DOI: 10.14529/jsfi230206

Neuromorphic Computing Based on CMOS-Integrated Memristive Arrays: Current State and Perspectives*

Alexey N. Mikhaylov[1], *Evgeny G. Gryaznov*[1], *Maria N. Koryazhkina*[1], *Ilya A. Bordanov*[2], *Sergey A. Shchanikov*[1,3], *Oleg A. Telminov*[4], *Victor B. Kazantsev*[1,3]

The paper presents an analysis of current state and perspectives of high-performance computing based on the principles of information storage and processing in biological neural networks, which are enabled by the new micro- and nanoelectronics component base. Its key element is the memristor (associated with a nonlinear resistor with memory or Resistive Random Access Memory (RRAM) device), which can be implemented on the basis of different materials and nanostructures compatible with the complementary metal-oxide-semiconductor (CMOS) process and allows computing in memory. This computing paradigm is naturally implemented in neuromorphic systems using the crossbar architecture for vector-matrix multiplication, in which memristors act as synaptic weights – plastic connections between artificial neurons in fully connected neural network architectures. The general approaches to the development and creation of a new component base based on the CMOS-integrated RRAM technology, development of artificial neural networks and neuroprocessors using memristive crossbar arrays as computational cores and scalable multi-core architectures for implementing both formal and spiking neural network algorithms are discussed. Technical solutions are described that enable hardware implementation of memristive crossbars of sufficient size, as well as solutions that compensate for some of the deficiencies or fundamental limitations inherent in emerging memristor technology. The performance and energy efficiency are analyzed for the reported prototypes of such neuromorphic systems, and a significant (orders of magnitude) gain in these parameters is highlighted compared to the computing systems based on traditional component base (including neuromorphic ones). Technological maturation of a new component base and creation of memristor-based neuromorphic computing systems will not only provide timely diversification of hardware for the continuous development and mass implementation of artificial intelligence technologies but will also enable setting the tasks of a completely new level in creating hybrid intelligence based on the symbiosis of artificial and biological neural networks. Among these tasks are the primary ones of developing brain-like self-learning spiking neural networks and adaptive neurointerfaces based on memristors, which are also discussed in the paper.

Keywords: memristor, CMOS integration, neuromorphic hardware, artificial intelligence.

Introduction

The fourth industrial revolution, on the brink of which the humanity stands, presents entirely new requirements for the hardware of artificial intelligence (AI) technologies, which should approach the capabilities of the human brain (natural intelligence). In addition to demands for compactness and energy efficiency, new AI hardware must be compatible with existing silicon microelectronics technology and with biological systems. Meeting these requirements will enable mass production of AI hardware systems and the implementation of new hybrid forms of AI. The second requirement implies that new electronic AI systems must not only replicate formally

[*] The paper is published as part of special issue “PST2023: Perspective Supercomputer Technologies”.

[1] Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, Russian Federation
[2] Murom Institute of Vladimir State University, Murom, Russian Federation
[3] Moscow Institute of Physics and Technology, Moscow, Russian Federation
[4] Joint-Stock Company Molecular Electronics Research Institute, Moscow, Russian Federation

(as they do now), but also reproduce functionally the elements of the nervous system and the brain.

Current paradigmatic changes in electronics are aimed at meeting these requirements associated with the transition from the traditional von Neumann architecture (in which memory and processing are separated in space) to analog computing in memory and massive parallelism in information processing similar to that in the brain. At the core of the new post-digital paradigm is a brain-like electronic component base represented by memristors (analog resistive memory devices) and memristive systems that mimic the functions of elements of the living nervous system (neurons and synapses). The diversity of possible computing architectures is ensured by the universal character of the memristive phenomenon, as it can be implemented in classical and quantum systems, in various artificial materials and structures (inorganic, organic, molecular, etc.), and in living systems.

The results of comprehensive research and diverse applications of memristive devices have become the subject of numerous publications in recent years (e.g., see [4, 12, 19, 20, 53, 55, 60], including roadmaps, reviews, and perspectives in top journals) showing the importance and relevance of this field at the global level, as well as the need to implement a master plan (coordinated and interdisciplinary efforts) in the field of bioinspired systems aimed at technological development of the new component base and creating prototypes of next-generation information-computing systems.

This paper presents the current state and prospects of high-performance computing based on memristors. We consider general approaches to the development of Resistive Random Access Memory (RRAM) integrated with complementary metal-oxide-semiconductor (CMOS) technology required for creating elements and functional blocks of a memristive neuroprocessor, as well as the application of new computing systems in artificial and hybrid intelligence technologies. To show the perspectives of memristor-based neuromorphic computing systems, their achieved parameters are compared to that of traditional computing systems.

The paper is organized as follows. Section 1 is devoted to a discussion of the relevance and prospects of research and development of memristors and memristor-based neuromorphic and neurohybrid systems. In Section 2, we discuss a multilevel and interdisciplinary approach to the development of neuromorphic systems based on CMOS-compatible memristive devices. Section 3 contains a consideration of various options for scaling up the CMOS-integrated memristive crossbars to increase the speed of signal transmission in artificial neural networks. Section 4 contains a comparison of neuromorphic computing systems based on traditional and new component base. Conclusion summarizes the study.

1. Memristor and Memristor-Based Neuromorphic and Neurohybrid Systems

Over the past five decades, global microelectronics has developed according to Moore's law, which predicts an exponential increase in the number of transistors on a chip, resulting in faster computing and reduced energy consumption for each new generation of technology. Currently, this trend has reached a physical limit – further increase in the number of transistors does not lead to an increase in clock speed or a reduction in energy consumption. The main bottleneck is the data exchange between the central processor and external memory, making digital processors based on traditional von Neumann architecture extremely inefficient in terms of

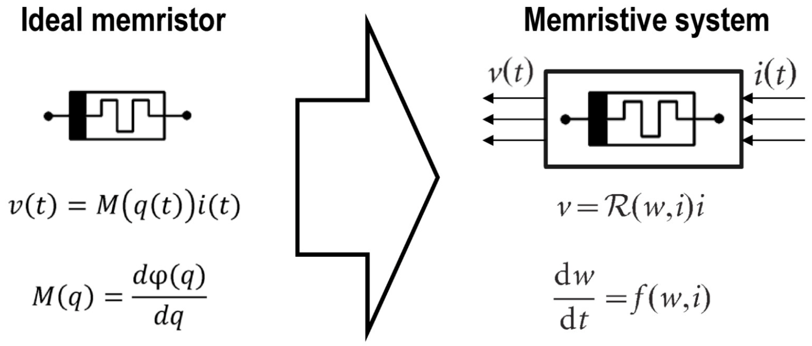


Figure 1. Original and generalized definitions of memristor [13, 14]

energy consumption and time delays. Meanwhile, the volume of digital data requiring processing continues to increase exponentially. Every two years, more data is created than in all of human history before that point. Unstructured data already comprises over 80 % of the total volume of data generated daily. Thus, the demand is growing faster than the performance of modern computers. Breakthrough technological solutions are required to address this von Neumann bottleneck. Currently, two main solutions are being explored in leading scientific centers around the world – combining computation and memory in a single functional unit, and transitioning from traditional von Neumann architectures to neuromorphic architectures that reproduce the principles of information storage and processing in the nervous system and brain.

The new paradigm in electronics, which is associated with a breakthrough in the hardware implementation of neuromorphic information-processing systems, is based on the use of memristors. The memristor (memory resistor) was theoretically described by Leon Chua in 1971 as a missing passive element of electrical circuits that relates the change in magnetic flux $\phi(t)$ to the electrical charge $q(t)$ [13] (Fig. 1). It can be shown that this element is equivalent to a nonlinear resistor that changes its resistance $M(q(t))$ depending on the history of the electrical charge flowing through it. This definition of an ideal memristor still causes doubts and disputes among scientists [15, 22, 50] and stimulates the search for materials and structures that exhibit a physical connection between magnetic and electrical properties [45]. However, in 1976 L. Chua and S. Kang proposed a generalized definition of memristors and memristive dynamical systems [14] that are described by a port equation equivalent to Ohm's law and a set of state equations that describe the dynamics of the internal state variables (w). This definition is universal and describes the change in resistance (memory effect) based on various phenomena in inorganic and organic nanomaterials (ion migration, redox reactions, phase transitions, spin and ferroelectric effects) [53], as well as in photonic [46] and superconducting [37, 41] circuits. Among them, it is necessary to highlight nanostructures of the metal-oxide-metal (MOM) type, which are ideal for creating compact (with nanometer-scale size) and energy-efficient (with femtojoules per switch) RRAM devices that can be integrated into the standard CMOS technological process. Such devices can not only store the logical value determined by conductivity, but also allow it to be changed in the same physical location implementing a non-von Neumann paradigm of in-memory computing. In addition, the simple structure of memristor enables the creation of ultra-dense and, in the future, three-dimensional arrays of crossbars that naturally (based on Ohm's and Kirchhoff's laws and in analog form) implement vector-matrix multiplication (VMM) operations, which underlies inference in traditional artificial neural networks with deep learning and new algorithms for training spiking neural networks [31].

The development of AI technologies relies on the development of neuromorphic computing systems according to the well-known forecast within the international technology roadmap: “The Future of AI is Neuromorphic”. Brain-like electronic components with memristors and memristive systems will provide timely diversification of hardware, which mainly imposes fundamental limitations on each cycle of AI development, and will prevent another “winter” of AI. Alternative neuromorphic technologies based on new component base are only just entering maturity, competing with currently dominant digital high-performance computing technologies. A detailed analysis and comparison of the achieved characteristics of neuromorphic computing systems based on memristors and traditional component base have been previously presented in the literature [4, 60], but every year new prototypes and records (see, e.g., [5, 51, 52, 61]) are reported, which are discussed in Section 4. According to the roadmap for brain-inspired computing chips [60], creating memristive general-purpose neuroprocessors is expected within the next 5–10 years. The prototypes of memristive computing systems demonstrated now already compete with the well-known neuromorphic processors based on traditional digital components and specialized architectures (ASIC) [4].

Despite all the successes in the development of AI technologies and the impressive progress in the development of specialized computing systems that implement neural network algorithms, more attention is being paid to the prospects for significantly deeper adaptation of neuromorphic principles than has been achieved so far [38]. In addition to being similar in form and essence to the functioning of the brain, neuromorphic systems (in their narrow understanding) implemented on the basis of memristive systems have significant potential for achieving a new level of cognitive abilities, primarily by means of efficient real-time processing of the electrical activity of biological neural systems as part of so-called bio- or neurohybrid systems [11, 17, 40]. At the same time, the first known examples from the literature in which memristive devices and arrays have been used to process bio-electrical activity only record the fact of communication between electronic and biological systems through individual memristive devices [43] or do so in isolation from the living systems (for example, in recently published papers [29, 30, 62], memristive chips are used to process emulated sequence of rectangular spikes or signals of neuronal activity taken from publicly available databases).

Remarkable progress in the development of memristive neurohybrid systems has been reported in the paper [44], which demonstrates the first bidirectional adaptive neurointerface based on advanced solutions in the field of memristive electronics and neuroengineering (Fig. 2).

A culture of hippocampal neuron cells with functional connections between neuron groups spatially ordered with the help of a microfluidic chip has been used on a multi-electrode array from the side of a living system. A memristive network is used not only to solve the problem of nonlinear classification of the spatial-temporal response of a cell culture to electrical stimuli, but also to control its functional state. Specifically, the output signals of the memristive network correspond to different stimuli and are used for adaptive stimulation control, which allows for the restoration of disrupted functional connections in the neural culture.

There is a great interest in the prospects of using such neurohybrid technologies for neurorehabilitation tasks, restoring or reorganizing biological neuronal functions after the development of a pathological condition [18].The perspective of creating cell cultures that highly reproduce brain architectural features is extremely attractive both from the standpoint of a convenient experimental model and from the standpoint of their use in real neurohybrid technology [10].

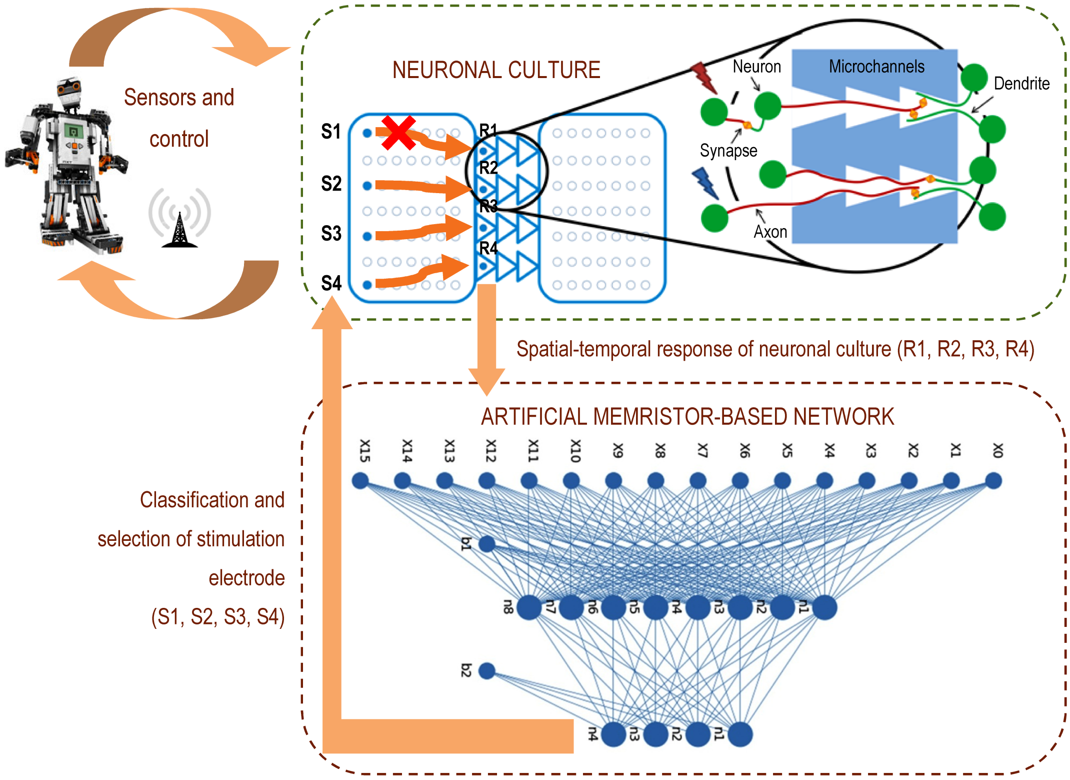

Sensors and control
NEURONAL CULTURE
S1
S2
S3
S4
R1
R2
R3
R4
Neuron
Microchannels
Dendrite
Synapse
Axon
Spatial-temporal response of neuronal culture (R1, R2, R3, R4)
ARTIFICIAL MEMRISTOR-BASED NETWORK
X15
X14
X13
X12
X11
X10
X9
X8
X7
X6
X5
X4
X3
X2
X1
X0
b1
n8
n7
n6
n5
n4
n3
n2
n1
b2
n4
n3
n2
n1
Classification and selection of stimulation electrode (S1, S2, S3, S4)

Figure 2. Bidirectional adaptive neurointerface between ordered neuronal culture and memristor-based artificial neural network [44]

Thus, the combination of high energy efficiency and unique scalability of memristive systems allows for a decisive step from neuromorphic computing systems to neurohybrid systems based on direct (physiological) and safe interaction between artificial electronic systems and living neuronal systems [33]. As a result, memristive neuromorphic systems will have a worthy place in AI medical technologies, providing not only efficient solutions to traditional AI tasks related to processing and analyzing biomedical data, but also creating compact and energy-efficient adaptive systems for replacing / restoring lost or improving existing brain and nervous system functions (neuroprosthetics and instrumental correction / support / enhancement of human cognitive abilities).

2. General Approach to Creating Memristor-Based Neuromorphic Computing Systems

According to recent perspectives [20, 32], research and development in the field of neuromorphic and brain-inspired computing systems are characterized by a complex (multi-level) and interdisciplinary nature. The first characteristic implies that new functional products are born from the co-optimization of solutions at the levels of materials, devices, and systems. The interdisciplinary nature not only requires the integration of different scientific communities (although this is already a big challenge in itself), but also the implementation of a coordinated plan, financing, and support (essentially, a master plan, as we have seen in the field of digital or quantum technologies, for example). In this section, let us consider how this combined approach is implemented in the case of developing neuromorphic and neurohybrid systems [33] based on CMOS-compatible MOM devices with resistive switching (Fig. 3).

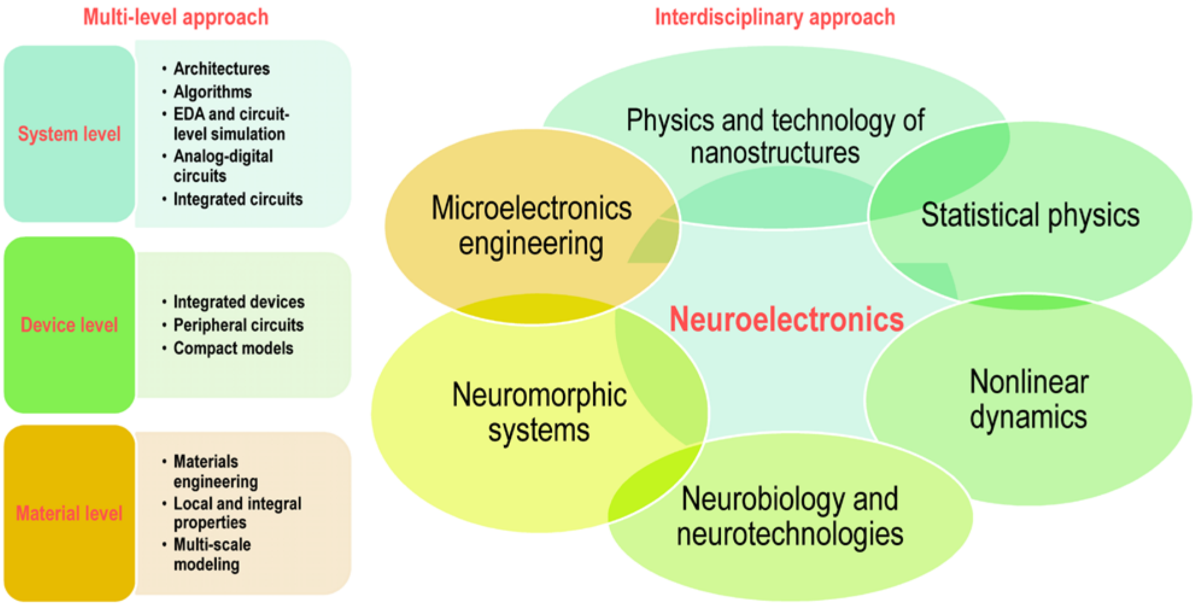


Figure 3. Illustration of complex (multi-level) and interdisciplinary approaches to designing neuromorphic and neurohybrid systems based on memristors

At the material level, MOM nanostructures are fabricated and studied, which exhibit resistive switching (one of the classic mechanisms of the memristive phenomenon). However, for understanding the regularities of memristive phenomenon and controlling its parameters, detailed study of physicochemical phenomena at the nano- or microlevel is insufficient. For example, the combination of different transport phenomena (phonons, electrons, ions) at different time scales makes even one memristor a complex nonlinear system with a rich dynamical response. In order to move further towards neuromorphic and neurohybrid systems, the same developed memristive structures are implemented as integrated devices and chips that are part of various functional circuits at the system level. Experimental work is always carried out in parallel with multiscale modeling: from models of physical phenomena at the micro-, meso-, and macrolevels to compact models of devices and circuit models required for automated design of electronic circuits. At the heart of such an approach lies the cross-cutting technology of memristive devices, compatible with traditional silicon technology and providing the creation of a component base for new brain-like information processing systems with a wide range of applications, including traditional and spiking neural network architectures, and neurointerfaces.

The interdisciplinary nature of the project is also illustrated in Fig. 3. Physics and technology of memristive nanostructures is one of the key areas that, based on traditional and new approaches in microelectronics, creates a technological platform for hardware implementation of memristor-based neuromorphic systems. To interpret, describe, and predict the memristive phenomenon, it is necessary to use the significant scientific knowledge in the fields of statistical physics and nonlinear dynamics. Based on the latest achievements in neurobiology and neurotechnology, the next step towards the symbiosis of artificial electronic and living biological systems can be taken.

To achieve the goal, interrelated tasks should be reached, including: 1) the investigation of new materials and devices, 2) the development of cross-cutting technology of a new component base, and 3) the development and hardware implementation of neural network architectures.

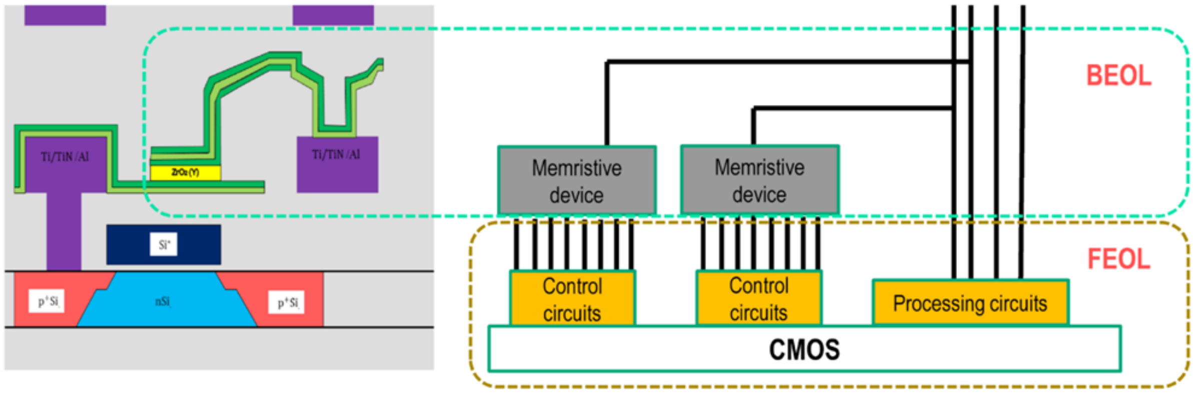


Figure 4. Illustration of memristive nanostructures integrated with CMOS circuitry using the BEOL process

Regarding memristors, reaching the first task is complicated by the fact that the complex nature of memristive phenomenon requires interconnected research at the micro- and macroscopic levels involving physics and chemistry of solid-state nanostructures, nonlinear dynamics, and statistical physics. The development of these interdisciplinary studies results in the discovery of new phenomena and the implementation of new methods to improve the characteristics of electronic devices based on memristive materials. Essentially, reaching this task means resolving fundamental issues associated with the correct description of the memristive phenomenon in various structures and materials and accompanying the design and creation of AI information and computing systems based on new component base.

The development of a cross-cutting technology based on resistive switching devices involves the development of scientific and technological solutions for creating elements and cells of non-volatile RRAM based on memristive nanostuctures with good yield, high endurance and retention parameters. The most important characteristic of memristive devices from the viewpoint of neuromorphic applications is their ability to store information at multiple levels, and significant progress is being made in this area now [39]. The main solution in the development of RRAM technology is the fabrication of functional RRAM blocks based on the integration of memristive structures, which are made at laboratory facilities in top metal layers (back-end-of-line – BEOL process), and the active layer of CMOS (front-end-of-line – FEOL process), which is made in industrial conditions (Fig. 4). Examples of images for the FEOL wafer, its fragment after the completion of the BEOL process, and the ready-made crossbar array of 1T1R (one memristor – one transistor) memristive cells are shown in Fig. 5. In the case of successful implementation, the cross-cutting technology for creating memristive microchips will provide a technological platform for a wide range of products, from RRAM microchips to neurochips, neurointerfaces, and neuroprosthesis for medical applications.

Research and development within this task results in the design and fabrication of test crystals with functional blocks of non-volatile resistive memory (memory cells and RRAM arrays) required to demonstrate the capabilities of new memory devices and basic principles of neuromorphic computing (VMM operations).

The main task within this scientific and technological field is the development of a neuromorphic processor with an array of synaptic weights based on memristors in a crossbar architecture (the most popular active RRAM crossbar is 1T1M). This processor should have digital-analog neurons of the leaky integrate and fire (LIF) type and other configurable parameters, with the

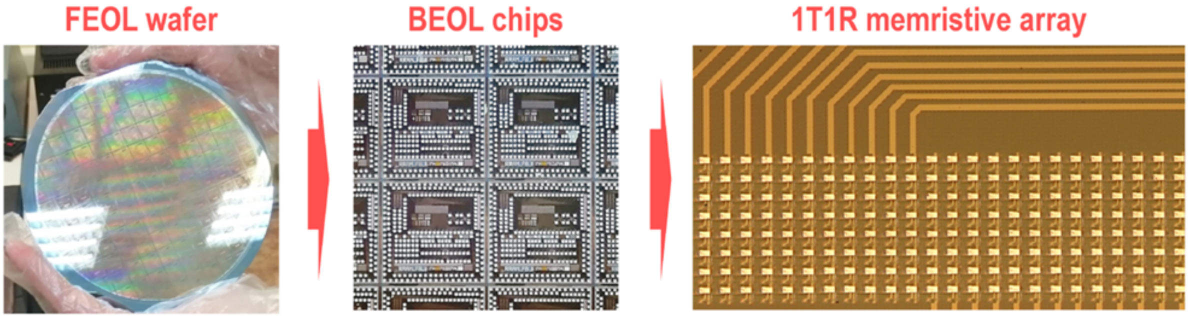


Figure 5. Images of the FEOL wafer, its fragment after completion of the BEOL process, and the final array of 1T1R memristors

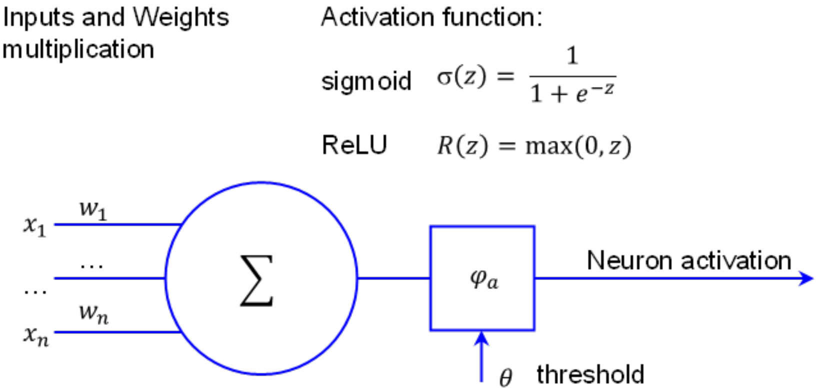


Figure 6. Formal neuron model – the weighted sum of inputs is fed into a basic nonlinear activation function, which can be either a sigmoid or a simpler Rectified Linear Unit (ReLU) transformation

ability to control and rewrite arbitrary memristive cells, supervised and unsupervised learning, including that based on local rules, and working in logical inference modes, as well as algorithms based on formal neural networks and spiking neural networks with spatio-temporal coding of multi-dimensional patterns of the solved problem. In the future, such a neuroprocessor should be able to solve various tasks in the field of AI: recognition of visual images, text and speech processing, analysis of various types of big data, prediction of temporal data series, sensorimotor control of mobile objects, optimization control of data flows in real-time, etc.

Let us take a closer look at the general approach to building an artificial neural network, which is based on a neuron model. There are two types of neuron models: formal (Fig. 6) and spiking one (Fig. 7) [34]. The main difference lies in the way the processed signals are represented: in a formal neuron, these signals have a continuous form, while in a spiking neuron, they are pulse-based. On the one hand, the hardware implementation of spiking neurons has an advantage of several orders of magnitude in terms of energy efficiency, but, on the other hand, the sharp fronts of the pulse signal make differentiation difficult, and as a result, the widely used backpropagation method fails when training a neural network. This situation leads to the necessity of developing new training algorithms for spiking neural networks based on bioplausible local plasticity rules [16].

The formal model of a neuron is widely used in various types of modern coprocessors, such as digital signal processors (DSP, digital signal processing), graphic processing unit (GPU), numerous neural accelerators and tensor accelerators (Google TPU (Google company), IVA

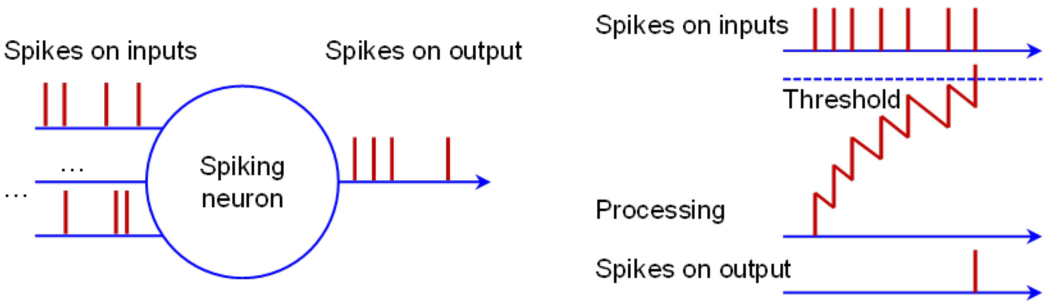
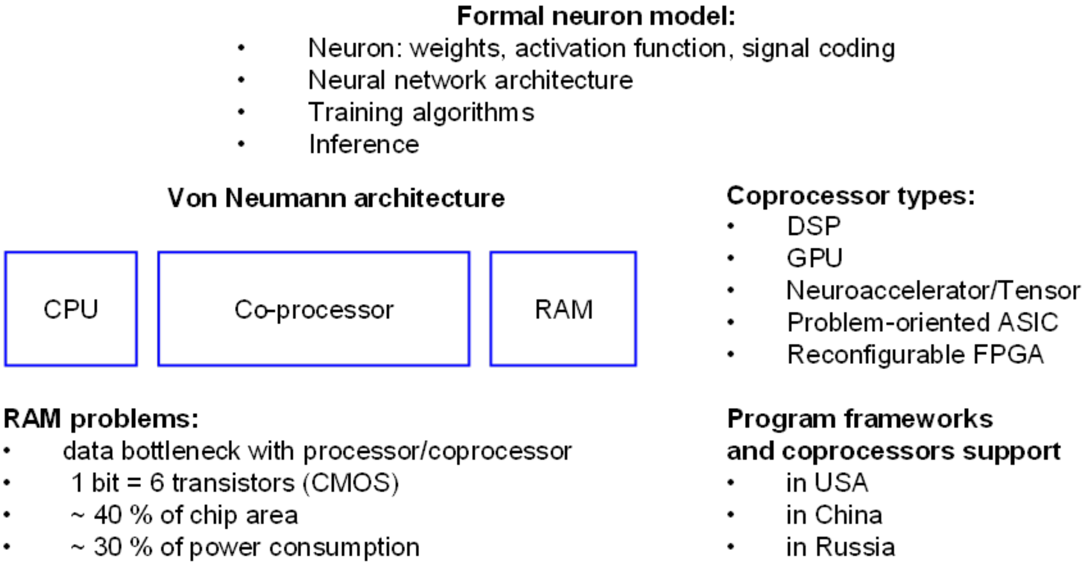


Figure 7. Spiking neuron receives sequences of spikes on its inputs and, under certain conditions, generates a spike at its output; for example, in the LIF model, each spike contributes to the neuron's status – its amplitude, which decays over time; if a sufficient number of spikes contributes to the status in a certain time window, the neuron amplitude exceeds a threshold, and the neuron generates an output spike. The electrical model of such a neuron can be implemented using an operational amplifier (OA) with an integrating RC circuit in the inverting input arm and a comparator



Formal neuron model:

- Neuron: weights, activation function, signal coding
- Neural network architecture
- Training algorithms
- Inference

Von Neumann architecture

CPU
Co-processor
RAM

Coprocessor types:

- DSP
- GPU
- Neuroaccelerator/Tensor
- Problem-oriented ASIC
- Reconfigurable FPGA

RAM problems:

- data bottleneck with processor/coprocessor
- 1 bit = 6 transistors (CMOS)
- ~ 40 % of chip area
- ~ 30 % of power consumption

Program frameworks and coprocessors support

- in USA
- in China
- in Russia

Figure 8. Software-hardware ecosystem for implementing neural networks on the formal neuron model: a significant foundation has been created, and best practices can be used for rapid development and testing of innovative neuromorphic systems

TPU (IVA Technologies company), NM6408 (Scientific and Technical Center "Module"), RoboDeus (Research and Development Center "ELVEES"), and many others), application-specific integrated circuit (ASIC), and field-programmable gate arrays (FPGA). Frameworks have been developed and widely used as software environments for developing neural networks, training them, and performing inference using the aforementioned processors. Thus, full hardware and software ecosystem has been developed for processing neural networks using the formal model of a neuron (Fig. 8). The further development of the formal model continues in the direction of improving processing algorithms and reducing the technology nodes of CMOS processors [23].

This background can be partially used for the hardware and software of new processors based on neuromorphic architectures and on a component base of new physical principles. Currently, the best neuromorphic model is based on the spiking model, but over time, neurophysiologists will discover and justify a more realistic model of neuron operation. At the moment, digital

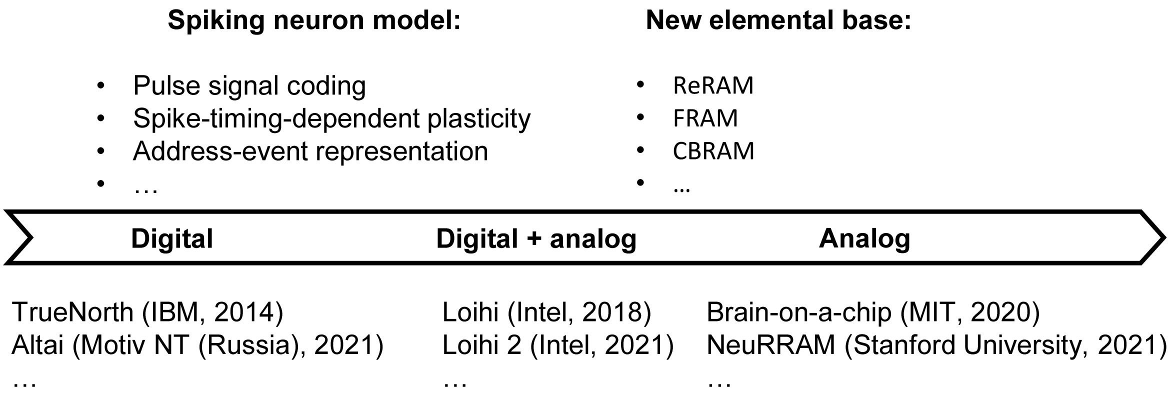


Figure 9. Modern neuromorphic systems based on the spiking neuron model

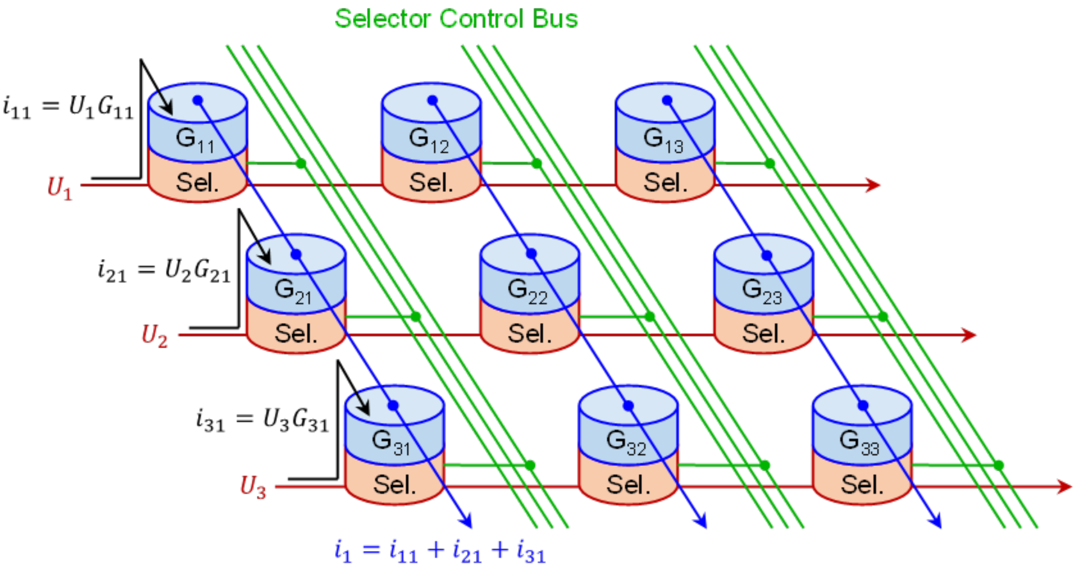


Figure 10. Memristive crossbar used in the VMM (inference) mode: input voltages are multiplied by the conductance G of the corresponding memristors in a certain column, and the resulting currents are summed up in the column. Selectors provide the connection of memristors to the crossbar lines. In some cases, selectors provide reverse connections, where inputs are swapped with outputs (blue lines receive voltages, and red lines extract currents) for the training process

(IBM, Motif NT), digital-analog (Intel), and analog (MIT and others) neuroprocessors have been developed on the spiking model. Naturally, there are many other developments not indicated in Fig. 9. The analog implementation of neurons is based on the use of operational amplifiers (OA) allowing a number of mathematical operations to be performed using currents and voltages in an electrical circuit.

The main operation carried out in neural network computation is VMM. As noted above, VMM is naturally and in analog form implemented in a memristive crossbar, which consists of a set of parallel metal lines in one plane and another set of parallel lines oriented perpendicular in another parallel plane. Memristors with programmable (self-adapting based on local rules) conductance values are placed at the crossbar nodes along with selectors – elements that provide correct addressing when accessing memristors (Fig. 10) [4].

On the one hand, analog representation and processing of information without the clocking characteristic of modern von Neumann architecture processors and coprocessors provides maximum speed and eliminates pipeline delays when obtaining results. On the other hand, digital

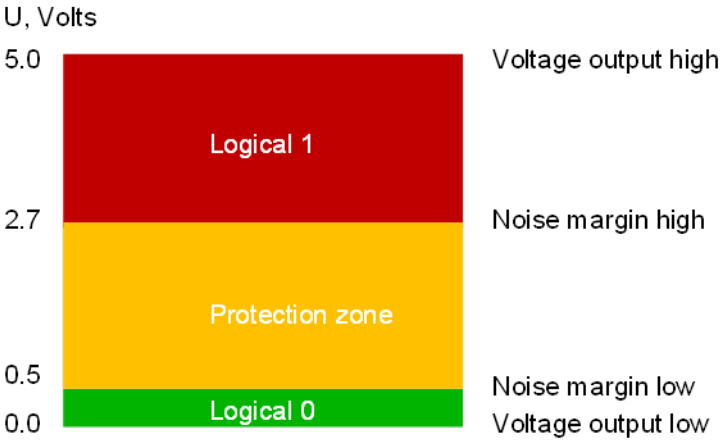


Figure 11. Levels of logical zero and one for a supply voltage range of 5V, ensuring high noise immunity for processed signals. The signal ranges V_O (output of the signal source) for the logical zero and one are wider than their corresponding ranges V_I (input of the signal receiver), compensating for possible voltage fluctuations during signal transmission through the interconnection lines between logical elements, signal source, and receiver

representation of information in the form of logical zeros and ones provides a high level of noise immunity due to the fact that the entire range of power supply voltage is divided into 3 zones (Fig. 11), the middle of which is not used and minimizes the number of possible errors. The use of analog, continuous amplitude scales for processing signals automatically imposes limitations on the dimensionality of the crossbar.

Memristive crossbars are the basis for hardware analog execution of mathematical operations inherent in various architectures of neuromorphic devices. Specifically, they allow for the execution of the VMM, which occupies the majority of data processing time in neuromorphic systems (inference), in parallel for several neurons in a single clock processor cycle with very low (picojoule) energy consumption. However, the potential for high performance and low energy consumption is not automatically realized – the computations in memristive crossbar-based systems must be organized in the most optimal way. Analogous to von Neumann architecture, the task of signal switching and control of a computing device can become a “bottleneck” in neuromorphic systems if not handled correctly.

A characteristic feature of neuromorphic systems is that, similar to biological networks of neurons, they contain a large number of interconnected nodes performing the same operations on the information being processed. For practical applications, the number of nodes (neurons) can be measured in thousands and the number of connections (synapses) – in millions. Individual memristive crossbars, having a specific number of memristive devices determined by the topology of the crystal and existing technological constraints, physically implement only a portion of the connections between neurons in different layers, with several crossbars possibly related to the same neurons. In these conditions, the developed architectures must be scalable.

The scalability of neuromorphic systems based on memristors logically should be implemented both at the neural model architecture level – “horizontally” (to provide the necessary number of neuron layers) and “vertically” (to provide the necessary number of neuron inputs), as well as at the level of parallel processing of data flows by multiple neuromorphic models with the same architecture. Moreover, physically, such scaling also has several levels – increasing the number of crossbars in a single neuroprocessor, increasing the number of neuroprocessors,

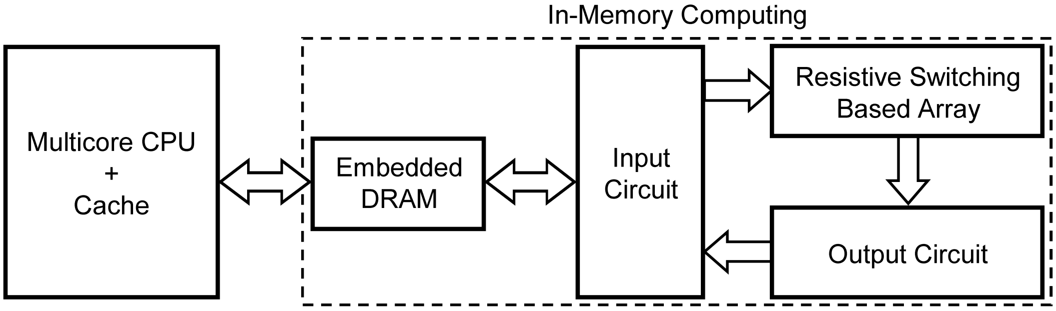


Figure 12. A separate in-memory computing chip based on the memristor-based array equipped with built-in dynamic RAM. Input and output circuits provide binary signal conversion into voltage and the reverse conversion of resulting currents into voltage. Computations are controlled by a multi-core processor device with cache memory implemented on another chip

and combining them into a cluster, and so on. The basic requirement for each level of scaling is to maintain a high level of parallelism in signal commutation for their simultaneous delivery to an equivalent crossbar (single crossbar or several crossbars combined "horizontally" and "vertically") and control of keys and selectors.

3. Approaches to Scaling Up Memristor-Based Neuromorphic Computing Systems

Let us consider various options for scaling active memristive crossbars in a CMOS-integrated form to increase the speed of signal transmission in memristive neural networks with static and spike coding [48].

In the classical von Neumann architecture, separate devices are used for data storage (random access memory, RAM) and computation (arithmetic logic unit, ALU). The operation principle of the slow dynamic memory DRAM limits the speed of reading/writing information of both initial and resulting data of the computational process. Therefore, when computing in memory, a separate chip is equipped with its own memory and computational cirtuit, which is controlled by the central processing unit (CPU) chip – Fig. 12 [4].

The computational process is organized as follows. The processor updates the weight coefficients in the memristive crossbar as needed, loads the input matrix into the embedded eDRAM memory in the chip for in-memory calculations, and issues the command to start the calculations. Data from eDRAM are transferred to the input circuit and converted into voltages required to operate the memristive crossbar. Each column of the memristive crossbar sums the products of input voltage and memristor conductivity in the form of current, performing an analog implementation of multiplication with accumulation in memory. In the output circuit, the results are converted into an output resulting matrix and stored in eDRAM for further use by the processor in the computing process.

The input and output circuits servicing the operation of the memristive crossbar are implemented using digital circuits with the use of analog-to-digital and digital-to-analog converters (ADCs and DACs) designed using CMOS technology – Fig. 13 [4].

The simplest binary neural networks require a relatively small percentage of CMOS processing circuits in the overall hardware implementation taking into account the memristive crossbar. The current level of development in the design and technology of memristive devices reflects the availability of devices with two levels of information storage. Binary networks are very energy-

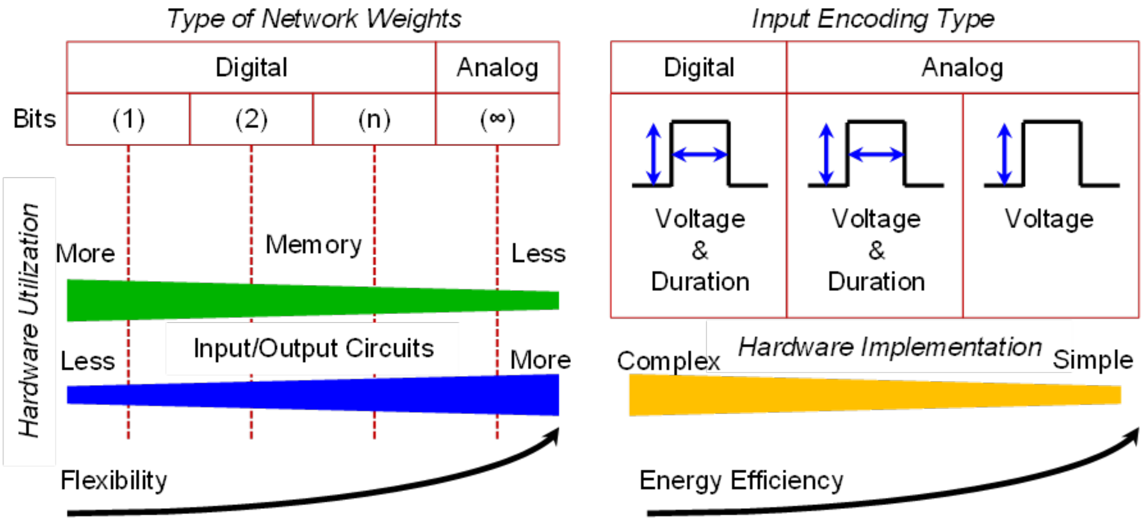


Figure 13. Flexibility and energy efficiency are maximized with analog signal processing. The hardware implementation of post-processing of the input signal (voltage) is also simplified for the transition from digital to analog representation

efficient but capable of solving relatively simple tasks, such as pre-processing and processing of sound and speech.

Moving to three or more binary digits on one side opens the possibility of using more complex neural networks, but also means an increase in the share of CMOS circuits in the overall hardware implementation. New technologies of such multi-level memristors are actively being developed [39].

The use of unlimited (analog) precision of weights requires the use of corresponding memristors, which are not widely available at present, as well as a significant volume of high-precision CMOS component base to provide digital-to-analog and analog-to-digital support for memristive crossbars. The undeniable advantage of neural networks in such implementation is the high degree of accuracy achieved during their operation due to the absence of the need to reduce the bit depth of weight coefficients during the conversion of the model into hardware implementation.

Each column is implemented in the simplest analog encoding circuit with amplitude encoding of the input signal and a 0T1R memristor cell in the crossbar node operated by an OA with a feedback resistor (Fig. 13, analog voltage encoding). The addition of duration to the input signal requires an integrating function in the OA (Fig. 13, analog voltage and duration encoding). To process signals in a full memristive crossbar with 1T1R cells and digitized amplitude and sampled duration of the input signal, the most complex CMOS circuit will be required, using a comparator and counter (Fig. 13, digital voltage and duration encoding).

The activation function is also implemented in circuits using OA (Fig. 14). A circuit containing 2 OA and a set of resistors serves one column of a memristive crossbar [28].

In well-designed CMOS circuits for memristive crossbars, the limiting factor for increasing their dimensionality is the presence of parasitic sneak paths in these crossbars. The problem is that current, in addition to the desired propagation path of row-column, also flows through adjacent undesirable paths. In [64], an analysis of this problem was carried out: the ratio of the voltage range in the crossbar to the voltage range in one memristor was calculated depending on the stored values in the crossbar and the grounding of rows and columns. The presence of parasitic paths depends significantly on the stored values in the memristive crossbar. The dependence of the parameter Δ', which is equal to the ratio of the power supply voltage and the

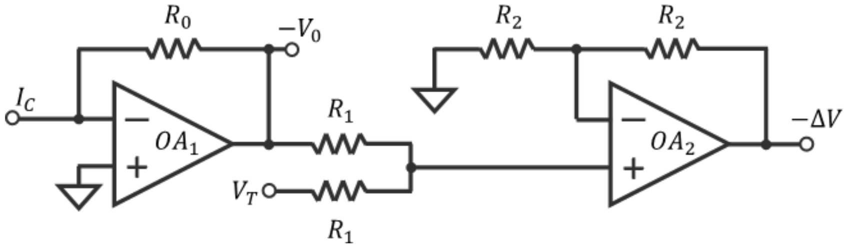


Figure 14. An example of implementing an activation function on an OA for a fully connected layer of a neural network when the total column current I_c is received: V_0 is the activated output, V_T is the target value for V_0, ΔV is the mismatch error between V_0 and V_T

zero voltage (ground) difference for the entire crossbar to the same difference for one memristor, was investigated. In the ideal case, the result is equal to one, in others – less than one. The simulation results show that a significant decrease in the analyzed parameter is observed even for relatively small dimensions of a 16 × 16 and 64 × 64 arrays.

To address the issue of parasitic sneak paths, several methods are being considered. The first method is called multi-stage reading and includes five steps: measuring the target cell current, setting the target cell to high-resistance state (HRS) and measuring the current, performing a similar operation for low-resistance state (LRS), comparing the measured currents, and returning the cell to its original state. The second method involves column separation architecture for each memristor. The third and fourth methods involve using a diode and a transistor as a selector (1D1R and 1T1R cells, respectively). The fifth method involves using complementary memristors that provide constant resistance $R_{\mathrm{LRS}}+R_{\mathrm{HRS}}$, significantly reducing parasitic currents. Although the 1T1R cell takes up more space and an additional line is required to control the transistor gate, this method is the most common.

In various crossbar circuits, duplication of elements is used to achieve the required functionality and increase performance, as well as multiplexing the component base for its subsequent reuse to perform various functions with time division. Thus, the HRS and LRS in binary ReRAMs are positive, so the XNOR operation is used for encoding signed weights, and the number of rows of memristive crossbars is doubled (Fig. 15) [44]. In Fig. 15, SL is the source line, BL is the bit line, WL is the word line: lines of source, bits and words; the input signal value "–1" is encoded by a pair of 1 and 0, the value "+1" – by a pair of 0 and 1; the weight "–1" is encoded by a pair of LRS and HRS, the weight "+1" – by a pair of HRS and LRS; eight bit lines are collected into a processing block; the next bit line is selected on the multiplexer and its value (128 levels) is digitized with the help of instrumentation amplifiers. High precision instrument amplifiers operating in voltage mode are used to process signals of the memristive array columns, which are separated by a multiplexer for processing the signal of one of the eight columns (bit lines). In the first case, there is a duplication of the hardware, in the second case, there are savings of the CMOS base by increasing the signal processing time. Various circuits of adaptive compensation of large or small current values in polled memristors are used, for example, the voltage clamp control circuit [59].

The active development of circuits on memristive crossbars is accompanied by the increase in the dimensionality of crossbars on one hand, and by proposals to map neural networks to the hardware implementation of such processors on the other. For example, in [51], the NeuRRAM processor with a multilevel organization of processor units is proposed. At the top level, the implemented hardware neural network is mapped onto such a processor consisting of 48 cores

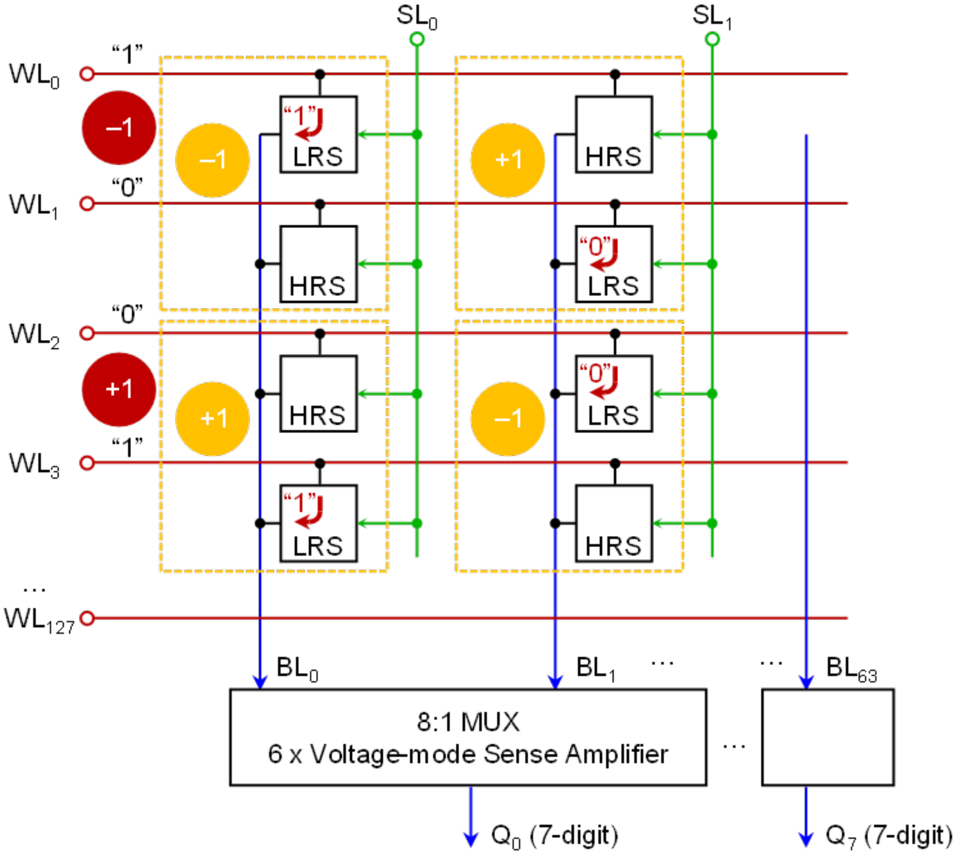


Figure 15. Example of implementing a memristive array with an effective size of 64 × 64

organized in an array of 8 rows by 6 columns (Fig. 16). To operate the neural network, it is mapped onto 48 cores of a chip in one of 6 ways: (1) 1 layer in 1 core, (2) duplication in multiple cores to increase throughput, (3) multiple layers in one core, (4) reordering in one core to increase utilization, (5) and (6) parallelization on multiple cores. Each core consists of an array of 16 × 16 corelets, each of which contains a 16 × 16 RRAM weights and a CMOS neuron. The single-bit BL and SL switches of the corelet can change the direction of the signal being processed by the CMOS neuron from BL to SL or vice versa. This configuration is called Bidirectional transposable neurosynaptic array (TNSA), meaning that the input signals can be fed to both rows and columns with the help of supporting CMOS circuits. At the stage of VMM input, the drivers convert the register inputs (REG) and PRN inputs into analog voltages and transmit them to TNSA. At the stage of VMM output, the drivers transmit digital outputs from neurons back to registers through REG. In addition, various activation functions, including stochastic ones, are implemented in the CMOS circuits.

At the lower level, the corelet consists of a 16 × 16 array of memristors and one CMOS neuron. The neuron is connected to one of 16 bit lines and one of 16 source select lines that pass through the corelet. It is responsible for integrating inputs from all 256 RRAMs connected to a single BL or SL: 16 RRAMs in the current corelet and 240 RRAMs in other corelets along the same row / column. Thanks to an advanced routing system, each core is capable of performing forward, backward, and recurrent VMM on all 256 rows.

The above-mentioned memristive crossbars with CMOS control circuits are implemented as monolithic microchips with 90 and 130 nm technology nodes. As noted above, the CMOS control circuits are located in the FEOL layer, while the memristive crossbar is located between

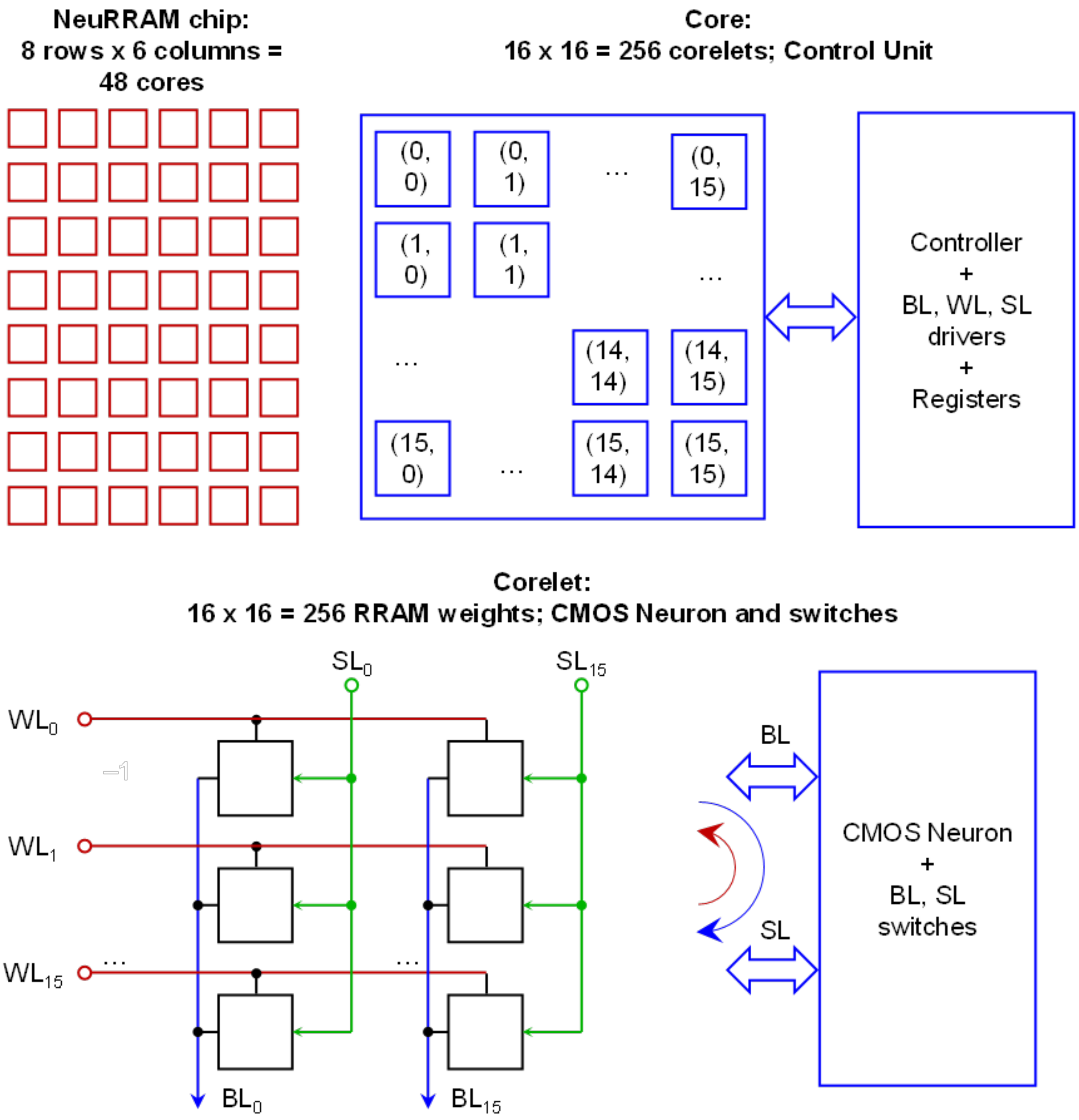


Figure 16. Architecture of the NeuRRAM project [51]

the metallization layers in the BEOL layer or on top of it (Fig. 4). However, there is another relatively new approach to implementing complex devices, including cases where their parts are made using different and possibly incompatible technologies. In the above example, the oxide layer in the memristor may be destroyed by the high temperature during the formation of the upper layers using CMOS technology – exceeding the temperature budget [63].

The idea of dividing a large chip by area into a set of separate chiplets (mini-chips) with their subsequent placement and side-by-side connection on the substrate-interposer plane (2.5D integration) or in the form of a stack (stepped structure, 3D integration) with connection by vertical conductors (TSV, through silicon via) originated in 2015 [25]. Each chiplet is usually a system module, implemented using incompatible technologies or implementing a complex functional block (IP, intellectual property). Pascal Vivet (LETI – Laboratory of Electronics and Information Technologies, European center for research in microelectronics) believes that "Chiplet-based ecosystems will deploy rapidly in high-performance computing and various other market segments, such as *embedded* HPC for the automotive and other sectors" [25]. LETI

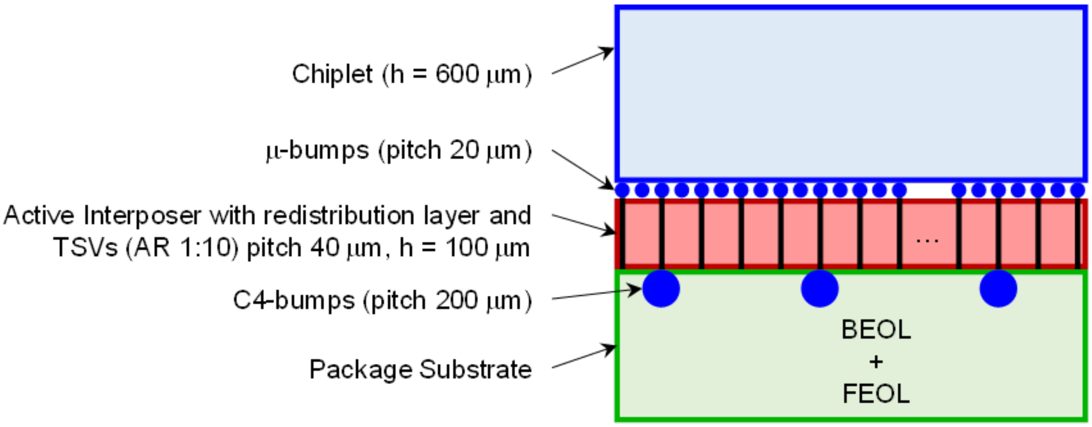


Figure 17. Active interposer presented by the LETI center [25] to combine 96 cores on 6 chiplets. The active interposer with RDL (redistribution layer) allows combining the interposer with a bump pitch of 200 μm (at the bottom) and micro-bump pitch of 20 μm (at the top of the chiplet)

presented an active interposer technology for chiplets, which was used to assemble a structure consisting of 6 chiplets with a total of 96 cores (Fig. 17).

The issues of chiplet assembly, testing, and yield, as well as CAD support, are not yet adequately addressed in the technology of chiplets. However, extensive work is being done to standardize interchip communication technologies, such as Intel's Advanced Interface Bus (AIB), the Optical Internetworking Forum's CEI-112G-XSR, and Open Domain-Specific Architecture's BoW (Bunch of Wires) and OpenHBI (High Bandwidth Interface).

The seriousness of chiplet technology is confirmed by the participation of well-known companies like Boeing, Cadence, Synopsys, Intel, Micron, and others in the project Common Heterogeneous Integration and IP Reuse Strategies (CHIPS, a program for integrating heterogeneous chips and reusing complex functional blocks since 2017), as well as GE, Intel, Keysight, Xilinx, and others in the project The State of The Art (SOTA) Heterogeneous Integrated Packaging (SHIP, an advanced program for packaging heterogeneous chips – to establish interface standards between chiplets and ensure the assembly of complex functional blocks since 2019). Both projects are being implemented by the American agency DARPA.

An actual example of using such technology for in-memory computing on memristive crossbars is the SIAM project – Chiplet-based Scalable In-Memory Acceleration with Mesh for Deep Neural Networks [24], a chiplet-based scalable in-memory computing accelerator for deep neural networks (Fig. 18). At the initial stage, a transition from a neural network to an architecture is made, taking into account: the IMC (In-Memory Computing) chip mode, the network frequency in the package (NoP), the size and number of chiplets, IMC mapping, the number of tiles per chiplet, the size of the crossbar, memory cell type, technology node, and accumulator size; the "engine" for partitioning and mapping: internal chiplet planning, chiplet placement, "engines" for NoP and DRAM; mapping to IMC tiles, external chiplet planning, routing and placement, "engine" for electrical circuit and chip network (NoC, network on chiplet); obtaining a chip partitioning as shown in Fig. 18 on the left.

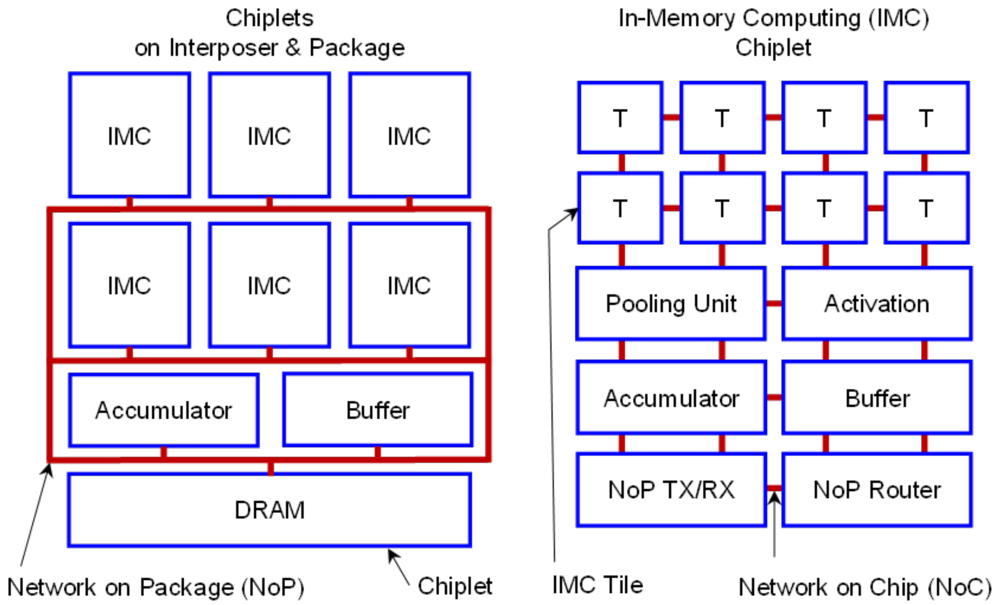


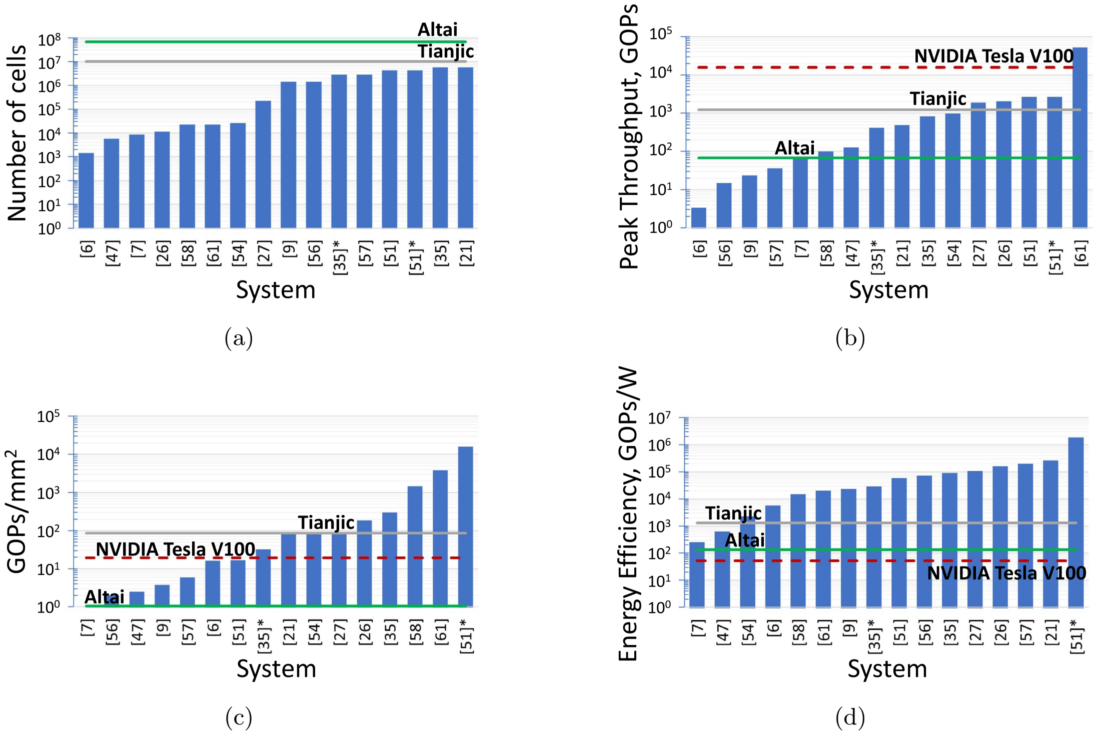
Figure 18. SIAM project [24]: implementing computational functions through chiplets, global accumulator and buffer, DRAM memory placed and connected on the interposer in a package and connected to the NoP (on the left). Each IMC chiplet consists of IMC tiles, calculation modules, communication and routing (on the right). Each tile consists of multiple processing elements (PE), a multiplexer, ADC, instrumentation OA, shift and adder device, buffer; each PE contains a memristive crossbar (not shown)

4. Comparison of Computational Systems Based on Traditional and New Component Base

Let us take a closer look at the results of comparing known GPU and neuromorphic processors based on traditional digital components with prototypes of memristor-based neuromorphic processors. For comparison, we will use two absolute criteria – the number of cells and peak performance (gigaoperations per second, GOPs), and two relative criteria – performance per chip area (GOPs/mm^2) and performance per watt of energy consumption (energy efficiency, GOPs/W). These criteria are calculated for the inference of neural networks, where the basic operation is a VMM. The comparison results are shown in Fig. 19.

For this comparison, specialized neuromorphic processors Altai [1] and Tianjic [36], optimized for spiking neural networks, and the most powerful GPU from NVIDIA – Tesla V100 [2], which is more universal than the previous ones, as it allows solving a wide range of tasks in the field of data processing, were chosen. The performance metrics were taken from open sources (references at the horizontal axis) as indicated by the authors. All prototypes of memristor-based processors selected for comparison are made using CMOS-compatible technology and have a device layer with transistor selectors (except [7]) and other electronics required for operation.

As seen in Fig. 19a, computing systems based on memristive devices have significantly fewer cells than existing processors. However, this is not a disadvantage and is explained by the fact that the presented developments are still prototypes created as a result of research and development. Nevertheless, even such relatively small processors, with up to 4 million cells, demonstrate sufficiently high performance, surpassing Altai and Tianjic processors with 67 and 10 million synapses, respectively (see Fig. 19b).



(a)

(b)

(c)

(d)

Figure 19. Comparison results of memristor-based computing systems (bar chart columns) with neuromorphic processors (horizontal solid lines) and GPU (horizontal dashed line) based on traditional digital components according to the following criteria: the number of cells (a), peak performance (b), performance per chip area (c) and performance per watt of energy consumption (d)

The advantages of computing systems based on memristive devices are most clearly demonstrated when compared according to relative criteria. The high potential for miniaturization of memristive devices (down to a few nanometers) and RRAM cells (requiring only 1–2 transistors) allows for more efficient use of chip space, as shown in Fig. 19c. For example, the RAND chip (Resistive Analog Neuro Device [35]) made using 40 nm technology has an area of 2.71 mm^{2} at a density of 1.48M synapses per mm^{2} with drivers, controllers, and multiplexers while providing three times higher relative performance than the Tianjic processor and 12.6 times higher performance than NVIDIA Tesla V100. In turn, the energy efficiency of memristor-based computing systems is 2–3 orders of magnitude better than existing processors (see Fig. 19d). For example, the nvCIM macro chip [21] made using 22 nm technology node demonstrates 12–150 times lower power consumption than Tianjic and 300–3700 times lower consumption than NVIDIA Tesla V100.

With the advancement of technology in creating memristor-based neural processors, the number of cells will increase, meaning that with higher computing density, peak performance will exceed the performance parameters of neuromorphic processors based on traditional digital electronics and specialized architectures presented in Fig. 19b. Of course, this growth cannot be indefinitely large, and potential high performance and energy efficiency will be more influenced by design solutions at the processor and computing system architecture levels, especially the growing overhead costs of routing and input/output data in digital form (see also Section 3). For example, when it comes to processing signals of different nature, performance will be limited by the characteristics of sensors and information transmission interfaces, so devices for computing in sensors with direct transmission of information in analog form to a memristor-based computing device are currently being developed for such tasks [31, 49].

A number of other notable examples of memristor-based computing systems were not included in this comparison, as authors in publications often do not provide the values of the criteria used in Fig. 19. In addition to these, there are more specialized criteria for assessing performance and energy efficiency in relation to the peculiarities of the processor architecture or the specific problem being solved. These criteria include the number of synaptic giga- or teraoperations per second (GSOPs, TSOPs) [3, 36], the number of giga- or teraoperations per second computed per 1 Mb of ReRAM (GOPs/Mb, TOPs/Mb) [51], AiMC TOPS/W [8], the number of processed frames per watt (frames/W) [61], and energy-delay-product (EDP, j·s) [51]. Furthermore, some authors use software simulators (such as XPEsim [58]) to evaluate the performance and energy efficiency characteristics of ReRAM-based chips due to the high cost of prototyping. In the future, a valuable criterion for comparing in-memory computing systems will be the cost of 1 k/M/G byte of memory.

Neuromorphic computing accelerators (standard digital ASICs based on CMOS, system solutions, and memristor-based microchips) presented in Fig. 19 were compared for performance and energy efficiency taking into account their high (comparable to software emulation) accuracy in inference of neural network models for specific tasks such as pattern recognition, classification, segmentation, etc. Table 1 shows the numerical values of characteristics of memristor-based computing systems, including the task, neural network model architecture and achieved accuracy metrics.

From Tab. 1, it can be seen that the considered processors perform at a high level on commonly accepted test tasks for image classification from the MNIST dataset with an accuracy range of 90.8 [35] to 99 % [51], CIFAR-10 – from 85.7 [51] to 95.19 % [21], CIFAR-100 – 65.71 % [57], recognize Google voice commands with an 84.7 % probability [51], and successfully solve other tasks whilst implementing well-known neural network architectures such as MLP (multilayer perceptron), DNN (deep neural network), CNN (convolutional neural network), LSTM (long short-term memory) and ResNet-20, ResNet-50, VGG16 models.

It should be noted that, among the considered prototypes, the most versatile in terms of the ability to launch different architectures of neural networks is the NeuRRAM chip [51]. As can be seen from Fig. 19 and Tab. 1, NeuRRAM already has 33–800 times better energy efficiency at technology node of 130 nm than Tianjic, Altai, and NVIDIA Tesla V100 processors, and provides high relative performance compared to them. At the same time, a many orders of magnitude gain in the mentioned and other parameters is expected when scaling the technology node to 7 nm from the current level of 90–130 nm, which are currently used in creating prototypes of multi-core processors based on memristive devices in the structure of MOM.

Thus, in-memory computing is currently the only way to increase the performance and reduce the energy consumption of AI computing systems, as it is the most bioplausible information processing principle from a functional point of view, and it allows for a significant reduction in data transfer distance and required memory volume (model parameters are constantly stored in the processor), as well as energy consumption required for VMM. For in-memory computing, different types of memory can be used [42]: SRAM, DRAM, Flash, however, the most suitable one is RRAM, as other types of memories have disadvantages (such as low scalability, high cost and volatility for SRAM, poor process compatibility with CMOS for processors and the need for regeneration tens of times per second for DRAM, difficulties in implementing write at arbitrary address for Flash, etc.) and impose significant limitations on the creation of neuromorphic chips.

Table 1. Numerical characteristics of memristor-based computing systems

Ref.	CMOS techn.	Unit cell	Cell numb.	Crossbar size	Peak Throughput (GOPS)	Energy Efficiency (GOPS/W)	Area Efficiency (GOPS/mm^2)	Accuracy on Applications Demonstrated
[51]*	7 nm	1T1R	3M	16×16	2,135	1,360,000	12,800	99 % MNIST, 85.7 % CIFAR-10, 84.7 % Google speech
[21]	22 nm	1T1R	4M	1024×512	394	194,000	65.7	92.01–95.19 % CIFAR-10
[35]	40 nm	1T1R	4M	n/a	660	66,500	240	90.8 % MNIST (MLP)
[51]	130 nm	1T1R	3M	16×16	2,135	43,000	13.4	see row [51]*
[27]	130 nm	2T2R	159k	n/a	1,500	78,400	71	94.4 % MNIST (MLP)
[61]	130 nm	1T1R	16k	128×16	41,900	14,900	3,100	40.21 dB PSNR and 22.38 dB SNR for MRI and CT images
[26]	2 μm	1T1R	8k	128×64	1,640	119,700	150	n/a
[57]	22 nm	1T1R	2M	512×512	29	146,000	4.8	90.88 % CIFAR-10 (ResNet-20), 65.71 % CIFAR-100 (ResNet-20)
[35]*	180 nm	1T1R	2M	n/a	330	21,000	26	see row [35]
[54]	130 nm	1T1R	18k	256×16	780	1,650	69	n/a
[58]	130 nm	1T1R	16k	128×16	81	11,000	1,160	96.92 % MNIST (CNN)
[56]	55 nm	1T1R	1M	512×256	12	53,170	1.6	88.52 % CIFAR-10 (CNN)
[9]	65 nm	1T1R	1M	512×256	19	16,950	3	98.8 % MNIST (LeNet DNN)
[47]	150 nm	1T1R	4k	32×32	101	462	2	n/a
[6]	130 nm	2T2R	1k	32×32	2.7	4,200	13	98.4 % MNIST (MLP), 87 % CIFAR-10 (CNN)
[7]	180 nm	0T1R	6k	54×108	57	187.6	0.9	94.6 % breast cancer screening dataset

Conclusion

Memristors are very simple devices and at the same time very smart and complex nonlinear systems promising a wide range of applications from memory chips and in-memory neuromorphic computing systems to adaptive neural interfaces. The implementation of neuromorphic computing systems based on this new component base requires coordinated and interdisciplinary research and development at various levels. The basis of the corresponding scientific and technological direction is the cross-cutting technology of memristive devices and circuits, providing for the creation of a new brain-like information and computing system base with a wide range of applications. The currently demonstrated perspectives are associated with the monolithic integration of memristive devices and arrays with CMOS circuits, as well as co-optimization of materials, devices, and architectures necessary for creating demonstration prototypes of information and computing systems based on memristors.

Various scaling options of active memristive crossbars in integrated implementation provide an increase in signal transmission speed in memristive neural networks with both static and spike coding. The analysis of circuit solutions based on CMOS component base, which ensure efficient operation of the memristive crossbar during training and inference, demonstrates an increase in effective crossbar dimensions in recent years. An alternative solution to monolithic integrated implementation is also presented in the paper through various examples of chiplet technology-based implementations.

Comparison of neuromorphic computing systems based on traditional and new component bases shows that existing prototypes already significantly (by orders of magnitude) outperform known computing systems based on traditional component base in terms of performance and energy efficiency without reducing precision in vector-matrix multiplication and artificial neural network inference.

Acknowledgements

This study was conducted within the framework of the scientific program of the National Center for Physics and Mathematics, section No. 9 “Artificial intelligence and big data in technical, industrial, natural and social systems”.

References

1. Neurochip “Altai”. `https://motivnt.ru/neurochip-altai/`, accessed: 2023-05-15

2. NVIDIA Tesla V100 GPU architecture the world’s most advanced data center GPU. `https://www.nvidia.cn/content/dam/en-zz/Solutions/Data-Center/tesla-product-literature/volta-architecture-whitepaper.pdf`, accessed: 2023-05-15

3. Akopyan, F., Sawada, J., Cassidy, A., *et al.*: TrueNorth: Design and Tool Flow of a 65 mW 1 Million Neuron Programmable Neurosynaptic Chip. IEEE Transactions on Computer-

Aided Design of Integrated Circuits and Systems 34(10), 1537–1557 (oct 2015). https://doi.org/10.1109/TCAD.2015.2474396

4. Amirsoleimani, A., Alibart, F., Yon, V., *et al.*: In-Memory Vector-Matrix Multiplication in Monolithic Complementary Metal-Oxide-Semiconductor-Memristor Integrated Circuits: Design Choices, Challenges, and Perspectives. Advanced Intelligent Systems 2(11), 2000115 (nov 2020). https://doi.org/10.1002/AISY.202000115

5. Bianchi, S., Muñoz-Martin, I., Covi, E., *et al.*: A self-adaptive hardware with resistive switching synapses for experience-based neurocomputing. Nature Communications 14(1), 1–14 (mar 2023). https://doi.org/10.1038/s41467-023-37097-5

6. Bocquet, M., Hirztlin, T., Klein, J.O., *et al.*: In-Memory and Error-Immune Differential RRAM Implementation of Binarized Deep Neural Networks. Technical Digest - International Electron Devices Meeting, IEDM 2018-December, 20.6.1–20.6.4 (jan 2019). https://doi.org/10.1109/IEDM.2018.8614639

7. Cai, F., Correll, J.M., Lee, S.H., *et al.*: A fully integrated reprogrammable memristor-CMOS system for efficient multiply-accumulate operations. Nature Electronics 2(7), 290–299 (jul 2019). https://doi.org/10.1038/s41928-019-0270-x

8. Cai, F., Yen, S.H., Uppala, A., *et al.*: A Fully Integrated System-on-Chip Design with Scalable Resistive Random-Access Memory Tile Design for Analog in-Memory Computing. Advanced Intelligent Systems 4(8), 2200014 (aug 2022). https://doi.org/10.1002/AISY.202200014

9. Chen, W.H., Dou, C., Li, K.X., *et al.*: CMOS-integrated memristive non-volatile computing-in-memory for AI edge processors. Nature Electronics 2(9), 420–428 (aug 2019). https://doi.org/10.1038/s41928-019-0288-0

10. Chiaradia, I., Lancaster, M.A.: Brain organoids for the study of human neurobiology at the interface of in vitro and in vivo. Nature Neuroscience 23(12), 1496–1508 (nov 2020). https://doi.org/10.1038/s41593-020-00730-3

11. Chiolerio, A., Chiappalone, M., Ariano, P., Bocchini, S.: Coupling resistive switching devices with neurons: State of the art and perspectives. Frontiers in Neuroscience 11(FEB), 70 (feb 2017). https://doi.org/10.3389/FNINS.2017.00070/BIBTEX

12. Christensen, D.V., Dittmann, R., Linares-Barranco, B., *et al.*: 2022 roadmap on neuromorphic computing and engineering. Neuromorphic Computing and Engineering 2(2), 022501 (may 2022). https://doi.org/10.1088/2634-4386/AC4A83

13. Chua, L.O.: MemristorThe Missing Circuit Element. IEEE Transactions on Circuit Theory 18(5), 507–519 (1971). https://doi.org/10.1109/TCT.1971.1083337

14. Chua, L.O., Kang, S.M.: Memristive Devices and Systems. Proceedings of the IEEE 64(2), 209–223 (1976). https://doi.org/10.1109/PROC.1976.10092

15. Demin, V.A., Erokhin, V.V.: Hidden symmetry shows what a memristor is. International Journal of Unconventional Computing 12, 433–438 (2016)

16. Demin, V.A., Nekhaev, D.V., Surazhevsky, I.A., *et al.*: Necessary conditions for STDP-based pattern recognition learning in a memristive spiking neural network. Neural Networks 134, 64–75 (feb 2021). https://doi.org/10.1016/J.NEUNET.2020.11.005

17. George, R., Chiappalone, M., Giugliano, M., *et al.*: Plasticity and Adaptation in Neuromorphic Biohybrid Systems. iScience 23(10), 101589 (oct 2020). https://doi.org/10.1016/J.ISCI.2020.101589

18. Guggisberg, A.G., Koch, P.J., Hummel, F.C., Buetefisch, C.M.: Brain networks and their relevance for stroke rehabilitation. Clinical Neurophysiology 130(7), 1098–1124 (jul 2019). https://doi.org/10.1016/J.CLINPH.2019.04.004

19. Ham, D., Park, H., Hwang, S., Kim, K.: Neuromorphic electronics based on copying and pasting the brain. Nature Electronics 4(9), 635–644 (sep 2021). https://doi.org/10.1038/s41928-021-00646-1

20. Huang, Y., Kiani, F., Ye, F., Xia, Q.: From memristive devices to neuromorphic systems. Applied Physics Letters 122(11), 110501 (mar 2023). https://doi.org/10.1063/5.0133044/2880793

21. Hung, J.M., Xue, C.X., Kao, H.Y., *et al.*: A four-megabit compute-in-memory macro with eight-bit precision based on CMOS and resistive random-access memory for AI edge devices. Nature Electronics 4(12), 921–930 (dec 2021). https://doi.org/10.1038/s41928-021-00676-9

22. Kim, J., Pershin, Y.V., Yin, M., *et al.*: An Experimental Proof that Resistance-Switching Memory Cells are not Memristors. Advanced Electronic Materials 6(7), 2000010 (jul 2020). https://doi.org/10.1002/AELM.202000010

23. Krasnikov, G.Y.: The capabilities of microelectronic processes with 5 nm critical dimension and less. Nanoindustry Russia 13(S5-1(102)), 13–19 (2020)

24. Krishnan, G., Mandal, S.K., Pannala, M., *et al.*: SIAM: Chiplet-based Scalable In-Memory Acceleration with Mesh for Deep Neural Networks. ACM Transactions on Embedded Computing Systems (TECS) 20(5s) (sep 2021). https://doi.org/10.1145/3476999

25. LaPedus, M.: Chiplet Momentum rising. Semiconductor Engineering. https://semiengineering.com/chiplet-momentum-rising/ (2020), accessed: 2022-10-28

26. Li, C., Hu, M., Li, Y., *et al.*: Analogue signal and image processing with large memristor crossbars. Nature Electronics 1(1), 52–59 (dec 2017). https://doi.org/10.1038/s41928-017-0002-z

27. Liu, Q., Gao, B., Yao, P., *et al.*: A Fully Integrated Analog ReRAM Based 78.4TOPS/W Compute-In-Memory Chip with Fully Parallel MAC Computing. Digest of Technical Papers - IEEE International Solid-State Circuits Conference 2020-February, 500–502 (feb 2020). https://doi.org/10.1109/ISSCC19947.2020.9062953

28. Liu, X., Zeng, Z.: Memristor crossbar architectures for implementing deep neural networks. Complex and Intelligent Systems 8(2), 787–802 (apr 2022). https://doi.org/10.1007/S40747-021-00282-4/TABLES/7

29. Liu, Z., Tang, J., Gao, B., *et al.*: Multichannel parallel processing of neural signals in memristor arrays. Science Advances 6(41) (oct 2020). `https://doi.org/10.1126/SCIADV.ABC4797/SUPPL_FILE/ABC4797_SM.PDF`

30. Liu, Z., Tang, J., Gao, B., *et al.*: Neural signal analysis with memristor arrays towards high-efficiency brain-machine interfaces. Nature Communications 11(1), 1–9 (aug 2020). `https://doi.org/10.1038/s41467-020-18105-4`

31. Makarov, V.A., Lobov, S.A., Shchanikov, S., *et al.*: Toward Reflective Spiking Neural Networks Exploiting Memristive Devices. Frontiers in Computational Neuroscience 16, 62 (jun 2022). `https://doi.org/10.3389/FNCOM.2022.859874/BIBTEX`

32. Mehonic, A., Kenyon, A.J.: Brain-inspired computing needs a master plan. Nature 604(7905), 255–260 (apr 2022). `https://doi.org/10.1038/s41586-021-04362-w`

33. Mikhaylov, A., Pimashkin, A., Pigareva, Y., *et al.*: Neurohybrid memristive cmos-integrated systems for biosensors and neuroprosthetics. Frontiers in Neuroscience 14, 358 (apr 2020). `https://doi.org/10.3389/FNINS.2020.00358/BIBTEX`

34. Miranda, E., Suñé, J.: Memristors for Neuromorphic Circuits and Artificial Intelligence Applications. Materials 13(4), 938 (feb 2020). `https://doi.org/10.3390/MA13040938`

35. Mochida, R., Kouno, K., Hayata, Y., *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. Digest of Technical Papers - Symposium on VLSI Technology 2018-June, 175–176 (oct 2018). `https://doi.org/10.1109/VLSIT.2018.8510676`

36. Pei, J., Deng, L., Song, S., *et al.*: Towards artificial general intelligence with hybrid Tianjic chip architecture. Nature 572(7767), 106–111 (jul 2019). `https://doi.org/10.1038/s41586-019-1424-8`

37. Pfeiffer, P., Egusquiza, I.L., DI Ventra, M., *et al.*: Quantum memristors. Scientific Reports 6(1), 1–6 (jul 2016). `https://doi.org/10.1038/srep29507`

38. Pulvermüller, F., Tomasello, R., Henningsen-Schomers, M.R., Wennekers, T.: Biological constraints on neural network models of cognitive function. Nature Reviews Neuroscience 22(8), 488–502 (jun 2021). `https://doi.org/10.1038/s41583-021-00473-5`

39. Rao, M., Tang, H., Wu, J., *et al.*: Thousands of conductance levels in memristors integrated on CMOS. Nature 615(7954), 823–829 (mar 2023). `https://doi.org/10.1038/s41586-023-05759-5`

40. Roy, K., Jaiswal, A., Panda, P.: Towards spike-based machine intelligence with neuromorphic computing. Nature 575(7784), 607–617 (nov 2019). `https://doi.org/10.1038/s41586-019-1677-2`

41. Schegolev, A.E., Klenov, N.V., Soloviev, I.I., *et al.*: Superconducting Neural Networks: from an Idea to Fundamentals and, Further, to Application. Nanobiotechnology Reports 16(6), 811–820 (nov 2021). `https://doi.org/10.1134/S2635167621060227/METRICS`

42. Sebastian, A., Le Gallo, M., Khaddam-Aljameh, R., Eleftheriou, E.: Memory devices and applications for in-memory computing. Nature Nanotechnology 15(7), 529–544 (mar 2020). https://doi.org/10.1038/s41565-020-0655-z

43. Serb, A., Corna, A., George, R., *et al.*: Memristive synapses connect brain and silicon spiking neurons. Scientific Reports 10(1), 1–7 (feb 2020). https://doi.org/10.1038/s41598-020-58831-9

44. Shchanikov, S., Zuev, A., Bordanov, I., *et al.*: Designing a bidirectional, adaptive neural interface incorporating machine learning capabilities and memristor-enhanced hardware. Chaos, Solitons and Fractals 142, 110504 (jan 2021). https://doi.org/10.1016/J.CHAOS.2020.110504

45. Shen, J., Shang, D., Chai, Y., *et al.*: Nonvolatile Multilevel Memory and Boolean Logic Gates Based on a Single Ni/ [Pb (Mg1/3Nb2/3) O3] 0.7 [PbTiO3] 0.3 /Ni Heterostructure. Physical Review Applied 6(6), 064028 (dec 2016). https://doi.org/10.1103/PHYSREVAPPLIED.6.064028/FIGURES/5/MEDIUM

46. Spagnolo, M., Morris, J., Piacentini, S., *et al.*: Experimental photonic quantum memristor. Nature Photonics 16(4), 318–323 (mar 2022). https://doi.org/10.1038/s41566-022-00973-5

47. Su, F., Chen, W.H., Xia, L., *et al.*: A 462GOPs/J RRAM-based nonvolatile intelligent processor for energy harvesting IoE system featuring nonvolatile logics and processing-in-memory. Digest of Technical Papers - Symposium on VLSI Technology pp. C260–C261 (jul 2017). https://doi.org/10.23919/VLSIT.2017.7998149

48. Telminov, O., Gornev, E.: Possibilities and Limitations of Memristor Crossbars for Neuromorphic Computing. Proceedings - 6th Scientific School "Dynamics of Complex Networks and their Applications", DCNA 2022 pp. 278–281 (2022). https://doi.org/10.1109/DCNA56428.2022.9923302

49. Vasileiadis, N., Ntinas, V., Sirakoulis, G.C., Dimitrakis, P.: In-Memory-Computing Realization with a Photodiode/Memristor Based Vision Sensor. Materials 14(18), 5223 (sep 2021). https://doi.org/10.3390/MA14185223

50. Vongehr, S., Meng, X.: The Missing Memristor has Not been Found. Scientific Reports 5(1), 1–7 (jun 2015). https://doi.org/10.1038/srep11657

51. Wan, W., Kubendran, R., Schaefer, C., *et al.*: A compute-in-memory chip based on resistive random-access memory. Nature 608(7923), 504–512 (aug 2022). https://doi.org/10.1038/s41586-022-04992-8

52. Wang, S., Li, Y., Wang, D., *et al.*: Echo state graph neural networks with analogue random resistive memory arrays. Nature Machine Intelligence 5(2), 104–113 (feb 2023). https://doi.org/10.1038/s42256-023-00609-5

53. Wang, Z., Wu, H., Burr, G.W., *et al.*: Resistive switching materials for information processing. Nature Reviews Materials 5(3), 173–195 (jan 2020). https://doi.org/10.1038/s41578-019-0159-3

54. Wu, T.F., Le, B.Q., Radway, R., *et al.*: 14.3 A 43pJ/Cycle Non-Volatile Microcontroller with 4.7μs Shutdown/Wake-up Integrating 2.3-bit/Cell Resistive RAM and Resilience Techniques. Digest of Technical Papers - IEEE International Solid-State Circuits Conference 2019-February, 226–228 (mar 2019). https://doi.org/10.1109/ISSCC.2019.8662402

55. Xia, Q., Yang, J.J.: Memristive crossbar arrays for brain-inspired computing. Nature Material 18(4), 309–323 (mar 2019). https://doi.org/10.1038/s41563-019-0291-x

56. Xue, C.X., Chen, W.H., Liu, J.S., *et al.*: 24.1 A 1Mb Multibit ReRAM Computing-In-Memory Macro with 14.6ns Parallel MAC Computing Time for CNN Based AI Edge Processors. Digest of Technical Papers - IEEE International Solid-State Circuits Conference 2019-February, 388–390 (mar 2019). https://doi.org/10.1109/ISSCC.2019.8662395

57. Xue, C.X., Chiu, Y.C., Liu, T.W., *et al.*: A CMOS-integrated compute-in-memory macro based on resistive random-access memory for AI edge devices. Nature Electronics 4(1), 81–90 (dec 2020). https://doi.org/10.1038/s41928-020-00505-5

58. Yao, P., Wu, H., Gao, B., *et al.*: Fully hardware-implemented memristor convolutional neural network. Nature 577(7792), 641–646 (jan 2020). https://doi.org/10.1038/s41586-020-1942-4

59. Yin, S., Sun, X., Yu, S., Seo, J.S.: High-Throughput In-Memory Computing for Binary Deep Neural Networks with Monolithically Integrated RRAM and 90-nm CMOS. IEEE Transactions on Electron Devices 67(10), 4185–4192 (oct 2020). https://doi.org/10.1109/TED.2020.3015178

60. Zhang, W., Gao, B., Tang, J., *et al.*: Neuro-inspired computing chips. Nature Electronics 3(7), 371–382 (jul 2020). https://doi.org/10.1038/s41928-020-0435-7

61. Zhao, H., Liu, Z., Tang, J., *et al.*: Energy-efficient high-fidelity image reconstruction with memristor arrays for medical diagnosis. Nature Communications 14(1), 1–10 (apr 2023). https://doi.org/10.1038/s41467-023-38021-7

62. Zhu, X., Wang, Q., Lu, W.D.: Memristor networks for real-time neural activity analysis. Nature Communications 11(1), 1–9 (may 2020). https://doi.org/10.1038/s41467-020-16261-1

63. Zhuk, M., Zarubin, S., Karateev, I., *et al.*: On-Chip TaOx-Based Non-volatile Resistive Memory for in vitro Neurointerfaces. Frontiers in Neuroscience 14, 94 (feb 2020). https://doi.org/10.3389/FNINS.2020.00094/BIBTEX

64. Zidan, M.A., Fahmy, H.A.H., Hussain, M.M., Salama, K.N.: Memristor-based memory: The sneak paths problem and solutions. Microelectronics Journal 44(2), 176–183 (feb 2013). https://doi.org/10.1016/J.MEJO.2012.10.001