in designing an asynchronized system, such as those based on SNNs. The 2D scalability and 3D stack-ability of the memristor arrays are also limited by the addition of a transistor in each cell.

Secondly, linear analog conductance modulation using identical pulses is a key requirement for on-chip training of ANNs. Although it has been demonstrated in small prototype arrays, uniform linear conductance modulation across large-scale arrays is still challenging to achieve because device-to-device variations of switching voltages, conductance range, response time, etc. can all affect the conductance modulation process. In addition, it is also highly desirable to have symmetric programming for potentiation and depression, which, however, is not well supported by most of the SET and RESET memristive switching mechanisms.

Lastly, the variability issue in memristors is also a key limiting factor for the implementation of SNNs since dynamical functions such as STDP and LIF usually have lower tolerance for variations. For example, a standard STDP response of synapses is nonlinearly related to the timing differences

between the pre- and post-synaptic spikes. As a result, variations in the time domain (such as inaccurate firing delay from the LIF neuron) would be nonlinearly magnified in synaptic response, causing large errors during learning. Finding a solution to improve the scalability of these SNN functions, whether through more precise control of ionic motions, or through other compensation strategies, are highly desired for the development of SNN based neuromorphic system.

Of course, challenges are also existed in other aspects of a neuromorphic systems. Developing a global training algorithm for large scale SNNs is among the top of these challenges. Moreover, implementing neuromorphic systems still requires more efficient peripheral circuit designs, more sophisticated system architecture, and would require a better understanding of the working principles of SNNs.

Advances in Science and Technology to Meet Challenges

First, we would like to see continuous efforts in material and device engineering to better understand the mechanisms of existing devices and develop new device concepts with breakthroughs in device performance. This idea is supported by rich switching behaviors from different types of memristors, which are dictated by a combination of factors including material compositions, fabrication processes, device morphologies and many others, therefore provide a high degree of design freedom to tailor a device under specific application requirements. For example, it is known that the filamentary switching mechanism of memristor is a major source of variability. Alternative memristor design based on non-filamentary switching mechanisms may be developed to achieve improved switching uniformity. A new type of memristor with interfacial switching mechanism was reported, and demonstrated better switching uniformity and improved analog switching characteristics, though data retention may be a potential issue as normally observed in non-filamentary types of devices [16].

Meanwhile, we hope to see new fabrication technologies or material synthesis methods for memristors, such as using sophisticated tools from commercial foundries. It is known that the use of ion implantation for CMOS process provided dramatic improvement to the doping profile of MOSFET. We believe that a disruptive improvement may also be achieved in a similar effort. For example, in a preliminary study, epitaxial tool was used to grow single crystal SiGe film with nanometer wide dislocation channels, which acted as predefined ion channel for switching. Owing to better control over the ion transport, improvement in switching uniformity of memristors was achieved [17].

Finally, challenges at device and hardware level may also be overcome through complementary research effort in computer science and neuroscience domains. For example, more robust, hardware-friendly algorithms and computational models could be developed to mitigate the variability issues of memristors and utilize some unexpected properties discovered in new device exploration. Meanwhile, co-optimizations of the parameters for both memristive devices and neural networks could be achieved through comprehensive modeling and simulations. Lastly, our understanding of the brain is still at its infant state, it is also possible that new findings in neuroscience could help to establish new training methods and design new network architectures.

Concluding Remarks

Neuromorphic computing is a disruptive technological solution to future AI, and memristor is one of the leading candidates to implement parallel, analog and in-memory computing as well as rich dynamics inside neuromorphic computing systems. At the current stage, large-scale memristor based neuromorphic systems are mainly based on ANN algorithms, while SNN based demonstrations are far behind, primarily due to lack of appropriate training algorithms and the challenges to reliably obtain the desirable dynamical functions at large scale.

As more technical challenges described in this roadmap being resolved, it is expected to see a much more substantial progress made in neuromorphic computing. A large-scale memristor system

based on a comprehensive SNN design can lead to significant improvements in energy efficiency, performance, and functionality over existing AI hardware. Meanwhile, the SNN hardware could, in return, inspire the development of SNN algorithms or even the understanding of biological neural networks, which have been inefficient to simulate using conventional computers.

Acknowledgements

This work was partially supported by the National Science Foundation under contract No. 2023752, Natural Science Foundation of China (No. 61925603) and The Key Research and Development Program of Zhejiang Province in China (2020C03004).

References

[1] Y. Yang *et al.*, "Electrochemical dynamics of nanoscale metallic inclusions in dielectrics," *Nat. Commun.*, vol. 5, no. 1, p. 4232, 2014, doi: 10.1038/ncomms5232.

[2] Z. Wang *et al.*, "Memristors with diffusive dynamics as synaptic emulators for neuromorphic computing," *Nat. Mater.*, vol. 16, no. 1, pp. 101–108, 2017, doi: 10.1038/nmat4756.

[3] Z. Wang *et al.*, "Resistive switching materials for information processing," *Nat. Rev. Mater.*, vol. 5, no. 3, pp. 173–195, 2020, doi: 10.1038/s41578-019-0159-3.

[4] J. J. Yang, D. B. Strukov, and D. R. Stewart, "Memristive devices for computing," *Nature Nanotechnology*, vol. 8, no. 1. pp. 13–24, 2013, doi: 10.1038/nnano.2012.240.

[5] M. Hu *et al.*, "Memristor-based analog computation and neural network classification with a dot product engine," *Adv. Mater.*, vol. 30, no. 9, p. 1705914, 2018, doi: 10.1002/adma.201705914.

[6] P. Lin *et al.*, "Three-dimensional memristor circuits as complex neural networks," *Nat. Electron.*, vol. 3, no. 4, pp. 225–232, 2020, doi: 10.1038/s41928-020-0397-9.

[7] Q. Liu *et al.*, "A Fully Integrated Analog ReRAM Based 78.4TOPS/W Compute-In-Memory Chip with Fully Parallel MAC Computing," *Dig. Tech. Pap. - IEEE Int. Solid-State Circuits Conf.*, vol. 2020-Febru, pp. 500–502, 2020, doi: 10.1109/ISSCC19947.2020.9062953.

[8] H. Yeon *et al.*, "Alloying conducting channels for reliable analog computing," *Nat. Nanotechnol.*, 2020, doi: 10.1038/s41565-020-0694-5.

[9] W. Maass, "Networks of spiking neurons: The third generation of neural network models," *Neural Networks*, vol. 10, no. 9, pp. 1659–1671, 1997, doi: 10.1016/S0893-6080(97)00011-7.

[10] S. H. Jo, T. Chang, I. Ebong, B. B. Bhadviya, P. Mazumder, and W. Lu, "Nanoscale memristor device as synapse in neuromorphic systems," *Nano Lett.*, vol. 10, no. 4, pp. 1297–1301, 2010, doi: 10.1021/nl904092h.

[11] M. Prezioso, F. Merrikh-Bayat, B. D. Hoskins, G. C. 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, doi: 10.1038/nature14441.

[12] J. Moon *et al.*, "Temporal data classification and forecasting using a memristor-based reservoir computing system," *Nat. Electron.*, vol. 2, no. 10, pp. 480–487, 2019, doi: 10.1038/s41928-019-0313-3.

[13] P. Yao *et al.*, "Fully hardware-implemented memristor convolutional neural network," *Nature*, vol. 577, no. 7792, pp. 641–646, 2020, doi: 10.1038/s41586-020-1942-4.

[14] Z. Wang *et al.*, "Fully memristive neural networks for pattern classification with unsupervised learning," *Nat. Electron.*, vol. 1, no. 2, pp. 137–145, 2018, doi: 10.1038/s41928-018-0023-2.

[15] J.-M. Hung *et al.*, "An 8-Mb DC-Current-Free Binary-to-8b Precision ReRAM Nonvolatile Computing-in-Memory Macro using Time-Space-Readout with 1286.4-21.6TOPS/W for Edge-AI Devices," pp. 1–3, 2022, doi: 10.1109/isscc42614.2022.9731715.

[16] L. Gao *et al.*, "Fully parallel write/read in resistive synaptic array for accelerating on-chip learning," *Nanotechnology*, vol. 26, no. 45, 2015, doi: 10.1088/0957-4484/26/45/455204.

[17] S. Choi *et al.*, "SiGe epitaxial memory for neuromorphic computing with reproducible high performance based on engineered dislocations," *Nat. Mater.*, vol. 17, no. 4, pp. 335–340, 2018, doi: 10.1038/s41563-017-0001-5.

2.2 – MemComputing: an opportunity for nanotechnology

Massimiliano Di Ventra[1] and Yuriy V. Pershin[2]

[1] Department of Physics, University of California, San Diego, La Jolla, CA 92093, USA, E-mail: diventra@physics.ucsd.edu

[2] Department of Physics and Astronomy, University of South Carolina, Columbia, South Carolina 29208, USA, E-mail: pershin@physics.sc.edu

Status

Any useful computing technology must satisfy one important goal: to aid in the computation of problems that are particularly challenging for us, the users. It is with this goal in mind that MemComputing was first suggested [1].

MemComputing is a new computing paradigm in which *time non-locality* (memory) and massive parallelism play the main role in the processing of information [2]. Time non-locality is the ability of a physical system to remember its past dynamics. The machine can then exploit it to solve the necessary tasks. The concept is radically different from the way our traditional computers, based on the Turing paradigm of computation [3], operate and even how quantum computers manipulate information [4]. In addition, time non-locality is a feature shared by both quantum and non-quantum dynamical systems. This is not a minor point, since non-quantum dynamical systems offer substantial advantages for computing compared to quantum systems, in terms of both fabrication and simulation [5].

MemComputing machines have been mathematically defined in [6], where it was formally shown that they are Turing-complete, namely any problem solved by a Turing machine can be solved by a MemComputing one. In addition, it was shown that a MemComputing machine with specific features can solve NP-complete problems in polynomial time [7]. This result then begs the question: is this just a theoretical, albeit interesting outcome of their mathematical definition, or such a machine can be actually built in hardware? To answer the above question, a practical realization of *digital* MemComputing machines (DMMs) was proposed [7], which maps finite strings of symbols, such as 0 and 1, into a finite string of symbols, and relies on a new type of `self-organizing gates' (SOGs), namely gates which satisfy their logical truth table or algebraic relation irrespective of whether the signal is fed to the traditional input or output terminals (Figure 1).

Up to now, DMMs have been only simulated. Some of the results include the efficient solution of several optimization problems [8], acceleration of deep learning [9], and efficient solution of Boolean satisfiability problems [10], see Figure 2. These results were obtained by simply simulating the ordinary differential equations of DMMs [2]. In fact, for certain types of industrial problems such an `off-line' solution is sufficient [11]. These results then show that a physics-based approach to computation, like MemComputing, offers advantages not easily achievable by traditional algorithms.

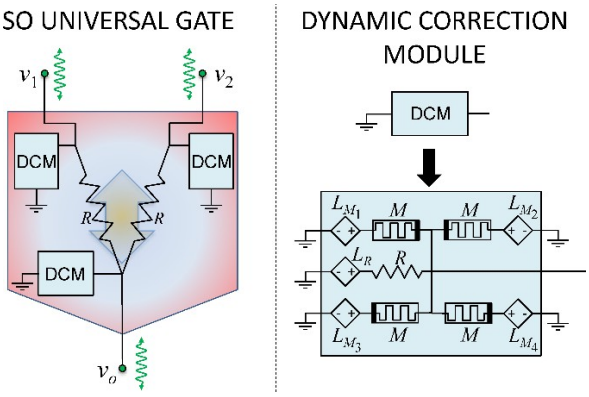

Figure 1. Left: Schematic of a self organizing gate: the gate attempts to satisfy its logical proposition irrespective of whether the signal comes from the traditional input or output. This is possible thanks to the dynamics of the internal state variables employed in DCM modules (right). From [7].

Current and Future Challenges

However, for certain types of problems that are prominent in, e.g., autonomous vehicles, robotics, cryptography, and so on, a `real-time' solution is necessary. In turn, this requires a hardware realization. It is then the practical, hardware realization of these SOGs of MemComputing machines which is emerging as an important research direction, and we believe, will be a major focus of future research as well.

For instance, in [7] resistive memories and active elements were suggested as a way to implement SOGs and circuits built out of them. Since resistive memories can be emulated using complementary metal-oxide semiconductor (CMOS) technology [12], a full CMOS implementation of these machines is doable. In fact, since the size of problems that are relevant to industry and academia can easily reach millions of variables and constraints, a CMOS-based implementation seems the most reasonable first step towards hardware. Such realizations could be based either on field-programmable gate arrays (FPGAs) or application-specific integrated circuits (ASICs). We expect that these technologies will deliver a real-time MemComputing solution for some relevant industrial problems (providing at least 10x-100x speedup).

However, CMOS may not be ideal for low-power applications. In view of this, in [13], nanomagnetic SOGs have been suggested. In particular, a NAND gate (which is functionally complete) has been proposed that employs two main properties. First, by appropriately tailoring stray-field interactions between magnetized nanomagnetic islands one can enforce the logic proposition of the gate with equal population of all correct states. Second, a local dynamic error suppression scheme can be applied to limit the time spent in excursions between logically correct states, as a result of thermal fluctuations.

Another interesting research direction is to use spintronic resistive memories to build SOGs and their circuits [14]. For instance, magnetic tunnel junctions, controlled by an electric current or a magnetic field, can be employed as a low-power realization of these gates. In addition, since antiferromagnets have been shown to support resistive memory features [15], one can envision their use in MemComputing for very fast (THz range) operations.

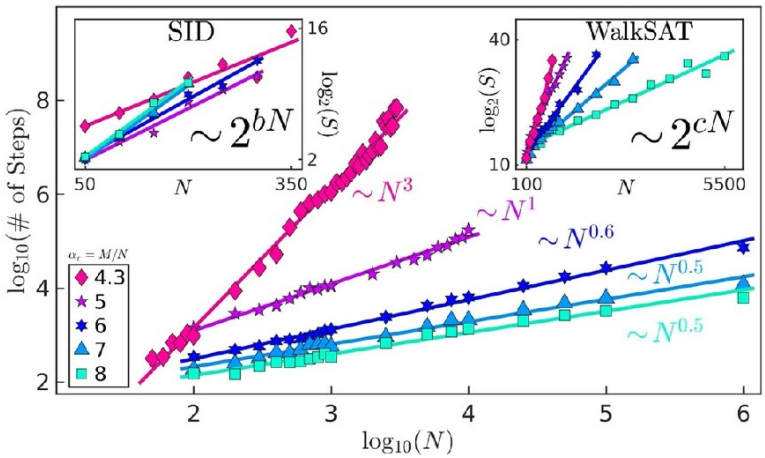

Figure 2. Polynomial scalability of time to solution of 3 SAT instances at fixed clause to variable ratio found by simulating DMMs. Insets: exponential scalability of classical algorithms (stochastic local search algorithm, WalkSAT, and a survey inspired decimation procedure, SID) on the same instances. From [10].

Advances in Science and Technology to Meet Challenges

The technological advances needed to realize the MemComputing paradigm in hardware are defined by various factors such as the selected technological platform, type of applications these machines will be used for, and the computing environment they will need to operate in. For instance, the CMOS and hybrid realizations require advancements in circuit theory that would allow the efficient implementation of the differential equations in terms of binary electronic circuits.

Moreover, the CMOS realizations of DMMs will require the development of a different theory: the theory of MemComputing maps. As the changes of states in digital electronic circuits are discrete (at times defined, e.g., by the clock cycle or cycles), the evolution of binary MemComputing circuits is a map [16]. In this case, particular care needs to be given to two important aspects of maps that may appear in the transition from continuous dynamical systems [2]: *i*) maps may introduce extra critical points in addition to the ones of the original dynamical system. These are called ghost critical points. *ii*) The basin of attraction of the equilibrium points may shrink for maps. This direction of study is then very important.

Regarding the type of applications of these machines, we need to stress that typical problem instances of interest in both academia and industry involve hundreds of thousands or even millions of variables and constraints. Such problems then would require a level of integration that is not easily achievable outside of CMOS technology. Therefore, emerging realizations of SOGs and circuits built out of them using nanotechnology components must also satisfy the high bar of being scalable and possibly compatible with CMOS.

Spintronics seems to have many features that would allow such a hybrid CMOS-based realization of MemComputing [14]. In fact, magnetic tunnel junctions, one of the basic ingredients of spintronics, can now be integrated with CMOS [17].

Concluding Remarks

We have started this article with an important point worth repeating. The usefulness of *any* technology should be first addressed in terms of its end goal. In the case of computing, the goal is to solve problems that are challenging for us, the users. If a computing machine does not accomplish such a goal, even if academically interesting, it is not practically useful [2]. In this respect, MemComputing has already

shown several advantages compared to our conventional computing model and other paradigms of current interest, such as quantum computing. These advantages reflect first in the possibility of emulating DMMs in software, thus allowing a direct comparison with traditional algorithms. Second, these machines do not rely on quantum phenomena, like entanglement, to work. Therefore, the path towards hardware is considerably less challenging than for quantum computers. In fact, DMMs can even be realized using our standard CMOS technology, providing an opportunity for very-large-scale integration. Finally, we expect MemComputing, or any other `unconventional computing' paradigm, to extend the reach, and enhance the functions of our modern computational fabric, not to replace it. In other words, we expect MemComputing machines to play the role of *co-processors* specialized to tackle particularly challenging problems.

Acknowledgements
M.D. acknowledges support from the National Science Foundation under Grant No. 2034558.

References

[1] M. Di Ventra and Y. V. Pershin, "The parallel approach," *Nature Physics*, vol. 9, p. 200, 2013.
[2] M. Di Ventra, *MemComputing: Fundamentals and Applications*. Oxford University Press, Oxford, UK, 2022.
[3] M. R. Garey, D. S. and Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman & Co., New York, 1990.
[4] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*. Cambridge University Press, Cambridge, UK, 2010.
[5] M. Di Ventra, "*MemComputing vs. Quantum Computing: some analogies and major differences"*, arXiv:2203.12031v2, 2022.
[6] F. L. Traversa and M. Di Ventra, "Universal MemComputing machines,*" IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, p. 2702, 2015.
[7] F. L. Traversa and M. Di Ventra, "Polynomial-time solution of prime factorization and NP-complete problems with digital MemComputing machines," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 27, p. 023107, 2017.
[8] F. L. Traversa, P. Cicotti, F. Sheldon, and M. Di Ventra, "Evidence of Exponential Speed-Up in the Solution of Hard Optimization Problems," *Complexity*, article ID 7982851, 2018.
[9] H. Manukian and M. Di Ventra, "Mode-assisted joint training of deep Boltzmann machines," *Scientific Reports*, vol. 11, pp. 1-8, 2021.
[10] S. R. B. Bearden, Y. R. Pei, and M. Di Ventra, "Efficient solution of Boolean satisfiability problems with digital memcomputing," *Scientific Reports*, vol. 10, p. 19741, 2020.
[11] MemComputing, Inc., https://www.memcpu.com/.
[12] Y. V. Pershin and M. Di Ventra, "Experimental demonstration of associative memory with memristive neural networks," *Neural Networks*, vol. 23, pp. 881-886, 2010
[13] P. Gypens, J. Leliaert, J., M. Di Ventra, B. Van Waeyenberge, and D. Pinna "Nanomagnetic self-organizing logic gates," *Physical Review Applied*, vol. 16, p. 024055, 2021.
[14] G. Finocchio, M. Di Ventra, K.Y. Camsari, K. Everschor-Sitte, P. Khalili-Amiri, and Z. Zeng, "The promise of spintronics for unconventional computing," *Journal of Magnetism and Magnetic Materials*, vol. 48, pp. 167506, 2020.
[15] T. Jungwirth, X. Marti, P. Wadley, J. Wunderlich, "Antiferromagnetic spintronics," *Nature Nanotechnology*, vol. 11, p. 231, 2016.
[16] O. Galor, *Discrete dynamical systems*. Springer Science & Business Media, 2007.
[17] A. Raychowdhury, ``MRAM and FinFETs team up,'' *Nat. Electron*. vol 1, p. 618, 2018.

3.1- Spin wave-based computing

Florin Ciubotaru[1], Andrii Chumak[2], Azad J. Naeemi[3], Sorin D. Cotofana[4]

[1]IMEC, Belgium (florin.ciubotaru@imec.be)
[2] University of Vienna, Faculty of Physics, Vienna, Austria (andrii.chumak@univie.ac.at)
[3]Georgia Institute of Technology, USA (an42@gatech.edu)
[4]Delft University of Technology, The Netherlands (S.D.Cotofana@tudelft.nl)

Status

Magnonics [1,2] is an emerging field of solid-state physics in which Spin Waves (SW) – the collective excitations of the magnetic orders – and their quanta magnons are utilized instead of electrons for information transport and processing [3]. SW characteristics, e.g., frequency range from GHz to THz, wavelengths down to the atomic scale, pronounced non-linear and non-reciprocal phenomena, tunability, low-energy data transport and processing, offer a variety of advantages towards building SW based nanotechnologies. As a result, the applied magnonics field is intensively growing. While spin-wave sensing [4] or spin-wave radio frequency applications [5], which are becoming increasingly topical in view of the requirements of 5G technology, are still in early stages of development, Boolean and unconventional SW computing have reached many milestones in recent years, and are experiencing constant growth [1-3]. Moreover, quantum magnonics [6] attracts increasing attention within the community [2,3] and potentially offers adding an additional entanglement-related degree of freedom to the quantum computing. Among the most important Boolean computing relevant achievements are the experimental realization of the inline majority gate [7], the directional coupler, and the magnetic half adder [8], as well as the demonstration of basic circuits by means of micromagnetic simulations [9] (see Fig. 1 (a)-(c)). Magnon-based unconventional computing [3] is primarily associated with neuromorphic computing [10-12] although versatile approaches of wave-based computing including spectrum analysis [13] or pattern recognition with magnonic holographic memory devices [14] can be placed in the same category (see Fig.1 (d)). Moreover, the first steps towards reservoir or stochastic magnon-based computing systems (e.g., using deeply nonlinear excitation in out-of-plane magnetized waveguides [15]) were recently reported.

The concept of inverse-design magnonics, which given a certain functionality makes use of a feedback-based computational algorithm to obtain the corresponding device design [16], has been successfully utilized for radio-frequency applications [16] and neural networks [12]. Such an approach results in a device with a rectangular ferromagnetic functional region patterned with square-shaped voids, as depicted in Fig. 1 (e). To demonstrate the universality of this approach, linear, nonlinear, and nonreciprocal magnonic functionalities were explored and a magnonic (de-)multiplexer (see Fig. 1 (e)), a nonlinear switch, and a circulator have been designed. Inverse design has significant potential for any kind of data processing, including the realization of complex multibit Boolean logic gates.

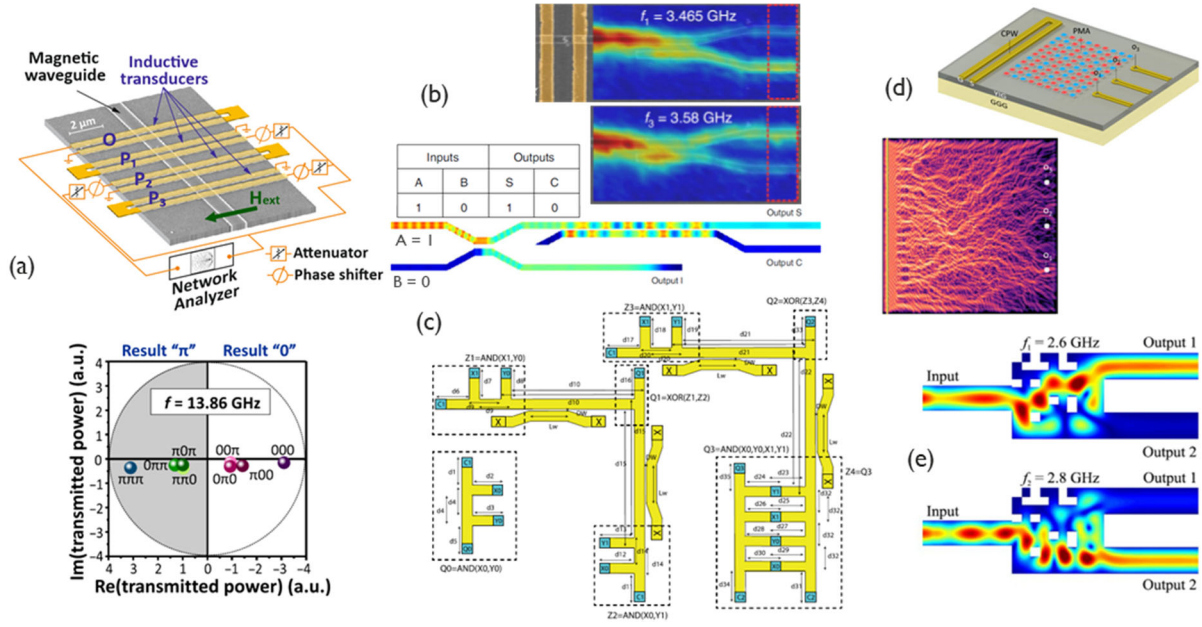

Figure 1.2 (a) Scanning electron micrograph of a sub-micron scaled spin wave majority gate (top) and the polar plot of the transmitted power for different input phases demonstrating strong and weak majority signals (bottom). Figure adapted from Ref. [7]. (b) 2D Brillouin light scattering spectroscopy maps of the spin wave intensity recorded in a directional coupler for two excitation frequencies (f = 3.465 GHz and f = 3.58 GHz) (top). Operational principle of a magnonic half adder demonstrated by micromagnetic simulations. Figure adapted from Ref. [8]. (c) 2 bit Inputs Spin Wave Multiplier. Reprint from Ref. [9]. (d Schematic of a nanomagnet based spin wave scatterer (top), and a spin wave intensity pattern (bottom) for neural network applications. Reprint from Ref. [12]. (e) Magnonic demultiplexer device simulated by inverse design micromagnetic simulations. Reprint from Ref. [16].

Current and Future Challenges

The experiments and simulations have clearly demonstrated that SW could serve as information carriers and their interaction for data processing. However, the design of a fully magnetic computing system is far away from being possible until effective solutions are found for the (i) combination of magnonic gates into circuits and (ii) realization of magnonic memories. Typical circuit construction challenges include fanout achievement and gate cascading, and they have been recently addressed in [9]. However, while the proposed approaches [9] enable magnonic circuit realization by enabling up to 4 gate fanout and direct gate cascading within the magnonic domain, they are expensive in terms of area and delay. Thus, more effective solutions are required in order to unleash magnonic computing full potential. Unfortunately, with the exception of the Holographic Memory concept introduced in [14], which is not a real magnonic memory as it doesn't store data in magnons, very little progress has been made towards the conceptual realization of magnonic memories.

Moreover, building a magnonic computing chip would require many more components apart of simple magnetic waveguides used for spin wave propagation. To pass the input information from CMOS to the magnonic circuitry for computing would require high energy efficient scaled transducers at the interface. The typical spin-wave transducers, e.g., inductive antennas, spin-transfer- or spin-orbit-torque based magnetic tunnel junctions, are scalable but very inefficient in terms of energy consumption. Their energy-delay characteristics need to be improved by few orders in magnitude to compete with the CMOS computing circuitry counterpart. Voltage driven transducers, e.g., based on magnetoelectric effects, might reach the required energy efficiency. Yet, such performances have to be demonstrated at nano-scale. Next to the transducers, the magnetic conduits should transport the information with minimum losses and delay. Recent studies demonstrated spin-wave propagation lengths in scaled waveguides in a several micrometers range. However, not all the studied materials (e.g., Y3Fe5O12 –

YIG) are CMOS compatible. Alloys based on CoFe(B) are widely used in the MRAM technology and are promising candidates for magnonic conduits. Nevertheless, a further material development and optimization, especially for the voltage driven transducers, will be required. In addition, spin-wave amplifiers might be needed to restore the amplitude losses during the propagation if the circuit length exceeds the wave mean free path. These amplifiers also need to operate at very low energies (towards aJ), which suggest that the amplification process should also rely on voltage driven mechanisms, e.g., using Voltage Control of Magnetic Anisotropy (VCMA) or other magnetoelectric effects.
Magnonic circuit layout design is much more challenging than the one of a charge-based counterpart. SW propagation and interaction are quite sensitive to waveguide dimensions and geometries, e.g., spin wave behaviour in straight waveguides and around corners are rather, and, as such, layout can significantly influence circuit performance and even make it malfunctions. Moreover, due to SW amplitude decay and dephasing phenomena the circuit size (real-estate) should be minimized by enabling 2D signal crossing and/or 3D interconnect.

Advances in Science and Technology to Meet Challenges

Much progress has been achieved in material development, realization and characterization of nano-scaled spin-wave conduits below 100, as well as in the understanding the underlaying physical mechanism of spin-wave generation and propagation both in the linear and non-linear regimes. Most of the studies focus on spin-wave properties (2D or 3D systems), and rely on optical characterization and/or micromagnetic simulations. However, the main challenge for building efficient magnonic computing circuits is related to physical realization of an efficient energy coupling interface for the information transfer between electric and magnetic domains. The research progress on heterogeneous integration of multiferroic materials, magnetoelectric composite [17] and VCMA [18] stacks for the generation and detection of SW could allow for the demonstration of nanoscaled cascaded magnonic logic gates in a full electric experiment. Furthermore, the coupling of phonons or photons to SW could bring additional functionalities and enhance or control their characteristics, e.g., the group velocity or amplitude.
State of the art SW-based computing assumes phase encoding of information and is performed via not input-output format coherent majority gates. The direct cascading of such gates results in input data dependent circuit malfunctions. Despite of the fact that gate cascading solutions have been proposed they are rather expensive and induce large gate delay overhead, e.g., the gate delay is increased from few ns to more than 20 ns [9]. As such, to build competitive spin wave circuits more effective cascading schemes are required and potential solutions might be found by means of inverse-design [12,16] and/or by investigating alternative information encodings that may result in directly cascadable gates, and potentially enable the realization of more computation within one single gate. Given that information can be encoded in spin wave phase, amplitude, frequency, and any combination of those, a plethora of alternative and more effective spin wave computation paradigms could be potentially developed. Moving from charge-based circuits, e.g., CMOS, to spin-wave circuits requires changes into the computer aided design framework. The traditional Boolean algebra-based logic synthesis needs to be able to accommodate a new universal gate set composed of majority gate and inverter. Given that magnonic circuits are expected to be hybrid (have to be interfaced with the environment within which they operate and may include CMOS parts) a not yet existing mixed micomagnetic-SPICE simulation paradigm needs to be proposed and developed. The spin-wave circuit layout design is governed by completely different principles and it has fundamental implications on the circuit performance and behaviour. To this end a novel approach to produce correct by construction spin-wave circuit layouts is essential for the proliferation of the spin-wave computing paradigm.

Concluding Remarks

Spin waves demonstrated to possess a high potential for computing but also for other emerging sensing or radio-frequency applications. The deep understanding of the associated physical phenomena and the development of materials and device fabrication techniques will allow the transition from the fundamental research to engineering devices in a near future. It is clear that for computing applications the spin-wave devices have to be integrated in hybrid magonic – CMOS architectures. In this quest, several important components are still to be developed**,** as explained in this roadmap. Independent on the computing paradigm, an efficient information transfer interface between CMOS and magnonic circuits will be one of the enabling factors towards real technologies. Development of multi-physics and SPICE design tools for device simulation as well as for circuit layouts will further pave the way for applications. Last but not least, novel computing architectures as (spiking) neuronal networks based on interference, non-linear effects or non-reciprocity of spin waves could be developed for special applications.

Acknowledgements

The work of F. Ciubotaru and S.D. Cotofana was supported by the European Union's Horizon 2020 Research and Innovation Programme through the FET-OPEN project CHIRON under Grant Agreement 801055. F. Ciubotaru acknowledge the support from imec's industrial affiliate program on Exploratory logic. A. Chumak acknowledge the financial support by the European Research Council (ERC) Starting Proof of Concept Grant 101082020 5G-Spin.

References

[1] A.N. Mahmoud, F. Ciubotaru, F. Vanderveken, A.V. Chumak, S. Hamdioui, C. Adelmann, and S.D. Cotofana, "An introduction to spin wave computing", *J. Appl. Phys.,* 128, 161101, 2020, doi.org/10.1063/5.0019328.

[2] A. Barman, G. Gubbiotti, S. Ladak, A.O. Adeyeye , M. Krawczyk , J. Gräfe, C. Adelmann, S.D. Cotofana, A. Naeemi, V.I. Vasyuchka, B. Hillebrands, S.A. Nikitov, H. Yu, D. Grundler, A.V. Sadovnikov, A.A. Grachev, S.E. Sheshukova, J.-Y. Duquesne, M. Marangolo, G. Csaba, W. Porod, V.E. Demidov, S. Urazhdin, S.O. Demokritov, E. Albisetti, D. Petti, R. Bertacco, H. Schultheiss, V.V. Kruglyak, V.D. Poimanov, S. Sahoo, J. Sinha, H. Yang, M. Münzenberg, T. Moriyama, S. Mizukami, P. Landeros, R.A. Gallardo, G. Carlotti, J.-V. Kim, R.L. Stamps, R.E. Camley, B. Rana,Y. Otani, W. Yu, T. Yu, G.E.W. Bauer, C. Back, G.S. Uhrig, O.V. Dobrovolskiy, B. Budinska, H. Qin, S. van Dijken, A.V. Chumak, A. Khitun, D.E. Nikonov, I.A. Young, B.W. Zingsem, and M. Winklhofer, "The 2021 Magnonics Roadmap", *J. Phys.: Condens. Matter*, 33, 413001, 2021, doi.org/10.1088/1361-648X/abec1a.

[3] A.V. Chumak, P. Kabos, M. Wu, C. Albert, C. Adelman, A. Adeyeye, J. Åkerman, F.G. Aliev, A. Anane, A. Awad, C.H. Back, A. Barman, G.E.W. Bauer, M. Becherer, E. N. Beginin, V.A.S.V. Bittencourt, Y.M. Blanter, P. Bortolotti, I. Boventer, D.A. Bozhko, S.A. Bunyaev, J.J. Carmiggelt, R. R. Cheenikundil, F. Ciubotaru, S.D. Cotofana, G. Csaba, O. V. Dobrovolskiy, C. Dubs, M. Elyasi, K.G. Fripp, H. Fulara, I. A. Golovchanskiy, C. Gonzalez-Ballestero, P. Graczyk, D. Grundler, P. Gruszecki, G. Gubbiotti, K. Guslienko, A. Haldar, S. Hamdioui, R. Hertel, B. Hillebrands, T. Hioki, A. Houshang, C.-M. Hu, H. Huebl, M. Huth, E. Iacocca, M. B. Jungfleisch, G.N. Kakazei, A. Khitun, R. Khymyn, T. Kikkawa, M. Kläui, O. Klein, J. W. K ȸs, S. Knauer, S. Koraltan, M. Kostylev, M. Krawczyk, I. N. Krivorotov, V. V. Kruglyak, D. Lachance-Quirion, S. Ladak, R.Lebrun, Y. Li, M. Lindner, R. Macêdo, S. Mayr, G.A. Melkov, S. Mieszczak, Y. Nakamura, H.T. Nembach, A.A. Nikitin, S.A. Nikitov, V. Novosad, J.A. Otalora, Y. Otani, A. Papp, B. Pigeau, P. Pirro, W. Porod, F. Porrati, H. Qin, B. Rana, T. Reimann, F. Riente, O. Romero-

Isart, A. Ross, A. V. Sadovnikov, A.R. Safin, E. Saitoh, G. Schmidt, H. Schultheiss, K. Schultheiss, A.A. Serga, S. Sharma, J.M. Shaw, D. Suess, O. Surzhenko , K. Szulc, T. Taniguchi, M. Urbánek, K. Usami, A.B. Ustinov, T. van der Sar, S. van Dijken, V.I. Vasyuchka, R. Verba, S. Viola Kusminskiy, Q. Wang, M. Weides, M. Weiler, S. Wintz, S.P. Wolski, X. Zhang, and H. Qin, "Advances in Magnetics Roadmap on Spin-Wave Computing", *IEEE Trans. Magn.*, 58, 0800172, 2022, doi: 10.1109/TMAG.2022.3149664.

[4] O. Haas, B. Dufay, S. Saez, and C. Dolabdjian, "Sensitivity and Noise of a Magnetic Field Sensor Based on Magnetostatic Spin Wave YIG Device and Its Integrated Electronics", *IEEE Sensors Journal*, vol. 20, no. 23, pp. 14148-14156, 2020, doi: 10.1109/JSEN.2020.3008555.

[5] F. Heussner, G. Talmelli, M. Geilen, B. Heinz, T. Brächer, T. Meyer, F. Ciubotaru, C. Adelmann, K. Yamamoto, A.A. Serga, B. Hillebrands, and P. Pirro, "Experimental realization of a passive gigahertz frequency-division demultiplexer for magnonic logic networks", *Phys. Status Solidi Rapid Res. Lett.*, 14, 1900695, 2020, doi.org/10.1002/pssr.201900695.

[6] D. Lachance-Quirion, Y. Tabuchi, A. Gloppe, K. Usami, and Y. Nakamura, "Hybrid quantum systems based on magnonics", *Appl. Phys. Express*, 12, 070101, 2019, doi.org/10.7567/1882-0786/ab248d.

[7] G. Talmelli, T. Devolder, N. Träger, J. Förster, S. Wintz, M. Weigand, H. Stoll, M. Heyns, G. Schütz, I.P. Radu, J. Gräfe, F. Ciubotaru, and C. Adelmann, "Reconfigurable submicrometer spin-wave majority gate with electrical transducers", Sci. Adv., 6: eabb4042, 2020, DOI: 10.1126/sciadv.abb4042.

[8] Q. Wang, M. Kewenig, M. Schneider, R. Verba, F. Kohl, B. Heinz, M. Geilen, M. Mohseni, B. Lägel, F. Ciubotaru, C. Adelmann, C. Dubs, S.D. Cotofana, O.V. Dobrovolskiy, T. Brächer, P. Pirro, and A.V. Chumak, "A magnonic directional coupler for integrated magnonic half-adders", *Nat. Electronics*, 3, 765-774, 2020, doi.org/10.1038/s41928-020-00485-6.

[9] A.N. Mahmoud, F. Vanderveken, C. Adelmann, F. Ciubotaru, S.D. Cotofana, and S. Hamdioui, "Spin wave normalizationtoward all magnonic circuits", IEEE Trans. Circuits Syst I Regul Pap., 68, pp. 536-549, 2021, doi: 10.1109/TCSI.2020.3028050.

[10] T. Brächer and P. Pirro, "An analog magnon adder for all-magnonic neurons", *J. Appl. Phys.*, 124, 152119, 2018, doi.org/10.1063/1.5042417.

[11] Q. Wang, A. Hamadeh, R. Verba, V. Lomakin, M. Mohseni, B. Hillebrands, A.V. Chumak, and P. Pirro, "A nonlinear magnonic nano-ring resonator", *Npj Comput. Mater.*, 6, 192, 2020, doi.org/10.1038/s41524-020-00465-6.

[12] Á. Papp, W. Porod, and G. Csaba, "Nanoscale neural network using non-linear spin-wave interference" *Nat. Commun.*, 12, 6422, 2021, doi.org/10.1038/s41467-021-26711-z.

[13] Á. Papp, W. Porod, Á.I. Csurgay, and G. Csaba, "Nanoscale spectrum analyzer based on spin-wave interference", *Sci. Rep.*, 7, 9245, 2017, doi.org/10.1038/s41598-017-09485-7.

[14] A. Kozhevnikov, F. Gertz, G. Dudko, Y. Filimonov, and A. Khitun, "Pattern recognition with magnonic holographic memory device", *Appl. Phys. Lett.*, 106, 142409, 2015, doi.org/10.1063/1.4917507.

[15] Q. Wang, R. Verba, B. Heinz, M. Schneider, O. Wojewoda, K. Davídková, K. Levchenko, C. Dubs, N. J. Mauser, M. Urbánek, P. Pirro, and A. V. Chumak, "Deeply nonlinear excitation of self-normalised exchange spin waves", arXiv:2207.01121, 2022.

[16] Q. Wang, A.V. Chumak, and P. Pirro, "Inverse-design magnonic devices", *Nat. Commun.*, 12, 2636, 2021, doi.org/10.1038/s41467-021-22897-4.

[17] X. Liang, H. Chen, and N.X. Sun, "Magnetoelectric materials and devices", *APL Mater.*, 9, 041114, 2021, doi.org/10.1063/5.0044532.

[18] K. Miura, S. Yabuuchi, M. Yamada, M. Ichimura, B. Rana, S. Ogawa, H. Takahashi, Y. Fukuma, and Y. Otani, “Voltage-induced magnetization dynamics in CoFeB/MgO/CoFeB magnetic tunnel junctions”, *Sci. Rep*., 7, 42511, 2017, doi.org/10.1038/srep42511.

3.2 Computing with skyrmions

Riccardo Tomasello, Department of Electrical and Information Engineering, Politecnico di Bari, 70125 Bari, Italy (riccardo.tomasello@poliba.it)
Christos Panagopoulos, Division of Physics and Applied Physics, School of Physical and Mathematical Sciences, Nanyang Technological University, S637371, Singapore (christos@ntu.edu.sg)
Mario Carpentieri, Department of Electrical and Information Engineering, Politecnico di Bari, 70125 Bari, Italy (mario.carpentieri@poliba.it)

Status

Magnetic skyrmions are localized whirling spin textures with topological properties and particle-like characteristics[1]–[3]. Research has exploded in the last decade with proposals for new materials and device concepts[4], [5], offering intriguing functionalities ranging from memory to computing applications. Originally observed at low temperatures in B20 compounds, today, skyrmions can be found in many thin film materials (ferro-, ferri-, and synthetic antiferro-magnets (SAFs)) at room temperature[3]. Their small size and the possibility to easily manipulate them electrically make these topologically protected chiral spin configurations attractive as information carriers in compact, and energetically efficient devices. Furthermore, the anticipated insensitivity to defects and potential for low energy consumption[1], [2] have accelerated efforts to understand their formation, stability, and dynamics sufficiently well, already extending the interest towards unconventional applications. In particular, the topological properties of magnetic skyrmions offer new paradigms in reservoir, stochastic, neuromorphic, and quantum operations[6], [7].

Reservoir computing is a concept of self-organized recursive neural networks to emulate the highly non-linear and recursively interconnected architectures of the brain[8]. Single skyrmions[9] or a skyrmion fabric[10], [11] are promising because the non-linear dynamics of magnetic skyrmions can increase the systems nonlinearity and therefore the efficiency of the reservoir. Stochastic computing, on the other hand, is based on the concept of p-value, i.e. the probability of a certain sequence of bits known as bitstream. Here, the essence for optimal operation is the decorrelation of the bitstreams[12]. A reshuffle chamber, a missing element in current implementations of stochastic computing, has been demonstrated taking advantage of the diffusion properties of skyrmions [13] (Fig. 1a).

Neuromorphic computing aims at mimicking the functionalities of the human brain, which performs very complex operations using only tens of Watts[14]. The brain neural network is composed of many neurons densely connected through a plethora of synapses. Neurons act as computing units with a non-linear activation function, while synapses act as memory elements with memristive behavior. While skyrmion-based neurons are still only theorized[6], a skyrmion-based synapsis has already been experimentally demonstrated[15] (Fig. 1b), with the weight represented by the Hall resistivity.

Magnetic skyrmions could also serve as a source for non-Abelian statistics. Imprinting skyrmions on superconductors may trigger the formation of special Majorana quasiparticles, granting unrivalled resilience to the decoherence problem that plagues other quantum computing platforms[16]. Nano-skyrmions are also of interest for their potential as a logical element of a quantum processor[17]. They develop quantized eigenstates with distinct helicities and out-of-plane magnetizations. In a skyrmion qubit, information is stored in the quantum degree of helicity, and the logical states can be adjusted by electric and magnetic fields, offering an operation rich regime with high anharmonicity.

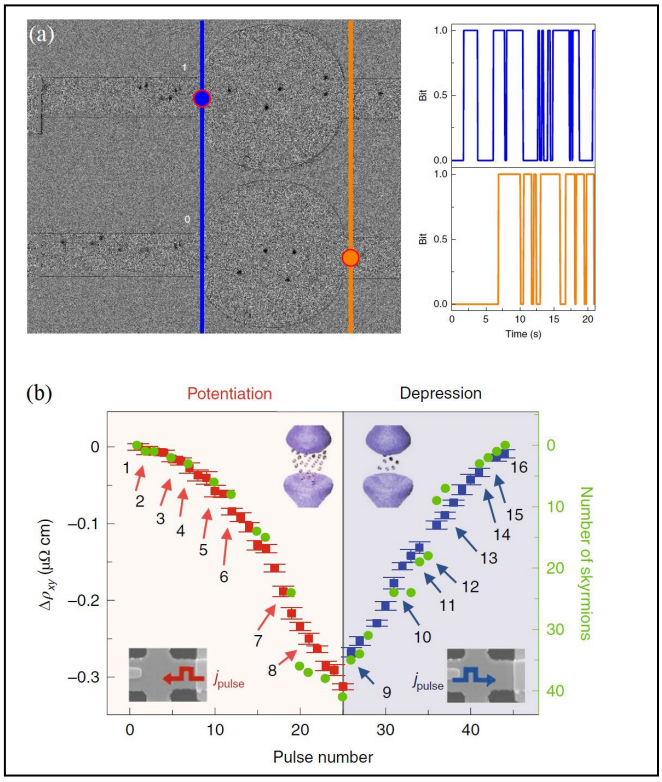

Figure 1.2 (a) example of a skyrmion reshuffler. Reproduced with permission from Ref. [13]. (b) memristive behavior of skyrmions for synapsis applications. Reproduced with permission from Ref. [15].

Current and Future Challenges

Fundamental challenges for utilizing magnetic skyrmions in technology include controlling their size, sustaining positional stability, enhancing electrical readout, deterministic nucleation and annihilation, reducing/suppressing the skyrmion Hall angle while maintaining high velocity, but also achieving an overall integration with digital circuits and associated circuit overhead.

Skyrmion position can be controlled by engineering pinning sites with lithography or ion irradiation[18]. Improvements in electrical readout calls for the use of optimized large Tunneling Magnetoresistance Magnetic Tunnel Junctions (TMR MTJs)[19]. Nucleation/annihilation protocols enable deterministic functionalities[20], but should be optimized to decrease energy consumption. The skyrmion Hall angle can be reduced in compensated ferrimagnets[21], and suppressed in SAFs[22]. However, the skyrmion velocity[23] is still far from theoretical predictions, calling for new and/or optimized materials and device architectures.

On the side of unconventional applications, Reservoir computing based on magnetic textures has not been realized experimentally yet. In stochastic computing, the first proof of concept of an algebraic computation based on skyrmions is still missing and hence a full skyrmion based implementation. Furthermore, energy efficiency, velocity and accuracy in computation should be evaluated and compared with standard CMOS systems. In neuromorphic computing, some proposals rely on the use of a skyrmion based spin transfer torque (STT) oscillator which has yet to be demonstrated. Whereas proposals involving memristive skyrmion behavior, such as skyrmion synapsis, should be based on large TMR MTJs for improved detection and compact solutions. Specifically for the results in [15], the accuracy of the network could be improved, compared to the current estimate of 89%.

Beyond non-interventional creation or observation studies, skyrmions promise to dramatically improve quantum operations. Skyrmion-vortex interaction in device architectures of imprinted magnetic skyrmions on conventional superconductors can assist topological quantum computing by operations carried out on Anyons. A novel quantum hybrid architecture composed of Néel skyrmions and Niobium grants realistic hope[24]. Moving forward, in a homogeneous chiral magnet and superconductor stack, a skyrmion-vortex pair – and hence a Majorana zero mode – would be intrinsically mobile. This allows for non-perturbative, non-contact braiding operations by moving skyrmion-vortex pairs around the surface.

Quantum skyrmions in frustrated magnets offer a new element for the construction of qubits based on the energy-level quantization of the helicity degree of freedom. The skyrmion state, energy-level spectra, transition frequency, and qubit lifetime are configurable and can be engineered by adjusting external electric and magnetic fields, offering a rich operation regime with high anharmonicity.

Advances in Science and Technology to Meet Challenges

Over the next two decades, a concerted experimental effort will promote skyrmions to future devices, hence, realizing their technological potential for information processing transcending existing limits. Challenges, such as decreasing energy consumption of skyrmion devices, controlling the skyrmion size, suppressing the skyrmion Hall angle, could be achieved by exploring new materials architectures. We can reduce dimensionality with emphasis on 2D van der Waals materials. At the same time, we can explore 3D bulk materials that, thanks to improved imaging techniques, demonstrate the presence of static 3D structures, such as vortex rings and hopfions[25]. Whereas, on the dynamical side, so far the community has been relying on theoretical predictions of the current-driven Hopfion motion.

Emphasis on frustrated magnetism, intrinsic to triangular or hexagonal lattices with antiferromagnetic spin correlations, can utilize the induced non-collinear magnetic order, which itself breaks spatial inversion symmetry for the formation and manipulation of skyrmions only a few nanometers wide. We can also focus on the design of hybrid systems by combining ferro-, ferri-, and antiferro-magnets. Specifically for unconventional skyrmion applications, experimental realization of reservoir computing needs efficient and precise electrical measurements among multiple contacts for a reliable resistance evaluation linked to the magnetic texture distribution. Stochastic computing demands an integration with CMOS technology via stacks compatible with STT-MRAM technology already integrated and commercialized. Neuromorphic computing calls for the enhancement of skyrmion detection (Fig. 2a) beyond the 30% TMR threshold for MTJs. This could be achieved by combining state-of-the-art CoFeB-MgO MTJ with conventional skyrmion-hosting magnetic multilayers. Another direction could be combinatorial, including topological magnetism and acoustic waves, already promising for skyrmion-based synaptic behavior [26], [27].

For quantum computing operations in skyrmion-superconductor hybrids, major tasks need to be performed for device capability. Firstly, ensuring that the magnetic interactions can spin-polarize the superconductor and cause a topological phase transition. Secondly, for Majorana braiding, the magnetic homogeneity of the architecture needs to be better than commonly achieved using magnetron-sputtering. In qubit-technology hardware, the applicability of nano-skyrmions can be further improved with the development of cleaner magnetic samples and interfaces in engineered architectures, without trading off qubit anharmonicity and scalability. Notably, demonstrating tunable macroscopic quantum tunneling, coherence, and oscillations for magnetic nano-skyrmions will also establish helicity in topologically protected chiral spin configurations as a quantum variable; much like macroscopic quantum tunneling and energy quantization in Josephson junctions, thus the fundamental physical step for developing a practical skyrmion qubit.

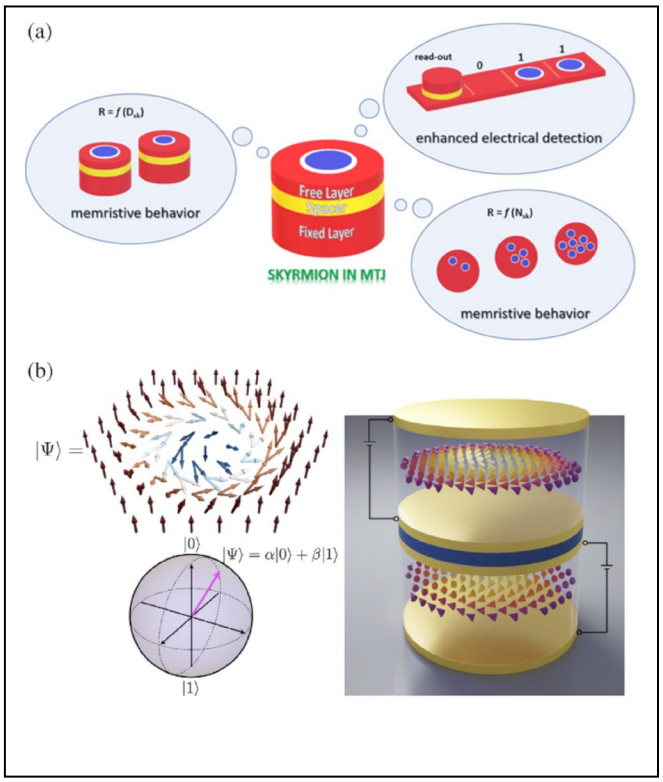

Figure 2. (a) Sketch of the improvements of skyrmion detectivity through the use of a skyrmion based MTJ. R is the electrical resistance of the MTJ, D_{sk} is the skyrmion diameter, and N_{sk} is the number of skyrmions. (b) Skyrmion qubit concept. Reproduced with permission from Ref. [17].

Concluding Remarks

Magnetic skyrmions are fascinating topological magnetization configurations with realistic potential for a beneficial impact on computing paradigms. Their small size, particle like behavior, topological properties, manipulability by electrical current, as well as memristive features promise disruptive advances in reservoir, stochastic, neuromorphic, and quantum computing. Stabilizing skyrmions in large TMR MTJ will enhance the skyrmion detectivity with unprecedented effects on unconventional operations. Extending the investigation to 3D bulk magnetic systems and 2D materials will lead to major breakthroughs and new functionalities associated with complex topology and multiple degrees of freedom. Remarkable advancements in unconventional computing can be driven by the combination of different physical systems, such as topological magnetism and acoustic waves which already promise to control skyrmion based synaptic behavior. Skyrmions interacting with superconductors can lead to chiral superconductivity and Majorana braiding platforms. Whereas, nano skyrmions stabilized in magnetic disks bounded by electrical contacts will allow static fields to control the quantized energy spectra, enabling changes in the helicity between energetically favored levels. It is expected that skyrmions will be a major building block for the next generation of low power computing architectures, transcending from the classical to the quantum regime.

Acknowledgements

This work was supported by the Project No. PRIN 2020LWPKH7 funded by the Italian Ministry of University and Research. C. Panagopoulos acknowledges support from the National Research Foundation (NRF) Singapore Competitive Research Programme NRF CRP21 2018 0001, and the Singapore Ministry of Education (MOE) Academic Research Fund Tier 3 Grant MOE2018 T3 1 002.

References

[1] G. Finocchio, F. Büttner, R. Tomasello, M. Carpentieri, and M. Kläui, "Magnetic skyrmions: from fundamental to applications," *J. Phys. D. Appl. Phys.*, vol. 49, no. 42, p. 423001, 2016, doi: 10.1088/0022 3727/49/42/423001.

[2] A. Soumyanarayanan, N. Reyren, A. Fert, and C. Panagopoulos, "Emergent phenomena induced by spin–orbit coupling at surfaces and interfaces," *Nature*, vol. 539, no. 7630, pp. 509–517, 2016, doi: 10.1038/nature19820.

[3] G. Bonanno *et al.*, "Contributors," in *Magnetic Skyrmions and Their Applications*, G. Finocchio and C. Panagopoulos, Eds. Elsevier, 2021, pp. ix–xi.

[4] X. Zhang *et al.*, "Skyrmion electronics: writing, deleting, reading and processing magnetic skyrmions toward spintronic applications," *J. Phys. Condens. Matter*, vol. 32, no. 14, p. 143001, Apr. 2020, doi: 10.1088/1361 648X/ab5488.

[5] C. H. Marrows and K. Zeissler, "Perspective on skyrmion spintronics," *Appl. Phys. Lett.*, vol. 119, no. 25, p. 250502, Dec. 2021, doi: 10.1063/5.0072735.

[6] S. Li *et al.*, "Magnetic skyrmions for unconventional computing," *Mater. Horizons*, vol. 8, no. 3, pp. 854–868, 2021, doi: 10.1039/D0MH01603A.

[7] H. Vakili *et al.*, "Skyrmionics—Computing and memory technologies based on topological excitations in magnets," *J. Appl. Phys.*, vol. 130, no. 7, p. 070908, Aug. 2021, doi: 10.1063/5.0046950.

[8] G. Tanaka *et al.*, "Recent advances in physical reservoir computing: A review," *Neural Networks*, vol. 115, pp. 100–123, Jul. 2019, doi: 10.1016/j.neunet.2019.03.005.

[9] D. Prychynenko *et al.*, "Magnetic Skyrmion as a Nonlinear Resistive Element: A Potential Building Block for Reservoir Computing," *Phys. Rev. Appl.*, vol. 9, no. 1, p. 014034, Jan. 2018, doi: 10.1103/PhysRevApplied.9.014034.

[10] G. Bourianoff, D. Pinna, M. Sitte, and K. Everschor Sitte, "Potential implementation of reservoir computing models based on magnetic skyrmions," *AIP Adv.*, vol. 8, no. 5, p. 055602, May 2018, doi: 10.1063/1.5006918.

[11] D. Pinna, G. Bourianoff, and K. Everschor Sitte, "Reservoir Computing with Random Skyrmion Textures," *Phys. Rev. Appl.*, vol. 14, no. 5, p. 054020, Nov. 2020, doi: 10.1103/PhysRevApplied.14.054020.

[12] D. Pinna *et al.*, "Skyrmion Gas Manipulation for Probabilistic Computing," *Phys. Rev. Appl.*, vol. 9, no. 6, p. 064018, Jun. 2018, doi: 10.1103/PhysRevApplied.9.064018.

[13] J. Zázvorka *et al.*, "Thermal skyrmion diffusion used in a reshuffler device," *Nat. Nanotechnol.*, vol. 14, no. 7, pp. 658–661, Jul. 2019, doi: 10.1038/s41565 019 0436 8.

[14] J. Grollier, D. Querlioz, K. Y. Camsari, K. Everschor Sitte, S. Fukami, and M. D. Stiles, "Neuromorphic spintronics," *Nat. Electron.*, vol. 3, no. 7, pp. 360–370, Jul. 2020, doi: 10.1038/s41928 019 0360 9.

[15] K. M. Song *et al.*, "Skyrmion based artificial synapses for neuromorphic computing," *Nat. Electron.*, vol. 3, no. 3, pp. 148–155, Mar. 2020, doi: 10.1038/s41928 020 0385 0.

[16] G. Bihlmayer, "A Chiral Magnet Induces Vortex Currents in Superconductors," *Physics (College. Park. Md).*, vol. 14, p. 39, Mar. 2021, doi: 10.1103/Physics.14.39.

[17] C. Psaroudaki and C. Panagopoulos, "Skyrmion Qubits: A New Class of Quantum Logic Elements Based on Nanoscale Magnetization," *Phys. Rev. Lett.*, vol. 127, no. 6, p. 067201, Aug. 2021, doi: 10.1103/PhysRevLett.127.067201.

[18] R. Juge *et al.*, "Helium Ions Put Magnetic Skyrmions on the Track," *Nano Lett.*, vol. 21, no. 7, pp. 2989–2996, Apr. 2021, doi: 10.1021/acs.nanolett.1c00136.

[19] S. Li *et al.*, "Experimental demonstration of skyrmionic magnetic tunnel junction at room temperature," *Sci. Bull.*, no. xxxx, Jan. 2022, doi: 10.1016/j.scib.2022.01.016.

[20] S. Finizio *et al.*, "Deterministic Field Free Skyrmion Nucleation at a Nanoengineered Injector Device," *Nano Lett.*, vol. 19, no. 10, pp. 7246–7255, Oct. 2019, doi: 10.1021/acs.nanolett.9b02840.

[21] S. Woo *et al.*, "Current driven dynamics and inhibition of the skyrmion Hall effect of

ferrimagnetic skyrmions in GdFeCo films," *Nat. Commun.*, vol. 9, no. 1, p. 959, Dec. 2018, doi: 10.1038/s41467-018-03378-7.

[22] R. Tomasello *et al.*, "Role of magnetic skyrmions for the solution of the shortest path problem," *J. Magn. Magn. Mater.*, vol. 532, no. March, p. 167977, Aug. 2021, doi: 10.1016/j.jmmm.2021.167977.

[23] T. Dohi, S. DuttaGupta, S. Fukami, and H. Ohno, "Formation and current-induced motion of synthetic antiferromagnetic skyrmion bubbles," *Nat. Commun.*, vol. 10, no. 1, p. 5153, Dec. 2019, doi: 10.1038/s41467-019-13182-6.

[24] A. P. Petrović *et al.*, "Skyrmion-(Anti)Vortex Coupling in a Chiral Magnet-Superconductor Heterostructure," *Phys. Rev. Lett.*, vol. 126, no. 11, p. 117205, Mar. 2021, doi: 10.1103/PhysRevLett.126.117205.

[25] P. Fischer, D. Sanz-Hernández, R. Streubel, and A. Fernández-Pacheco, "Launching a new dimension with 3D magnetic nanostructures," *APL Mater.*, vol. 8, no. 1, p. 010701, Jan. 2020, doi: 10.1063/1.5134474.

[26] T. Yokouchi *et al.*, "Creation of magnetic skyrmions by surface acoustic waves," *Nat. Nanotechnol.*, vol. 15, no. 5, pp. 361–366, May 2020, doi: 10.1038/s41565-020-0661-1.

[27] C. Chen *et al.*, "Surface acoustic wave controlled skyrmion-based synapse devices," *Nanotechnology*, vol. 33, no. 11, p. 115205, Mar. 2022, doi: 10.1088/1361-6528/ac3f14.

4.1 – Nanomaterials for unconventional computing

Aida Todri Sanial, Gabriele Boschetto, and Kremena Makasheva
[aida.todri@lirmm.fr, Gabriele.boschetto@lirmm.fr and kremena.makasheva@laplace.univ tlse.fr]

Status

With the emergence of nonconventional computing paradigms, one has the means to overcome the fundamental limitations of von Neumann architecture and perform highly complex functions with extremely low power. This perception prompts for materials and devices that can emulate the biological functions of neurons and synapses. Combining memory and resistor, memristors have become the most important electronic component for brain - inspired neuromorphic computing. The device has the ability to control resistance with multiple states by memorizing the history of previous electrical inputs – allowing it to mimic biological synapses and neurons of the biological neural networks. The switching in memristor devices is a reversible and controllable change of resistance induced by different stimuli, such as current, voltage, or light, with different physical processes such as ionic/electronic motion and redox reactions. Thus, the material selection plays a key role in the conductive path formation and modulation of the resistive switching behavior, and here we review them based on material properties. Owing to the dependence of their resistance states in the history of the applied electrical bias, memristors can store information in the form of electrical resistance and are typically driven by one of the four main mechanisms: electrochemical reactions (namely, redox and ion migration), phase changes (such as thermally activated amorphous crystalline transitions), tunnel magneto resistance (as such as spin dependent tunnel resistance) or ferroelectricity (namely, tunneling or domain wall transport). In addition, memristors can allow to process information intrinsically through the "let physics compute" (namely, perform complex signal transformation with physical dynamics), which are beneficial beyond neuromorphic computing, such as solving NP hard optimization problems and hardware security. To uncover their potentiality, we summarize here the nanomaterials used for unconventional computing and their different types of switching mechanisms (Figure 1).

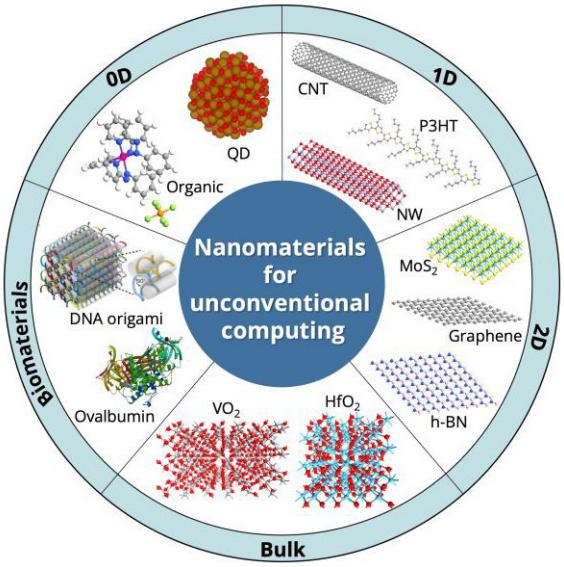

Figure 1. Illustration of different classes of nanomaterials used for unconventional computing, based on their dimensionality.

0D Nanomaterials – a variety of 0D materials have been investigated for memristors, mainly metal nanoparticles (NPs) and semiconductor quantum dots (QDs). Metal NPs are typically used to modify the resistive layer or electrodes to lower the charge injection barrier and interfacial potential drop of the electrode. Semiconductor QDs have also been investigated as promising candidates for developing electronic synaptic devices due to their electrical and optical properties. For example, charge accumulation in QD floating gate has obtained linearly programmable conductance states by controlling applied gate voltages. The optical properties of QDs enable electro photoactive synaptic devices, offering a promising candidate for new electronic synaptic devices. In addition, pairs of QDs are explored to serve as a single basic element in a quantum logic device such as quantum bit or qubit. At the same time, tunable small organic molecules (e.g., azo aromatics) and devices coupling ionic with electronic currents are investigated for emulating biological synapses behavior such as plasticity for continual learning. Resistance switching in most of these devices relies on either electrochemical doping, ion migration, or charge trapping mechanisms.

1D Nanomaterials – Carbon nanotubes (CNT) are among the most widely studied 1D nanomaterial with metallic or semiconducting behavior depending on their chiral vector. CNT network based transistors have been demonstrated as a synaptic transistor with physical mechanisms, including charge trapping in oxide dielectrics, ion migration, and electric double layer effects in polymer dielectrics. CNT networks have been used to show simple realizations of plasticity learning for spiking neural networks. In addition, CNT based transistors have been pursued for post silicon digital logic and memory implementation. Semiconductor nanowires (NW) possess attractive attributes for neuromorphic devices. A wide range of inorganic 1D nanomaterials (metal oxides, Ag, and Cu nanowires) have shown both volatile and non volatile resistive switching. In addition, organic polymer nanowires (e.g., P3HT) can emulate synapses' morphology and possess a learning mechanism similar to biological ion channels.

2D Nanomaterials – 2D materials, including graphene, transition metal dichalcogenides (TMDs), and hexagonal boron nitride (h BN), have been widely investigated as emerging materials for low power transistors, sensors, and memristive devices. 2D material based memristors can be further categorized based on their device geometry as lateral, vertical, and heterojunction structures. Depending on their geometry and materials, the switching mechanisms of these devices are based on phase transition, filament formation, charge trapping, defect migration, vacancy migration, or direct tunneling. Such devices are promising candidates for energy efficiency as artificial synapses while emulating plasticity for short term and long term memory.

Bulk Materials – Metal oxides exhibit electrical, optical, and semiconducting properties suitable for memristive devices. Interestingly, the crystal structure of some of these metal oxides undergoes changes under external stimuli such as thermal field, strain energy, surface energy, external force, magnetic field, and applied electric bias. This makes such materials suitable for memristor devices. For instance, resistive switching behavior from the repeatable formation of conductive filaments (i.e., oxygen deficient or oxygen vacancy rich) with electric bias yields the low and high resistance states in oxide memristors such as hafnia (HfO_2). In addition, vanadium dioxide (VO_2), which shows a reversible metal insulator transition (MIT) at near room temperature, and a reversible large conductance change, has been used to emulate biological neurons (i.e., oscillating behavior) and synapses. VO_2 has also been used in field effect transistors and gas sensors.

Biomaterials – These materials have attracted attention due to their long time natural evolution and well defined "structure function" relation. Proteins, in particular egg albumen and ferritin, have been shown suitable for developing memristors, with high plasticity in synaptic networks. In addition, the self assembly of nucleic acids (DNA and RNA) has provided a powerful and effective approach for constructing synthetic molecular structures, tiles, 2D lattices, 3D crystals, finite 2D shapes, DNA origami, and more complex 3D nanostructures. Such DNA structures enable the engineering of molecular structures with programmable shapes and properties for applications such as drug delivery and biological computing.

Current and Future Challenges

In the Table below, we provide the current advancements in nanomaterials and their switching mechanisms, which lead to unique electrical, optical, magnetic, or quantum properties that are the basis for nonconventional computing paradigms.

Material Dimensionality	Material Synthetic Route	Switching Mechanism	Materials (a few examples)	Switching Energy (fJ)	I_{ON}/I_{OFF}	Reference
0D	Wet chemistry	Charge-trapping	Organic, *mer*-$[Ru(L)_3](PF_6)_2$	10pJ	-	[1] [2]
	Molecular beam epitaxy, ion implantation, wet chemistry, vapor-phase	Hybrid charge-trapping filament formation, multi-band emission	BPQD; QD	-	10^1 - 10^7	[3] [4]
1D	Wet chemistry	Electrochemical doping	P3HT / PEO	10 fJ	-	[1]

	CVD, hydro-thermal growth	Ion migration	ZnO NW	-	<6	[5]
	CVD, arc discharge	Charge trapping	CNTFET with DMC	-	$5x10^4$	[6]
2D	CVD, PVD, exfoliation	Defect migration	Au-Mos_2-Au	-	$<10^4$	[7]
		Phase transition	Au-MoS_2-Au/Ti-TaS_2-Ti	-	10^1 - 10^7	[7]
		Vacancy migration	Pd/WSe_2/Pt; Pd/WS_2/Pt	>30fJ	2 - 10^{10}	[7]
		Filament formation	Cu/MoS_2/Au; Metal/*h*-BN/Metal	-	2 - 10^4	[7]
		Schottky/ Direct tunneling	Au/MoS_2/Au	-	10^8	[7]
3D	Sputtering, CVD, hydrothermal growth	Electrochemical Redox	Metal oxides (TiO_2, HfO_2, TaO_x)	>10fJ	2 - 40	[8]
		Phase change	Metal-insulator transition (VO_2)	>100fJ	$<10^3$	[8]
		Magnetic tunneling	Magnetoresistive materials (MgO)	>10fJ	2 - 3	[8]
		Ferroelectric polarisation	Ferroelectric materials ($BiFeO_3$, $BaTiO_3$, $PbZrTiO_x$)	>100fJ	45-300	[8]
Biomaterials	Natural way and Wet chemistry for synthetic growth	Generation of conductive filament	Proteins (silk fibroin, ferritin, collagen, egg albumen)	-	3 - 10^7	[9] [10]
		Biochemical operations; Self-assembly of nucleic acids	2D and 3D DNA nanostructures	-	-	[11] [12]

Advances in Science and Technology to Meet Challenges

Material and device variability and stability – these characteristics rely on the intrinsic quality of the active materials and manufacturing process, which are yet to be mastered. This leads to some variability in the material properties due to, for instance, defect density and grain size. A major challenge is to reduce variability and enhance the reliability of material growth with high crystallinity and uniform thickness and effective passivation techniques without deteriorating device performances. The device characteristics can be precisely controlled to realize functional circuits based on the switching mechanism and material properties.

Plasticity – as in biological neural networks, communication between neurons is dynamic and occurs at different time scales. Communication strength depends on the history of synapse activity, also known as plasticity. Short term plasticity facilities computation, while long term plasticity is attributed to learning and memory. To enable on chip learning, it is important that artificial synaptic devices display long term memory and architectures to allow for local learning rules.

Large scale integration – to provide commercially available unconventional computing paradigms using nanomaterials (2D, QD, organic, etc.), it is important to achieve large scale device array integrated with other circuits to show entire system operation while being CMOS compatible process.

Energy consumption – Reducing the energy consumption of devices and electronics is an important endeavor for future low power computing. It is not only important to develop low power devices but also the architecture of the full system in which it is implemented to achieve an energy efficient computing system.

Biofriendly materials The use of biocompatible and biodegradable materials in electronic devices can be an important trend in the development of green electronics. Compared to metal oxide semiconductors, biofriendly polymers and/or natural materials are attracting interest for their suitability on flexible neuromorphic platforms.

Collaborations and training – a close cooperation between neuroscientists, device physicists, computer scientists, computer architecture, and material scientists is of utmost importance to design and fabricate integrated circuits based on these new devices and realize the full potential of novel computing paradigms. The rapidly growing knowledge in each of these domains provides new insights and concepts to design energy efficient computing, necessitating collaborative efforts to train the new generation of students and scientists.

Concluding Remarks

Unconventional computing paradigms have propelled the research into novel devices that have led to a wide variety of solutions in terms of device physics, materials, and information processing. Despite the recent successes and advancements in novel devices and materials, more research is necessary to overcome the current limitations of devices to lower their variabilities and increase long term operations, state retention, and modulation for enabling both short term and long term plasticity.

Acknowledgements

ATS and GB acknowledge the support from the EU H2020 SmartVista project with grant agreement No. 825114 and EU H2020 NeurONN project with grant agreement No. 871501. KM acknowledges the support from the Agence Nationale de la Recherche in France, project ANR BENDIS (ANR 21 CE09 0008).

References

[1] Y. v. d. Burgt, A. Melianas, S. T. Keene, G. Malliaras, A. Salleo, Organic Electronics for Neuromorphic Computing, Nature Electronics, vol. 1, issue 7, 2018, https://doi.org/10.1038/s41928 -018 010

[2] V. K. Sangwan, M. C. Hersam, Neuromorphic Nanoelectronic Materials, Nature Nanotechnology, vol. 15, pp. 517 528, 2020, https://doi.org/10.1038/s41565 020 0647 z

[3] Z. Lv, Y. Wang, J. Chen, J. Wang, Y. Zhou, S T. Han, Semiconductor Quantum Dots for Memories and Neuromorphic Computing Systems, ACS Chemical Reviews, 2020, 120, 3941 4006.

[4] S. T. Han, L. Hu, X. Wang, Y. Zhou, Y. J. Zeng, S. Ruan, C. Pan, Z. Peng, Black Phosphorus Quantum Dots with Tunable Memory Properties and Multilevel Resistive Switching Characteristics, Adv. Sci., 4 (2017) 1600435, 10.1002/advs.201600435

[5] G. Milano, M. Luebben, Z. Ma, R. Dunin Borkowski, L. Boarino, C. F. Pirri, R. Waser, C. Ricciardi, I. Valov, Self limited single nanowire systems combining all in one memristive and neuromorphic functionalities, Nat. Commun., 9 (2018) 5151, 10.1038/s41467 018 07330 7

[6] J. L. Xu, R. X. Dai, Y. Xin, Y. L. Sun, X. Li, Y. X. Yu, L. Xiang, D. Xie, S. D. Wang, T. L. Ren, Efficient and Reversible Electron Doping of Semiconductor Enriched Single Walled Carbon Nanotubes by Using Decamethylcobaltocene, Sci. Rep., 7 (2017) 6751, https://doi.org/10.1038/s41598 017 05967 w

[7] W. Huh, D. Lee, Ch H. Lee, Memristors Based on 2D Materials as an Artificial Synapse for Neuromorphic Electronics, Advanced Materials, 2020, 32, 2002092, DOI: 10.1002/adma.202002092

[8] J. Tang, F. Yuan, X. Shen, Zh. Wang, M. Rao, Y. He, Y. Sun, Z. Li, W. Zhang, Y. Li, B. Gao, H. Qian, G. Bi, S. Song, J. J. Yang, H. Wu, Bridging Biological and Artificial Neural Networks with Emerging Neuromorphic Devices: Fundamentals, Progress, and Challenges, Advanced Materials, 2019, 31, 1902761, DOI: 10.1002/adma.201902761.

[9] G. Zhou, Z. Ren, L. Wang, B. Sun, S. Duan, and Q. Song, Artificial and wearable albumen protein memristor arrays with integrated memory logic gate functionality, Mater. Horiz., 6 (2019), 1877, 10.1039/c9mh00468h

[10] J. Wang, F. Qian, S. Huang, Z. Lv, Y. Wang, X. Xing, M. Chen, S. T. Han, and Y. Zhou, Recent Progress of Protein Based Data Storage and Neuromorphic Devices, Adv. Intell. Syst., 3 (2021), 2000180, 10.1002/aisy.202000180

[11] Y. Ke, L. L. Ong, W. M. Shih, P. Yin, Three Dimensional Structures Self Assembled from DNA Bricks, Science, 338 (2012) 1177 1183, 10.1126/science.1227268

[12] A. S. Perumal, Z. Wang, G. Ippoliti, F. v. Delft, L. Kari, D. V. Nicolau, As good as it gets: A scaling comparison of DNA computing, network biocomputing, and electronic computing approaches to an NP complete problem, New J. Phys, 23 (2021) 125001, https://doi.org/10.1088/1367 2630/ac3883

4.2 Neuromorphic Computing with Emerging Two-Dimensional Nanomaterials

Vinod K. Sangwan,[1] Amit Ranjan Trivedi[2], Mark C. Hersam[1,3,4]

1. Department of Materials Science and Engineering, Northwestern University, Evanston, IL, USA 60208
2. Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, IL, USA 60607
3. Department of Chemistry, Northwestern University, Evanston, IL, USA 60208
4. Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA 60208

Status

The emergence of two-dimensional (2D) and van der Waals (vdW) materials has invigorated fundamental research at the device level for brain-inspired computing hardware.[1, 2] A critical component of neuromorphic circuits is an analog non-volatile memory (NVM) that is not only fast, reliable, and high-density but also possesses multiple states and internal temporal dynamics to mimic the spike-based learning rules of biological synapses. Crossbars of NVM technologies based on conventional bulk materials, such as memristors, phase change memories, and magnetic and ferroelectric tunnel junctions, can outcompete CMOS counterparts for neural network performance metrics. All of these NVMs have also been realized using 2D materials with unprecedented functionalities (e.g., gate tunability) that translate into improved performance as a result of simplified circuit architectures. For example, 2D materials have been integrated into atomically thin vertical memristors with femtojoule switching energies (Fig. 1a,b).[3,4] The most promising vertical memristors are based on 2D transition metal dichalcogenides (TMDCs) or hexagonal boron nitride (hBN) where resistive switching has been achieved with intrinsic defects or metal cations. Although the constituent 2D materials can be grown over a wafer-scale, most of

the demonstrations thus far have been limited to 10 x 10 crossbar arrays without a selector (Fig. 1c,d).[3] A particularly promising approach is a self-selective crossbar based on two hBN memristors with volatile and non-volatile switching in an Au/hBN/graphene/hBN/Ag stack (Fig. 1c).[5] Although some applications have been proposed for 2D vertical memristors (e.g., RF switches, encryption circuits), their characteristics and functions are similar to conventional two-terminal memristive systems.[3,4]

To gain more unique functionality, semiconducting 2D materials (e.g., MoS_2) can also be integrated into lateral memtransistors where nonvolatile switching is tuned by a third gate electrode (Fig. 1d,e).[6] In addition, 2D channels enable dual-gated control where one of the gates can achieve tunable learning behavior, while the other gate can be used as a selector in a manner analogous to a one-transistor-one-memristor (1T1M) crossbar (Fig. 1f,g).[7] Lateral memtransistors are also compatible with multiple electrodes to realize heterosynaptic learning behavior.[6] Memtransistors have been generalized to a wider class of heterojunctions using charge trap, floating gate, ferroelectric, conducting bridge, and phase change memories.[2] Crossbars consisting of 10 x 10 memtransistors have been experimentally demonstrated, achieving the same level of complexity as 2D vertical memristors (Fig. 1d,g).

Solution-processed 2D and vdW materials are also promising for printed and flexible neuromorphic circuits. For instance, femtojoule vertical memristors and memcapacitors have been demonstrated using printed films on flexible substrates.[8] However, most of these devices use electrochemical filaments such as Ag and Cu, and thus the role of the layered materials is unclear. Recently, a new thermally activated volatile switching mechanism has been reported for a range of solution-processed 2D materials that can be exploited for artificial spiking neurons (Fig. 1i,j).[9] Here, the morphology of the 2D nanoflakes plays a vital role to produce non-linear behavior that

can be used for high-order oscillator circuits. However, the lack of an effective selector has limited the integration of printed memristors in crossbar architectures (Fig. 1k,l). Recently, neuromorphic applications have also been proposed for 2D magnets, 2D charge density wave switches, and 2D moiré heterostructures, suggesting further opportunities in this space.

Current and Future Challenges

The main challenge facing vertical 2D memristors is competition with conventional metal oxide memristors that outperform their 2D counterparts in nearly all relevant metrics. Furthermore, wafer-scale 2D materials are generally polycrystalline, and spatial variations in grain boundaries are likely to lead to device-to-device variability, unlike the relatively high device-to-device homogeneity of amorphous metal oxide films.[3,4] This spatial inhomogeneity is further exacerbated by the inherent variability arising from stochastic switching that is common to all filamentary switches. While 2D memristors provide atomically thin channels, the lateral dimensions of metal lines are the more relevant scaling parameters for high-density crossbars, which also may be complicated by the finite grain sizes in 2D films. While single-crystal 2D flakes have also shown stable memristive switching arising from partially oxidized layers, wafer-scale growth of layered single crystals has not been shown. Thus, one immediate challenge in vertical memristors is to scale N x N crossbar arrays from N = 10 to N = 1000. Another key challenge is to integrate vertical memristors with a selector to avoid sneak current issues. A 2D transistor selector may be possible, although integration of a functional 1T1M crossbar has not yet been demonstrated.

Lateral memtransistors are faced with similar scaling challenges where the device footprint and operating power are not yet competitive with conventional vertical memristors. Since grain boundaries are believed to be essential for resistive switching in memtransistors, polycrystalline grain size likely dictates the ultimate scaling limits. Furthermore, since the operating mechanism

of memtransistors relies upon the modulation of Schottky injection at the contacts, the operating voltage is not expected to scale linearly with channel length. Despite these challenges, the state-of-the-art complexity of lateral memtransistor crossbars (channel < 1 μm) and operating voltages (<1 V) approach that of vertical memristors (Fig. 1c,f).[3,5,7] Moreover, dual-gated lateral memtransistors achieve 1T1M functionality within the same device without requiring integration with another selector technology.[5,7] On the other hand, the switching speed of lateral memtransistors is significantly slower than vertical memristors, and gradual soft switching is likely to reduce the dynamic range of resistance under fast operating conditions.[2] Nevertheless, dual-gated ferroelectric and floating gate memtransistors have the potential to reduce the switching power and increase the switching speed.[2] Another challenge is integrating 2D NVM devices into circuits such as spiking neurons, activation circuits, and analog-to-digital converters (ADCs) for complete neural network chips.[1] Solution-processed 2D material devices also need to be scaled to sub-micron length scales for practical applications such as wearable electronics for off-grid classification and medical diagnostics.[8,9] In addition, the integration of printed NVM circuits with flexible logic circuits for full data processing has not been demonstrated. Overall, the grand challenges for 2D neuromorphic computing are centered on materials control and device engineering to achieve comparable metrics to conventional NVMs but with additional functionalities that yield improved efficiency in hardware computation.

Advances in Science and Technology to Meet Challenges

The last few years have seen significant advances in wafer-scale growth of 2D semiconductors and insulators that are directly relevant to the unique challenges of neuromorphic circuits.[3] Current growth advances are focused on achieving large grain sizes and minimizing lattice defects for wafer-scale uniformity of conventional transistor technology. For memristive

devices using intrinsic defects, growth also needs to be optimized to yield small gain sizes (< 10 nm) and well-controlled defect densities.[1] Although wafer-scale 2D transistors have been used to realize neural network chips consisting of > 800 devices, this technology does not yet compete effectively with existing Si CMOS-based neural network chips. In this context, self-aligned vdW anti-ambipolar Gaussian transistors have been shown to significantly simplify the circuit architecture of spiking neurons with a smaller number of elements than conventional CMOS circuits. These Gaussian transistors could also be integrated with a non-volatile memory (e.g., floating gate or ferroelectric gate) to achieve Bayesian neural networks for machine learning predictions with confidence bounds. In terms of advances in fabrication, the self-aligned scheme also provides an opportunity for highly scaled lateral memtransistor crossbars. While efforts are underway to improve the performance of individual devices, the existing neuromorphic paradigms also need to be revisited to identify unique opportunities enabled by the unique characteristics of 2D devices. For example, recent algorithmic innovations in deep neural network architectures require higher-order processing where, along with inputs and model parameters (i.e., weights), the application context should also be considered in making predictions. For these higher-order neural networks, the additional gate electrode layer in dual-gated memtransistor crossbars presents a promising pathway to dynamic weight selection.[10] In this manner, 2D neuromorphic computing has the potential to provide efficient hardware accelerators for emerging artificial intelligence and machine learning algorithms.[11]

Figures

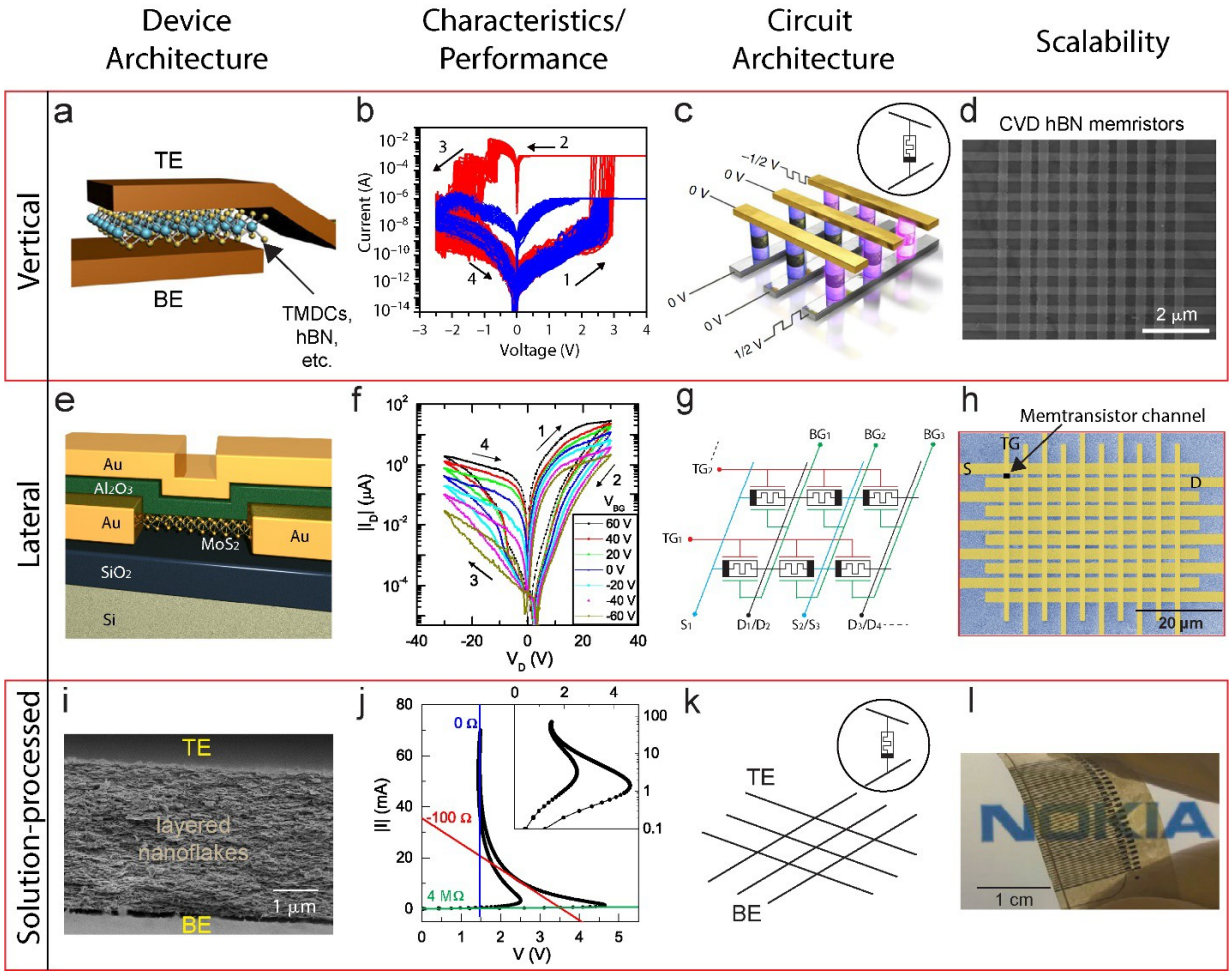

Figure 1. (a) Device architecture of a vertical memristor using layered materials such as TMDCs, hBN, and other insulators. (b) Typical current-voltage (I-V) characteristics of a vertical hBN memristor for two current compliances (red and blue curves). Arrows show the voltage sweep direction. (c) Crossbar architecture for vertical memristors where the desired node is selected by a V/2 biasing scheme. (d) Scanning electron microscopy (SEM) image of a 10 x 10 crossbar array of vertical memristors on a wafer-scale hBN film. (e,f) Device architecture and gate bias (V_{BG})-dependent I-V characteristics of a dual-gated lateral MoS_2 memtransistor, respectively. (g,h) Architecture and SEM image (false color) of a dual-gated memtransistor crossbar array,

respectively. (i,j) Cross-sectional SEM image and current-controlled I-V characteristics of a solution-processed MoS_2 memristor, respectively. (k) Schematic of a 3 x 3 crossbar using memristors from solution-processed 2D materials. (l) Optical image of a 50 x 1 crossbar array of MoS_2 memristors on printed Ag electrodes. (a) Reproduced with permission.[4] Copyright 2018, Springer Nature. (b,d) Reproduced with permission.[3] Copyright 2020, Springer Nature. (c) Reproduced with permission.[5] Copyright 2019, Springer Nature. (e-h) Reproduced with permission.[7] Copyright 2020, Wiley-VCH. (i,j) Reproduced with permission.[9] Copyright 2021, Wiley-VCH. (l) Reproduced with permission.[8] Copyright 2015, Springer Nature.

Acknowledgments

This work was primarily supported by National Science Foundation (NSF) Grant Number CCF-2106964. V.K.S. and M.C.H. also acknowledge support from the Department of Energy (DOE) Threadwork Program under Grant Number 8J-30009-0032A and the Laboratory Directed Research and Development Program at Sandia National Laboratories (SNL). SNL is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. DOE National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. DOE or the United States Government.

References

1. Sangwan V. K. & Hersam M. C. Neuromorphic nanoelectronic materials. *Nat Nanotechnol* **15**, 517 (2020).

2. Yan X., Qian J. H., Sangwan V. K. & Hersam M. C. Progress and challenges for memtransistors in neuromorphic circuits and systems. *Adv Mater*, **34**, 2108025 (2022).

3. Chen S.*, et al.* Wafer-scale integration of two-dimensional materials in high-density memristive crossbar arrays for artificial neural networks. *Nat Elect* **3**, 638-645 (2020).

4. Kim M.*, et al.* Zero-static power radio-frequency switches based on MoS_2 atomristors. *Nat Commun* **9**, 2524 (2018).

5. Sun L.*, et al.* Self-selective van der Waals heterostructures for large scale memory array. *Nat Commun* **10**, 3161 (2019).

6. Sangwan V. K.*, et al.* Multi-terminal memtransistors from polycrystalline monolayer molybdenum disulfide. *Nature* **554**, 500 (2018).

7. Lee H.-S.*, et al.* Dual-gated MoS_2 memtransistor crossbar array. *Adv Funct Mater* **30**, 2003683 (2020).

8. Bessonov A. A.*, et al.* Layered memristive and memcapacitive switches for printable electronics. *Nat Mater* **14**, 199 (2015).

9. Sangwan V. K.*, et al.* Visualizing thermally activated memristive switching in percolating networks of solution-processed 2D semiconductors. *Adv Funct Mater* **31**, 2107385 (2021).

10. Rahimifard L.*, et al.* Higher-order neural processing with input-adaptive dynamic weights on MoS_2 memtransistor crossbars. *Front Elect Mater*, DOI: 10.3389/femat.2022.950487 (2022).

11. M. E. Beck, A. Shylendra, V. K. Sangwan, S. Guo, W. A. Gaviria Rojas, H. Yoo, H. Bergeron, K. Su, A. R. Trivedi, and M. C. Hersam, "Spiking neurons from tunable Gaussian heterojunction transistors," *Nature Communications*, **11**, 1565 (2020)

5.1- Computing with p-bits: A case study in the new era of electronics

Kerem Y. Camsari[1], Peter L. McMahon[2], Giovanni Finocchio[3] and Supriyo Datta[4]

[1] Department of Electrical and Computer Engineering, University of California, Santa Barbara, Santa Barbara, USA

[2] School of Applied and Engineering Physics, Cornell University, Ithaca, USA

[3] Department of Mathematical and Computer Sciences, Physical Sciences and Earth Sciences, University of Messina, Messina, Italy

[4] Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana, USA

camsari@ece.ucsb.edu, pmcmahon@cornell.edu, giovanni.finocchio@unime.it, datta@purdue.edu

Status

The slowing down of Moore's Law led to the emergence of an exciting new era of electronics. Following decades of continuous improvements in transistor technology, the new era is marked by blurred layers of abstraction in the computing stack, creative combinations of CMOS technology with emerging technologies and the development of domain-specific hardware and architectures. In this short piece, we describe probabilistic computing with p-bits, as a representative example of the many promising directions in the new era. We describe recent developments of p-bits; starting from their energy-efficient realization in different hardware substrates, their physics-inspired and parallel architectures, all the way up to their use in high-level algorithms and applications. A recurring theme in this piece is that of *co-design*; where algorithms and applications are modified to naturally conform to the underlying physics of hardware. Along with domain-specificity, co-design will play an increasingly important role in the new era which will be marked by various CMOS + X type heterogeneous architectures.

Origins of probabilistic computing with p-bits

In a celebrated talk delivered at a conference in May 1981, Richard Feynman introduced the first clear vision of a quantum computer [1]. The main idea of Feynman's talk, appropriately titled "Simulating Physics with Computers", can be summarized by the credo "Let physics do the computing." In other words, simulating physical phenomena is efficient when the simulating "computer" itself is made of the building blocks it is trying to simulate. This profound connection between physics and computing Feynman emphasized has since been used to develop quantum computers built out of quantum mechanical bits. This part of the story is very well-known and often discussed, see, for example, "Quantum Computing: 40 years later" by John Preskill for more details [1]. What is less appreciated is that before getting on to quantum computing, Feynman talked about a vision of a probabilistic computer with essentially the same idea: a probabilistic *Nature* should be efficiently simulated by a machine that itself makes probabilistic decisions. In a few lines, Feynman laid out the main idea that is used in many probabilistic models today: in an interacting system with many degrees of freedom, if we need to compute correlations between small parts of the system, all we have to do is to observe those parts. What is otherwise an intractable summation over the exponentially large "rest of the system states" then becomes approximately tractable[1].

Driven by the nearing end of Moore's Law, a few years ago, we took Feynman's vision a step further. Intrigued by the large degrees of inherent noise in magnetic nanodevices, we imagined a truly probabilistic computer down to its most basic building block. Early work involved using the

[1] A technical note for the expert: the observed correlation is *still* an approximation unless we observe the system with time $T \rightarrow \infty$ which amounts to computing the intractable sum exactly, for an ergodic system.

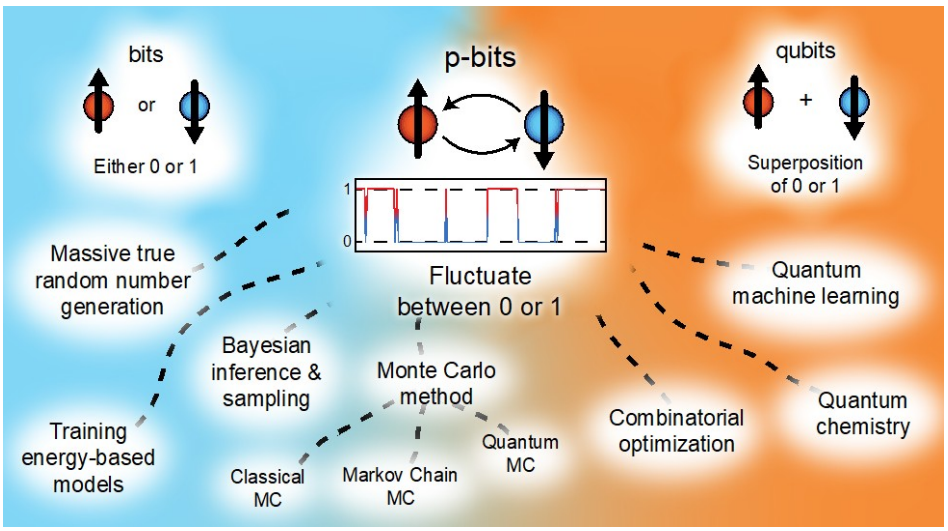

FIG. 1. Various applications of probabilistic computing with p bits, in the space of applications between bits and qubits.

probabilistic switching behavior of stable magnets [2], but gradually the temporal noise of low barrier nanomagnets (LBM) became a more natural choice. LBMs offered the possibility of a compact realization of the basic building block of a probabilistic computer, which we named the "p-bit" [3], and we experimentally demonstrated a prototype "p-computer" shortly after [4].

The ubiquitous nature of probabilistic methods and randomized algorithms allows p-bits to be applied to a broad range of applications (FIG. 1). Examples include massively parallel true random number generation, solving combinatorial optimization problems using powerful algorithms such as simulated annealing and parallel tempering, probabilistic sampling for Bayesian inference and learning, training energy-based and variational classical and quantum models, accelerating Monte Carlo (MC), Markov Chain Monte Carlo (MCMC), Quantum Monte Carlo methods, computational biology and protein folding among others. The wide-ranging application space for p-bits makes them potential candidates for domain-specific computing, with overlapping applications envisioned for near-term quantum computers, particularly for Machine Learning and AI applications.

Device-circuit co-design of p-bits. In essence, a p-bit is the abstraction of a tunable Bernoulli variable with many different physical implementations. p-bits are closely related to the basic unit of Boltzmann Machines, the binary stochastic neuron, pioneered by Hinton and Sejnowski [5]. The p-bit, when defined as a mathematical abstraction [3], has a much wider range of applications than just Boltzmann Machines (FIG. 1). Therefore, finding the most energy-efficient, technologically scalable, and robust p-bit is an active area of research. In addition to magnetic p-bits with stochastic magnetic tunnel junctions, stochastic resistors (e.g., diffusive memristors, perovskite nickelates), diodes (e.g., Zener, single photon avalanche) and even analog or digital CMOS (e.g., RTN in silicon transistors, LFSRs) can make compact and energy-efficient p-bits. Similarly, connecting p-bits to one another can be achieved in many ways: resistive (or capacitive) crossbar arrays, digital or mixed-signal CMOS-based interconnections are examples, to name a few. Key metrics in designing good p-bits are the energy (E) and delay ( ) to produce a random bit. Like novel switches, minimizing the energy-delay product of a truly random bit with the minimum area footprint guides the development of novel p-bits. Exciting new experimental developments with stochastic magnetic tunnel junctions [6] have shown that can be a few nanoseconds or less. Combining various possible MTJ designs (e.g., in-plane, circular disk, perpendicular, double-free-layer) with CMOS transistors by deliberate co-design, energy-efficient circuits with E < 1 fJ/rng could be obtained in monolithically integrated p-computers with tens of millions of p-bits. The key advantage of a nanodevice-based p-bit comes from the large savings in energy, area and the quality of randomness over digital CMOS. Even when compared to low quality

pseudo-random number generators in CMOS, an MTJ-based p-bit is at least 1000X smaller in area and 100X smaller in energy to produce a random bit. The compactness in area and energy efficiency opens up the potential to a high degree of scalability with MTJ-based p-bits, beyond what is accessible with present-day technology.

Architecture-algorithm co-design of p-bits. A key step in mapping algorithms to hardware is to find an efficient architecture co-designed with the algorithm. Designing probabilistic computers starting from single devices to systems allows imagining completely new architectures with suitably modified algorithms. To illustrate this point, consider a simple MC algorithm for calculating the number ``π'': Imagine a square with a circle in the centre and divide each side of the square into 2^{20} segments. Then, use a (20 + 20)-bit RNG to generate a random coordinate (x, y). Calculate an output s {0, 1} indicating whether the random coordinate (x, y) lies inside the circle or not (s = [$x^2 + y^2 < 1$]). Perform N trials and obtain an average to estimate π This is clearly a parallelizable algorithm but not trivially: avoiding significant delays while calculating the sum of squares requires a carefully pipelined architecture such that a new sample is obtained at every clock cycle [7]. Similarly, accelerating Markov Chain MC algorithms requires deliberate designs since for directly connected p-bits, parallel updates are not allowed. One such design is exploiting the idea that not directly connected (conditionally independent) p-bits can be updated in parallel. This allows reaching high levels of parallelism if the connecting graphs are sparse and can be divided into large segments that are be updated in parallel [15]. More intriguingly, p-computers can have entirely asynchronous architectures where each p-bit updates with its randomly ticking internal clock in physics-inspired, massively parallel architectures. Preliminary results indicate the promise of such architectures [8], however, our main point is to illustrate the wide range of possibilities that exist for architecture-algorithm co-design.

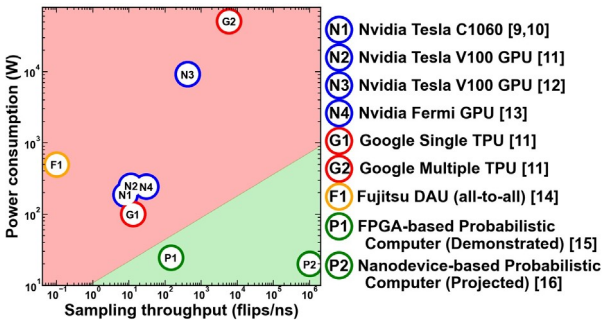

FIG. 2 Published data on power consumption (W) vs. sampling throughput (flips/ns) on state of the art probabilistic samplers.

Benchmarking probabilistic computers. Virtually all applications shown in FIG. 1 benefit from one key metric in probabilistic computers, namely the sampling throughput or flips per second [9-16]. Sampling throughput is a commonly reported in specially designed probabilistic samplers and FIG. 2 shows power consumption vs. sampling throughput for highly optimized implementations. GPU and TPUs often use highly regular graphs (typically 2D nearest neighbor grids) to achieve scalability in their architectures. Others, such as Fujitsu's digital annealer uses all-to-all connected graphs, taking fewer samples per second but compensating this by means of powerful algorithms such as parallel tempering and population annealing [14]. For example, the Google TPU can take more than 5,000 flips / ns but at the expense of 50,000 W power dissipation to achieve this feat! On the other hand, FPGA-based p-computers take 100 flips / ns using around 20 W. But more importantly, projections based on nanodevice based p-bits indicate the possibility for 1,000,000 flips /ns at only 20 W of power! This number can be reached in designs where each p-bit dissipates 20 μW with a million of them flipping every nanosecond in an asynchronous setup. All of these pieces have been individually demonstrated: the magnetic memory industry have scaled MTJs up to billion-bit densities and stochastic magnets have been shown to fluctuate every nanosecond. The potential for growth and acceleration by p-bits seem highly promising if challenges for integration and co-design can be surmounted in the future.

Future Directions. Similar in spirit to many other promising domain-specific computing paradigms [17,18], there are several important areas requiring further attention. From the physics end, identifying the best possible mixed-signal p-bit design is still a work in progress. Controlling device-to-device variations or overcoming them through algorithm hardware co-design is also critical. From the systems side, identifying and adopting powerful algorithms and applications conforming to p-bits require expert algorithmic understanding. Optimizing the necessary bit precision, design modes (synchronous vs. asynchronous) with the right architecture, while being amenable to monolithic integration all require a concerted effort and the widest possible expertise in the computing stack. Overall, this exciting, full-stack research program is simply one example of a powerfully emerging trend in the new era of electronics where domain-specific hardware and architectures will play an increasingly important role.

Acknowledgment

KYC is grateful to A. Grimaldi and N. A. Aadit for assistance in the preparation of the figures. KYC acknowledges support through National Science Foundation (CCF 2106260), Samsung GRO program and Office of Naval Research Young Investigator Program. The work of GF was supported under the project PRIN 2020LWPKH7 funded by the Italian Ministry of University and Research.

References

[1] Feynman, Richard P. "Simulating physics with computers." Feynman and computation. CRC Press, 2018. 133-153, Preskill, John. "Quantum computing 40 years later." *arXiv preprint arXiv:2106.10522* (2021).

[2] Behin-Aein, Behtash, Vinh Diep, and Supriyo Datta. "A building block for hardware belief networks." Scientific reports 6.1 (2016): 1-10.

[3] Camsari, Kerem Yunus, et al. "Stochastic p-bits for invertible logic." Physical Review X 7.3 (2017): 031014.

[4] Borders, William A., et al. "Integer factorization using stochastic magnetic tunnel junctions." Nature 573.7774 (2019): 390-393.

[5] Hinton G E, Sejnowski T J et al. 1986 Parallel distributed processing: Explorations in the microstructure of cognition 1 2

[6] Hayakawa K, Kanai S, Funatsu T, Igarashi J, Jinnai B, Borders W, Ohno H and Fukami S 2021 Physical review letters 126 117202

[7] Kaiser, Jan, and Supriyo Datta. "Probabilistic computing with p-bits." Applied Physics Letters 119.15 (2021): 150503.

[8] Aadit, Navid Anjum, et al. "Physics-inspired Ising Computing with Ring Oscillator Activated p-bits." arXiv preprint arXiv:2205.07402 (2022).

[9] Block, Benjamin, Peter Virnau, and Tobias Preis. "Multi-GPU accelerated multi-spin Monte Carlo simulations of the 2D Ising model." Computer Physics Communications 181.9 (2010): 1549-1556

[10] Preis, Tobias, et al. "GPU accelerated Monte Carlo simulation of the 2D and 3D Ising model." Journal of Computational Physics 228.12 (2009): 4468-4477.

[11] Yang, Kun, et al. "High performance Monte Carlo simulation of Ising model on TPU clusters." Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. 2019.

[12] Romero, Joshua, et al. "High performance implementations of the 2D Ising model on GPUs." Computer Physics Communications 256 (2020): 107473

[13] Fang, Ye, et al. "Parallel tempering simulation of the three-dimensional Edwards–Anderson model with compact asynchronous multispin coding on GPU." Computer Physics Communications 185.10 (2014): 2467-2478.

[14] Aramon, Maliheh, et al. "Physics-inspired optimization for quadratic unconstrained problems using a digital annealer." Frontiers in Physics 7 (2019): 48.

[15] Aadit, Navid Anjum, et al. "Massively parallel probabilistic computing with sparse Ising machines." Nature Electronics (2022): 1-9.

[16] Sutton, Brian, et al. "Autonomous probabilistic coprocessing with petaflips per second." IEEE Access 8 (2020): 157238-157252.

[17] M Mohseni, Naeimeh, Peter L. McMahon, and Tim Byrnes. "Ising machines as hardware solvers of combinatorial optimization problems." Nature Reviews Physics 4.6 (2022): 363-379.

[18] Finocchio, Giovanni, et al. "The promise of spintronics for unconventional computing." Journal of Magnetism and Magnetic Materials 521 (2021): 167506.

6.1- Simulated Bifurcation

Kosuke Tatsumura, Hayato Goto, Toshiba Corporation
[kosuke.tatsumura@toshiba.co.jp, hayato1.goto@toshiba.co.jp]

Status

Simulated bifurcation (SB) [1] is a quantum-inspired heuristic algorithm for finding the exact or approximate ground states of Ising spin models and is expected to be useful for various practical combinatorial optimization. Many combinatorial optimization problems are classified as non-deterministic polynomial-time (NP)-hard, where the computational complexity scales exponentially with the problem size, and can be converted to Ising problems [2,3]. Special-purpose hardware devices for solving Ising problems are called Ising machines, including SB-based machines.

The SB algorithm was found as a classical counterpart of bifurcation-based adiabatic quantum computation with a nonlinear oscillator network [4,5]. In SB, we numerically simulate the adiabatic evolution of a classical Hamiltonian dynamical system (a nonlinear oscillator network) with bifurcations (Fig. 1). Two branches of a bifurcation in each nonlinear oscillator represent two states of each Ising spin. The operational mechanism of SB is based on an adiabatic and ergodic search (Fig. 1) [1,6]. Recently two other variants of SB called the ballistic simulated bifurcation (bSB) and the discrete simulated bifurcation (dSB) [7] have been proposed and demonstrated to outperform the original adiabatic SB (aSB) in terms of both speed and solution accuracy. These algorithms exploit new effects, such as a quasi-quantum tunneling effect [7].

The SB algorithms are highly parallelizable and thus can be accelerated with massively parallel processors such as FPGAs (field-programmable gate arrays) and GPUs (graphics processing units) [1,7,8,9]. SB allows us to simultaneously update N coupled-oscillator variables for N-spin problems at each time step. This is in contrast to simulated annealing (SA, a conventional heuristic algorithm), which involves sequential updates of spins, with simultaneous updates allowed only for isolated spins. For N-spin Ising problems with full connectivity, the maximum numbers of parallelizable operations in SB and SA are, respectively, N^2 and N [8,9]. Custom-circuit implementations of SB [1,8] have demonstrated a higher degree of computational parallelism than the problem size N (8,192 parallel processing elements for 2,048-spin Ising problems).

Various Ising machines based on different principles such as SA, quantum annealing, and dynamical-system evolution have been implemented with a variety of technologies including superconducting circuits, optics, emerging nanodevices, parallel digital processors, etc [2,3]. SB-based machines have been evaluated for various benchmark problems and compared with other Ising machines [2,3,7], demonstrated to be highly competitive, especially showing the highest performance for Ising problems with full connectivity [known as the Sherrington-Kirkpatrick (SK) model].

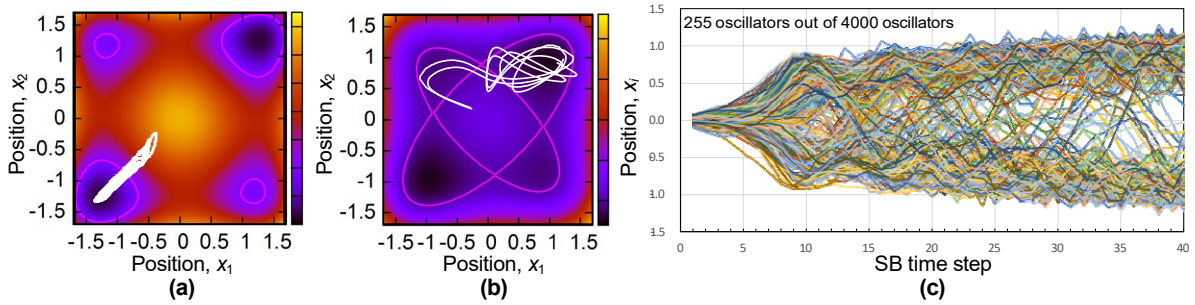

Figure 1. Dynamics in simulated bifurcation. (a) and (b) show trajectories indicating adiabatic and ergodic searches, respectively, for a two spin problem (N=2) [1]. (c) Time evolution of oscillators exhibiting bifurcations (N=4000) [8].

Current and Future Challenges

SB is theoretically new (published in 2019 [1]) and there are many challenges and opportunities for further enhancement and wider applicability.

While quantum adiabatic optimization is based on the quantum adiabatic theorem [4,5], the operational mechanism of SB (adiabatic and ergodic search), which implies the classical adiabatic theorem, has been only empirically understood [1]. The mathematically rigorous proof of the operational principle of SB as well as the convergence property have been left for future work [1]. Potential theoretical studies include extending SB to polynomial unconstrained binary optimization (PUBO), relating SB with nonequilibrium statistical mechanics, and combining SB with techniques for complex constraints. Since the comparison between various Ising machines in terms of performance depends on problem instances, figure-of-merits and physical implantations [2,3,7], comprehensive and systematic comparisons should be continued.

Building larger Ising machines while avoiding speed degradation is challenging. In SB, the matrix-vector multiplication (MM) of the coupling matrix J and the position vector $\boldsymbol{x}$ of nonlinear oscillators (many-body interaction) is the most computationally intensive part [8]. To process the MM part in a massively parallel fashion, we have to prepare many processing elements (multiply-accumulators) and supply the J and $\boldsymbol{x}$ data to the processing elements at a sufficient transfer rate (the transfer rate needed increases with increasing the processing elements). As an example, the FPGA implementations of SB [1,7,8] were equipped with optimized memory subsystems to supply the J and $\boldsymbol{x}$ data by using on-chip memory (having larger bandwidth than external memory) and thus were allowed to fully utilize the computation resources in the time domain. However, the machine size (maximum problem size) of such a single-chip implementation is limited by the on-chip memory resource. Hence enlarging the machine size while fully utilizing the computation resources is of importance. The possible two approaches are scale-up (making a chip larger or denser) and scale-out (increasing the number of networked chips). The SB-based machines would benefit from emerging nanodevices for processing, memory, and communication in conjunction with in-memory computing, stochastic computing, and cluster computing architectures.

By implementing not only SB processing circuits but also interface/control circuits on a single chip, we can shorten the system-wide latency, enabling real-time systems based on combinatorial optimization that make the optimal responses to ever-changing situations. SB-based systems are thus expected to realize innovative applications.

Advances in Science and Technology to Meet Challenges

SB has been receiving increasing attention because of both the high performance and high practicability. Several advances in theoretical extension [10], custom-circuit architecture [9], and applications [11-14] are as follows.

Kanao *et al.* introduced a heating process to the SB Hamiltonian dynamics to assist the system during the search to escape from local minima, leading to improved performance [10]. This method was inspired by the Nosé-Hoover method for simulating Hamiltonian dynamics at finite temperature. The heated SB does not use random numbers, unlike SA, and thus is as deterministic and simple for parallel implementation as the original SBs (aSB, bSB and dSB).

A larger Ising machine can, in principle, be built by partitioning a spin system into multiple subsystems. In this case, the spin-spin couplings over the subsystems must be incorporated, and the partitioned subsystems also have to evolve in a single time domain. Communication and synchronization between the partitioned subsystems can easily degrade the speed performance. Tatsumura *et al.* proposed and demonstrated a scale-out architecture for SB-based Ising machines that enables continued scaling of both the machine size and speed performance by connecting multiple FPGAs as shown in Fig. 2 [9]. To maintain time consistency between multiple chips and a sufficiently small stall rate for every SB time

step, the architecture relies on an autonomous synchronization mechanism that is implemented in the information exchange processes between neighbouring chips.

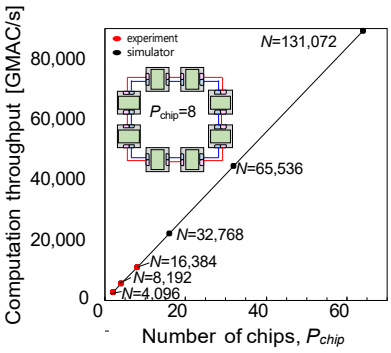

Figure 2. Scale out architecture for SB [9]. Constant efficiency scaling characteristic. (Inset) Connection of multiple chips in a bidirectional ring topology.

As an example of the application of SB accelerators for real-time systems, Tatsumura *et al.* presented an ultrafast financial transaction machine with a total response time of about 30 μs, including not only the detection of the most profitable cross-currency arbitrage opportunities by SB but also issuing order packets [11]. The detection problem of currency arbitrage opportunity was reduced to an optimal path search in a directed graph called a market graph, further formulated as an Ising problem, then solved with an SB accelerator. Steinhauer *et al.* used SB, in the financial field, for solving the integer portfolio and trading trajectory problem [12]. Zhang *et al.* applied bSB to traveling salesman problems and reported better solution accuracy and higher speed than an SA implementation [13]. Matsumoto *et al.* presented a hybrid method that iteratively uses a general-purpose processor (CPU) and an SB-based Ising machine for solving a discrete optimization problem (a distance-based clustering) with a complicated cost function (fractional-type) [14]. The complicated discrete problem is reformulated to an iterative algorithm including a step that solves an Ising problem. To minimize the communication overhead between the CPU and Ising machine, a low-latency implementation of SB was realized.

Concluding Remarks

Simulated bifurcation is a recently proposed, quantum-inspired, and highly parallelizable algorithm for combinatorial optimization. The high parallelism with massively parallel implementation technologies leads to high speed and scalability. The FPGA-based and GPU-based SB machines have been competitive against other cutting-edge Ising machines and have shown the highest performance for Ising problems with full connectivity. Massively parallel implementations of SB need many multiply-accumulators, large-capacity on-chip memory, and low-latency communication interfaces, and would best benefit from emerging nanodevices in conjunction with in-memory computing, stochastic computing, and cluster computing architectures. Integrating SB accelerators with other system components on a processing chip enables combinatorial optimization in real-time systems, and will offer new innovative applications.

References

[1] Hayato Goto, Kosuke Tatsumura, Alexander R. Dixon, "Combinatorial optimization by simulating adiabatic bifurcations in nonlinear Hamiltonian systems," Science Advances **5**, eaav2372, 2019. https://doi.org/10.1126/sciadv.aav2372

[2] Naeimeh Mohseni, Peter L. McMahon, Tim Byrnes, "Ising machines as hardware solvers of combinatorial optimization problems," Nature Reviews Physics, 2022. https://doi.org/10.1038/s42254-022-00440-8

[3] Hiroki Oshiyama, Masayuki Ohzeki, “Benchmark of quantum-inspired heuristic solvers for quadratic unconstrained binary optimization,” Scientific Reports **12**, 2146, 2022. https://doi.org/10.1038/s41598-022-06070-5

[4] Hayato Goto, "Bifurcation-based adiabatic quantum computation with a nonlinear oscillator network," Scientific Reports **6**, 21686, 2016. https://doi.org/10.1038/srep21686

[5] Hayato Goto, "Quantum Computation Based on Quantum Adiabatic Bifurcations of Kerr-Nonlinear Parametric Oscillators," Journal of the Physical Society of Japan **88**, 061015, 2019. https://doi.org/10.7566/JPSJ.88.061015

[6] Hayato Goto, Taro Kanao, "Chaos in coupled Kerr-nonlinear parametric oscillators," Physical Review Research **3**, 043196, 2021. https://doi.org/10.1103/physrevresearch.3.043196

[7] Hayato Goto, Kotaro Endo, Masaru Suzuki, Yoshisato Sakai, Taro Kanao, Yohei Hamakawa, Ryo Hidaka, Masaya Yamasaki, Kosuke Tatsumura, "High-performance combinatorial optimization based on classical mechanics," Science Advances **7**, eabe7953, 2021. https://doi.org//10.1126/sciadv.abe7953

[8] Kosuke Tatsumura, Alexander R. Dixon, Hayato Goto, "FPGA-Based Simulated Bifurcation Machine," Proc. of IEEE International Conference on Field Programmable Logic and Applications (FPL), pp. 59-66, 2019. https://doi.org/10.1109/FPL.2019.00019

[9] Kosuke Tatsumura, Masaya Yamasaki, Hayato Goto, "Scaling out Ising machines using a multi-chip architecture for simulated bifurcation," Nature Electronics **4**, pp. 208-217, 2021. https://doi.org/10.1038/s41928-021-00546-4

[10] Taro Kanao, Hayato Goto, “Simulated bifurcation assisted by thermal fluctuation,” arXiv preprint arXiv:2203.08361, 2022. https://doi.org/10.48550/arXiv.2203.08361 [to be published in Communications Physics]

[11] Kosuke Tatsumura, Ryo Hidaka, Masaya Yamasaki, Yoshisato Sakai, Hayato Goto, "A Currency Arbitrage Machine based on the Simulated Bifurcation Algorithm for Ultrafast Detection of Optimal Opportunity," Proc. of IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1-5, 2020. https://doi.org/10.1109/ISCAS45731.2020.9181114

[12] Kyle Steinhauer, Takahisa Fukadai, Sho Yoshida, “Solving the Optimal Trading Trajectory Problem Using Simulated Bifurcation,” arXiv preprint arXiv:2009.08412, 2020. https://doi.org/10.48550/arXiv.2009.08412

[13] Tingting Zhang, Qichao Tao, Jie Han, "Solving Traveling Salesman Problems Using Ising Models with Simulated Bifurcation," Proc. of International SoC Design Conference (ISOCC), pp.288-289, 2021. https://doi.org/10.1109/ISOCC53507.2021.9613918

[14] Nasa Matsumoto, Yohei Hamakawa, Kosuke Tatsumura, Kazue Kudo, "Distance-based clustering using QUBO formulations,” Scientific Reports **12**, 2669, 2022. https://doi.org/10.1038/s41598-022-06559-z

7.1 -Compute-in-Memory with Nanoscale CMOS Technologies

Amit Ranjan Trivedi, Saibal Mukhopadhyay, and Kaushik Roy

1 Status

Deep learning algorithms have shown that the growing volume and variety of data can be leveraged for highly accurate predictions and decision-making in many complex problems. A deep neural network (DNN) typically utilizes tens of thousands to millions of weights (i.e., model parameters). Operating over such a large parametric space facilitates the network with robust inductive biases. Yet, it also presents critical inference constraints for real-time or low power applications. Especially, DNN's extensive model size induces excessive memory accesses to read model weights from off-chip memories and to read/write operands to off-chip memory and intermediate memory hierarchy. Thus, on a conventional digital hardware, the inference performance of DNN succumbs to limited processor-memory bandwidth. A radical approach gaining attention to address the performance challenge is to design alternate non-von Neumann computing modules that can not only store model weights but also locally process the majority of inference operations within the same structure. Therefore, using such "compute-in-memory" processing of DNN, high volume data traffic between processor and memory units can be a verted. Compute-in-memory using conventional CMOS-based memory structures is especially more promising. Prior works have shown that CMOS-based conventional memory structures such as SRAM, DRAM, embedded-DRAM, SONOS, and NAND-Flash, *etc.*, can be adapted for compute-in-memory, thus enabling a rapid and cost-effective adoption of the scheme in commercial computing substrates.

In Figure 1, matrix-Vector multiplications (MVM) constitute the dominant computations in a DNN. To leverage CMOS memories for the storage of model weights and MVM computations, most compute-in-memory schemes employ a mixed-signal processing. Digital inputs to a DNN layer are converted to analog representation such as charge [1], current [2], or time [3]. The input vectors are loaded in parallel to the memory array where the memory cells multiply them with the stored weights in an analog fashion. The analog output of all memory cells within a column is summed to produce the output of MVM. Especially, the accumulation of products in many schemes simply reduces to current/charge summation over a wire, thus further minimizing the necessary workload. The analog MVM outputs are subsequently digitized for storage and routing to the other processing units.

Among early works on CMOS-based compute-in-memory, Zhang et. al. presented the processing using an array of standard six-transistor (6T) SRAM cells [2]. The resulting memory array, however, was vulnerable to instability under process variability. The challenges were resolved in [1] using 10T SRAM cells which separated ports for inference and write. While early adoptions of CMOS-based compute-in-memory focused on binary weights, the schemes were later enhanced for multibit processing to improve the accuracy of DNN. Detailed survey of compute-in-memory processors was compiled in [4].

2 Current and Future Challenges

Most compute-in-memory schemes employ a mixed-signal processing which raises critical challenges to integrate analog circuits such as analog-to-digital converter (ADC), digital-to-analog converter (DAC), and comparator within the memory structures. In Figure 1, using CONV-SRAM [1] as a motivating example, we highlight these limitations. To compute the inner product of *l*-element weight and input vectors **w** and **x**, *l*-DACs and one ADC are required. Since DACs are concurrently active, they lead to both high area and power. Since most memory modules are designed using advanced nanometer node CMOS technologies for energy and area efficiency, designing memory-integrated analog circuits at the same technology node is challenging. At such advanced technology nodes, the analog circuits are susceptible to failure under process variability and require complicated calibration processes. Due to such processing challenges of mixed-signal operations in computing-in-memory, a significant research in the past many years has focused on exploring alternatives to alleviate the implementation complexity, especially at advanced CMOS technology nodes.

In [5], time-domain DACs were used to replace analog circuits for DAC implementation. However, with increasing input precision, either operating time increases exponentially, or complex analog-domain voltage scaling is necessitated. All digital compute-in-memory processing with SRAM was shown for binary neural networks, e.g., in [6]. However, for more complex deep learning applications such as object detection and autonomous navigation, networks with binary-weighted inputs and weights have very low accuracy. The accuracy of SRAM-implemented binary networks was improved using supported-BinaryNet architecture in [7] and by leveraging peripherals DACs to implement the support parameters. A novel approach to mitigate the challenges of multi-bit inference with SRAM

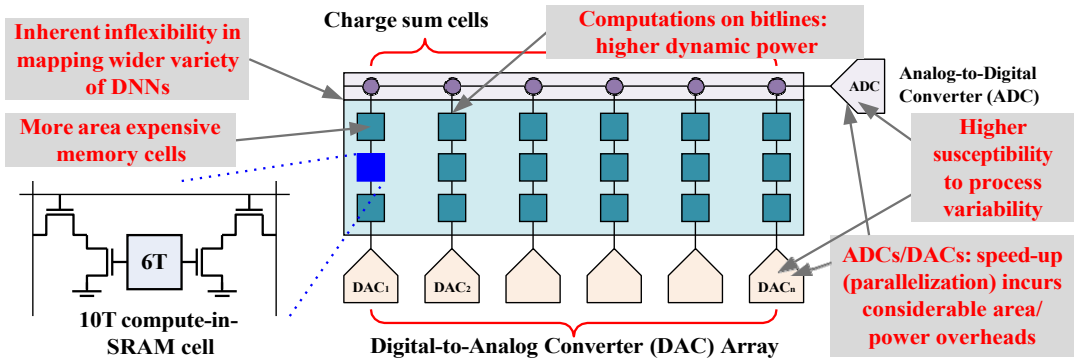

Figure 1: Overview of compute-in-SRAM and critical challenges being studied by the researchers.

was discussed in [8] by adapting deep learning's inference operator such that the multiplications between multibit precision weight and input vectors was not necessary. Novel adaptations of SRAM-based compute-in-memory were discussed where the memory structure was employed for non-classical inference schemes such as Markov chain Monte Carlo (MCMC) in [9] and Monte Carlo Dropout (MC-Dropout) in [10].

3 Advances in Science and Technology to Meet Challenges

Integrating computations and storage invariably demands more area per cell in compute-in-memory. Meanwhile, state-of-the-art DNNs continue to increase in model-size, thereby demanding higher energy and area-efficiency of the memory structures. In the future, several complementary efforts must be pursued in cohesion to improve the area efficiency of compute-in-memory. Compute-in-memory inference architectures that can robustly operate in more advanced CMOS nodes, such as 7 nm or below, will be imperative. Compute-in-memory in monolithic and vertically-integrated memory structures need to be pursued. Low and mixed precision DNNs, better suited for compute-in-memory processing, will be needed. Pruning and compression methods of DNN will be critical. Almost or completely digital architectures will be needed that maintaining multibit precision operations as well as the advantages of analog mode processing such as minimizing workload by exploiting physics for computations. In parallel, DNN architectures themselves are going through a dramatic evolution to improve their computational efficiency. In the last few years, novel layers such as inception, residual layers, dynamic gating, polynomial layers, self-attention, and Hypernetworks have been added to the repository of DNN building blocks. Therefore, a critical challenge for the next generation compute-in-memory accelerators for DNN is to exhibit high versatility in their processing flow for efficient mapping of these diverse DNN layers into hardware circuits. Especially, many emerging layers, unlike classical layers, simultaneously correlate multiple variables to enhance computational efficiency and representation capacity. Therefore, novel compute-in-memory schemes will be needed to map higher-order processing of the emerging layers within simplified cells.

References

[1] A. Biswas and A. P. Chandrakasan, “Conv-sram: An energy-efficient sram with in-memory dot-product computation for low-power convolutional neural networks,” *IEEE Journal of Solid-State Circuits*, vol. 54, no. 1, pp. 217–230, 2018.

[2] J. Zhang, Z. Wang, and N. Verma, “In-memory computation of a machine-learning classifier in a standard 6t sram array,” *IEEE Journal of Solid-State Circuits*, vol. 52, no. 4, pp. 915–924, 2017.

[3] M. Kang, S. K. Gonugondla, and N. R. Shanbhag, “Deep in-memory architectures in sram: An analog approach to approximate computing,” *Proceedings of the IEEE*, vol. 108, no. 12, pp. 2251–2275, 2020.

[4] N. Verma, H. Jia, H. Valavi, Y. Tang, M. Ozatay, L.-Y. Chen, B. Zhang, and P. Deaville, “In-memory computing: Advances and prospects,” *IEEE Solid-State Circuits Magazine*, vol. 11, no. 3, pp. 43–55, 2019.

[5] S. K. Gonugondla, M. Kang, and N. R. Shanbhag, “A variation-tolerant in-memory machine learning classifier via on-chip training,” *IEEE Journal of Solid-State Circuits*, vol. 53, no. 11, pp. 3163–3173, 2018.

[6] A. Agrawal, A. Jaiswal, D. Roy, B. Han, G. Srinivasan, A. Ankit, and K. Roy, “Xcel-ram: Accelerating binary neural networks in high-throughput sram compute arrays,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 66, no. 8, pp. 3064–3076, 2019.

[7] S. Nasrin, S. Ramakrishna, T. Tulabandhula, and A. R. Trivedi, “Supported-binarynet: Bitcell array-based weight supports for dynamic accuracy-energy trade-offs in sram-based binarized neural network,” in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2020, pp. 1–5.

[8] S. Nasrin, D. Badawi, A. E. Cetin, W. Gomes, and A. R. Trivedi, “Mf-net: Compute-in-memory sram for multibit precision inference using memory-immersed data conversion and multiplication-free operators,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 5, pp. 1966–1978, 2021.

[9] P. Shukla, A. Shylendra, T. Tulabandhula, and A. R. Trivedi, “Mc^{2}ram: Markov chain monte carlo sampling in sram for fast bayesian inference,” in *2020 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2020, pp. 1–5.

[10] P. Shukla, S. Nasrin, N. Darabi, W. Gomes, and A. R. Trivedi, “Mc-cim: Compute-in-memory with monte-carlo dropouts for bayesian edge intelligence,” *arXiv preprint arXiv:2111.07125*, 2021.

7.2– In-memory computing using non-volatile memories

I-Ting Wang[1], National Yang Ming Chiao Tung University (itwang@nycu.edu.tw)
Wang Kang[2], Beihang University (wang.kang@buaa.edu.cn)
Yao Zhu[3], Institute of Microelectronics (zhuya@ime.a-star.edu.sg)
Brajesh Kumar Kaushik[4], Indian Institute of Technology-Roorkee (brajesh.kaushik@ece.iitr.ac.in)

[1]Institute of Electronics, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
[2]School of Microelectronics, Beihang University, Beijing, China
[3]Institute of Microelectronics, Agency for Science, Technology and Research (A*STAR), Singapore
[4]Department of Electronics and Communication Engineering, Indian Institute of Technology-Roorkee, Uttarakhand, India

Status

In-memory computing (IMC) using emerging non-volatile memories (NVMs) has successfully opened up new opportunities for future computing paradigm. The NVMs, including resistive random access memory (RRAM), ferroelectric RAM (FRAM), and magnetoresistive RAM (MRAM), resemble artificial synapses with adjustable conductance as synaptic weight. In particular, NVM-based synaptic device is provided with computing and storage functions simultaneously, which eliminates the inefficient data movement between physically separated processor and memory units in conventional von Neumann computing architectures (Fig. 1 a) [1]. Besides, NVM-based IMC provides massively parallel computation in a crossbar configuration to further boost the computing efficiency. Therefore, an enormous amount of research has been devoted to developing NVM-based synaptic devices for building a high energy- and area-efficient computing hardware.

RRAM is the first found memristor that adjusts conductance through the migration of mobile ions and charged defects [2]. In addition, its inherently two-terminal structure ensures a compact and high-density synaptic array that simply performs the multiplication of the input signal and synaptic weight state with high computing parallelism in the neural network. Therefore, fruitful results from the device-, circuit-, and system-level demonstrations have been extensively presented [3].

The revival of FRAM has rapidly attracted increasing attention since the unprecedented discovery of ferroelectricity in hafnium oxide (HfO_2) [4], where the ferroelectric HfO_2 successfully solved the limitations of complementary metal-oxide-semiconductor (CMOS) process incompatibility and scalability in conventional perovskite oxides. In particular, the HfO_2-based ferroelectric field-effect transistor (FeFET) not only promises fast operating speed and low energy consumption due to the field-driven domain switching, it also provides stable multi-state with partially switched domains in the ferroelectric. These superior properties make HfO_2-based FRAMs stand out from the currently developed synaptic devices.

MRAM utilizes the spin property of electrons and holds the promise of low power, high speed and high endurance for in-memory computing. Very recently, implementation of in-memory computing artificial neural networks (ANNs) based on a crossbar array of MRAM has been demonstrated, offering a potential platform for downloading biological neuronal networks to mimic the brain [5]. After the commercialization of spin transfer torque (STT) driven MRAM (STT-MRAM), recently, the spin orbit torque (SOT) driven MRAM (SOT-MRAM) and the voltage controlled magnetic anisotropy (VCMA) driven MRAM (VCMA-MRAM) as well as the combination of the two effects, i.e., voltage controlled SOT (VC-SOT) driven MRAM (VC-SOT-MRAM), are under intensive investigations, targeting further reduction of power and latency.

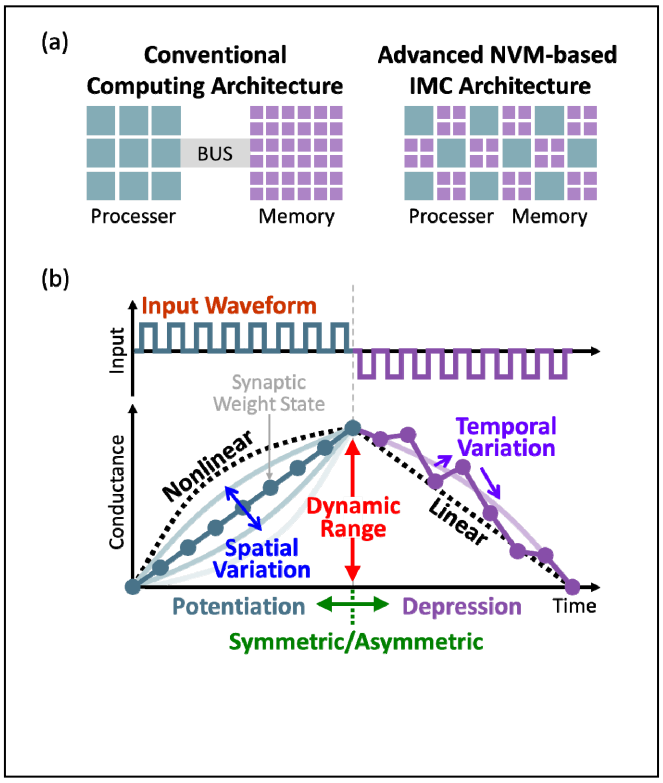

Figure 1. (a) Comparison between conventional computing architecture and NVM based IMC architecture, where the former suffers from the inefficient data transfer between physically separated processing and memory units while the latter realizes computing and storage functions in the same location to boosts the computing efficiency. (b) General behaviors in the NVM based synaptic device with bidirectionally adjustable synaptic weight states through the given input waveform. The non-ideal device properties such as nonlinear/asymmetric weight modulation and temporal/spatial variation are indicated accordingly.

Current and Future Challenges

Figure 1b summarizes the behaviors commonly found in the NVM-based synaptic devices. Generally speaking, an ideal synaptic device should have a bidirectional modulation (i.e., potentiation and depression) in synaptic states that fulfils the requirements of computing and storage. First, the synaptic states should be non-volatile and free from spatial variation among device-to-device. Second, the synaptic modulation should be linear and symmetric and without temporal variation from cycle-to-cycle. Besides, high endurance in synaptic modulation by identical input waveform is important not only to guarantee a sufficient training epochs but also prevent a time-consuming read-before-write process. Moreover, an ideal synaptic device should expand in an adequate dynamic range compatible with that of peripheral circuit because the higher conductance results in additional energy consumption while the lower increases sensing latency. However, even though a tremendous amount of work has been carried out for searching the holy grail, the truly ideal synaptic device is still lacking, which has left ample room for improvement and compromise.

Although RRAM-based synaptic device is relatively matured for realizing hardware neural network, it still inevitably suffers from non-ideal device properties, which significantly impact on accuracy degradation [6, 7]. Improvements and optimizations from material/device engineering are therefore important. However, adopting novel materials such as two-dimensional transition metal dichalcogenides (2D TMDs) [8] is still under scrutiny and thus lacks of statistic data to support its practicability. Moreover, selecting elements must be implemented to suppress the unwanted leakage current and interference from the unselected cells, but the process complexity is increased. Developing a self-selecting and self-rectifying synaptic device without the need of selecting element is still challenging.

As for the novel HfO_2-based FeFET, integrating the ferroelectric gate stack at font-end-of-line process compromises the write efficiency and performance. By adding an additional floating gate between ferroelectric and gate insulator, the FeMFET [9] and *m*-MFMFET [10] not only solve the above-mentioned issues, but they also provide back-end-of-line (BEOL) fabrication flexibility to ease the hardware design. However, scaling down the ferroelectric in both vertical (thickness) and horizontal (cell dimension) directions under BEOL-compatible process temperature while maintaining sufficient remanent polarization is still challenging. Besides, although FeFET-based synaptic device may improve the variability issue due to the relatively stable spontaneous polarization, achieving linear synaptic modulation using identical input waveform is still challenging.

Regarding MRAM, despite its practical advantages in power, endurance and technology maturity, the difficulty for high-performance IMC hardware stems from the low absolute resistance (~several Kohm) and the low on/off resistance ratio (~300%) of MRAM devices, which bring challenges in implementing large-scale multi-bit computing, e.g., analogue multiply–accumulate operations. Techniques from devices/circuits co-engineering to design new computing paradigms or architectures are therefore important.

Advances in Science and Technology to Meet Challenges

Although IMC with NVMs is promising for future computing paradigm, it still remains rooms for improvement. Several possible directions that we anticipate are described as follows:

First, continued optimization and innovation in material/device engineering are the keys, where the inevitably intrinsic variation and non-ideal device properties could be greatly improved. For instance, 2D TMDs with ferroelectricity are recently found and reported with promising scalability and reliability [11], which may shed some light on the hardware neural network. Besides, the recent discovery of aluminum scandium nitride (AlScN) with superior ferroelectric properties such as high remanent polarization and tightly distributed coercive field may become another relevant candidate [12]. Moreover, alternative device structures for different computing paradigms need to be explored. Skyrmionic MRAM devices for reservoir computing and probabilistic/stochastic computing are good examples to further exploit the device features [13].

Meanwhile, a neural network evaluating platform with holistic optimizations from device-, circuit-, and system-level for ANN design guideline and performance prediction is especially crucial to continuously take pre-emptive actions for constructing future computing hardware. A general standard for measuring the synaptic device is usually based on that for the NVMs [14]. However, their characteristics for application-specific criteria are found to be much relaxed for ANN applications [15]. A more practical evaluating methods suitable for the synaptic devices is thus required.

Finally, by leveraging the strengths in the NVMs and the matured CMOS components, developing novel computing architecture with both IMC and digital hardware designs may further relive the device requirements since the non-ideal properties in the NVM-based synaptic device unavoidably exist. By pulling together different devices with their superiorities, the hybrid architecture promises the best trade-off that is surely worth developing.

Concluding Remarks

In-memory computing with the emerging NVMs is gaining great momentum in research as the data-centric tasks no longer be affordable in conventional computing architectures. By taking advantage of the NVM crossbar array, the NVM-based IMC is promising for massively parallel computation, which successfully improves the computing efficiency. However, each NVM device has its own issues that are mostly attributed to the intrinsic and non-ideal device properties, and it therefore remains rooms for improvement. It is worth mentioning that to pursue a more practical IMC hardware, investigation involving with device, circuit, and system co-optimization is more desirable compare to that of merely

focusing on a single angle. Therefore, with these driving forces for resolving current challenges in multiple aspects, we anticipate that NVM-based IMC to be more energy- and area-efficient and continuously pave the way for leading-edge computing paradigm.

Acknowledgements

*This work was partially supported by the Science and Engineering Research Council of A*STAR (Agency for Science, Technology and Research) Singapore, under Grant No. A20G9b0135.*

References

[1] P. A. Merolla, J. V. Arthur, R. Alvarez-Icaza, A. S. Cassidy, J. Sawada, F. Akopyan, B. L. Jackson, N. Imam, C. Guo, Y. Nakamura, B. Brezzo, I. Vo, S. K. Esser, R. Appuswamy, B. Taba, A. Amir, M. D. Flickner, W. P. Risk, R. Manohar, and D. S. Modha, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, pp. 668–673, 2014.

[2] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams, "The missing memristor found," *Nature*, vol. 453, no. 7191, pp. 80–83, 2008.

[3] D. Ielmini and S. Ambrogio, "Emerging neuromorphic devices," *Nanotechnology*, vol. 31, 092001, 2019.

[4] T. S. Böscke, J. Müller, D. Bräuhaus, U. Schröder, and U. Böttger, "Ferroelectricity in hafnium oxide thin films," *Appl. Phys. Lett.*, vol. 99, 102903, 2011.

[5] S. Jung, H. Lee, S. Myung, H. Kim, S. K. Yoon, S.-W. Kwon, Y. Ju, M. Kim, W. Yi, S. Ham, B. Kwon, B. Seo, K. Lee, G.-H. Koh, K. Lee, Y. Song, C. Choi, D. Ham, and S. J. Kim, "A crossbar array of magnetoresistive memory devices for in-memory computing," *Nature*, vol. 601, no. 7892, pp. 211–216, Jan. 2022.

[6] P.-Y. Chen, B. Lin, I-T. Wang, T.-H. Hou, J. Ye, S. Vrudhula, J.-s. Seo, Y. Cao, and S. Yu, "Mitigating effects of non-ideal synaptic device characteristics for on-chip learning," in *Proc. Int. Conf. Comput. Aided Design (ICCAD)*, 2015, pp.194–199.

[7] C.-C. Chang, P.-C. Chen, T. Chou, I-T. Wang, B. Hudec, C.-C. Chang, C.-M. Tsai, T.-S. Chang, and T.-H. Hou, "Mitigating symmetric nonlinear weight update effects in hardware neural network based on analog resistive synapse," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems (JETCAS)*, vol. 8, no.1, pp. 116–124, 2017.

[8] Y. Shi, X. Liang, B. Yuan, V. Chen, H. Li, F. Hui, Z. Yu, F. Yuan, E. Pop, H.-S. P. Wong, and M. Lanza, "Electronic synapses made of layered two-dimensional materials," *Nature Electronics*, vol. 1, no. 8, pp. 458–465, 2018.

[9] K. Ni, J. A. Smith, B. Grisafe, T. Rakshit, B. Obradovic, J. A. Kittl, M. Rodder, and S. Datta, "SoC logic compatible multi-bit FeMFET weight cell for neuromorphic applications," in *IEDM Tech. Dig.*, 2018, pp. 296–299.

[10] M.-H. Yan, M.-H. Wu, H.-H. Huang, Y.-H. Chen, Y.-H. Chu, T.-L. Wu, P.-C. Yeh, C.-Y. Wang, Y.-D. Lin, J.-W. Su, P.-J. Tzeng, S.-S. Sheu, W.-C. Lo, C.-I Wu, and T.-H. Hou, "BEOL-compatible multiple metal-ferroelectric-metal (*m*-MFM) FETs designed for low voltage (2.5V), high density, and excellent reliability," in *IEDM Tech. Dig.*, 2020, pp. 75–78.

[11] K. C. Kwon, Y. Zhang, L. Wang, W. Yu, X. Wang, I.-H. Park, H. S. Choi, T. Ma, Z. Zhu, B. Tian, C. Su, and K. P. Loh, "In-plane ferroelectric tin monosulfide and its application in a ferroelectric analog synaptic device," *ACS Nano*, vol. 14, no. 6, pp. 7628–7638, 2020.

[12] S. Fichtner, N. Wolff, F. Lofink, L. Kienle, and B. Wagner, "AlScN: A III-V semiconductor based ferroelectric," *J. Appl. Phys.*, vol. 125, 114103, 2019.

[13] S. Li, W. Kang, X. Zhang, T. Nie, Y. Zhou, K. L. Wang, and W. Zhao, "Magnetic skyrmions for unconventional computing," *Mater. Hotiz.*, vol. 8, no. 3, pp. 854–868, 2021.

[14] M. Lanza, H.-S. P. Wong, E. Pop, D. Ielmini, D. Strukov, B. C. Regan, L. Larcher, M. A. Villena, J. J. Yang, L. Goux, Attilio Belmonte, Y. Yang, F. M. Puglisi, J. Kang, B. Magyari-köpe, E. Yalon, A. Kenyon, M. Buckwell, A. Mehonic, A. Shluger, H. Li, T.-H. Hou, B. Hudec, D. Akinwande, R. Ge, S. Ambrogio, J. B. Roldan, E. Miranda, J. Suñe, K. L. Pey, X. Wu, N. Raghavan, E. Wu, W. D. Lu, G. Navarro, W. Zhang, M. Liu, S. Long, Q. Liu, H. Lv, A. Padovani, P. Pavan, I. Valov, X. Jing, T. Han, K. Zhu, S. Chen, F. Hui, and Y. Shi, "Recommended methods to study resistive switching devices," *Adv. Electron. Mater.*, vol. 5, 1800143, 2019.

[15] C.-C. Chang, S.-T. Li, T.-L. Pan, C.-M. Tsai, I-T. Wang, T.-S. Chang, and T.-H. Hou, "Device quantization policy in variation-aware in-memory computing design," *Scientific Reports*, vol. 12, no. 112, pp. 1–12, 2022.

8.1- Computing with Dynamical Systems

Davi Röhe Rodrigues, Politecnico di Bari, Italy
Satoshi Sunada, Kanazawa University, Japan
Karin Everschor-Sitte, University of Duisburg-Essen, Germany
davi.rodrigues@poliba.it
sunada@se.kanazawa-u.ac.jp
karin.everschor-sitte@uni-due.de

Status

Dynamical systems are ubiquitous in nature and can be used for computing when the system's spatio-temporal dynamics is engineered to model a computational task. The use of dynamical systems for computer applications dates back to analog computers which were replaced in the mid-1990s by digital computers. In the last decades, digital computers offered a greater versatility, lower susceptibility to errors, and better scalability made possible by the rapid development of transistors [1]. Dynamical systems are used in a plethora of computing applications, including cellular automata, random-Boolean networks, Ising models, Lindenmayer systems and neuromorphic computing. In particular in the context of the latter dynamical systems offer several advantages over digitized circuits, and thus are making a comeback. Dynamical systems can be designed to have the required structural similarities of neural networks, such as hierarchy, approximate symmetries, memory, redundancy, and nonlinearity [2]. Therefore, they provide a natural hardware implementation of neural networks that overcomes the von Neumann bottleneck at much lower power consumption and higher scalability compared to transistor-based digital technology, which only artificially emulates the required properties. Moreover, analog information processing in dynamical systems allows to directly process sensor signals and, thus, offers energy efficient and low latency processing. The use of dynamical systems for computing has been fostered by recent advances in material science and, for example, ground-breaking discoveries in the field of photonics [3] and spintronics [4], [5], which have succeeded in developing low-power consuming proof-of-concept devices compatible with CMOS technology.

More specifically, recent proposals have shown that dynamical systems can be employed to emulate neuron synapses and firing for synaptic neural networks [6], [7], as well as to perform weight calculations [8] or completely substitute the hidden layers in a neural network [2], [9]–[11]. Two promising applications that consolidate the use of dynamical systems for spatio-temporal pattern recognition are reservoir computing [12]–[14] and physical neural networks [2], [10], [15]. While both computational paradigms allow to learn and extract patterns from data, they rely on different learning schemes. In reservoir computing, the training is performed only at the output level and, thus, requires a sufficiently complex response of the physical reservoir to distinguish small variations in the input. In contrast, in physical neural networks the training is performed directly on the physical system which, often implies the need for complex learning algorithms. In general, a physical realization of a reservoir computer benefits from a system with highly nonlinear dynamics with short-term memory [16], while a physical neural network requires a well-modelled nonlinear system with controllable dynamical parameters [2].

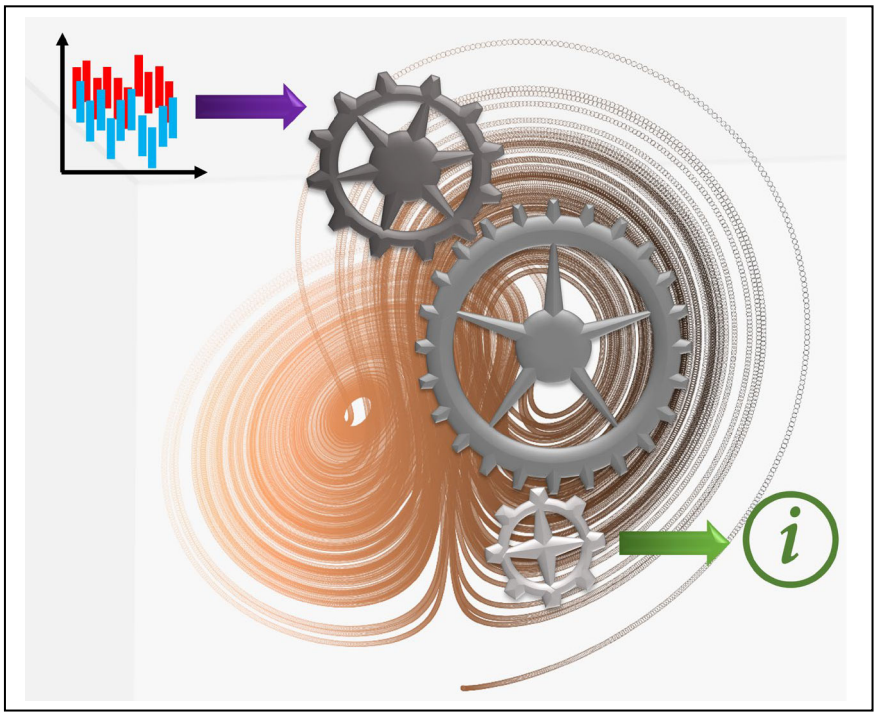

Figure 1.3 Computing with dynamical systems: A dynamical system computes based on the input signal and generates information as output.

Current and Future Challenges

Computing with dynamical systems typically implies the encoding of computer tasks into the functional response of the device. Thus, current and future challenges in this field are associated to design devices with tailored functionalities and to ensure the reliable encoding of information in terms of inputs and outputs of the dynamical system. While dynamical systems can be tailored to perform complex calculations with higher efficiency compared to digital computers, they often are a single purpose device with analog outputs. Designing dynamical systems as multipurpose devices is a key challenge. Additionally training and learning strategies must become more efficient when computing with dynamical systems. In the future, they need to extend beyond simple supervised learning models.

Furthermore, for the commercial implementation and widespread use of dynamical system-based computers, for example, in Internet of Things and Industry 4.0 applications as well as real-time computing, a major challenge is to produce devices that are scalable, inexpensive and have a significant better efficiency compared to CMOS technology. At least on intermediate time scales it is also necessary to be integrable into the market-dominating CMOS technology. This demands great efforts in material science and engineering in device development. Although the number of proof-of-concept devices is rapidly increasing, most of them still face major obstacles in their large-scale realization and production.

A fundamental issue concerns the general device design. It can be realized either with an assembled network of individual simple components, or directly with a large complex system. Both strategies have advantages and disadvantages and will most likely be used for different types of applications. For example, a large complex system is often less tunable, but it can circumvent the challenge of creating a highly dense connected network of individual components, which only within the network structure will allow for high computational performance.

Another great challenge is that so far computers based on dynamical systems do not fully work in-situ. For example, the input data need to be pre-processed by an external device to generate a significant non-linear response of the dynamical system, which typically reacts only to certain time and length scales [17], [18]. Analogously, dynamical systems often provide a continuous

set of outputs which then need to be interpreted or even learned (as in reservoir computing) by an external computer. An autonomously working device with a reliable map between the input, the functional response of the device, and the correct output is still missing.

Advances in Science and Technology to Meet Challenges

The field of computing with dynamic systems is steadily growing with many proposals to address the challenges, ranging from the development of novel computational algorithms [2], [15], [19]–[21] to efficient manufacturing techniques. The main goal is to develop an efficient algorithm-and-hardware codesign to fully exploit dynamical features assisted by state-of-the-art nanotechnologies. The interdisciplinary approach is of crucial importance and ensures knowledge transfer in particular within computer science, mathematics, biology and physics. A key example that connects all disciplines is the development of novel brain-inspired algorithms that are then transferred to computing in materials and devices.

The development of techniques to quantify, tailor and exploit the nonlinear response and short-term memory of dynamical systems provides means to reduce the complexity and energy cost of conventional learning algorithms and to approximate the biologically inspired behavior of the brain, which is optimized by evolution over millions of years. Structured material studies based on machine learning are used to achieve advances in material properties. A particularly successful example in material science is the recent progress in manufacturing heterostructures from 2d materials and metamaterials [4], [22]. These multiphysics systems with potentially different natural timescales offer a variety of physical properties that make them attractive as highly scalable and tunable platforms for novel unconventional computing schemes.

In addition, there are many approaches in device development to implement neuromorphic functionalities in dynamical systems including targeted studies to improve device topologies and device designs. The development of scalable multifunctional systems is crucial for the various design concepts, such as considering a network of coupled individual units or a single large complex system. Current proposals target the use of scalable systems based on time-delay structures [15], [23], spatial parallelism [24] of photonic systems, spin nano-oscillators, and patterned magnetic samples [9]. Furthermore, there are focused efforts to integrate as many functionalities as possible directly in-situ into the device. These include the implementation of new learning algorithms based on physical properties [15], [19] and the engineering of the functional response of the systems [10], [20]. Fully autonomous devices, that can perform both calculations and learning, however, will still demand great efforts.

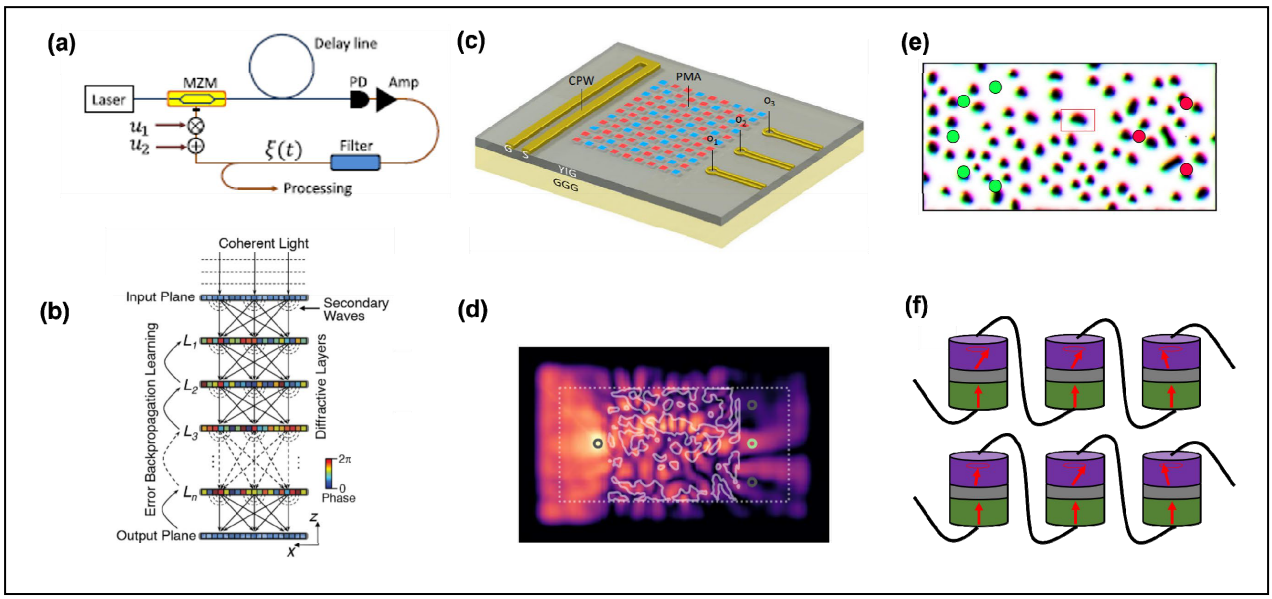

Figure 2. Examples of computing with dynamical systems. a) Optoelectronic delay system from Ref. [14], b) Diffractive deep neural network from Ref [23], c) Nanomagnet based spinwave scatterer from Ref. [11], d) Recurrent Neural network based on wave physics from Ref. [10] e) Reservoir Computing based on Skyrmion fabrics from Ref. [9], and f) Coupled spin torque nano oscillators proposed on Ref. [16].

Concluding Remarks

Computing with dynamical systems is an exciting field of research, which is experiencing a revival especially driven by the great advances in neuromorphic computing. The demand for efficient and scalable hardware implementations of neuromorphic systems, which can naturally be emulated in dynamic systems, is bringing the field from a niche to the forefront of research. Rapidly advancing major developments in material science promise low-cost, easy-to-manufacture and highly efficient task-oriented devices in the future. The ever-expanding capabilities to directly manipulate physics at the nanoscale and at ultra-high frequencies expand the possibility of employing physical principles for novel algorithms and computational schemes that fully embody the functionalities of the brain.
In summary, computing with dynamical systems has the potential to overcome the high-power consumption as well as the scalability limitations imposed by current CMOS technology, and to actively shape the development of Industry 4.0 and the Internet of Things.

Acknowledgements

[Please include any acknowledgements and funding information as appropriate.]
We thank Jake Love for discussions. DRR acknowledges funding from the Ministerio dell'Università e della Ricerca, Decreto Ministeriale n. 1062 del 10/08/2021 (PON Ricerca e Innovazione). SS acknowledges supports from JSPS KAKENHI (20H04255) and JST PRESTO (JPMJPR19M4). KES acknowledges funding from the German Research Foundation (DFG) Project No. 320163632 and the Emergent AI Center funded by the Carl-Zeiss-Stiftung.

References

[1] B. J. Maclennan, "A Review of Analog Computing," pp. 1–45, 2007, [Online]. Available: www.cs.utk.edu/~mclennan.

[2] L. G. Wright *et al.*, "Deep physical neural networks trained with backpropagation," *Nat. 2022 6017894*, vol. 601, no. 7894, pp. 549–555, Jan. 2022, doi: 10.1038/s41586-021-04223-6.

[3] G. Wetzstein *et al.*, “Inference in artificial intelligence with deep optics and photonics,” *Nature*, vol. 588, no. 7836. pp. 39–47, 2020, doi: 10.1038/s41586-020-2973-6.
[4] C. Felser, G. H. Fecher, and B. Balke, “Spintronics: A challenge for materials science and solid-state chemistry,” *Angewandte Chemie - International Edition*, vol. 46, no. 5. John Wiley & Sons, Ltd, pp. 668–699, Jan. 22, 2007, doi: 10.1002/anie.200601815.
[5] J. M. Hu, C. G. Duan, C. W. Nan, and L. Q. Chen, “Understanding and designing magnetoelectric heterostructures guided by computation: Progresses, remaining questions, and perspectives,” *npj Comput. Mater.*, vol. 3, no. 1, pp. 1–21, May 2017, doi: 10.1038/s41524-017-0020-4.
[6] L. Chua, “Memristor, Hodgkin–Huxley, and Edge of Chaos,” *Nanotechnology*, vol. 24, no. 38, p. 383001, Sep. 2013, doi: 10.1088/0957-4484/24/38/383001.
[7] D. Markovi ć A. Mizrahi, D. Querlioz, and J. Grollier, “Physics for neuromorphic computing,” *Nat. Rev. Phys.*, vol. 2, no. 9, pp. 499–510, Sep. 2020, doi: 10.1038/s42254-020-0208-2.
[8] L. Mazza *et al.*, “Computing with Injection-Locked Spintronic Diodes,” *Phys. Rev. Appl.*, vol. 17, no. 1, p. 014045, Jan. 2022, doi: 10.1103/PhysRevApplied.17.014045.
[9] D. Pinna, G. Bourianoff, and K. Everschor-Sitte, “Reservoir Computing with Random Skyrmion Textures,” *Phys. Rev. Appl.*, vol. 14, no. 5, p. 054020, Nov. 2020, doi: 10.1103/PHYSREVAPPLIED.14.054020/FIGURES/8/MEDIUM.
[10] T. W. Hughes, I. A. D. Williamson, M. Minkov, and S. Fan, “Wave physics as an analog recurrent neural network,” *Sci. Adv.*, vol. 5, no. 12, Dec. 2019, doi: 10.1126/SCIADV.AAY6946/SUPPL_FILE/AAY6946_SM.PDF.
[11] Á. Papp, W. Porod, and G. Csaba, “Nanoscale neural network using non-linear spin-wave interference,” *Nat. Commun.*, vol. 12, no. 1, p. 6422, Dec. 2021, doi: 10.1038/s41467-021-26711-z.
[12] J. Dambre, D. Verstraeten, B. Schrauwen, and S. Massar, “Information Processing Capacity of Dynamical Systems,” *Sci. Reports 2012 21*, vol. 2, no. 1, pp. 1–7, Jul. 2012, doi: 10.1038/srep00514.
[13] M. Lukoševi čus and H. Jaeger, “Reservoir computing approaches to recurrent neural network training,” *Comput. Sci. Rev.*, vol. 3, no. 3, pp. 127–149, Aug. 2009, doi: 10.1016/J.COSREV.2009.03.005.
[14] G. Tanaka *et al.*, “Recent advances in physical reservoir computing: A review,” *Neural Networks*, vol. 115, pp. 100–123, Jul. 2019, doi: 10.1016/J.NEUNET.2019.03.005.
[15] G. Furuhata, T. Niiyama, and S. Sunada, “Physical Deep Learning Based on Optimal Control of Dynamical Systems,” *Phys. Rev. Appl.*, vol. 15, no. 3, p. 034092, Mar. 2021, doi: 10.1103/PHYSREVAPPLIED.15.034092/FIGURES/9/MEDIUM.
[16] G. Bourianoff, D. Pinna, M. Sitte, and K. Everschor-Sitte, “Potential implementation of reservoir computing models based on magnetic skyrmions,” *AIP Adv.*, vol. 8, no. 5, p. 055602, Jan. 2018, doi: 10.1063/1.5006918.
[17] J. Torrejon *et al.*, “Neuromorphic computing with nanoscale spintronic oscillators,” *Nat. 2017 5477664*, vol. 547, no. 7664, pp. 428–431, Jul. 2017, doi: 10.1038/nature23011.
[18] C. Mead, “Neuromorphic Electronic Systems,” *Proc. IEEE*, vol. 78, no. 10, pp. 1629–1636, 1990, doi: 10.1109/5.58356.
[19] E. Martin *et al.*, “EqSpike: spike-driven equilibrium propagation for neuromorphic implementations,” *iScience*, vol. 24, no. 3, 2021, doi: 10.1016/j.isci.2021.102222.
[20] M. Hermans, M. Burm, T. Van Vaerenbergh, J. Dambre, and P. Bienstman, “Trainable hardware for dynamical computing using error backpropagation through physical media,” *Nat. Commun.*, vol. 6, pp. 1–8, 2015, doi: 10.1038/ncomms7729.
[21] D. R. Rodrigues, K. Everschor-Sitte, S. Gerber, and I. Horenko, “A deeper look into

natural sciences with physics-based and data-driven measures," *iScience*, vol. 24, no. 3, p. 102171, Mar. 2021, doi: 10.1016/j.isci.2021.102171.

[22] W. Ma, F. Cheng, and Y. Liu, "Deep-Learning-Enabled On-Demand Design of Chiral Metamaterials," *ACS Nano*, vol. 12, no. 6, pp. 6326–6334, Jun. 2018, doi: 10.1021/acsnano.8b03569.

[23] A. Uchida *et al.*, "Compact reservoir computing with a photonic integrated circuit," *Opt. Express, Vol. 26, Issue 22, pp. 29424-29439*, vol. 26, no. 22, pp. 29424–29439, Oct. 2018, doi: 10.1364/OE.26.029424.

[24] X. Lin *et al.*, "All-optical machine learning using diffractive deep neural networks," *Science (80-.).*, vol. 361, no. 6406, pp. 1004–1008, Sep. 2018, doi: 10.1126/science.aat8084.

8.2– Computing with Ising Machines realized through coupled nano-oscillators

Vito Puliafito,
Politecnico di Bari, via E. Orabona 4, 70125 Bari, Italy, vito.puliafito@poliba.it
Johan Åkerman,
University of Gothenburg, Fysikgränd 3, 41296 Göteborg, Sweden, johan.akerman@physics.gu.se
Hiroki Takesue,
NTT Corporation, 3-1 Morinosato-Wakamiya, Atsugi, Kanagawa, 243-0198 Japan, hiroki.takesue@ntt.com

Status

The solution of Combinatorial Optimization Problems (COPs) is currently of great interest for industrial applications, especially considering problems which are NP-hard, and their complexity scales exponentially with the number of the variables defining the phase space.

Ising Machines (IMs) are hardware solutions for the minimization of the cost function defined by the Ising model. This model describes the dynamics of spins (عى1) through the fol lowing Hamiltonian:

$$\text{طي سي} \quad , \quad \text{طي} \quad , \quad 1$$
$$, \quad \text{سي} 1 \qquad \text{سي} 1$$

where is the matrix of coupling among the spins and is a local bias field. This research field is important because the minimization of is NP-hard and several COPs with direct impact in logistics, manufacturing, financial management and artificial intelligence can be mapped into Ising model [1], [2].

Several physical approaches have been used for implementing IMs and can be broadly divided into two categories: annealers and dynamical solvers. The former are physical systems that can reach the minimization of their energy (corresponding to the minimization of *IH*) by means of a gradual decrease of the temperature, through different thermal equilibrium states. They have been realized with optical systems, magnetic devices [3], memristors [4], CMOS circuits, and FPGAs, to cite a few. Dynamical solvers are characterized by a temperature-independent evolution towards the minimization state where a supplementary annealing process can speed up the process. They are mostly based on the coupling of oscillators, giving rise to the so-called Oscillator-based IMs (OIMs), and implementations include analog electronic [5], [6], integrated CMOS [7], [8], VO_2-based [9], [10], spintronic [11], [12], spinwave [13] and optical coupled oscillators [14]–[19] (see Fig. 1 for a few examples).

The most important aspect of OIMs is their scalability, which can guarantee the possibility to solve COPs with large number of spins densely connected. The investigation of this aspect can be easily performed through software solvers for the corresponding Ising models, which are of great support for a practical use of hardware IMs on the market.

Here, we concentrate on the two more promising and unconventional solutions of OIMs, those based on spintronic and optical oscillators. From a theoretical point of view, those machines can be simulated in software by using the well-established Kuramoto model. Calculations show a great potential in creating arrays as large as million nodes.

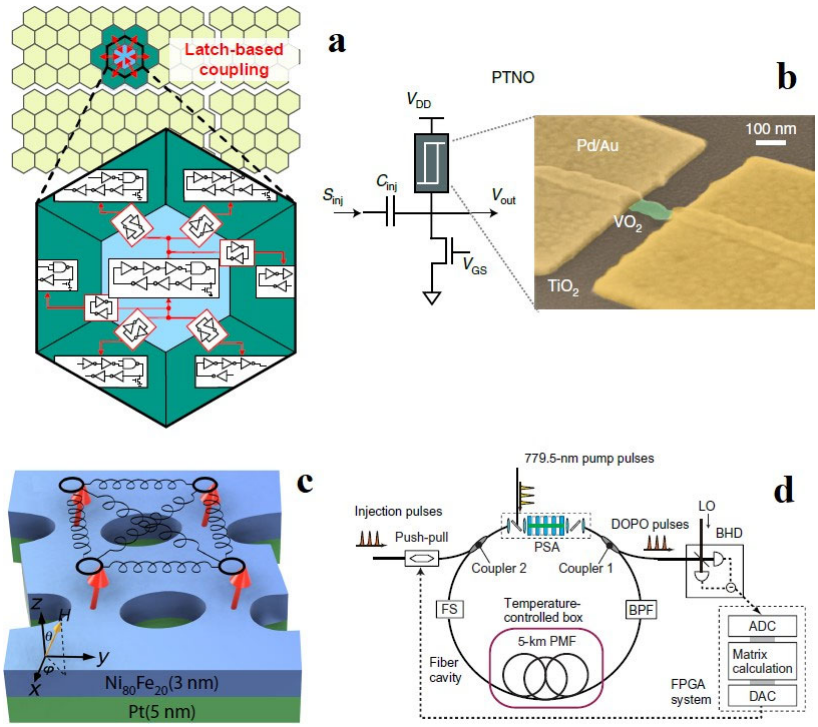

Figure 1.2 Sketches of different implementations for Ising Machines realized through coupled oscillators of different types: a) Exagonal ring oscillators [8] b) phase transition oscillators based on VO_2 [10] c) nanoconstriction based spin Hall nano oscillators [12] d) optical parametric oscillators [19].

Current and Future Challenges

IMs have been studied and tested to challenge the most important limits of conventional computing, such as computational time, scalability and high integration, and possibility to approach the optimal solution of a large size COP, with a particular reference to the Max Cut Problem (MCP).

Spin-torque and spin Hall nano-oscillators (STNOs and SHNOs) have an attractive combination of properties, such as easy tunability, GHz frequency operation, and nanoscale size. They have been proposed for realizing OIMs in a theoretical approach making use of a universal model for a non-linear oscillator where sub-harmonic injection locking (SHIL) is implemented [11]. More recently, an experimental demonstration of a 2x2 array of nano-constriction SHNOs was realized showing binarization of their phases (fig. 1c) [12]. In the former study, the probability to solve a MCP remains very close to 98% up to about 180 nodes in Mobius graph, whereas in the latter, an estimation of 5000 SHNOs highlighted better properties with respect to a reference quantum solution. Spintronic oscillators, therefore, are very promising, but practical large-scale coupling between them requires additional development if complete and programmable all-to-all connections are required.

The use of degenerate optical parametric oscillators is the key-point of coherent IMs (CIMs), where each spin is encoded in the phase of light in an optical mode and oscillators are either in-phase or out-of-phase with respect to pump light (fig. 1d) [14], [15]. Spin connections can be realized through a network of optical delay lines, but a solution for a fully programmable all-to-all connections has been realized through an architecture that uses measurement-feedback [16], [17]. In this case, a Mobius graph of 100 nodes has been used to test the MCP with a 21% of success probability [16], and fair solutions have been obtained for 2000-node MCPs [17]. CIMs have been compared to D-wave quantum annealers showing a more efficient performance in case of dense COPs [18]. More recently, the MCP for a huge number of 100000-node graph has been solved with a CIM providing very good solution, comparable with those obtained by standard algorithms and annealers, in a shorter time to solution [19].

Software approaches take advantage of analytical models for the oscillators used in OIMs to predict their properties. The most famous model of mutually coupled oscillators was defined by Kuramoto [20]. It predicts that stability occurs when the phase difference between the oscillators is 0 or π, which can be obtained through an external signal at double frequency, introducing what was later called SHIL. The model has been developed and tested for solving COPs as well as for image processing, and it has been used as a reference for realizing physical implementations [5]. It has been tested for problem sizes ranging from 800 to 3000. Challenges include the possibility to extend the model for oscillators with frequency-phase coupling and additive annealing techniques to speed up the time-to-solution.

Advances in Science and Technology to Meet Challenges

SHNOs have been demonstrated down to 20 nm, can operate at about 100 uA of current and 26 GHz, have been mutually synchronized in two-dimensional arrays of up to 64 oscillators, and individual SHNOs and their coupling to nearest neighbours can be controlled by voltage gates. While all these numbers need further improvement, the most fundamental limitation is the planar topology of nearest-neighbour coupling, which must be overcome. However, it can be shown that if next-nearest neighbour interactions are included, with or without control, e.g. along the diagonal in square arrays, this results in a non-planar topology. Fundamental research and experimental demonstrations in this direction are therefore needed.

CIMs seem to have a great potentiality to solve graphs of larger and larger number of nodes. Here, an important future challenge is to clarify how quantum nature of degenerate optical parametric oscillators contributes to the computational performance of CIMs.

The optimization of annealing techniques will be the advancement required not only for hardware IMs but also for algorithms. The latter ones will take advantage also from parallelization methods, while the inclusion of phase-power coupling in the model should be investigated to optimize the time to solution.

Concluding Remarks

IMs are on the crest of the wave nowadays, and they will surf it for the next years. Many solutions are on the table, and it is not trivial to compare them due to the several requested properties, such as a large number of nodes and a short time to solution, to cite the most important ones. In this scenario, IMs based on coupled oscillators guarantee advancements in technology and wide research activity in the upcoming future.

Acknowledgements

This work was also supported by Project No. PRIN 2020LWPKH7 funded by the Italian Ministry of University and Research, the Swedish Research Council Framework Grant no. 2016-05980, and the Horizon 2020 research and innovation programme (ERC Advanced Grant No.~835068 "TOPSPIN").

References

[1] S. Rudich and A. Wigderson, *Computational complexity theory*. American Mathematical Soc., 2004.

[2] N. Mohseni, P. L. McMahon, and T. Byrnes, “Ising machines as hardware solvers of combinatorial optimization problems,” *Nat. Rev. Phys.*, pp. 1–23, May 2022, doi: 10.1038/s42254-022-00440-8.

[3] W. A. Borders, A. Z. Pervaiz, S. Fukami, K. Y. Camsari, H. Ohno, and S. Datta, “Integer factorization using stochastic magnetic tunnel junctions,” *Nature*, vol. 573, no. 7774, pp. 390–393, Sep. 2019, doi: 10.1038/s41586-019-1557-9.

[4] F. Cai *et al.*, “Power-efficient combinatorial optimization using intrinsic noise in memristor Hopfield neural networks,” *Nat. Electron.*, vol. 3, no. 7, pp. 409–418, Jul. 2020, doi: 10.1038/s41928-020-0436-6.

[5] T. Wang and J. Roychowdhury, “OIM: Oscillator-based Ising Machines for Solving

Combinatorial Optimisation Problems," Mar. 2019, [Online]. Available: http://arxiv.org/abs/1903.07163.

[6] J. Vaidya, R. S. Surya Kanthi, and N. Shukla, "Creating electronic oscillator-based Ising machines without external injection locking," *Sci. Rep.*, vol. 12, no. 1, p. 981, Dec. 2022, doi: 10.1038/s41598-021-04057-2.

[7] M. K. Bashar, A. Mallick, D. S. Truesdell, B. H. Calhoun, S. Joshi, and N. Shukla, "Experimental Demonstration of a Reconfigurable Coupled Oscillator Platform to Solve the Max-Cut Problem," *IEEE J. Explor. Solid-State Comput. Devices Circuits*, vol. 6, no. 2, pp. 116–121, Dec. 2020, doi: 10.1109/JXCDC.2020.3025994.

[8] I. Ahmed, P. W. Chiu, W. Moy, and C. H. Kim, "A Probabilistic Compute Fabric Based on Coupled Ring Oscillators for Solving Combinatorial Optimization Problems," *IEEE J. Solid-State Circuits*, vol. 2, pp. 2019–2020, 2021, doi: 10.1109/JSSC.2021.3062821.

[9] N. Shukla *et al.*, "Synchronized charge oscillations in correlated electron systems," *Sci. Rep.*, vol. 4, pp. 1–6, 2014, doi: 10.1038/srep04964.

[10] S. Dutta *et al.*, "An Ising Hamiltonian solver based on coupled stochastic phase-transition nano-oscillators," *Nat. Electron.*, vol. 4, no. 7, pp. 502–512, Jul. 2021, doi: 10.1038/s41928-021-00616-7.

[11] D. I. Albertsson, M. Zahedinejad, A. Houshang, R. Khymyn, J. Åkerman, and A. Rusu, "Ultrafast Ising Machines using spin torque nano-oscillators," *Appl. Phys. Lett.*, vol. 118, no. 11, p. 112404, Mar. 2021, doi: 10.1063/5.0041575.

[12] A. Houshang *et al.*, "Phase-Binarized Spin Hall Nano-Oscillator Arrays: Towards Spin Hall Ising Machines," *Physical Review Applied*, vol. 17, no. 1. 2022, doi: 10.1103/PhysRevApplied.17.014003.

[13] A. Litvinenko *et al.*, "A spinwave Ising machine," Sep. 2022, [Online]. Available: http://arxiv.org/abs/2209.04291.

[14] A. Marandi, Z. Wang, K. Takata, R. L. Byer, and Y. Yamamoto, "Network of time-multiplexed optical parametric oscillators as a coherent Ising machine," *Nat. Photonics*, vol. 8, no. 12, pp. 937–942, Dec. 2014, doi: 10.1038/nphoton.2014.249.

[15] T. Inagaki, K. Inaba, R. Hamerly, K. Inoue, Y. Yamamoto, and H. Takesue, "Large-scale Ising spin network based on degenerate optical parametric oscillators," *Nat. Photonics*, vol. 10, no. 6, pp. 415–419, Jun. 2016, doi: 10.1038/nphoton.2016.68.

[16] P. L. McMahon *et al.*, "A fully programmable 100-spin coherent Ising machine with all-to-all connections," *Science (80-.).*, vol. 354, no. 6312, pp. 614–617, Nov. 2016, doi: 10.1126/science.aah5178.

[17] T. Inagaki *et al.*, "A coherent Ising machine for 2000-node optimization problems," *Science (80-.).*, vol. 354, no. 6312, pp. 603–606, Nov. 2016, doi: 10.1126/science.aah4243.

[18] R. Hamerly *et al.*, "Experimental investigation of performance differences between coherent Ising machines and a quantum annealer," *Sci. Adv.*, vol. 5, no. 5, pp. 1–11, May 2019, doi: 10.1126/sciadv.aau0823.

[19] T. Honjo *et al.*, "100,000-spin coherent Ising machine," *Sci. Adv.*, vol. 7, no. 40, Oct. 2021, doi: 10.1126/sciadv.abh0952.

[20] Y. Kuramoto, "Self-entrainment of a population of coupled non-linear oscillators," in *International Symposium on Mathematical Problems in Theoretical Physics*, Berlin/Heidelberg: Springer-Verlag, pp. 420–422.

9.1–Brain-Inspired Unconventional Computing

Jennifer Hasler, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 303332, USA. Jennifer.hasler@ece.gatech.edu
Samiran Ganguly, Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22904, USA. sg7wr@virginia.edu
Avik Ghosh, Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22904, USA. ag7rq@virginia.edu
William Levy, Department of Neurological Surgery, University of Virginia, Charlottesville, VA 22904, USA. wbl@virginia.edu
Vwani Roychowdhury, Department of Electrical Engineering, University of California at Los Angeles, CA 90095, USA. vwani@ee.ucla.edu
Supriyo Bandyopadhyay, Department of Electrical and Computer Engineering, Virginia Commonwealth University, Richmond, VA 23284, USA. sbandy@vcu.edu

Status. *[(350 words of 400 max)]*

[This section provides a brief history and status, why the field is still important, what will be gained with further advances. (400 words max)]

The first roadmap to develop a human cortex and a human brain appeared nearly a decade ago [1] arising from the neuromorphic tradition using close connections between biological and synthetic (e.g., Si CMOS) hardware (e.g. [2]). Neuromorphic and analog (Fig. 1) physical computing techniques, computing through real-valued representations [3], together developed energy efficient computation. Mead's hypothesis (1990) [4] predicted a 1000x factor improvement in energy efficiency by replacing digital with analog computation, as well as more improved factors using neuromorphic techniques. The 1000x improvement in analog computing has been well established (e.g. Fig. 1) over the last two decades (e.g. [5]), enabled through the first learning crossbar arrays (1994-1995, Single-transistor learning synapses [6]) employing analog programmable and adaptable CMOS Floating-Gate (FG) devices (Fig. 1). These techniques enabled end-to-end (Fig. 1), sensor-to-refined or classified output signal in custom or reconfigurable hardware (e.g. large-scale Field Programmable Analog Arrays (FPAA) [5]) innovating Computing-in-memory (2001, [7]) and computing in reconfigurable routing (e.g. [5]). FPAA scaling (e.g. [8]) follows a consistent path with predicted roadmap scaling [1].

Although the roadmap of building human cortex with a careful mapping between biological and Si CMOS physics is still as relevant today as when published a decade ago, including the capability of building human cortex in existing production technologies in $1m^2$ in under 100W [1], neuromorphic techniques still have many untapped opportunities, primarily in using the highly energy efficient temporal encoding of events with parallels to engineering applications. Hardware demonstrations show tight neuroscience models of biological channels (Fig. 1), computing and learning synapses (Fig. 1), dendrites, and neurons, as well as networks (Fig. 1) of these devices (e.g. [1]); additional approaches whether in CMOS or in other technologies that integrate with Si add to the available opportunities. A lower bounds on the energy efficiency improvement through using neuromorphic techniques based on neuron energy consumption is roughly 100,000x improvement over established analog computing techniques [1]. The challenge is unlocking these opportunities using the efficient use of temporal neural encoding as well as modelling and computationally abstracting the fundamental computations of 100s to 10,000s of cortical neurons (e.g. neural columns).

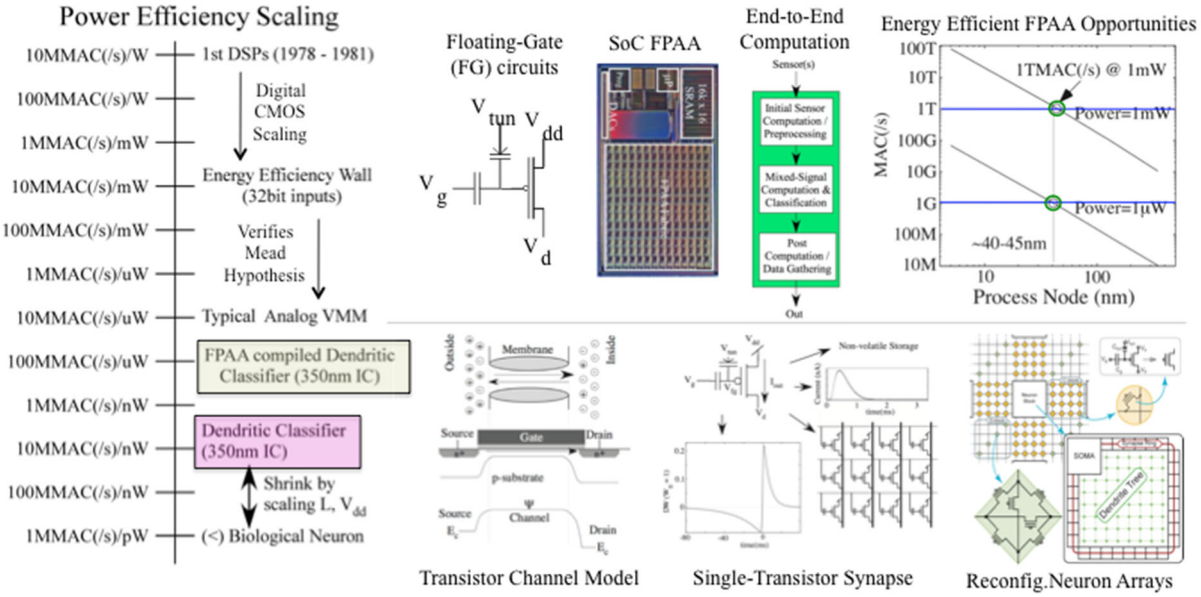

Figure 1: Brain inspired computing enables energy efficient capabilities, first enabling programmable and configurable analog computation, and then enabling neuromorphically inspired computing starting at similarities of biological channels into even more energy efficient computation. The improvement (1000xt in digital synchronous CMOS transistor scaling energy efficiency from the first DSP devices (1978) to our current limits is roughly similar to the improvements in analog computing as first predicted by Mead (1990), demonstrated frequently for nearly two decades, and will continue to improve with CMOS transistor scaling. Neuromorphic computing enables potentially an 100,000x or further improvement over these analog computing techniques, where existing CMOS implementations of channel implementations, synapses, and neuron arrays have been demonstrated. A few initial techniques demonstrate the potential opportunities.

Current and Future Challenges

[This section discusses the big research issues and challenges. (425 of 400 words max)]

Building energy efficient brain-like neural systems requires rethinking models of neural computation utilizing the timing of neural events rather than simply encoding values in number of spikes, not just in single neurons, but in networks of thousands of neuron components (Fig. 2). Transmitting events between neurons is energetically expensive as compared to other operations (dendritic & synaptic processes) [1]; each event arises from a large number of physical computing processes. The issue is primarily in developing computing and learning algorithms in bidirectionally connected neuron layers to utilize the already existing physical computing capabilities.

Learning and computation requires using neural architectures that might include input sensors refining their input data (Fig. 2), processing through bidirectional layers of neurons, and a hippocampus layer (and related layers) for encoding and training. Hippocampus is a dense, relatively small, recurrent network of neurons enabling sequence-processing [9,10] that integrates events from a number of cortical layers to create a short-term sophisticated multi-sensory spatial "World Model" for the organism (Fig. 2). During awake activity, sensory signals are processed through subcortical layers in the cortex and the refined outputs reach the hippocampus. During the sleep cycle, these memory events are replayed to the neocortex where sensory signals cannot disrupt the playback. During sleep, hippocampal interactions strengthen the memory representations in the neocortex by strengthening some synapses and even establishing new synapses. By efficiently encoding new concepts as a composition of already encoded sub-concepts, these techniques are a radical departure from typical Artificial Neural Networks (ANN), the core concept in traditional Machine Learning (ML), and potentially can reduce the effect of *Catastrophic forgetfulness* [11,12] exhibited in Deep Learning where incremental learning (as new sensory inputs arrive) leads to an indiscriminate erasure of old memory. While Deep Learning hardcodes memory to solve one-off tasks such as classification or generation of specific datasets, and additional hardware units are added for each new scenarios [13], neural memory approaches allow to learn incrementally and delete memory when needed.

Expansion of learning mechanisms to include the growth of neurons (neurogenesis), synapses (synaptogenesis), and dendrites (dendritogenesis) remain a challenge in physical hardware. Although reconfigurable systems can provide additional resources to bring into a computation, including new hardware incorporated into existing networks, biological networks seem to directly enable growing new hardware. New synaptic connections (***synaptogenesis***) between two neurons are formed based on averaged event dynamics between those two neurons and the resulting local dendritic activity in the receiving neuron. The growing of new dendrites (***dendritogensis***) works to further enable that neuron to fire given the space of axons in its area.

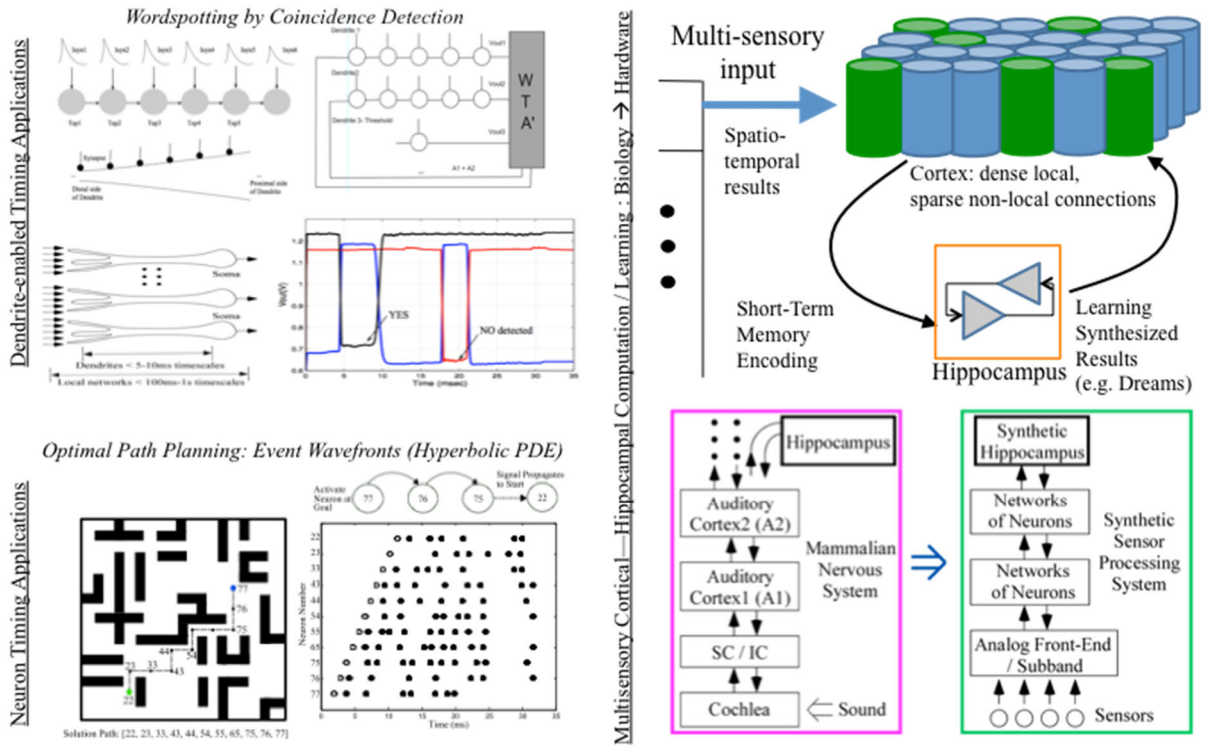

Figure 2: Neural computation requires spatio-temporal processing involving many modes of processing neuron output events that includes event-timing between neurons (e.g. optimal path planning) as well as event-timing (e.g. coincidence detection) within neuron dendrite (e.g. wordspotting). Timing is utilized throughout the neural infrastructure and sensory inputs, between the bidirectional computation of cortical layers, as well as with hippocampus. The cortex--hippocampus architecture performs multi-sensory data fusion by constructing meaningful semantics from episodic memories. Integrating this entire computing and learning biological model into a synthetic system becomes the challenge for next-generation energy-efficient neuromorphic systems.

Advances in Science and Technology to Meet Challenges

[This section discusses the advances in science technology needed to address the challenges. (356 words of 400 max)]

Building efficient event-timing networks becomes the primary challenge for synthetic neuromorphic systems. One example uses event-timing for predicting optimal paths through an array of neurons, where an optimal path is found by the first arriving events in a polynomial resource algorithm (Fig. 2). Another example uses coincidence detection of event timing in efficient dendritic processes between a cluster of dendritic-enabled neurons (Fig. 2). These techniques require physical computation to efficiently model the ordinary differential equations (ODEs) as well as the dendritic partial differential equations (PDEs) of cortical neurons (e.g. pyramidal cells) that includes the large dendritic arborization and as well as networks (e.g. cortical columns) of these pyramidal cells [1]. Both of these techniques result in significantly higher energy efficient computations (Fig. 1) compared to analog CMOS operations.

Event-timing networks will require additional techniques to handle the range of timescales and morphological changes. Neurobiological systems operate over many orders of magnitude in timing and the learning on that timing, and utilize structures like glial (and other) cells for the timing modulation. These challenges will be enabled by scaling existing design techniques to cutting edge CMOS IC processes (e.g. 7 and 10nm) as well as developing high-level code based synthesis techniques, issues that are resource, rather than technology, constrained.

Approaches that expand the capability of standard CMOS enable opportunities for even greater energy efficiencies. Nano-magnets driven through strain or spin transfer/orbit torques can implement synaptic weights [15], as well as spiking neurons (energy budget ~ 100aJ) [15,16]. Nanomagnets are now integrated on CMOS and commercialized as memory technologies. Dipole interaction between nanomagnets implement synaptic connections between nanomagnetic neurons, consuming no chip area and dissipating no energy, as it involves no current flow. The synapse's weight (strength of dipole interaction) can be modulated locally with voltage-generated strain (applied via gate electrodes) if the nanomagnets are magnetostrictive and delineated on a piezoelectric substrate [17,18]. Two dipole coupled magnetic tunnel junctions can generate joint probabilities and conditional probabilities [17,18] while a single magnetic tunnel junction (MTJ) can generate any desired probability distribution [19]. These attributes can be leveraged to mimic many features of brain inspired learning involving synaptogenesis and dendritogenesis.

Concluding Remarks

[Include brief concluding remarks. This should not be longer than a short paragraph. (200 words max)]

The neural roadmap [1] continues to show a path towards building human brain-structures. Unlocking these opportunities requires the efficient use of event-timing and temporal neural encoding as well as modelling and computationally abstracting the fundamental computations of 100s to 10,000s of cortical neurons (e.g. neural columns). These components organize into bidirectional interconnected cortical layers interacting with other neural layers (e.g. hippocampus) for computation and learning, where the learning process requires both parameter updates (e.g. synaptic weights) as well as new topologies and components (e.g. neurogenesis, synaptogenesis, dendritogenesis). These opportunities could utilize CMOS design as well as technologies that integrate with CMOS infrastructure. These techniques greatly expand the current computing research focused on deep neural networks, recurrent networks, and similar techniques that do not utilize the rich and highly energy efficient neural system computing capabilities. These physical techniques not only can drastically improve the energy efficiency of current deployed neural networks (e.g. [20]), but enable an entirely new energy-efficient computing ecosystem.

Acknowledgements

[Please include any acknowledgements and funding information as appropriate.]

The work of S. B. in this field is currently supported by the National Science Foundation under grants CCF 2001255 and CCF 2006843.

References

[(Separate from the two page limit) Maximum 20 References. Please provide the full author list, and article title, for each reference to maintain style consistency in the combined roadmap article. Style should be consistent with all other contributions -use IEEE style] Maximum number of references allowed is 20.

[1] J. Hasler and H.B. Marr,"Finding a roadmap to achieve large neuromorphic hardware systems," *Frontiers in Neuromorphic Engineering*, pp. 1–29, 2013.

[2] C. Mead, *Analog VLSI and Neural Systems*, Addison Wesley, 1989.

[3] J. Hasler and E. Black, "Physical computing: Unifying real number computation to enable energy efficient computing," *Journal of Low-Power Electronics Applications*, pp. 1–21, 2021.

[4] C. Mead, "Neuromorphic electronic systems," *Proceedings of the IEEE*, no. 78, pp. 1629–1636, 1990.

[5] J. Hasler, "Large-Scale Field Programmable Analog Arrays," *IEEE Proceedings*, vol. 108. no. 8. August 2020. pp. 1283-1302.

[6] Hasler, C. Diorio, B. A. Minch, and C. A. Mead, "Single transistor learning synapses," in *Advances in Neural Information Processing Systems*, 1994, pp. 817–824.

[7] M. Kucic, Hasler, J. Dugger, and D. Anderson, "Programmable and adaptive analog filters using arrays of floating-gate circuits," *Advanced Research in VLSI*, 14-16 March 2001, pp. 148 – 162.

[8] J. Hasler, "The Rise of SoC FPAA Devices," *IEEE CICC*, April 2022.

[9] Levy, W. B, Hocking, A. B., and Wu, X. B. Interpreting hippocampal function as recoding and forecasting. Neural Networks 18, 2005, 1242-1264

[10] Levy, W. B, A sequence predicting CA3 is a flexible associator that learns and uses context to solve hippocampal-like tasks, Hippocampus, 6, 1996, 579-590.

[11] French RM (1999), Catastrophic forgetting in connectionist networks. Trends Cognit Sci 3(4):128–135.

[12] McCloskey M, Cohen NJ (1989), Catastrophic interference in connectionist networks: The sequential learning problem. The Psychology of Learning and Motivation, ed. G. H. Bower (Academic, New York), Vol 24, pp 109–165.

[13] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, Raia Hadsell, Overcoming catastrophic forgetting in neural nets, Proceedings of the National Academy of Sciences Mar 2017, 114 (13) 3521-3526

[14] Jäkel Sarah and Dimou Leda, "lial Cells and Their Function in the Adult Brain: A Journey through the History of Their Ablation," *Frontiers in Cellular Neuroscience*, vol. 11, 2017.

[15] Samiran Ganguly, Kerem Y. Camsari, Yunfei Gu, Mircea Stan, Avik W. Ghosh, "A Complete Set of Spintronic Hardware Building Blocks for Low Power, Small Footprint, High Performance Neuromorphic Architectures". Invited paper to Proceedings of SPIE Spintronics XII, San Diego, 2019

[16] Samiran Ganguly, Nikhil Shukla, Avik W. Ghosh, "Ultra-Compact, Scalable, Energy-Efficient VO2 Insulator-Metal-Transition Oxide Based Spiking Neurons for Liquid State Machines", Proceedings of 28th IFIP/IEEE International Conference on Very Large Scale Integration (VLSI-SoC 2020), Oct 2020.

[17] M. T. McCray, M. A. Abeed and S. Bandyopadhyay. "Electrically programmable probabilistic bit anticorrelator on a platform", Sci. Rep., vol. 10, 12361 (2020).

[18] S. Nasrin, J. Drobitch, P. Shukla, T. Tulabandhula, S. Bandyopadhyay and A. R. Trivedi, "Bayesian reasoning machine on a magneto-tunneling junction network", Nanotechnology, vol. 31, 484001 (2020).

[19] S. Nasrin, J. Drobitch, S. Bandyopadhyay and A. R. Trivedi, "Low power restricted Boltzmann machine using mixed mode magneto-tunneling junctions", IEEE Elec. Dev. Lett., vol. 40, 345 (2019).

[20] J. Hasler, "The Potential of SoC FPAAs for Emerging Ultra-Low-Power Machine Learning," *J. Low Power Electron. Appl, May* 2022.

10.1- Quantum computers with spin-based qubits in silicon

Belita Koiller[1], Gabriel H. Aguilar[1] and Guilherme P. Temporão[2]

[1] Instituto de Física, Universidade Federal do Rio de Janeiro, 21941-972 Rio de Janeiro, RJ, Brasil
[2] Centro de Estudos em Telecomunicações, Pontifícia Universidade Católica do Rio de Janeiro, Rua Marquês de São Vicente 225, 22451-900 Rio de Janeiro, RJ, Brasil
bk@if.ufrj.br; gabo@if.ufrj.br; temporao@puc-rio.br

Status

Universal Quantum Computers (QCs) can potentially solve open problems that are not only relevant to science and technology but also likely to assist current social demands and expectations, e.g. regarding food supplies and environment issues.

The formalism for quantum information processing is substantially simplified by the following result [1]: A universal set of gates, consisting of all one-qubit quantum gates and a single two-qubit gate, e.g. the controlled-NOT (C-NOT) gate, may be combined to perform any logic operation on arbitrarily many qubits (thus pointing a clear path towards universal QC). In addition, few requirements must be fulfilled: 1) Proposals should eventually provide a prototype performing any operation compatible with universality; 2) The fidelity of one- and two-qubit operations must be above 99%;
and 3) The QC architecture is expected to fit within a few centimeters chip.

Among a few existing quantum technologies compatible with these criteria are electron spins in semiconductors, which we briefly review here. Semiconductor-based QC started as a promising candidate for implementation of QC [2], by confining electrons within electrostatically defined quantum wells in a GaAs/AlAs interface 2DEG. Changing the height of the barrier separating two electrons in neighboring wells, 2-qubit quantum operations can be performed (top Fig.1a). These ideas were adapted to qubits defined by spins of electrons bound to shallow donors in Si [3]. Even though the original proposal suggested impurity nuclear spins-1/2 as qubits, such as P in Si, it was soon recognized that nuclear spins are much too isolated from the environment for fast, efficient control by external fields. A more promising route considered the spin of the extra electron in P in a Si host, which is loosely bound to the donor (lower Fig.1a). The long-lived P nuclear spins have been considered appropriate for quantum memories [4].

Two decades after the first proposals, silicon qubits still hold the status of a great promise. They have been shown to exhibit long enough coherence times, high-fidelity gates, fast operation capabilities and a huge potential for a scalable solution. Indeed, the search for a scalable and universal silicon-based quantum computer has so far attracted full attention from active research groups in academic institutions, major industrial labs and start-up companies. Moreover, spin-based QC uses the same technology adopted in fabrication of transistor-based electronics, benefiting from the established know-how and huge investment in the silicon technology. Notwithstanding all advantages, there are many challenges yet to be overcome, as described in the next sessions.

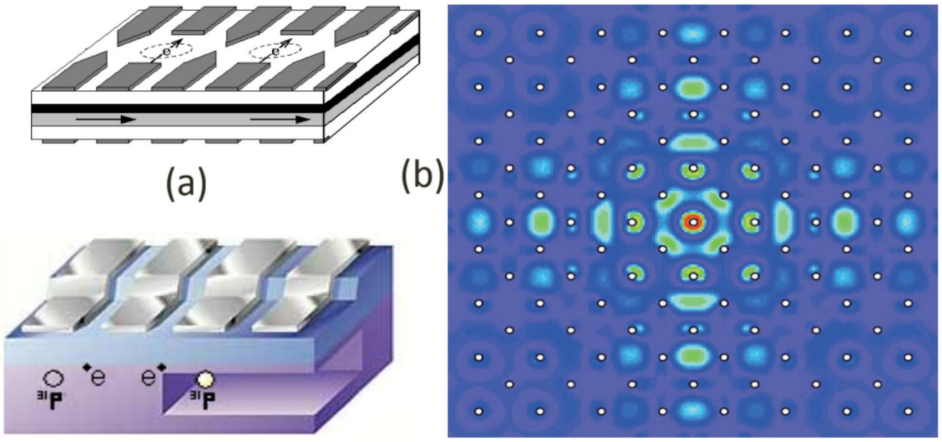

Figure 1. (a) Schematic representation of early (1998) QC architectures with spin qubits in semiconductors : upper scheme represents Loss–DiVincenzo's proposal on GaAs (qubits bound to quantum wells [2]); bottom scheme represents Kane's proposal involving P substitutional donors in Si [3]. (b) Electron probability density around a P substitutional donor on the (001) plane of bulk Si for the donor ground state. The white dots give the in-plane atomic sites. Obtained from ref. [5].

Current and Future Challenges

Silicon qubits are currently undergoing a major transition, from lab prototypes featuring a few qubits to massive scale production. Scalability is a key ingredient in demand for progress in any QC platform. Recent progress reported by a joint Intel – QuTech partnership [6] demonstrated a fully optical lithography technique - similar to those in use in current integrated circuit fabrication methods - in order to obtain more than 10,000 arrays of quantum dot (QD) spin qubits on a single 300-mm wafer (Fig. 2). This leads to a device yield above 98% and good QD uniformity, with a normalized standard deviation in the gate threshold voltages around 7%. The spin relaxation and dephasing times $_1 \sim 1$ and $_2 \sim 20$ - which can be improved to the 3 ms range, or even higher, by dynamical decoupling. Such indicators stress the huge potential of QC using QDs in silicon.

There are many challenges, however, remaining to be addressed. Spin qubits require individual control, involving a connection to a classical auxiliary electronic device. This creates serious difficulties to implement the required individual wiring when the number of qubits is in the order of magnitude of millions. Qubit readout is also a sensitive aspect, as not only the readout bandwidth must be much faster than the spin decoherence time but the readout of a single qubit must not interfere with the neighboring qubits [7]. Concerning donor spins in Si, fabrication and control of multi-qubit arrays are among the most critical limitations.

Remaining challenges concern adapting the large-scale manufacturing process for two-dimensional spin arrays, which are important for two main reasons: being able to implement surface codes [8] for robust quantum error correction and optimizing the efficiency of cryogenic cooling of the circuit [9]. Moreover, two-qubit gates still need to be fully implemented and characterized; fault-tolerant quantum computing requires two-qubit fidelities of at least 99%, a threshold which has been overcome very recently for QDs in silicon [10]. In fact, circuit characterization and benchmarking - including defining figures of merit for assessing the quality of the QC along all the manufacturing steps, which are very different from the classical counterpart - are among the main challenges that need to be tackled to keep any expectations of achieving a QC with millions of qubits.

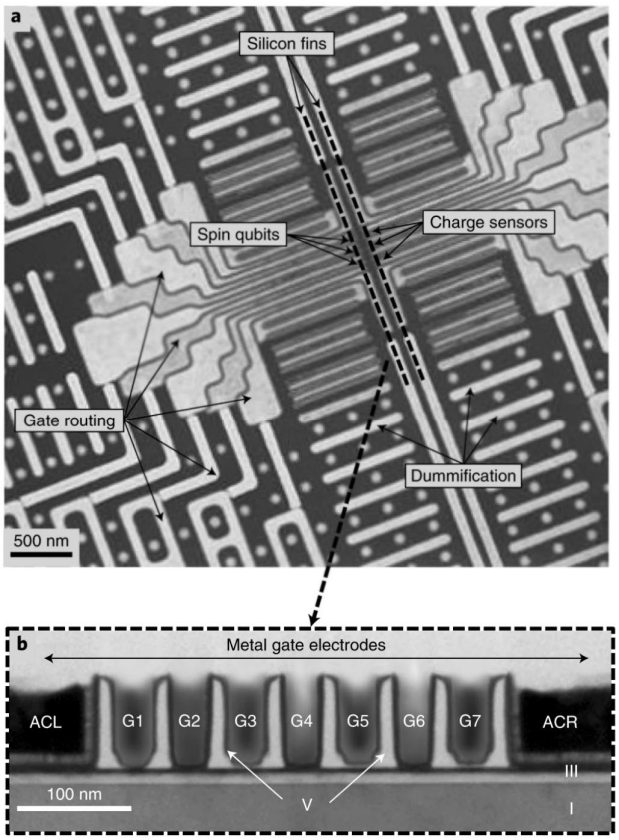

Figure 2. (a) Electron microscopy image of an industrially fabricated QD device by Intel and QuTech, showing two parallel silicon fins, one hosting the qubits (left) and the other hosting the sensing QDs (right). Gate routing and dummification (required for maintaining a constant metal density on the surface) are also clearly shown. (b) Image along a Si fin of a QD array showing seven metallic silver gates (G1–G7) between two accumulation gates (ACL and ACR). Figure adapted from ref. [6].

Advances in Science and Technology to Meet Challenges

In order to scale up the spin system to reach a fully-fledged QC, we need to avoid the heat generated by every qubit added. This is especially critical in the case of superconducting devices, limiting the number of qubits per dilution-refrigerator. This is not the case for spin qubits, which can operate in an environment of up to 2K [11]. Therefore, the cost on the refrigeration can be reduced significantly in spin QCs, and the scaling-up is likely also be favored. The answer is unknown, as current simulation software employed by the electronics industry is not designed to properly deal with low temperature behaviour. Promising candidates for overcoming this limitation include Contact Block Reduction-based Quantum Computer Aided Design (QCAD) [12].

Problems with spin initialization and readout are usually mitigated by employing a conversion spin-charge. This can be implemented by using a reservoir or by Pauli spin blockade (reservoir-independent). Dispersive read-out has also been considered for single and few electrons as well as silicon nanowire transistors. However, the spin blockade has been advantageous so far because it allows higher readout fidelities and lower qubit operation frequency (~ 1GHz), opening the possibility of reading-out a qubit array in a time smaller than the decoherence time of the single qubit (millisecond range) [13].

To overcome the engineering challenge of simultaneous addressing many qubits in a large-scale spin-based QC, electron-spin-resonance techniques in conjunction with Kane's proposal ideas may be utilized. This technique could include a three-dimensional dielectric resonator that acts as a single global source that can deliver multiple control signals to the qubits. Recent advances show that such resonators, constructed from potassium tantalate (KTO), could be manufactured within a compact surface area of 0.7x0.55 mm^2, allowing its integration to nanoelectronic circuits and performing large-scale control over millions of spin qubits [14].

Finally, high volume fabrication of spin qubits requires a proportional capacity for characterization and tests. This is currently delaying the evolution of spin-based QC, due to lengthy tests in low-temperature environments, which require cooling down the devices in dilution refrigerators. This limitation could be avoided by a thorough characterization of the correlation between classical semiconductor device metrics (such as mobility) and spin qubit performances, but no definitive results have been obtained so far. Alternative characterization procedures, possibly performed at higher temperatures, are still missing, but some effort concerning this huge challenge can already be found in the literature [15,16].

Concluding Remarks

Scalability is one of the most important challenges that any architecture needs to achieve in order to obtain a quantum computer that is able to solve relevant problems. Spin qubits are possibly the most suited candidates for building a universal QC, especially because of its potential scalability which has already been demonstrated experimentally. Other ongoing debates relevant for the future development of spin-based QC include: finding the most effective way of encoding quantum information in silicon (e.g. electrons vs. holes, quantum dots vs. donors, Loss-DiVincenzo vs. Singlet-Triplet qubits) [17-19], which figures of merit need to be measured for proper device characterization (and how they should be measured), the best way to perform spin readout, which quantum error correction codes should be employed, and so on. In any case, QC is an inherent multidisciplinary field, and as such, it is expected that physicists, electrical engineers and computer scientists participate in discussions, working together towards a large-scale QC. Otherwise, questions that cannot be solved by the available classical computers will keep increasing the number of problems that remain open.

Acknowledgements

The authors thank André Saraiva for the interesting and very fruitful conversations. The authors acknowledge financial support from the Brazilian agencies CNPq (PQ Grants No. 307058/2017-4, INCT-IQ 246569/2014-0 and 307910/2019-9). G.H.A. and G.P.T. also acknowledge FAPERJ (Grants No. 210.069/2020 and 211.094/2019, respectively) and FAPESP (Grant No. 2021/96774-4).

References

[1] A. Barenco, C. H. Bennett, R. Cleve, D. P. DiVincenzo, N. Margolus, P. Shor, T. Sleator, J. A. Smolin and H. Weinfurter, *Elementary gates for quantum computation,* Physical Review A. **52**, pp. 3457, 1995.
[2] D. Loss and D. P. DiVincenzo, *Quantum computation with quantum dots,* Physical Review A. **57**, pp. 120–126, 1998.
[3] B. E. Kane, *A silicon-based nuclear spin quantum computer,* Nature **393**, pp. 133, 1998.
[4] L. Dreher, F. Hoehne, M. Stutzmann, M. S. Brandt, *Nuclear spins of ionized phosphorous donors in silicon*. Phys. Rev. Lett. **108**, 027602, 2012.
[5] B. Koiller, R. Capaz, X. Hu, S. das Sarma, *Shallow-donor wave functions and donor-pair exchange in silicon: Ab initio theory and floating-phase Heitler-London approach*. Phys. Rev. B 70, 115207, 2004.
[6] A. M. J. Zwerver et al., *Qubits made by advanced semiconductor manufacturing,* Nature Electronics **5**, pp. 184-190, 2022.
[7] A. Saraiva, W. Han Lim, C. H. Yang, C. C. Escott, A. Laucht and A. S. Dzurak, *Materials for Silicon Quantum Dots and their Impact on Electron Spin Qubits,* Adv. Funct. Mater. **32**, pp. 2105488, 2022.
[8] A. G. Fowler, M. Mariantoni, J. M. Martinis, A. N. Cleland, *Surface codes: towards practical large-scale quantum computation*. Physical Review A **86**, 032324, 2012.
[9] M. Vinet, *The path to scalable quantum computing with silicon spin qubits*, Nature Nanotechnology 16, 1296-1298, 2021.
[10] A. Noiri, K. Takeda, T. Nakajima, T. Kobayashi, A. Sammak, G. Scappucci, S. Tarucha. *Fast universal quantum gate above the fault-tolerance threshold in Silicon*. Nature **601**, 338-442, 2022.
[11] C. H Yang, R. C. C. Leon, J. C. C. Hwang, A. Saraiva, T. Tanttu, W. Huang, J. Camirand Lemyre, K. W. Chan, K. Y. Tan, F. E. Hudson, K. M. Itoh, A. Morello, M. Pioro-Landrière, A. Lauch, A. S. Dzurak., *Operation of a silicon quantum processor unit cell above one kelvin,* Nature **580**, pp. 350-354, 2020.
[12] X. Gao, E. Nielsen, R. P. Muller, R. W. Young, A. G. Salinger, N. C. Bishop, M. S. Carroll, *The QCAD framework for quantum device modeling*. 2012 15th International Workshop on Computational Electronics, 2012, pp. 1–4.
[13] J. Yoneda et al., *Coherent spin qubit transport in silicon,* Nature Communications **12**, pp. 1-9, 2021.
[14] E. Vahapoglu, J. P. Slack-Smith, R. C. C. Leon, W. H. Lim, F. E. Hudson, T. Day, T. Tanttu, C. H. Yang, A. Laucht, A. S. Dzurak, J. J. Pla, *Single-electron spin resonance in a nanoelectronic device using a global field,* Science Advances **7**, eabg9158, 2021.
[15] K. W. Chan, W. Huang, C. H. Yang, J. C. C. Hwang, B. Hensen, T. Tanttu, F. E. Hudson, K. M. Itoh, A. Laucht, A. Morello and A. S. Dzurak, *Assessment of a silicon quantum dot spin qubit environment via noise spectroscopy*, Phys. Rev. Appl. **10**, pp. 044017, 2018.

[16] R. Pillarisetty, H.C. George, T.F. Watson, L. Lampert, N. Thomas, S. Bojarski, P. Amin, R. Caudillo, E. Henry, N. Kashani, P. Keys, R. Kotlyar, F. Luthi, D. Michalaek, K. Millard, J. Roberts, J. Torres, O. Zietz, T. Krähenmann, A. M. Zwerver, M. Veldholst, G. Scappucci, L. M. K. Vandersypen, J. S. Clarke, *High Volume Electrical Characterization of Semiconductor Qubits*, 2019 IEEE International Electron Devices Meeting (IEDM), 2019, pp. 31.5.1-31.5.4
[17] R. M. Jock, N. Tobias Jacobson, M. Rudolph, D. R. Ward, M. S. Carroll, D. R. Luhman, *A silicon singlet-triplet qubit driven by spin-valley coupling.* Nature Communications 13, 641, 2022.
[18] A. Noiri, N. Takajima, J. Yoneda, M. R. Delbecq, P. Stano, T. Otsuka, K. Takeda, S. Amaha, G. Allison, K. Kawasaki, Y. Kojima, A. Ludwig, A. D. Wieck, D. Loss, S. Tarucha, *A fast quantum interface between different spin qubit encodings*. Nature Communications 9, 5066, 2018.
[19] R. Maurand, X. Jehl, D. Kotekar-Patil, A. Corna, H. Bohuslavskyi, R. Laviéville, L. Hutin, S. Barraud, M. Vinet, M. Sanquer, S. de Franceschi, *A CMOS silicon spin qubit*. Nature Communications 7, 13575, 2016.