Accelerating neuromorphic workloads through novel devices, architectures and algorithms

Jonas Weiss, Valeria Bragaglia, Antonio La Porta, Jean Fompeyrine, Stefan Abel, Folkert Horst, Bert Jan Offrein

IBM Research – Zurich, Säumerstrasse 4, 8803 Rüschlikon, Switzerland

Over the last decade, countless machine learning and AI algorithms have gradually penetrated into everyday applications. Most prominently, deep learning neural networks and their derivatives, fuelled by the abundant availability of compute power through GPUs and the cloud, have become the work-horse for large scale classification and prediction problems.
To fully unlock the potential of deep neural networks and deploy them, both in high-performance as well as in mobile and edge-devices, the power-efficiency of their underlying compute-platform requires orders of magnitudes improvements. On the short-term, enhanced parallelism and moving from *full accuracy floating point* to *reduced resolution integer* arithmetic provides initial improvements. For the long run, moving from digital to analog arithmetic, using memristive memory elements promises huge potential in accelerating these and similar workloads.

At IBM Research - Zurich, we are contributing to a larger, global effort establishing materials and devices for analog crossbar fabrics. These compute entire matrix vector multiplications, core of many AI and scientific computation algorithms, at a fraction of the energy and time as required by today's digital computers. However, the use of such devices and architectures requires understanding of both, physical limitations of the devices as well as algorithmic and computational constraints.

We will discuss some recent trends and results on memristive devices, e.g. on filamentary oxide memristors, will elaborate on implementation options and constraints in crossbar arrays and by what means of algorithmic adjustments, these structures will be able to fulfil their promise of highly energy efficient AI workload acceleration. Furthermore, we will also peek into less conventional concepts using optical and reservoir computing.

Neuromorphic systems in consumer electronics: near-term and mid-term prospects.

Slava Chesnokov

ARM Holdings, 1 Summerpool Rd, Loughborough LE11 5R, United Kingdom

ARM is one of big players in area of consumer electronics and aims to increase its role in the future consumer devices, based on AI and robots. Traditionally the main differentiation of ARM systems (most famous is ARM CPU, which could be found in any mobile phone): high power efficiency. Therefore ARM is keen to retain the leadership in power-efficient computation systems (especially for consumer devices). That's why ARM would be interested to be well-informed and even involved into new HW developments, which would help making the computational systems like AI/Neural Networks, etc. more power efficient. ARM would be particularly interested in new computational components (e.g. arithmetic blocks such as addition/multiplication in analog domain, new and highly efficient elements of memory (analog memory) and way to modify the values in memory, etc.)

The presentation will outline the current state of Neuromorphic systems (mainly, what is called: Machine Learning) in consumer devices: CPU/GPU/NPU (NPU – Neural Processing Unit).It will also outline the likely near-term and mid-term future stages in semiconductor industry towards highly efficient NNs:

- ✓ Digital Fixed Neural Networks (specialized for tasks) with multiple methods of improving power efficiency and silicon area: such as Approximate Arithmetic, optimal data width for each data path, pruning, etc.
- ✓ Hybrid Digital and Analog systems: Neural Accelerators: where Analog Arithmetic is used to improve the power efficiency;
- ✓ Fully Analog Fixed Neural Systems (most power efficient).

In all the above cases only the inference of the Neural Network will be implemented by the HW of a consumer device, whereas the training will be on server or by SW on CPU of the device. In the more remote future the training of the NNs will be done on the device itself (self-learning machines). The fundamental advantage of Digital Inference of a Neural Network: the result of training of a NN will be reproducible on multiple devices. Whereas in case of Analog inference: at least there should be an additional retraining (to adapt to differences of ideal NN from real analog NN, which would have off-sets and imperfections. There is likelihood that the most complicated analog NNs would require a local (not cloud server) training system and process, thus the artificial brain will be unique, not reproducible.

Advances in experimental unconventional computing

Andrew Adamatzky

Unconventional Computing Group, University of the West of England, Bristol, BS16 1QY, UK

The unconventional computing is a niche for interdisciplinary science aimed to exploit principles of information processing in and functional properties of physical, chemical and living systems to develop efficient algorithms, design optimal architectures and manufacture working prototypes of future and emergent computing devices. Theoretical models and constructs of unconventional computers flourish in abundance while experimental laboratory prototypes are rare and precious. We will overview our prototypes of reaction-diffusion chemical processors, slime mould sensing and computing devices, electric current pathfinders and liquid marble computers. Unusual neuromorphic architectures will be addressed via examples of a nervous system made of a slime mould, computing with action potential like spikes in fungal networks and venation of a plant leaf, Belousov-Zhabotinsky liquid marble oscillators

Neuromorphic devices using organic materials

Paschalis Gkoupidenis

Max Planck Institute for Polymer Research, Ackermannweg 10, D-55128 Mainz, Germany

Neuromorphic devices and architectures offer novel ways of data manipulation and processing, especially in data intensive applications. At a single device level, various forms of neuroplasticity have been emulated over the past years, mainly with inorganic devices. The implementation of neuroplasticity functions with these devices also enabled applications at a circuit level related to machine learning such as feature or pattern recognition. Although the field of organic-based neuromorphic devices and circuits is still at its infancy, organic materials may offer attractive features for neuromorphic engineering. Over the past years for example, a few simple neuromorphic functions have been demonstrated with biological substances and bioelectronic devices.

Here, various neuromorphic devices will be presented that are based on organic mixed conductors, materials that are traditionally used in organic bioelectronics. A prominent example of a device in bioelectronics that exploits mixed conductivity phenomena is the organic electrochemical transistor (OECT). Devices based on OECTs show volatile and tunable dynamics suitable for the emulation of short-term synaptic plasticity functions. Chemical synthesis allows for the introduction of non-volatile phenomena suitable for long-term memory functions. The device operation in common electrolyte permits the definition of spatially distributed multiple inputs at a single device level. The presence of a global electrolyte in an array of devices also allows for the homeostatic or global control of the array. Global electrical oscillations can be used as global clocks that frequency-lock the local activity of individual devices in analogy to the global oscillations in the brain. Finally, “soft” interconnectivity through the electrolyte can be defined, a feature that paves the way for parallel interconnections between devices with minimal hard-wired connections.

A route to hierarchical control in artificial intelligent systems: memristors with optically tuneable STDP synaptic plasticity

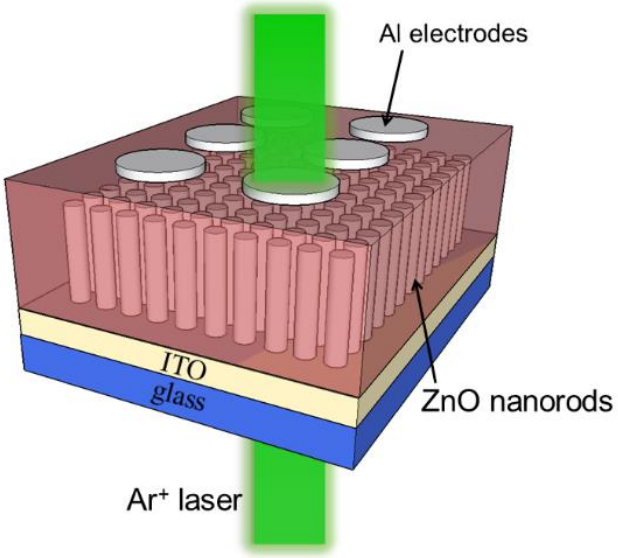
Neil Kemp

Department of Physics and Mathematics, University of Hull, Cottingham Road, Hull, HU6 7RX, UK

Modulation of resistive switching memory by light opens the route to new optoelectronic devices that can be switched optically and read electronically. Applications include integrated circuits with memory elements switchable by light and optically reconfigurable and tunable synaptic circuits for neuromorphic computing applications and brain-inspired artificial intelligent systems. A novel polarization and wavelength-specific optoelectronic memory device is presented that can be controlled purely optically, electronically or by both. Optical or electronic modulation switches the device between low and high conducting states whilst modulation by both facilitates fine tuning of the resistive memory properties and switching characteristics of the device.

In biological synapses the connection strength (plasticity) between two neurons is controlled by the ionic flow through the synaptic cleft and it is widely believed that the adaptation of synaptic weights enables biological systems to learn and function. Similarly, the conductance of a memristor depends on the history of the total charge that has travelled through it. A key feature of neuronal learning is habituation, whereby repeated stimuli strengthens the synaptic plasticity whilst a lack of stimuli results in weakening. Learning in biological systems also involves spike-timing-dependent plasticity (STDP). In STDP learning the synaptic efficacy governing potentiation and depression is determined instead by the temporal order of pre-synaptic and post-synaptic spikes.

Using our optical memristor device we demonstrate optical control of synaptic potentiation and depression, optical switching between short and long term memory and optical modulation of the synaptic efficacy via spike timing dependent plasticity. The work opens the route to dynamic patterning of memristor networks both spatially and temporally by light, thus allowing the development of new optically reconfigurable neural networks and adaptive electronic circuits.



Memristive devices and arrays for brain-inspired computing

Qiangfei Xia

Department of Electrical and Computer Engineering, University of Massachusetts, Amherst, MA 01003, USA

It becomes increasingly difficult to improve the speed-energy efficiency of traditional digital processors because of limitations in transistor scaling and the von Neumann architecture. To address this issue, computing systems augmented with emerging devices, in particular memristors, offer an attractive solution. A memristor, also known as a resistance switch, is an electronic device whose internal resistance state is dependent on the history of the current and/or voltage it has experienced. With their working mechanisms based on ion migration, the switching dynamics and electrical behaviour of memristors closely resemble those of biological synapses and neurons. Because of its small size and fast switching speed, a memristor consumes a small amount of energy to update the internal state (training). Built into large-scale crossbar arrays, memristors perform in-memory computing by utilizing physical laws, such as Ohm's law for multiplication and Kirchhoff's current law for accumulation. The current readout at all columns (inference) is finished simultaneously regardless of the array size, offering a huge parallelism and hence superior computing throughput. The ability to directly interface with analog signals from sensors, without analog/digital conversion, could further reduce the processing time and energy overhead.

I will introduce a high performance memristor that is the basis for our recent artificial neural networks, highlighting its two nanometer scalability and eight layer stackability. I will then showcase the integration of large memristor crossbar arrays for analog signal and image processing, and the implementation of multilayer memristor neural networks for machine learning applications. Finally, I will briefly introduce a diffusive memristor as a bio-realistic synapse and neuron emulator, and review further applications of memristors in reconfigurable radiofrequency systems and hardware security.

Finding the Mott memristor

R. Stanley Williams

Department of Electrical and Computer Engineering, Texas A&M University, USA

I had the great privilege to have many enjoyable discussions about physics and life with Shasha Alexandrov, and the honor to collaborate with him on five papers. One of those publications, the understanding of a mechanism for thermally activated negative differential resistance in a metal oxide, played a major role in the development of the material for this talk. I am doubly honored to present the Mott Lecture at Loughborough University, since the Mott insulator-to-metal transition is also strongly featured in my presentation. Both of these results are needed to understand the nonlinear dynamical behavior of electron and thermal transport in NbO_2. In 1963 Ridley postulated that, under appropriate biasing conditions, a system that exhibits a current-controlled negative differential resistance will bifurcate to form regions with different current densities in a single medium, i.e. there will be a spontaneous symmetry breaking. The ensuing discussions in the non-equilibrium statistical mechanics community, however, failed to agree on specific mechanisms causing such bifurcations or general predictive models. Using thermal and chemical spectro-microscopy, my group experimentally imaged current-density-bifurcations in NbO_2 and other transition metal oxides, and developed a simple memristor model that quantitatively agrees with the experimental results. The dynamical electronic behavior of NbO_2 mimics that observed for the action potential of neurons, and we have built a 'neuristor' that emulates the threshold, amplification and oscillations embodied by the Hodgkin-Huxley model of the neuron. These observations are inspiring further research in neuromorphic computing, i.e. utilizing concepts obtained from neurophysiology to create new devices and circuits for brain-inspired computers that are orders of magnitude more energy efficient than present or future semiconductor technology.

Memristive Technologies: a viable pathway for beyond Moore electronics and AI

Themis Prodromakis

Electronic Materials and Devices Research Group, Zepler Institute for Photonics and Nanoelectronics, University of Southampton, Southampton, SO17 1BJ, UK

In the not so far future, electronic devices will be everywhere – embedded within our physical world and even in our bodies – empowering modern societies with unprecedented capabilities. Yet, the technological progress that brought us the mobile revolution is not any more sustainable for allowing us reaching this point. Up until now, the processing of data in electronics has relied on assemblies of vast numbers of transistors – microscopic switches that control the flow of electrical current by turning it on or off. Transistors have got smaller and smaller in order to meet the increasing demands of technology, but have nowadays reached their physical limit, with – for example – the processing chips that power smartphones containing an average of five billion transistors that are only a few atoms wide.

A novel nano-electronic technology, known as the memristor, proclaims to hold the key to a new era in electronics, being both smaller and simpler in form than transistors, low-energy, and with the ability to retain data by 'remembering' the amount of charge that has passed through them – akin to the behaviour of synaptic connections in the human brain. In his lecture Themis Prodromakis will present a few examples on how memristive technologies can be exploited in practical applications ranging from neuromorphic systems to charge-based computing and even enabling bioelectronics medicines.

Universal mem-computing by Cellular Neural Networks

Ronald Tetzlaff[1], M. Weiher[1], A. Ascoli[1], M.Herzig[2], S. Slesazeck[2], T. Mikolaijck.[2,3], and L. O. Chua[4]

[1] *Institute of Circuits and Systems, TU Dresden, Dresden, Germany*
[2] *Nano-electronic Materials Laboratory (NaMLab) gGmbH, Dresden, Germany*
[3] *Institute of Semiconductors and Microsystems, TU Dresden, Dresden, Germany*
[4] *Department of Electrical Engineering and Computer Sciences, University of California Berkeley, Berkeley, CA 94720 USA*

The ubiquitous digitalization of future systems requires new concepts of computing in order to overcome the limits of classical technology caused by the so-called von-Neumann bottleneck. While different authors have shown that dot-product operations or correlation detection can be performed efficiently by memristive memory crossbar arrays, new results point on universal memristive computing structures known as Cellular Neural Networks (CNN) which offer high performance computing usually under real-time conditions but when endowed with memristors can be applied as pure memory systems within a program flow as well. Although, computing by these structures is often based on complex behaviour emerging in arrays of nonlinear identical systems, these networks can be implemented in a form to be resilient to device variability. The CNN paradigm represents nature inspired high-speed sensor-processor universal computing arrays with stored programmability. Fundamental results have been derived by Chua [1] proving that the emergence of complexity in these structures is based on local activity and especially on a parameter subset called the “Edge of Chaos (EOC)”.

An introduction to the theory of memristive CNN (M-CNN) will be provided in this contribution. Furthermore, by assuming a compact model based on a Frenkel-Poole like conduction mechanism with a thermal feedback [2,3], M-CNN with NbOx memristor cells and RC-bridge coupling circuits have been implemented and studied in the EOC parameter domain. Finally, a new method allowing the determination of M-CNN equilibrium points being necessary for programming these structures will be proposed and discussed in detail.

[1] L. O. Chua, “Local activity is the origin of complexity,” International journal of bifurcation and chaos, vol.15, no. 11, pp. 3435–3456, 2005

[2] A. Ascoli, S. Slesazeck, H. Mähne, R. Tetzlaff, and T. Mikolajick, “Nonlinear dynamics of a locally-active memristor,” IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 62, no. 4, pp. 1165–1174, 2015

[3] S. Slesazeck, H. Mähne, H. Wylezich, A. Wachowiak, J. Radhakrishnan, A. Ascoli, R. Tetzlaff, and T. Mikolajick, “Physical model of threshold switching in NbO2 based memristors,” RSC Advances, vol. 5, no. 124, pp. 102 318–102 322, 2015.

Silicon oxide neuromorphic elements – synaptic behaviour and beyond.

Anthony J. Kenyon, A. Mehonic, W. Ng, M. Buckwell, D. Joksas, D. Mannion, L. Zhao, S. Ibrahim, K. Zarudnyi, L. Montesi.

Department of Electronic & Electrical Engineering, UCL, Torrington Place, London WC1E 7JE, UK

Silicon oxide is a promising material for memristors, being fundamentally CMOS compatible and offering advantages of high stability and endurance. In this presentation I shall present a summary of our studies of silicon oxide neuromorphic elements. We have demonstrated spike-based synaptic responses (Long Term Potentiation, Long Term Depression, Spike Timing Dependent Plasticity), and have gone beyond this to show that silicon oxide memristors can also perform some of the key functions of the neuron – spiking, thresholding and integration. In addition, devices can be optically triggered. I shall review these results and relate them to fundamental mechanisms of resistance switching and to the role of material microstructure in defining memristive behaviour.

Random Telegraph Noise and its Impact on Pattern Recognition Accuracy of RRAM-Based Synaptic Neural Network

Wei Zhang

Department of Electronics and Electrical Engineering, Liverpool John Moores University, Liverpool L3 3AF, UK

Random Telegraph Noise (RTN) is the random fluctuation between discrete current levels caused by trapping and de-trapping in defects. RTN has become a critical issue in nanoscale logic and memory devices where the impact of a single defect becomes significant, as it can cause large threshold voltage variation, reduce the memory window and cause read errors. On the other hand, RTN can also be used as a useful technique for analysing the spatial location and energy level of the responsible defect. By utilizing our recently developed defect tracking and profiling technique based on RTN signals, defect movement and profile modulation during the resistive switching in several different types of RRAM devices are revealed and correlated to the mechanisms of switching, failure and instability. Both filamentary and non-filamentary RRAM can be used as synaptic devices in hardware neural networks. The amplitude and occurrence rate of RTN in both types of RRAM devices are evaluated and its impact on the pattern recognition accuracy of neural networks is analysed. It is revealed that the non-filamentary RRAM has a tighter RTN amplitude distribution and much lower RTN occurrence rate than its filamentary counterpart, which leads to negligible RTN impact on recognition accuracy, making it a promising candidate in synaptic application.

Modelling artificial neurons

Sergey Saveliev

Department of Physics, Loughborough University, Loughborough, LE11 3TU, UK

Brain emulators aim to develop software and hardware with cognitive abilities which are reminiscent of those of the animal or human brain. Even though we are now in a very early stage along this way, this research can stimulate development of brain engineering technology and neuromorphic computing mimicking brain operation. Next generation of brain emulators requires information processing and communication resembled the signal propagation in biological neural tissue. At Loughborough University, we are now working on establishing new neuromorphic laboratory combining expertise in memristor technology, neuroscience, physiology, material science, and memristor modelling.

One of the directions for such research is neuromorphic engineering dealing with electronic circuits of memristors, with a goal to mimic neuro-biological architectures present in the nervous system. I will discuss several models of memristive neurons, where conductive nanoclusters, heat and electric degrees of freedom are considered simultaneously. These models are able to predict and explain complex dynamics of the neuromorphic memristive structures.

Threshold switching in niobium oxide devices

Pavel Borisov

Department of Physics, Loughborough University, Loughborough, LE11 3TU, UK

A metastable form of niobium oxide, niobium dioxide (NbO_2) is a narrow band semiconductor which becomes metallic above the transition temperature of 808C. This makes it very interesting for a number of applications in form of thin film devices demonstrating threshold current switching behaviour at room temperature. The current can rise by several orders of magnitude when the applied voltage exceeds a certain threshold, typically between 1V and 2V. Niobium oxide devices can be combined with memristive non-volatile memory elements in order to reduce sneak-path leakage in crossbar arrays, or used for generating self-sustained current oscillations when integrated into Pearson-Anson circuits.
I will provide an overview of our studies of NbO_2 thin films devices of different structural quality, grown either in lateral or in vertical geometry and on a variety of substrates and electrodes, Al_2O_3, SiO_2, GaN, TiN and Pt.