DOI: 10.24412/2181-144X-2023-2-5-9

Sattorov O.U., Murodullayeva Sh.Sh.

NEYRON TARMOQLARNI BOSHQARISH TIZIMLARIGA MAVJUD YONDASHUVLARNI KO'RIB CHIQISH

Sattorov O.U. – PhD., dotsent, Navoiy davlat konchilik va texnologiyalar universiteti
Murodullayeva Sh.Sh. – magistr, Navoiy davlat konchilik va texnologiyalar universiteti

Annotatsiya. Maqolada formal neyron tarmoq algoritmlarini apparatni tezlashtirishning mavjud yondashuvlari ko'rib chiqiladi, bu esa ishlash va energiya iste'moli nuqtai nazaridan ushbu muammolarni hal qilishning eng istiqbolli vositalari sinaptik birikmalar sifatida memristor tuzilmalari bilan ixtisoslashgan arxitektura yechimlari ekanligini ko'rsatadi..
Kalit so'zlar: neyron tarmoq, hisoblash tizimlari, grafik ishlov berish, neyrochip.

ОБЗОР СУЩЕСТВУЮЩИХ ПОДХОДОВ К СИСТЕМАМ УПРАВЛЕНИЯ НЕЙРОННЫМИ СЕТЯМИ

Сатторов О.У. – PhD., доцент, Навоийский государственный горно-технологический университет
Муродуллаева Ш.Ш. – магистр, Навоийский государственный горно-технологический университет

Аннотация. В статье рассматриваются существующие подходы к аппаратному ускорению алгоритмов формальных нейронных сетей, которые показывают, что наиболее перспективными средствами для решения этого круга задач с точки зрения производительности и энергопотребления являются специализированные архитектурные решения с мемристорными структурами в качестве синаптических соединений.
Ключевые слова: нейронная сеть, вычислительные системы, обработка графики, нейрочип.

OVERVIEW OF EXISTING APPROACHES OF NEURAL NETWORK CONTROL SYSTEMS

Sattorov O.U. – PhD., Associate Professor, Navoi State University of Mining and Technology
Sh.Sh. Murodullayeva – Master, Navoi State University of Mining and Technology

Annotation. The article discusses existing approaches to hardware acceleration of formal neural network algorithms, which shows that the most promising means for solving this range of problems in terms of performance and energy consumption are specialized architectural solutions with memristor structures as synaptic connections.

Key words: Neural network, computing systems, processing graphic, neurochip.

Continuous development of solid-state electronics and computer systems it largely determined the success of science and technology in the second half of the twentieth and the beginning of the twenty-first century and is an increasingly strengthening trend at the present time. The free access of the scientific community to computing resources that are growing every year allows us to reach a qualitatively new level of research in many areas, for example, in decoding the human genome, processing medical data, developing new energy sources and ways to use it effectively, and even in astronomical observations and spacecraft control. One of these most important areas is the development of neural network algorithms – mathematical models of systems of interconnected nerve cells that perform information processing functions. Despite a fairly long history of development, the real heyday in this area came with the widespread use of powerful hardware systems for processing visual images – graphics processors (GPUs). Thanks to a multi-core architecture focused on the mass execution of vector-matrix multiplication operations, which simultaneously underlie both image construction and neural network algorithms, researchers have obtained a hardware simulation tool for hundreds and thousands of interconnected neurons. This allowed us to achieve breakthrough results in such areas as: recognition of visual images and speech, handwritten text, language machine translation, clustering and classification of big data, unmanned vehicle control, forecasting, planning, decision-making and many others.

Despite the considerable theoretical and practical groundwork of research in the field of neuromorphic systems, significant progress in this area was possible only with the widespread use of hardware acceleration devices for processing graphic images (GPU), which, in turn, is associated with the rapid development of the computer game industry in the late 1990s.

The close interaction of such distant regions is due to the specifics of the rendering process of continuously changing models of objects in the video frame. It consists in the need to perform a significant number of vector-matrix multiplication operations, since the object is defined by a set of coordinates in space, over which it is necessary to perform some transformations in the process of creating a new frame in order to visualize its movement or change.

The success of the development of neural network technologies in the 2000s, their active

the complication and introduction of voice recognition algorithms into search engines, in particular, led to the need for a significant increase in the cloud computing capacity on which they were performed, up to doubling the number of data centers, which, on the one hand, is unprofitable from an economic point of view, and on the other hand could solve the problem only for a while [1]. The answer to this challenge was the development by Google of a specialized digital integrated circuit, which is focused on processing tensor data of deep neural networks (TPU).

The choice of the architecture of the future device and the rejection of graphics accelerators was justified by two reasons (Figure 1):

− The first is the approach to the boundary of Moore's and Dennard's laws in 2006

year [2], according to which the maximum performance of single-core systems was achieved with the maximum possible and non-failure of the chip allocated per unit area of power;

− The second is a consequence of the solution of the first: approaches to borders

scaling the size of processor cores and reaching the limit of Amdahl's law, illustrating the limitations of system performance growth with an increase in the number of calculators [3].

Together, this leads to the fact that since the beginning of the epoch microprocessor devices simultaneously with an incredible increase in their performance, there is also a decrease in the rate of this growth: from 50%/year in the 1980s and 90s, to 3-4%/year in the 2010s. Thus, in order to gain an advantage in the performance of neural network computing, it is necessary to develop a chip specialized for a specific task.

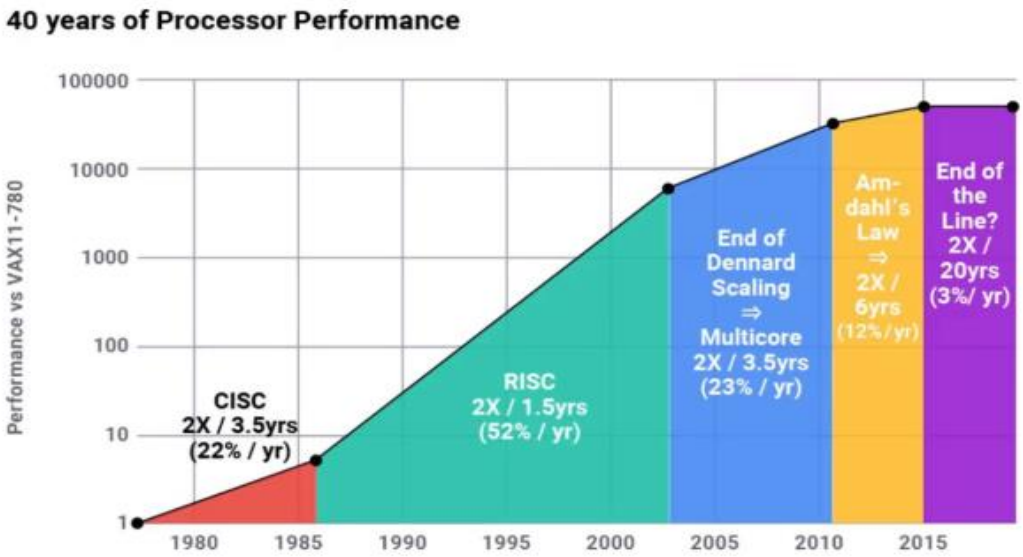

Figure 1 – Evolution of computing systems performance over time.

Nevertheless, the tensor processor is not a comprehensive solution, but only complements the CPU and GPU, optimizing the operation of large-scale convolutional and recurrent neural network algorithms [4].

Currently, PLCs are widely used in two directions:

− Prototyping and testing of new neuromorphic systems, as an intermediate stage in the creation of digital integrated circuits, such as TPU;

− Solving applied problems by implementing on the basis of a programmable logic of various neural network algorithms: the use of convolutional neural networks for recognizing musical notes, the paper describes the implementation of a multilayer perceptron for detecting various gas media (methane, hydrogen, carbon monoxide, etc.), analysis of encephalograms to search for arrhythmia in a patient, control of a humanoid robot arm, handwritten digit recognition from a mobile camera, speech recognition, etc.

The spread of this method of hardware acceleration is due to a wide choice of FPGAs from a hardware point of view and a developed software base, which allows researchers to design highly parallel data processing systems in a very short time with a multiple increase in computing speed, in comparison with central processors.

Unlike the digital systems described earlier and used for to accelerate algorithms for recognition and classification of various information, the SpiNNaker architecture was developed in 2012 to simulate the operation of the cortex column of the brain in real time based on anatomical data obtained from anatomical studies in order to determine the principles of its operation [5]. As in previous cases, the creation of a new architecture is due to the extremely slow modeling of large networks on the von Neumann architecture. For example, the paper [6] describes a thousandfold lag from the real-time mode when simulating the mouse cortex. According to the proposed approach, each SpiNNaker chip contains 2500 thousand neurons implemented according to the Izhikevich model [7], one million synaptic connections, a routing system for data exchange with other processors of the system. The results of the implementation showed a significant reduction in energy

consumption when performing similar tasks on computers with classical architecture (100 NJ per neuron, 43 NJ per synaptic connection).

The developers of the Neurocore neurochip have set themselves the task of performing similar functions to emulate the work of the brain [8]. The processor consists of an array of 256x256 silicon neurons, a transmitter, receiver, router and two blocks of RAM. The neuron circuit provides for the presence of a soma, a dendrite, four strobing and four synapses of the population.

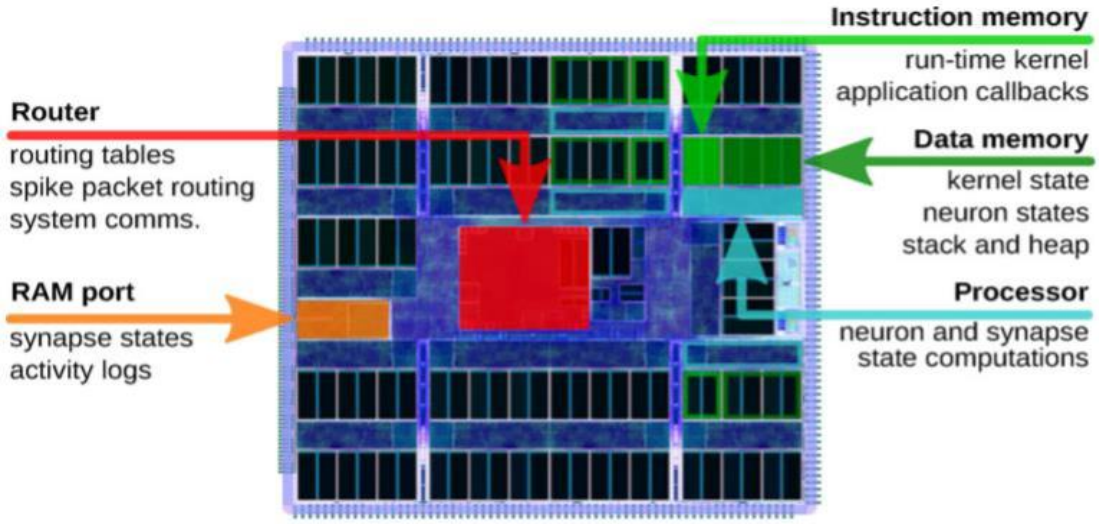

Figure 2 – Block diagram of the SpiNNaker neurochip.

Despite this, currently accelerators based on memristors have not yet become widespread and remain at the level of individual prototypes of microprocessor systems. First of all, this is due to the variations in the characteristics of memristor structures: the storage time of resistive states, resistance to a large number of switches from one boundary state in conductivity to another, low plasticity and the presence of a spread of parameters from device to device. In addition, existing hardware implementations are built around a crossbar-type architecture, which imposes a number of restrictions on the implementation of formal and, in particular, impulse neural network algorithms, for example, the need to introduce special blocks for calculating changes in synaptic weight into the scheme, following the example of Intel Loihi, which affects the overall performance of the system with a large number of synaptic weights in the neural network.

Other difficulties of working with arrays of memristors in crossbar geometry manifest themselves in the form of a number of parasitic effects, including attempts to minimize their influence. The analysis of the work showed that despite the property of the transistor to control the current passing through it, there is still a problem with shunt currents in large 1T1R type matrices (more than 256 elements in a row/column). In addition, there are other effects that can lead to incorrect operation of the system. These include transients and leakage currents associated with the discharge of capacitances of field-effect transistors and non-ideal properties of the gate dielectric, respectively.

References

1. Sejnowski T. J. The Computer and the Brain Revisited // IEEE Ann. Hist. Comput. 1989. Vol. 11. P. 197–201.

2. Mead C. Neuromorphic Electronic Systems Proceedings of the IEEE 1990. vol. 78. P. 1629–1636.

3. Chiang M. L., Lu T. G., Kuo J. B. Analogue adaptive neural network circuit // IEE Proceedings, Part G Circuits, Devices Syst. 1991. Vol. 138. P. 717–723.

4. Burr J. B. Digital Neural Network Implementations 1 Introduction 2 Classifying VLSI implementations // Neural networks, concepts, Appl. implementations. 1995. P. 1–48.

5. Monroe D. Neuromorphic computing gets ready for the (really) big time // Commun. ACM. 2014. Vol. 57. P. 13–15.

6. Nickolls J., Dally W. J. The GPU computing era // IEEE Micro. 2010. Vol. 30. P. 56–69.