Neuromorphic Technology

NEUROTECH

Deliverable D1.3

Neuromorphic Computing Technology (NCT) state of the art overview.

Grant Agreement number: 824103

Funding Scheme: H2020 – FET PROACT - CSA

PROJECT COORDINATOR: University of Zurich

PARTNERS:

University of Zurich

University of Manchester

Heidelberg university

Bielefeld University

Italian Institute of Technology

University of Bordeaux

University of Hertfordshire

THALES SA

IBM Research GmbH, Zurich Research Laborator

Consiglio Nazionale delle Ricerche

Inter-university microelectronic center

Commissariat a l'Energie Atomique et aux Energies Alternative

PROJECT START DATE: 01/11/2018

PROJECT DURATION: 36 months

DELIVERABLE DUE MONTH: 18

DELIVERABLE DATE OF ISSUE: 30/04/2020

DELIVERABLE LEAD BENEFICIARY: CNR, Italy

NEUROTECH

Deliverable D1.3: NCT state of the art overview.

- **Project**: NEUROTECH, Grant number 824103
- **Report dissemination level**: Public
- **Lead beneficiary**: CNR – (Sabina Spiga, IMM- Unit of Agrate Brianza).
- **Contributors**: UNIMAN, THALES, IMEC, UHEI, UBx, CEA, UZH, IBM
- Final revision of the document by:
  Steve Furber (UNIMAN); Melika Payvand, Elisa Donati, Giacomo Indiveri (UZH); Alice Mizrahi (THALES), Bert Jan Offrein (IBM), Björn Kindler (UHEI) and Sabina Spiga (CNR)

Index

INTRODUCTION.

The aim of the deliverable D1.3 " NCT state of the art overview" is to summarize the current status of all types of neuromorphic hardware technology from fully CMOS-based systems to solutions exploiting the use of memristive devices, advanced device concepts in the field of spintronic and photonics, and novel materials including 2D, nanowires and organic materials.

D1.3 has been then been divided into the following three sections to cover all the above topics:

Section I introduces the current state of the art of *large-scale neuromorphic computing systems based on digital CMOS or analogue/mixed-signal technologies*.

Section II introduces the *various types and physical mechanisms of memristive device technologies*, which include a broad class of two or three-terminal devices whose resistance can be modified upon electrical stimuli. Moreover, we describe the *proposed hardware implementation of synaptic and neuronal circuits* exploiting those memristive technologies, which are currently at high maturity level, namely resistive random access memory (RRAM), phase change memory (PCM), ferroelectric memory (FeRAM), and magnetoresistive random access memories (MRAM), metal-insulator-transition (MIT) devices, as well as more explorative and innovative concepts.

Section III introduces the current state of the art of *mixed CMOS-memristive device neuromorphic chips.* While in the previous section we discuss mainly the implementation of some specific neuromorphic function by exploiting single or small blocks of memristive devices, in this section we summarize the current state of monolithic integrated CMOS-memristive devices in a chip, or of large systems demonstrated at mixed software – hardware level. Currently this section includes mainly the hybrid CMOS-RRAM neuromorphic chip, and IBM work on PCM, but it will be updated in the future with other technologies and future advancements. The updated version of this document will be published on the Neurotech web site (https://neurotechai.eu/).

Section I. STATE OF THE ART OF FULLY-CMOS LARGE-SCALE NEUROMORPHIC PROCESSORS

1.1 Introduction.

Digital CMOS. The mainstay of the semiconductor manufacturing industry, **digital CMOS** is well understood and delivers very consistent performance in volume manufacture. It can access the most advanced semiconductor technologies, which helps offset its intrinsic energy-efficiency disadvantages compared with analogue circuits. When applied to neuromorphic architectures, asynchronous, clocked and hybrid approaches to circuit timing can be used, and algorithms can be mapped into fixed (albeit highly parameterised and configurable) circuits for efficiency, or into software for flexibility. Examples of the former include the DeepSouth, IBM TrueNorth, and Intel Loihi , while SpiNNaker and Tianjic are examples of the latter software-based approach.

Analogue and mixed-signal CMOS. Event-based **analogue/mixed-signal CMOS** based neuromorphic technology combines the compact and low power features of analogue circuits with the robustness and low-latency of digital event-based asynchronous circuits. The key feature of the mixed-signal design approach, compared to the pure digital approach, is the ability to build systems able to carry out processing with stringent resources in terms of power and memory. This goal is implemented by (i) only dissipating power when the data is present, and (ii) processing the data on-line, as it is sensed or streamed through the system, using circuits that have time constants matched to the dynamics of the sensory signals processed, and without needing to store data or state variables in memory. This technology is an enabler for applications requiring sub-mW always-on real-time processing of sensory signals, for example in edge computing, personalized medicine and Internet of Things domains. Examples of neuromorphic processors that follow this approach are the **DYNAP** (Dynamic Neuromorphic Asynchronous Processor) series of devices [Moradi, TBioCAS 2017], **BrainScaleS**, **Neurogrid, and MNIFAT.**

Recent review papers describing large scale neuromorphic processor can be found in refs [1,2].

1.2. Brief description of current state of the art large-scale neuromorphic systems.

IBM TrueNorth chip [3,4,5]. The IBM TrueNorth chip is based upon distributed digital neural models aimed at real-time cognitive applications. IBM's TrueNorth neuromorphic chip consists of 1 million digital neurons capable of various spiking behaviours. Each die holds 4096 cores, each core holding 256 digital neurons and 256 synapses per neuron. A single die consumes 72 mW of power. A board (NS16e) comprising 16 TrueNorth chips has been developed; it consumes 1W of power at 1KHz speed, making it ideal for energy-efficient applications. Although digital in its implementation, low power consumption results from fabrication in an aggressive, state-of-the-art 28 nm technology process.

Neurogrid [6]. The Stanford Neurogrid uses real-time sub-threshold analogue neural circuits. Neurogrid is a mixed-mode multichip system primarily used for large-scale neural simulations and visualization. Neurogrid uses subthreshold analogue circuits to model neuron and synapse dynamics in biological real time, with digital spike communication. The neuronal model uses shared leaky integrator dendritic structures whereby an input to one neuron affects neighbouring neurons through a resistive network. The neuron dynamics are defined by a quadratic integrate and fire model. The neuron circuits used in Neurogrid are closely correlated to the physical characteristics of neurons in the brain. It models the soma, dendritic trees, synapses, and axonal arbors. It consists of 16 neurocores/chips each with 65 k neurons (totalling 1M neurons) implemented in sub-threshold analog circuits. A single neurocore is fabricated on an 11.9 mm× 13.9mm die. A board of 16 neurocores is of size 6.5′′ × 7.5′′ and the complete board consumes roughly 3W of power (a single neurocore consumes ∼150 mW).

BrainScalesS [7,8]. BrainScaleS stands for mixed-signal accelerated neuromorphic computing based on above-threshold analogue neural circuits running up to 10,000 times faster than biological real time. It targets research in the fields of computational neuroscience, in particular long-term learning, and beyond-von-Neumann computing. The second generation systems add an embedded SIMD microprocessor allowing for, amongst others, programmable plasticity rules. The systems were developed at Heidelberg University over a series of projects funded by the European Union, including the FACETS and the BrainScaleS project. On-going support comes from the EU ICT Flagship Human Brain Project.

SpiNNaker [9-12]. The Manchester SpiNNaker machine is a real-time digital many-core system that implements neural and synapse models in software running on small embedded processors, again primarily aimed at modelling biological nervous systems. SpiNNaker was designed for scalability and energy-efficiency by incorporating brain-inspired communication methods. It can be used for simulating large neural networks and performing event-based processing for other applications. Each node comprises 18 ARM968 processor cores each with 32 Kbytes of local instruction memory and 64 Kbytes of local data memory, 128 Mbytes of shared memory, a packet router, and supporting circuitry. A single node can model up to 16,000 digital neurons with up to 16M synapses consuming 1W of power. There are two sizes of SpiNNaker circuit board, the smaller being a 4-node (64,000 neuron) board and the larger a 48-node (768,000 neuron) board. The 48-node board consumes up to 60W of power. The SpiNNaker HBP neuromorphic computing system incorporates a million processors on 1,200 48-node boards and is capable of simulating spiking networks up to the scale of a mouse brain in biological real time.

Loihi [13]. The Loihi is a neuromorphic chip introduced by Intel Labs in 2018 and fabricated in Intel's 14 nm FinFET process technology. It simulates 130K neurons and 130M synapses in real time. The chip consists of 128 neuromorphic cores that are capable of on-chip training and inference. A hierarchical mesh protocol is implemented to support communication between the neuromorphic cores. Loihi is said to be the first fully-integrated spiking neural network chip that supports sparse network compression, core-to-core multicast, variable synaptic format, and population-based hierarchical connectivity. Loihi incorporates an epoch-based synaptic modification architecture in addition to pairwise and triplet STDP. Loihi includes computation blocks such as stochastic noise, which might be added to a neuron's synaptic response current, membrane voltage, and refractory delay for solving probabilistic inference problems. Loihi can solve optimization problems such as LASSO, being over three orders of magnitude better in terms of energy delay- product as compared to a CPU-based solver.

MNIFAT [14]**.** This is a mixed-mode VLSI-based neural array with reconfigurable, weighted synapses/connectivity. The novel integrate-and-fire array transceiver (IFAT) neural array (MNIFAT) consists of 2,040 Mihalas–Niebur (M–N) neurons developed in the lab of Ralph Etienne-Cummings at the John Hopkins University. Each of these M–N neurons was designed to have the capability to operate as two independent integrate-and-fire (I&F) neurons. This resulted in 2,040 M–N neurons and 4,080 leaky I&F neurons. This neural array was implemented in 0.5μm CMOS technology with a 5V nominal power supply voltage (Lichtsteiner et al., 2008). Each I&F consumes an area of 1,495 μm2, while the neural array dissipates an average of 360 pJ of energy per synaptic event at 5V.

HiAER-IFAT [15,16]. The Hierarchical address-event routing integrate-and-fire array transceiver (HiAER-IFAT) provides a multiscale tree based extension of AER synaptic routing for dynamically reconfigurable long-range synaptic connectivity in neuromorphic computing systems, developed in the lab of Gert Cauwenberghs at the University of California San Diego.

DeepSouth [17,18]. DeepSouth is the cortex emulator designed for simulating large and structurally connected spiking neural networks in the lab of André van Schaik at the MARCS Institute, Western Sydney University, Australia.

DYNAP [19-21]. The DYNAP (Dynamic Neurmorphic Asynchronous Processor) family of neuromorphic chips consists of dynap-se [19] and dynap-sel. DYNAP-SE implements a multi-core neuromorphic processor with scalable architecture fabricated using a standard 0.18 μm 1P6M CMOS technology. It is a full-custom asynchronous mixed-signal processor, with a fully asynchronous inter-core and inter-chip hierarchical routing architecture. Each core comprises 256 adaptive exponential integrate-and-fire (AEI&F) neurons for a total of 1k neurons per chip. Each neuron has a Content Addressable Memory (CAM) block, containing 64 addresses representing the pre-synaptic neurons that the neuron is subscribed to. Four different synapse types can be chosen for each synapse: fast excitatory/inhibitory, slow excitatory/inhibitory. Each synapse type is modelled by a dedicated DPI circuit [21] with globally shared bias values per core that determine synaptic weights and time constants. These circuits produce EPSCs and IPSCs (Excitatory/Inhibitory Post Synaptic Currents), with time constants that can range from a few microseconds to hundreds of milliseconds. The analog circuits are operated in the sub-threshold domain, thus minimizing the dynamic power consumption, and enabling implementations of neural and synaptic behaviors with biologically plausible temporal dynamics. For each core, there is an on-chip programmable temperature-compensated bias-generator which supplies 25 different parameters to the analog circuits to govern the behavior and dynamics of the neurons and synapses.The asynchronous CAMs on the synapses are used to store the tags of the source neuron addresses connected to them, while the SRAM cells are used to program the address of the destination core/chip that the neuron targets.

DynapSEL is a five-core fully-asynchronous mixed-signal spiking neural network chip with on-chip learning (STDP) fabricated in 28nm FDSOI process with a silicon area of 2.8mm x 2.6mm. The 28 nm Dynaps-sel chip is a mixed-signal multi-core neuromorphic processor that comprises four neural processing cores, each with 16 × 16 AEI&F neurons and 64 4-bit programmable synapses per neuron, and a fifth core with 1 × 64 neurons and 64 × 128 plastic synapses featuring on-chip learning circuits. The learning core also includes 64 × 64 non-plastic synapses. All synaptic inputs in all cores are triggered by incoming Address Events (AEs), which are routed among cores and across chips by asynchronous Address-Event Representation (AER) digital router circuits. Neurons integrate synaptic input currents and eventually produce output spikes, which are translated into AEs and routed to the desired destination via the AER routing circuits. Both chips include a 3-

level hierarchical routing architecture for memory efficient routing of the events between multiple cores and chips which makes them scalable to much larger networks.

2IFWTA chip [22,23]**.** The 2DIFWTA (2D Integrate-and-Fire Winner-Take-All) chip was developed at the cluster of Excellence in Cognitive Interaction Technology CITEC and Bielefeld University, Germany in the lab of Pr. Elisabetta Chicca. The 2DIFWTA chip was implemented using a standard 0.35-µm four-metal CMOS technology (Figure 18). It comprises a two-dimensional array of 32 × 64 (2,048) I&F neurons. Each neuron) receives inputs from AER synapses (two excitatory and one inhibitory) and local excitatory synapses. The local connections implement recurrent cooperation for either a two-dimensional or 32 mono-dimensional WTA networks. Cooperation in 2D involves first-neighbour connections, while cooperation in 1D involves first- and second-neighbour connections. Competition has to be implemented through the AER communication protocol, and it is therefore flexible in terms of connectivity pattern.

Tianjic chip [24,25]. The Tianjic chip is a fully synchronous digital ASIC that integrates two approaches, namely the prevailing computer-science-based artificial neural network (ANN) and neuroscience-inspired (SNN) models and algorithms, to provide a hybrid, synergistic platform. The Tianjic chip adopts a many-core architecture, reconfigurable building blocks and a streamlined dataflow with hybrid coding schemes. A 28-nm prototype chip was fabricated in the in UMC 28-nm HLP CMOS process with >610-GB/s internal memory bandwidth. Tianjic is the first unified ASIC that covers most neural network models across neuromorphic computing and deep learning. The unified functional core (UFC) has a number of neurons N = 256 and 156 UFCs are integrated in one single chip. Tianjic requires 5050 clock cycles to complete a round of computation and communication, which reflects the minimum latency of the time phase. For a single chip, the effective peak power efficiency is 1.28 TOPS/W (ANN mode) and 649 GSyOPS/W (SNN mode), and the internal memory bandwidth could reach >610 GB/s.

ODIN [26]. Odin is a 28-nm digital neuromorphic chip by Catholic University Louvain in 2019 supporting simple forms of on-chip spike-driven synaptic plasticity [8]. The core sup- ports 256 neurons that can be configured to implement first- order LIF dynamics as well as second-order Izhikevich dy- namics. The neuronal parameters are stored in a 4-kilobyte SRAM array, and a global controller is used to time-multi- plex the neuron logic circuit to implement the dynamics of the neurons in a sequential fashion. The core also integrates 3-bit 2562 synapses, which are implemented as a 32-kilobyte SRAM array. An additional bit is used per synapse to enable or disable online learning.

Finally, **a group at Intel has recently reported in 2019 a paper** [27] describing a reconfigurable 4096-neuron, 1M-synapse chip in 10-nm FinFET CMOS . The **SNN f**eatures digital circuits for leaky integrate and fire neuron models, on-chip spike-timing-dependent plasticity (STDP) learning, and high-fan-out multicast spike communication. The SNN achieves a peak throughput of 25.2 GSOP/s at 0.9 V, peak energy efficiency of 3.8 pJ/SOP at 525 mV, and 2.3-µW/neuron operation at 450 mV. On-chip unsupervised STDP trains a spiking restricted Boltzmann machine to de-noise Modified National Institute of Standards and Technology (MNIST) digits and to reconstruct natural scene images with RMSE of 0.036. A binary-activation multilayer perceptron with 50% sparse weights is trained offline with error backpropagation to classify MNIST digits with 97.9% accuracy at 1.7-µJ/classification.

Finally, additional literature works are also related to accelerators, for completeness we report the related references of some of them: Eyeriss [28], ESE [29], EIE [30], DRISLA [31], DNA [32]

1.3. References of section I.

[1] Steve Furber, Large-scale neuromorphic computing systems, J. Neural Eng. 13, 051001 (2016)

[2] C. S. Thakur, J. L. Molin, G. Cauwenberghs, G. Indiveri, K. Kumar, N. Qiao, J. Schemmel, R. Wang, E. Chicca, J. Olson Hasler, J.-s Seo, S. Yu, Y. Cao, A. van Schaik, R. Etienne-Cummings, Large-scale neuromorphic spiking array processors: a quest to mimic the brain, Front. Neurosci. 12:891 (2018)

[3] P. Merolla, J. Arthur, F. Akopyan, N. Imam, R. Manohar, D. S. Modha, A digital neurosynaptic core using embedded crossbar memory with 45pJ per spike in 45nm, in 2011 IEEE Custom Integrated Circuits Conference (CICC) (San Jose, CA) (2011), doi: 10.1109/CICC.2011.6055294

[4] P. A. Merolla et al., A million spiking-neuron integrated circuit with a scalable communication network and interface, Science 345, 668-73 (2014)

[5] Akopyan, F. et al., TrueNorth: design and tool flow of a 65 mw 1 million neuron programmable neurosynaptic chip, IEEE Trans. Comput. Aided Des. Integrated Circ. Syst. 34, 1537–1557 (2015).

[6] B. V. Benjamin et al., Neurogrid: a mixed-analog-digital multichip system for large-scale neural simulations, Proc. IEEE 102, 699-716 (2014)

[7] Sebastian Schmitt et al., Neuromorphic Hardware In The Loop: Training a Deep Spiking Network on the BrainScaleS Wafer-Scale System, Proceedings of the 2017 IEEE International Joint Conference on Neural Networks, DOI 10.1109/IJCNN.2017.7966125

[8] Johannes Schemmel et al., Accelerated Analog Neuromorphic Computing, 2020, arXiv https://arxiv.org/abs/2003.11996

[9] S. B. Furber et al., The SpiNNaker project, Proc. IEEE 102, 652-665 (2014)

[10] L. A. Plana et al., SpiNNaker: design and implementation of a GALS multi-core system-on-chip, ACM J. Emerg. Technol. Comput. Syst. 7, 1–17 (2011)

[11] E. Painkras et al., SpiNNaker: a 1W 18-core system-on chip for massively-parallel neural network simulation, IEEE J. Solid-State Circuits 48, 1943–53 (2013)

[12] S. B. Furber et al., Overview of the SpiNNaker system architecture, IEEE Trans. Comput. 62, 2454–67 (2013)

[13] M. Davies, N. Srinivasa, T. H. Lin, G. Chinya, Y. Cao, S. H. Choday, , et al. , Loihi: a neuromorphic many core processor with on-chip learning, IEEE Micro 38, 82–99 (2018), doi: 10.1109/MM.2018.112130359

[14] Lichtsteiner, P., Posch, C., and Delbruck, T. (2008). A 128 × 128 120 dB 15 µs latency asynchronous temporal contrast vision sensor. IEEE J. Solid State Circuits 43, 566–576. doi: 10.1109/JSSC.2007.914337

[15] Park, J., Ha, S., Yu, T., Neftci, E., and Cauwenberghs, G. (2014). "A 65k-neuron 73-Mevents/s 22-pJ/event asynchronous micro-pipelined integrate-and-fire array transceiver," in 2014 IEEE Biomedical Circuits and Systems Conference (BioCAS) Proceedings (Lausanne), 675–678.

[16] Park, J., Yu, T., Joshi, S., Maier, C., and Cauwenberghs, G. (2017). Hierarchical address event routing for reconfigurable large-scale neuromorphic systems. IEEE Trans. Neural Networks Learn. Syst. 28, 2408–2422, doi: 10.1109

[17] Wang, R., Thakur, C. S., Cohen, G., Hamilton, T. J., Tapson, J., and van Schaik, A. (2017). Neuromorphic hardware architecture using the neural engineering framework for pattern recognition. IEEE Trans. Biomed. Circuits Syst. 11,574–584. doi: 10.1109/TBCAS.2017.2666883

[18] Wang, R., Hamilton, T. J., Tapson, J., and Van Schaik, A. (2014b). "A compact reconfigurable mixed-signal implementation of synaptic plasticity in spiking neurons," in 2014 IEEE International Symposium on Circuits and Systems (ISCAS) (Melbourne VIC), 862–865. doi: 10.1109/ISCAS.2014.6865272

[19] Moradi, S., Qiao, N., Stefanini, F., and Indiveri, G. (2018). A scalable multicore architecture with heterogeneous memory structures for Dynamic Neuromorphic Asynchronous Processors (DYNAPs). IEEE Trans. Biomed. Circuits Syst. 12, 106–122. doi: 10.1109/TBCAS.2017.2759700

[20] Qiao, N., and Indiveri, G. (2016). "Scaling mixed-signal neuromorphic processors to 28 nm FD-SOI technologies," in 2016 IEEE Biomedical Circuits and Systems Conference (BioCAS) (Shanghai), 552–555. doi: 10.1109/BioCAS.2016.7833854

[21] Chicca, E., Stefanini, F., Bartolozzi, C., & Indiveri, G. (2014). Neuromorphic electronic circuits for building autonomous cognitive systems. Proceedings of the IEEE, 102(9), 1367-1388. doi: 10.1109/JPROC.2014.2313954

[22] Rost, T., Ramachandran, H., Nawrot, M. P., and Chicca, E. "A neuromorphic approach to auditory pattern recognition in cricket phonotaxis," in 2013 European Conference on Circuit Theory and Design (ECCTD) (Dresden).

[23] Engelmann, J., Walther, T., Grant, K., Chicca, E., and Gómez-Sena, L. , Modeling latency code processing in the electric sense: from the biological template to its VLSI implementation. Bioinspir. Biomim. 11:055007 (2016), doi: 10.1088/1748-3190/11/5/055007

[24] J. Pei, L. Deng, S. Song, M. Zhao, Y. Zhang, S. Wu, G. Wang, Z. Zou, Z. Wu, W. He, F. Chen, N. Deng, S. Wu, Y. Wang, Y Wu, Z. Yang, C. Ma, G. Li, W. Han, H. Li, H. Wu, R. Zhao, Y. Xie & Luping Shi, Towards artificial general intelligence with hybrid Tianjic chip architecture, Nature 572, 106 (2019)

[25] L. Deng, G. Wang, G. Li, S. Li, L. Liang, M. Zhu, Y. Wu, Z. Yang, Z. Zou, J. Pei, Z. Wu, X. Hu, Y. Ding, W. He, Y. Xie, and L. Shi, Tianjic: A Unified and Scalable Chip Bridging Spike-Based and Continuous , Neural Computation, n IEEE Journal of Solid-State Circuits, pp 1-19 (2020) doi: 10.1109/JSSC.2020.2970709.

[26] C. Frenkel, M. Lefebvre, J. Legat, and D. Bol, "A 0.086-mm2 12.7-pJ/SOP 64k-synapse 256-neuron online-learning digital spiking neuromorphic processor in 28-nm CMOS," IEEE Trans. Biomed. Circuits Syst., vol. 13, no. 1, pp. 145–158, 2019.

[27] G. K. Chen, R. Kumar, H. E. Sumbul, P. C. Knag and R. K. Krishnamurthy, "A 4096-Neuron 1M-Synapse 3.8-pJ/SOP Spiking Neural Network With On-Chip STDP Learning and Sparse Weights in 10-nm FinFET CMOS," in IEEE Journal of Solid-State Circuits, vol. 54, no. 4, pp. 992-1002, April 2019, doi: 10.1109/JSSC.2018.2884901.

[28] Chen, Y.-H. et al., Eyeriss: an energy-efficient reconfigurable accelerator for deep convolutional neural networks. IEEE J. Solid-State Circuits 52, 127–138 (2017).

[29] Han, S. et al. ESE: efficient speech recognition engine with sparse LSTM on FPGA. In Proc. 2017 ACM/SIGDA Int. Symposium on Field-Programmable Gate Arrays 75–84 (ACM, 2017).

[30] Han, S. et al. EIE: efficient inference engine on compressed deep neural network. In 2016 ACM/IEEE 43rd Annual Int. Symposium on Computer Architecture 243–254 (IEEE, 2016).

[31] S. Li, D. Niu, K. T. Malladi, H. Zheng, B. Brennan, and Y. Xie, "DRISA: A DRAM-based reconfigurable in-situ accelerator," in Proc. 50th Annu. IEEE/ACM Int. Symp. Microarchitecture (MICRO), 2017, pp. 288–301.

[32] F. Tu, S. Yin, P. Ouyang, S. Tang, L. Liu, and S. Wei, "Deep convolutional neural network architecture with reconfigurable computation patterns," IEEE Trans. Very Large Scale Integr. (VLSI) Syst., vol. 25, no. 8, pp. 2220–2233, Aug. 2017.

Section II: EMERGING NEUROMORPHIC TECHNOLOGIES BEYOND CMOS

2.1 Memristive devices technologies: from physics of novel devices to the implementation of synaptic and neuron functionalities

The so-named *memristive device technologies* include a broad class of two or three-terminal devices whose resistance can be modified upon electrical stimuli. The resistance changes can last for short or long time scales, leading to a volatile or non-volatile memory effect, respectively. Memristive devices are based on a large variety of physical mechanisms, such as redox reactions and ion migration, phase transitions, spin-polarized tunnelling, and ferroelectric polarization [1-6], and they have the potential to meet the considerable demand for new devices that enable energy-efficient and area-efficient information processing beyond the von Neumann paradigm [1, 6-9]. The leading memristive technologies which are currently at high maturity level are those firstly developed as non-volatile memory devices for storage applications and then integrated in large arrays and in combination with CMOS, namely *resistive random access memory (RRAM), Phase change memory (PCM), Ferroelectric memory (FeRAM), and magnetoresistive random access memories (MRAM).* Recently, RRAM, PCM, FeRAM and spin-transfer torque MRAM have been receiving increasing interest for neuromorphic computing, and many hardware demonstration have been reported at device, but also circuit and systems level [10-12]. The results are promising and despite the system level integration is still not at the level of the fully CMOS-based one, the field is improving very fast, and driven by the parallel advancement of these technologies and their CMOS integration for storage or in-memory computing applications. In addition to the more consolidated technologies, many developments are underway towards new and less matures concepts which span from new materials (2D, nanowires) [13,14], devices based on metal-insulator transition (for instance VO_2-based devices) [15,16], organic material [17], advanced device concepts in the field of spintronics (domain wall, race-trace memory, skyrmions) [6,18] and photonics[19] . A recent review on emerging neuromorphic devices and architectures enabled by quantum dots, metal nanoparticles, polymers, nanotubes, nanowires, two-dimensional layered materials and van der Waals heterojunctions can be found in [20].

The interesting device features which can be exploited for neuromorphic computing are, in some extent, the same engineered for storage applications. In particular we can mention the capability to retain the information for extended time (i.e., their non-volatility), fast switching speed, low switching energy, long cycling switching endurance, compact size, low process temperature fabrication (down to < 400 °C), compatibility with CMOS integration, stackability on multi-layer to increase the density. Despite the differences and peculiarity of each technology, the listed properties can enable the use of memristive device technologies in complex circuits and systems, and the high device density decreases the cost of computing systems. Moreover, it is worth noting that these devices can exhibit additional interesting features which can be explored and optimized for neuromorphic computing, in particular the multilevel state or analogue operation, stochasticity and intrinsic variability, rich dynamics of the devices including the possibility to engineer their retention in different time scales [1,7-9] . Today, there is therefore a significant effort in the scientific and industrial community to take advantage of these new technologies to build a brain inspired computing hardware, mimicking key features of biological synapses and neurons, such non-volatility and plasticity, as well as oscillatory and stochastic behaviour. While it is not generally true that a single memristive device can implement at hardware level all the desired functionalities reproducing the synaptic or neural dynamics, memristive devices can enable the fabrication of small circuit blocks for synapses and neurons, bringing the additional advantage, with respect to standard CMOS solutions, of non-volatility and overall smaller size. Many solutions have been currently already proposed for the hardware implementation of synaptic and neuronal functionalities, as listed below.

2.2. Synapse Implementation.

The key features of artificial synapses are the ability to update their states given new information (learning, plasticity) and to store analogue information (memory). Two approaches have been mainly proposed to implement synapses: analogue synapses which exploit the multilevel or analog control of RRAM [1,2,21], PCM [1,22,23] and FeRAM [2,24-26] devices; and advanced spintronic devices storing analogue information in magnetic textures (as demonstrated through domain wall motion in magnetic tunnel junctions, or representing analogue information in the number of magnetic skyrmions). [6, 18] . A second approach relies on the use of binary stochastic device, as demonstrated for filamentary RRAM [9,27,28], and STT-MRAM [29]

2.3 Neuron function implementation.

Despite currently the neuron functionalities in hardware neural network can be implemented in CMOS by using transistors and capacitor, the stochastic, volatility and non-linear properties of memristive device technologies pave the way of building advanced low power and compact hardware neuronal blocks representing complex and biological inspired neural function. In particular, we can mention FeRAM [30] , VO_2 –based MIT devices [14,31,32], PCM [33], STT-MRAM [34], and spin-torque nano-oscillators (i.e. specific types of magnetic tunnel junctions, which can be driven into spontaneous microwave oscillations by an injected direct current) [6,35]

2.4 References of Section II.

[1] Z. Wang, H. Wu, G. W. Burr, C. S. Hwang, K. L. Wang, Q. Xia. J. J. Yang, Resistive switching materials for information processing, Nat. Rev. Mater. (2020). https://doi.org/10.1038/s41578-019-0159-3

[2] S. Slesazeck and T. Mikolajick, Nanoscale resistive switching memory devices: a review, Nanotechnology 30, 352003 (2019)

[3] D. Wouters, R. Waser, M. Wuttig, Phase-Change and Redox-Based Resistive Switching Memories, Proceedings of the IEEE 103(8):1274-1288, (2015)

[4] R. Waser, R. Dittmann, G. Staikov, K. Szot, Redox-Based Resistive Switching Memories – Nanoionic Mechanisms, Prospects, and Challenges, Adv Mater. 21, 2632–63 (2009)

[5] S. Bhatti, R. Sbiaa, A. Hirohata, H. Ohno, S. Fukami, N. Piramanayagam, Spintronics based random access memory: a review, Materials Today 20 (Issue 9) 530-548 (2017), https://doi.org/10.1016/j.mattod.2017.07.007

[6] J. Grollier, D. Querlioz, K. Y. Camsari, K. Everschor-Sitte, S. Fukami & M. D. Stiles, Neuromorphic spintronics, *Nat Electron* (2020). https://doi.org/10.1038/s41928-019-0360-9

[7] Y. Zhang, Z. Wang, J. Zhu, Y. Yang, M. Rao, W. Song, Y. Zhuo, X. Zhang, M. Cui, L. Shen, R. Huang, and J. Joshua Yang, Brain-inspired computing with memristors: Challenges in devices, circuits, and systems, Applied Physics Reviews 7, 011308 (2020); https://doi.org/10.1063/1.5124027

[8] A. Sebastian, M. Le Gallo , R. Khaddam-Aljameh and E. Eleftheriou, Memory devices and applications for in-memory computing, Nat. Nanotechnol. (2020). https://doi.org/10.1038/s41565-020-0655-z

[9] R. Carboni and D. Ielmini, Stochastic Memory Devices for Security and Computing, Adv. Electron. Mater. 2019, 1900198, DOI: 10.1002/aelm.201900198

[10] A. Valentian, F. Rummens, E. Vianello, T. Mesquida, C. Lecat-Mathieu de Boissac, O. Bichler, C. Reita, Fully Integrated Spiking Neural Network with Analog Neurons and RRAM Synapses, IEDM 2019, p.314-318

[11] P. Yao, H. Wu, B. Gao, J. Tang, Q. Zhang, W. Zhang, J. Joshua Yang,He Qian, Fully hardware-implemented memristor convolutional neural network, Nature 577, 641–646(2020)

[12] M. Ishii, S. Kim, S. Lewis, A. Okazaki, J. Okazawa, M. Ito, M. Rasch, W. Kim, A. Nomura, U. Shin, K. Hosokawa, M. BrightSky, and W. Haensch, On-Chip Trainable 1.4M 6T2R PCM Synaptic Array with 1.6K Stochastic LIF Neurons for Spiking RBM, *IEDM 2019, p.310-313*

[13] C.-Y. Wang, C. Wang, F. Meng, P. Wang, S. Wang, S.-J. Liang, and F. Miao, 2D Layered Materials for Memristive and Neuromorphic Applications, Adv. Electron. Mater. 6, 1901107 (2020)

[14] G. Milano, S. Porro, I. Valov, C. Ricciardi, Nanowire Memristors: Recent Developments and Perspectives for Memristive Devices Based on Metal Oxide Nanowires, Adv. Electron. Mater. 9, 2019: https://doi.org/10.1002/aelm.201970044

15] J. d. Valle, Y. Kalcheim, J.Trastoy, A. Charnukha, D. N. Basov, and I. K. Schuller, Electrically Induced Multiple Metal-Insulator Transitions in Oxide Nanodevices , Phys. Rev. Applied 8, 054041 (2017)

[16] Wei Yi, Kenneth K. Tsan, Stephen K. Lam, Xiwei Bai, Jack A. Crowell & Elias A. Flores, Biological plausibility and stochasticity in scalable VO2 active memristor neurons, NATURE COMMUNICATIONS | (2018) 9:4661 | DOI: 10.1038/s41467-018-07052-w

[17] Battistoni, V.Erokhin, S.Iannotta, Frequency driven organic memristive devices for neuromorphic short term and long term plasticity,Organic Electronics 65, 434-438 (2019), https://doi.org/10.1016/j.orgel.2018.11.033

[18] W. Kang *et al.*, "Magnetic skyrmions for future potential memory and logic applications: Alternative information carriers," *2018 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, Dresden, 2018, pp. 119-124.

[19] J. Feldmann, N. Youngblood, C. D. Wright, H. Bhaskaran & W. H. P. Pernice, All-optical spiking neurosynaptic networks with self-learning capabilities, Nature 569, pages208–214 (2019)

[20] Vinod K. Sangwan and Mark C. Hersam, Neuromorphic nanoelectronic materials, Nat. Nanotechnol. (2020). https://doi.org/10.1038/s41565-020-0647-z

[21] E. Covi, S. Brivio, A. Serb, T. Prodromakis, M. Fanciulli, and S. Spiga, Analog Memristive Synapse in Spiking Networks Implementing Unsupervised Learning, Front. Neurosci.10:482, (2016), doi: 10.3389/fnins.2016.00482

[22] I.Boybat, M. Le Gallo, S.R. Nandakumar, T. Moraitis, T. Parnell, T. Tuma, B. Rajendran, Y. Leblebici, A. Sebastian & E.vangelos Eleftheriou, Neuromorphic computing with multi-memristive synapses, NATURE COMMUNICATIONS 9, 2514 (2018); DOI: 10.1038/s41467-018-0493

[23] S. La Barbera, DRB Ly, G. Navarro, N. Castellani,O. Cueto, G. Bourgeois, et al., Narrow Heater Bottom Electrode-Based Phase Change Memory as a Bidirectional Artificial Synapse. Adv Electron Mater. 4(9):1800223 (2018)

[24] H. Mulaosmanovic, J. Ocker, S. Müller, M. Noack, J. Müller, P. Polakowski, T. Mikolajick and S. Slesazeck, Novel ferroelectric FET based synapse for neuromorphic systems, Symposium on VLSI Technology, pp. T176 - T177, 2017.

[25] S. Boyn et al., Learning through ferroelectric domain dynamicsin solid-state synapses, NATURE COMMUNICATIONS 8, 14736 (2016), DOI: 10.1038/ncomms147

[26] M. Halter, L. Bégon-Lours, V. Bragaglia, M. Sousa, B. J. Offrein, S. Abel, M. Luisier, J. Fompeyrine, Back-End, CMOS-Compatible Ferroelectric Field-Effect Transistor for Synaptic Weights, ACS Appl. Mater. Interfaces 2020, https://doi.org/10.1021/acsami.0c00877

[27] M. Suri, D. Querlioz, O. Bichler, G. Palma, E. Vianello, D. Vuillaume, et al., Bio-Inspired Stochastic Computing using Binary CBRAM Synapses, IEEE Trans Electron Devices.60, 2402 (2013).

[28] A. Yousefzadeh, E. Stromatias, M. Soto, T. Serrano-Gotarredona, and B. Linares-Barranco, On Practical Issues for Stochastic STDP Hardware With 1-bit Synaptic Weights, Front. Neurosci. 12:665 (2018), doi: 10.3389/fnins.2018.00665

[29] A. F. Vincent et al., "Spin-Transfer Torque Magnetic Memory as a Stochastic Memristive Synapse for Neuromorphic Systems," in IEEE Transactions on Biomedical Circuits and Systems, vol. 9, no. 2, pp. 166-174, April 2015, DOI: 10.1109/TBCAS.2015.2414423

[30] H. Mulaosmanovic, E. Chicca, M. Bertele, T. Mikolajick and S. Slesazeck, Mimicking biological neurons with a nanoscale ferroelectric transistor, Nanoscale, 10, 21755 (2018).

[31] Wei Yi, Kenneth K. Tsan, Stephen K. Lam, Xiwei Bai, Jack A. Crowell & Elias A. Flores, Biological plausibility and stochasticity in scalable VO2 active memristor neurons, NATURE COMMUNICATIONS | (2018) 9:4661 | DOI: 10.1038/s41467-018-07052-w

[32] M. Jerry, A. Parihar, B. Grisafe, A. Raychowdhury and S. Datta, Ultra-low power probabilistic IMT neurons for stochastic sampling machines, 2017 Symposium on VLSI Technology, Kyoto, 2017, pp. T186-T187.

[33] Tuma, T., Pantazi, A., Le Gallo, M., Sebastian, A. & Eleftheriou, E. Stochastic phase-change neurons. Nat. Nanotech. 11, 693–699 (2016).

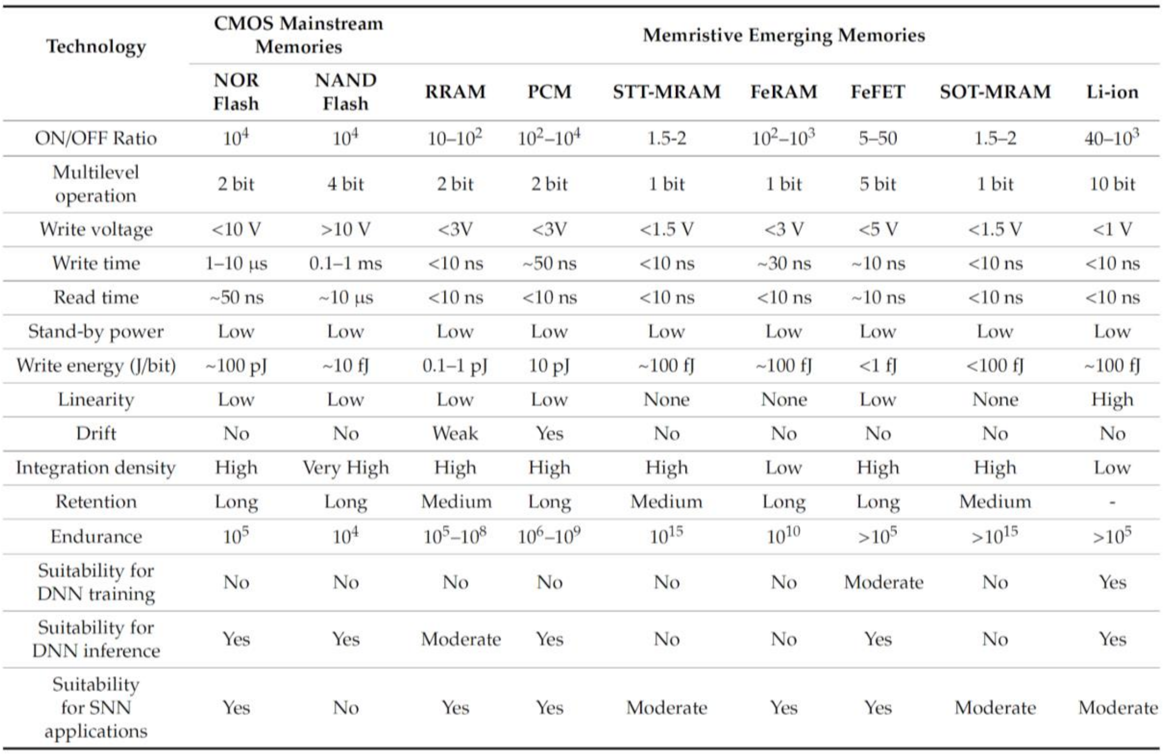
[34] M. Wu et al., "Extremely Compact Integrate-and-Fire STT-MRAM Neuron: A Pathway toward All-Spin Artificial Deep Neural Network," 2019 Symposium on VLSI Technology, Kyoto, Japan, 2019, pp. T34-T35.

[35] J. Torrejon, M. Riou, F. Abreu Araujo, S. Tsunegi, G. Khalsa, D. Querlioz, P. Bortolotti, V. Cros, A. Fukushima, H. Kubota, S. Yuasa, M. D. Stiles, and J. Grollier, Neuromorphic computing with nanoscale spintronic oscillators, Nature. 2017 Jul 26; 547(7664): 428–431. doi: 10.1038/nature23011

Section III. NEUROMORPHIC SYSTEMS BASED ON MIXED CMOS-MEMRISTIVE TECHNOLOGIES ARCHITECTURES

3.1. Introduction.

The aim of this section is to present the current state of the art of neuromorphic hardware based on *mixed CMOS-memristive device neuromorphic chips.* We will mainly focus on the demonstration of monolithic integrated CMOS-memristive devices in a chip. In addition we will list any relevant advancement on systems based on a mixed software – hardware level, or board demonstration. It is worth noting that currently most of the neuromorphic chip or chip for artificial intelligence still use external memories or costly embedded SRAM. Anyway, it more and more important to have memory embedded as close as possible to the processing element, and therefore embedded memory technology (or more in general memristive technologies) with CMOS is a hot topic for future neuromorphic chips. Then it is expected in the future that the number of proposed neuromorphic architectures which exploit new memory technologies will increase in the future. A summary and comparison of current emerging memories under investigation for neuromorphic computing can be found in the following Table (Figure 3.1) extracted from the review paper by V. Milo et al. [1]



Technology	CMOS Mainstream Memories		Memristive Emerging Memories						
	NOR Flash	NAND Flash	RRAM	PCM	STT-MRAM	FeRAM	FeFET	SOT-MRAM	Li-ion
ON/OFF Ratio	10^4	10^4	$10–10^2$	$10^2–10^4$	1.5-2	$10^2–10^3$	5–50	1.5–2	$40–10^3$
Multilevel operation	2 bit	4 bit	2 bit	2 bit	1 bit	1 bit	5 bit	1 bit	10 bit
Write voltage	<10 V	>10 V	<3V	<3V	<1.5 V	<3 V	<5 V	<1.5 V	<1 V
Write time	1–10 µs	0.1–1 ms	<10 ns	~50 ns	<10 ns	~30 ns	~10 ns	<10 ns	<10 ns
Read time	~50 ns	~10 µs	<10 ns	<10 ns	<10 ns	<10 ns	~10 ns	<10 ns	<10 ns
Stand-by power	Low	Low	Low	Low	Low	Low	Low	Low	Low
Write energy (J/bit)	~100 pJ	~10 fJ	0.1–1 pJ	10 pJ	~100 fJ	~100 fJ	<1 fJ	<100 fJ	~100 fJ
Linearity	Low	Low	Low	Low	None	None	Low	None	High
Drift	No	No	Weak	Yes	No	No	No	No	No
Integration density	High	Very High	High	High	High	Low	High	High	Low
Retention	Long	Long	Medium	Long	Medium	Long	Long	Medium	-
Endurance	10^5	10^4	$10^5–10^8$	$10^6–10^9$	10^{15}	10^{10}	$>10^5$	$>10^{15}$	$>10^5$
Suitability for DNN training	No	No	No	No	No	No	Moderate	No	Yes
Suitability for DNN inference	Yes	Yes	Moderate	Yes	No	No	Yes	No	Yes
Suitability for SNN applications	Yes	No	Yes	Yes	Moderate	Yes	Yes	Moderate	Moderate

Figure 3.1. Comparison of key features exhibited by CMOS mainstream memory devices and memristive emerging memory devices under investigation to implement neuromorphic computing in hardware. Reproduced from [1].

3.2 Mixed CMOS-memristive devices architecture: monolithic integration of memristive devices (RRAM and PCM) with CMOS.

In this paragraph we describe few examples of fully integrated CMOS-RRAM neuromorphic chips.

SPIRIT chip. SPIKING NEURAL NETWORK WITH ANALOG NEURONS AND RRAM SYNAPSES - CEA-LETI, France

The **Spirit chip** [2] is a chip featuring the complete integration of a Spiking Neural Network, combining analog neurons and Resistive RAM (RRAM)-based synapses. The implemented topology is a perceptron, aimed at performing MNIST classification. An existing framework was tailored for offline learning and weight quantization. The test chip, fabricated in 130nm CMOS, shows well-controlled integration of synaptic currents and no RRAM read disturb issue during inference tasks (at least 750M spikes). The number of RRAM synapses/mm2 is 16 kbit, The classification accuracy is 84%, with a 3.6 pJ energy dissipation per spike at the synapse and neuron level (up to 5x lower vs. similar chips using formal coding).

Moreover, additional chips are the one proposed by CNRS and CEA-LETI [3], and by CEA.LETI – Stanford [4]. At ISSCC 2019, CEA-LETI and Stanford University jointly presented a testchip integrating **18kB of ReRAM on top of 130nm silicon CMOS** with a MCU 16-bit with 8KB of SRAM [4]. The proof-of-concept chip was validated for a variety of applications including machine learning, control, security, AIoT.

RAND chip. Resistive Analog Neuro Device (RAND) chip from PANASONIC

The **RAND** chip proposed by Panasonic [5] is a low-power and high-accuracy neural-network (NN) processor using ReRAM to store weights as analog resistance. They proposed a ReRAM perceptron circuit for realizing large scale integration, highly accurate cell current controlled writing scheme, and flexible network architecture (FNA) in which any NNs can be configured. The fabricated 180nm test chip shows well-controlled analog cell current with linear 30µA dynamic range and 0.59µA variation of 1 sigma, results in 90.8% MNIST numerical recognition rate. Furthermore, 4M synapses integrated 40nm test chip achieves lower analog cell current and 66.5 TOPS/W power efficiency. The RAND chip fabricated by 180nm process consumes power of 15.8mW on a 1024 input inference-READ, achieving power efficiency of 20.7 TOPS/W. In addition, 40nm ReRAM reduces power consumption during an inference-READ to 9.9mW, thus achieving power efficiency of 66.5 TOPS/W

ReASOn chip – University of Zürich, Switzerland: 130 nm CMOS test-chip with integrated RRAM.

ReASOn (Resistive Array of Synapses with ONline learning) [6] is a test-chip featuring 2048 memristive devices (HfO_2 RRAM) that implement 1024 memory cells connected to 2 neurons via a programmable routing fabric. The neurons include circuits that implement online learning. This chip also features test circuits relevant for neuromorphic applications including an alpha synapse, shunting synapse, Hebbian/anti-Hebbian/stop learning, and a new neuron circuit. All the circuits are highly tunable using external biases to make testing and experimentation easier. The ReASOn chips was developed under the NeuRAM^3 project (http://www.neuram3.eu). The hardware realization is fully-integrated with CMOS by two-stage process: traditional CMOS process to the penultimate metal layer (foundry fabrication); and RRAM devices fabricated by CEA-LETI as a post processing.

Process	130 nm CMOS with an extra post-processing step
Power domains	1.2 V and 4.7 V
Number of IOs and packaging	104, CQFP128
Area of the chip	2705.16 um x 1905.16 um
Neuron type	Leaky Integrate and Fire with learning block
Number of Synapses and Neurons	1024 plastic synapses and 2 neurons

Specification of the ReASOn chip, as from the http://www.neuram3.eu web site

Non-volatile-memory for synaptic signal processing – IBM Research

At IBM Research, the work on integrating analog synaptic signal processing on digital CMOS is bundled in the AI Hardware Center [7]. The center is focused on enabling next-generation chips and systems that support the tremendous processing power and unprecedented speed that AI requires to realize its full potential. In Europe, the IBM Research lab located in Zurich is a strong contributor to this effort. Crossbar arrays containing memristive devices are integrated in the back-end-of-line of a CMOS chip. The array performs an analog computing of the synaptic interconnect between two layers of neurons in a deep neural network. The memristors represent the weights of the neural network. Requirements on the memristors depend on their application, inference or training. For deep neural network inference, the memristive devices must be tuned to the desired resistance values at up to 100 levels. Long term retention and stability are key properties for this application. Memristive elements based on phase change materials (PCM) are an excellent candidate [8]. High-end demonstrations showing the integration of these devices in the back-end-of-line of a CMOS process were realized [8-10]. For training neural networks on an analog synaptic signal processor, the controlled change of the resistance is essential. PCM devices can be tuned from higher to lower resistance by crystallizing the phase change element. Moving back to a high resistance requires a full amorphization. Resistive RAM devices based on oxides in which a filamentary conductive path is formed offer the potential for CMOS compatibility and enhanced resistance control [11, 12]. The technological efforts are complemented with improvements of DNN training algorithms to facilitate the application of analog synaptic signal processing [13].

3.3. References of section III.

[1] Valerio Milo , Gerardo Malavena, Christian Monzio Compagnoni and Daniele Ielmini, Memristive and CMOS Devices for Neuromorphic Computing, Materials 2020, 13, 166; doi:10.3390/ma13010166

[2] A. Valentian, F. Rummens, E. Vianello, T. Mesquida, C. Lecat-Mathieu de Boissac, O. Bichler, C. Reita, Fully Integrated Spiking Neural Network with Analog Neurons and RRAM Synapses, 2019 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, USA, 2019, pp. 14.3.1-14.3.4, doi: 10.1109/IEDM19573.2019.8993431.

[3] T. Hirtzlin *et al.*, "Hybrid Analog-Digital Learning with Differential RRAM Synapses," *2019 IEEE International Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2019, pp. 22.6.1-22.6.4, doi: 10.1109/IEDM19573.2019.8993555.

[4] Tony F. Wu, Binh Q. Le, Robert Radway, Andrew Bartolo, William Hwang, Seungbin Jeong, Haitong Li1 Pulkit Tandon, Elisa Vianello, Pascal Vivet, Etienne Nowak, Mary K. Wootters, H.-S. Philip Wong, Mohamed M. Sabry Aly, Edith Beigne, Subhasish Mitra, "14.3 A 43pJ/Cycle Non-Volatile Microcontroller with 4.7µs Shutdown/Wake-up Integrating 2.3-bit/Cell Resistive RAM and Resilience Techniques," 2019 IEEE International Solid- State Circuits Conference - (ISSCC), San Francisco, CA, USA, 2019, pp. 226-228, doi: 10.1109/ISSCC.2019.8662402.

[5] R. Mochida et al., "A 4M Synapses integrated Analog ReRAM based 66.5 TOPS/W Neural-Network Processor with Cell Current Controlled Writing and Flexible Network Architecture," 2018 IEEE Symposium on VLSI Technology, Honolulu, HI, 2018, pp. 175-176, doi: 10.1109/VLSIT.2018.8510676.

[6] ReaSONs. http://www.neuram3.eu web site

[7] IBM AI Hardware Center. https://www.research.ibm.com/artificial-intelligence/ai-hardware-center/

[8] Kersting, B., Ovuka, V., Jonnalagadda, V.P. et al. State dependence and temporal evolution of resistance in projected phase change memory. Sci Rep 10, 8248 (2020). https://doi.org/10.1038/s41598-020-64826-3

[9] Ambrogio, S., Narayanan, P., Tsai, H. et al. Equivalent-accuracy accelerated neural-network training using analogue memory. Nature 558, 60–67 (2018). https://doi.org/10.1038/s41586-018-0180-5

[10] E. Eleftheriou et al., "Deep learning acceleration based on in-memory computing," in IBM Journal of Research and Development, vol. 63, no. 6, pp. 7:1-7:16, 1 Nov.-Dec. 2019. https://doi.org/10.1147/JRD.2019.2947008

[11] Chen, P.-Y. et al. Mitigating effects of non-ideal synaptic device characteristics for on-chip learning. In2015 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)https://doi.org/10.1109/iccad.2015.7372570

[12] Gong, N., Idé, T., Kim, S. et al. Signal and noise extraction from analog memory elements for neuromorphic computing. Nat Commun 9, 2102 (2018). https://doi.org/10.1038/s41467-018-04485-1

[13] T Gokmen, W Haensch, Algorithm for Training Neural Networks on Resistive Device Arrays, Frontiers in Neuroscience, 2020. doi: 10.3389/fnins.2020.00103