Development of Neuromorphic Accelerator

Dmitry E. Ipatov[1], Alexey V. Zverev[2]
[1]A.V. Rzhanov Institute of Semiconductor Physics SB RAS, Novosibirsk, Russia
[2]"Motiv NT" LLC, Moscow, Russia

Abstract – **The paper shows the design of a neuromorphic accelerator, as well as unified design solutions for creating a scalable modular system of neuromorphic accelerators. The neuromorphic accelerator based on the FPGA was developed. It allows simulating up to 131 thousand neurons with 67 million synaptic connections.**

The neuromorphic cross-connection board, which is the universal platform for working with the neuromorphic accelerators and neural networks that they stimulate, was developed. One cross connection board allows placing up to 16 neuromorphic accelerators, which makes a simulation of up to 2 million neurons with a total number of synaptic connections of up to 1 billion possible.

An energy consumption analysis was made for every developed device.

Index Terms – **Neuromorphic, neuromorphic architecture, neural networks, accelerator, machine learning.**

I. INTRODUCTION

THE UPCOMING era of big data needs new computing tools that allow hearing, seeing and processing a huge amount of noisy data in real time [1]. The complexity of solving such problems becomes almost an insuperable obstacle for traditional software systems [2].

There are many examples of computational problems, in which the development of clear and effective algorithms is exceedingly complicated and often unfeasible (for example, optical symbol recognition; object, gesture, emotion recognition; network intrusion detection, spam filters; clustering problems and so on).

These fairly severe restrictions require to use a different approach that has another area of computer science – machine learning – to solve such problems. The machine learning of artificial neural networks (NN) allows us to achieve the best results and, sometimes, even to find solutions for problems that cannot be solved in a conventional way.

The conventional hardware implementation of neural networks on a Central Processing Unit (CPU) or Graphical Processing Unit (GPU) requires huge energy resources in the neural network inference. As a result, it is difficult to use neural networks in various fields of human activity and, in particular, neural networks in compact and autonomous intelligent devices. At the same time, the outstanding human brain with 89 billion parallel computational units reaches tremendous efficiency in perceptual and cognitive tasks with an energy consumption up to 25 W [3]. Modern advances in the field of microelectronics enable us to develop and manufacture application-specific integrated neuromorphic circuits (ASICs) with the biologically initiated architecture [4]. It tries to resemble the structure and working principles of biological neural systems in a simplified way. Such ASICs are significantly different from the CPU, GPU in terms of energy efficiency and density.

However, there is no single rational solution for the neuromorphic ASIC architecture design [5-10]. The designs presented to the public are mostly research projects. Current commercial products for neural network inference and training do not use the neuromorphic architecture. They are merely a tensor processor modification similar to the GPU (for example, Intel Movidius [11]).

Neuromorphic accelerator presented here is focused on testing and spiking neuromorphic architectures approbation to solve visual object recognition and real time manufacturing processes anomaly detection problems. Since spiking neural networks (SNN) training is currently underdeveloped, a pre-trained SNN is loaded in neuromorphic accelerator in order to inference it. The hardware development assumes close cooperation with companies involved in NN implementation in various areas of human activity to optimize neuron hardware model and mechanisms of interaction between them, with subsequent optimal architecture implementation in an ASIC chip.

Unlike tensor accelerators presented earlier (for example [11]), neuromorphic accelerator developed implements SNN model. The accelerators can be combined with each other to increase neural network scale dramatically and, accordingly, its performance and processing depth. The accelerator interface is unified so it can be used with the succeeding ASIC-based accelerator together.

II. PROBLEM STATEMENT

Therefore, for the development of commercial neuromorphic ASIC to be made, it is necessary to develop a hardware infrastructure for the study and trial operation of spike neural networks built with such ASIC. This infrastructure includes both specific software and a number of hardware solutions, comprising a prototype of neuromorphic ASIC based on a field-programmable gate array (FPGA). Moreover, we assume the possibility of using FPGA-based accelerators in solving practical problems where power consumption is not a limiting factor, but the use of NN is required.

978-1-7281-1753-9/19/$31.00 ©2019 IEEE

Thus, in this work, the main objectives are the following: developing common design solutions for the neuromorphic accelerator, including its overall dimensions, mechanical and electrical interface; developing FPGA-based neuromorphic accelerator for the approbation of solutions, which will be implemented later on a neuromorphic ASIC; developing neuromorphic cross connection board, i.e. a backplane, which is a universal platform for working with neuromorphic accelerators.

III. THEORY

A. Neuromorphic accelerator system architecture

The project of neuromorphic accelerator involves the development of a modular system that will support the scaling of a neural network size. The elementary block of a neural network is a neuron. It is a computational unit that has incoming signal lines (dendrites) and outcoming signal lines (axons). Neurons interconnection is made with synapses (interconnection between an axon and a dendrite, which is thought to be memory), and a developed architecture assumes that the function of a neuron is based on a rigid algorithm, which is implemented as a finite state machine (called core). This core approach allows a simulation of a set of neurons by the time multiplexing of one physical circuit. Part of an internal core structure is shown in Fig.1.

The neuromorphic system involves a global synchro-signal called tick. It initializes the neuron function simulation. Every neuron potential is updated sequentially before the next tick is issued. Some neurons emit a packet, called spike, to the pre-specified neurons if their potentials become sufficiently high.

Any neuron of any core can issue spikes to any target neuron inside of one core or one chip, as well as cores in other chips. When a spike is issued, it comes to the core router that reads the relative target destination address of a spike and transmits it to the nearest core in the x or y direction. The nearest core router reads a spike and, if the target is not in that core, it transmits it to the nearest core again. This process is continued repeatedly until the target is found. Peripheral chip controllers allow transmitting spikes seamlessly. Thus, they enable us to greatly increase the simulated neural network size. The intra- and inter-chip communication process is shown in Fig. 2.

By combining cores in a two-dimensional cross connected matrix with directions “North”, “East”, “West” and “South”, we can get an easily scalable neural network that is shown in Fig. 3. The further implementation of such modular system consists of two solutions: prototyping a FPGA-based neuromorphic accelerator and developing a neuromorphic accelerator based on ASIC. The results of the development of an FPGA-based accelerator module are presented in this paper. Neuromorphic accelerators are installed in a neuromorphic cross-connection board or backplane. Backplane supports an installation of up to 16

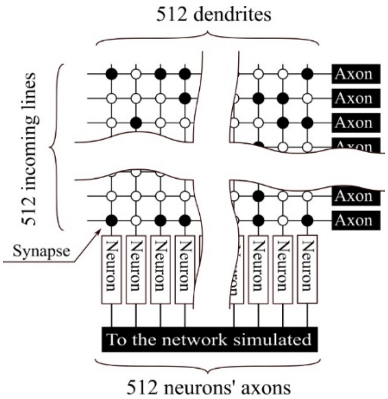

Fig. 1. Part of the internal core structure. Every neuron in a core has dendrites that can be connected to the other neurons axons with a synapse (black dot). When the neuron potential in a core becomes sufficiently high, the neuron emits a spike on its axon to a pre-specified neuron.

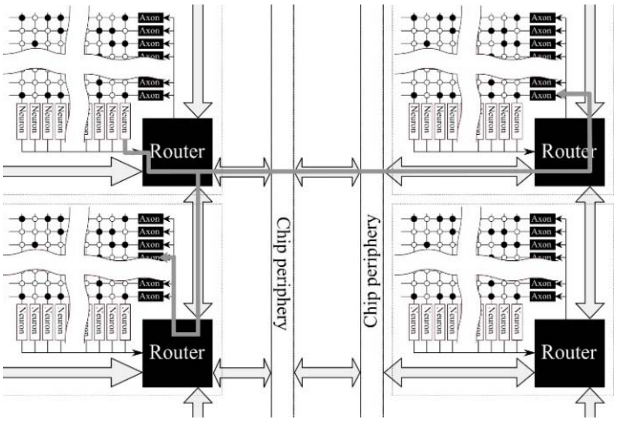

Fig. 2. The process of intra- and inter-chip communication. When a spike is issued, it comes to the core router that reads the relative destination address of a spike and transmits it to the nearest core in the x or y direction. The nearest core router reads the spike information and, if the spike target is not in that core, transmits it again. Seamless transitions of the spike between chips is made with the chip periphery, thus, allowing a big increase of the simulated neural network size.

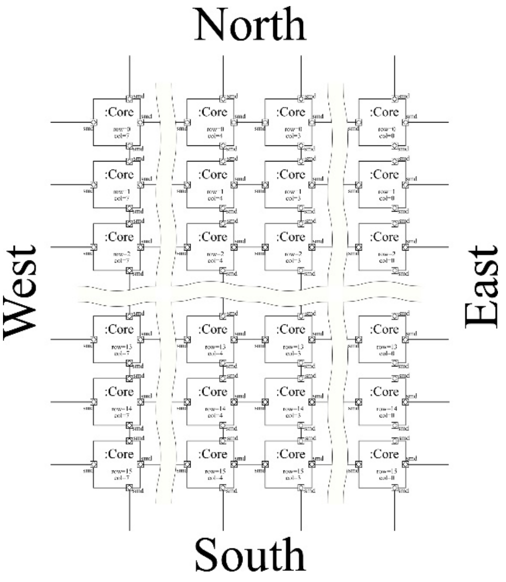

Fig. 3. Two-dimensional cross connected matrix of cores with directions “North”, “East”, “West” and “South”.

neuromorphic accelerators, both based on FPGA and ASIC. It provides the communication between a personal computer and simulated neural network, allows a connection of several others backplanes for increasing the simulated neural network size.

B. *Neuromorphic core structure*

The neuromorphic core structure of is determined by the requirements for its implementation in an ASIC. The FPGA-based neuromorphic accelerator must be designed in such a way that the neuromorphic cores in it could have all the functionality and characteristics of ASIC cores.

Having analyzed a number of typical tasks that the neural network based on a neuromorphic structure should solve, we decided that the optimal ratio between the number of simulated neurons and the number of axons in the core should be one-to-one [12]. It is known from the literature that the typical frequency of updating neuron state (tick) is 1 kHz [13].

From the viewpoint of effective memory access organization, optimal use of memory block capacity, dynamic range of core finite state machine registers and the dynamic range of a number of values in data structures, it is necessary that the number of stimulated neurons be a multiple of two.

Based on the above stated, the single neuromorphic core should simulate 512 neurons, each neuron is to have 512 synapses and each synapse is characterized with a 2-bit weight selector. In total with the neuron parameters, information about the structure of neural network and service information, the single core requires 640 Kb of memory.

C. *Backplane system structure*

In order to work with neuromorphic accelerators, an unified technical solution is required for building neural networks with one or several accelerators, which should have a user-friendly interface to PC.

The neuromorphic cross connection board or backplane is a printed circuit board with some connectors for installing several neuromorphic accelerators. The backplane is one of the higher-level modular components, which allow scaling the neural network size.

The following are the main technical requirements for the backplane developed: providing a high-speed and user-friendly interface to a PC, supporting multiple neuromorphic accelerators, ability of configuring accelerator modules and simulated neural network, providing an interface for scaling the neural network, neural network activity monitoring, supplying voltage to all consumers.

The maximum number of installed accelerators depends on the following: overall permissible dimensions of a backplane, maximum number of backplane output signals, maximum data transfer rate, total power consumption of a neuromorphic accelerator.

IV. FPGA-BASED NEUROMORPHIC ACCELERATOR

A. *Physical device structure*

The developed FPGA-based neuromorphic accelerator is a printed circuit board sized 133.35 mm x 62.2 mm with 12 copper layers. An FPGA-based neuromorphic accelerator is shown in Fig. 4.

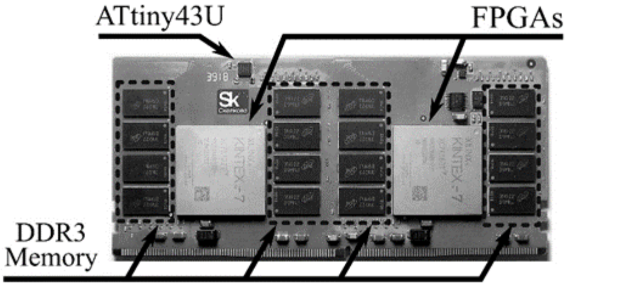

Fig. 4. FPGA-based neuromorphic accelerator. This 12-layer module contains two Xilinx FPGAs, each with two DDR3 64-bit memory controllers. In total, this accelerator allows simulating up to 256 cores. An ATtiny43U is used for the identification and FPGA configuration processes.

During one tick all neurons must be sequentially processed, namely, 640 Kb of the memory of each core must be read. Based on the theoretical assumption, a core must process 512 synapses in each of 512 neurons. In case of optimal core finite state machine implementation, only one clock cycle is needed for processing one neuron synapse. Thus, the core frequency must be at least 263 MHz, and the required data transfer rate to the core should be at least 640 Mbps.

A preliminary analysis during the development of a neuromorphic ASIC has shown that the memory required for storing information amounts to about 70% of the total chip area. In case of an FPGA-based accelerator, it is possible to use the internal FPGA memory. However, the solutions presented on the market do not have any sufficient memory capacity [14].

Therefore, we decided to provide neuromorphic cores with a high bandwidth external memory. A conventional DDR3 memory interface is the preferable one [15], and it is more energy-efficient, compared to DDR2, and cheaper (in terms of FPGA resources) than DDR4. The choice of FPGA was made in favor of the Xilinx Kintex 7 XC7K160T-1FFG676 chip with the optimal price/performance ratio. This FPGA can implement two 64-bit DDR3 memory controllers, with the total bandwidth of 100 Gbps at 400 MHz controllers clock frequency. Each memory controller is supported with four Micron 16-bit data bus width memory chips. The neuromorphic accelerator construction allows placing two FPGAs on a single printed circuit board, only slightly increasing the overall dimensions of the accelerator (comparably to ASIC one). As a result, the FPGA-based accelerator allows simulating up to 256 cores. In addition, it was developed signal compatible with the ASIC-based accelerator. Thus, both can be used on one backplane. The neuromorphic accelerators use the 288-pin edge connector DDR4 DIMM format. This common edge connector is focused on high speed interfaces and is very convenient in design and use.

The accelerator identification on the backplane and the FPGA configuration routine are carried out with the ATtiny43U microcontroller.

B. Logical device structure

The main requirement for signal transmission lines is the data transfer rate of 1200 Mbps, which corresponds to frequency 600 MHz using the Double Data Rate (DDR). In case of the FPGA-based accelerator, the memory controller frequency is limited to 400 MHz or 800 Mbps due to the architecture features.

The developed module has several interfaces: spike transmission interface, memory interface and service interfaces.

The signal lines that correspond to cardinal directions “North”, “South”, “East” and “West” are the spike transmission interface elements. Each spike is a packet that contains 40 bits of useful data, encrypted with the 4b/5b encoding and transmitted with the NRZI line code.

The memory interface, as said before, is implemented on four 64-bit DDR3 memory controllers.

The FPGA-based accelerator also contains the data transmission lines of I2C standard [16] for obtaining diagnostic information and configuring the FPGAs on the accelerator. The FPGA configuration process is performed by loading the data into the internal memory from external sources through a two-wire serial interface.

The ATtiny43U microcontroller is responsible for the FPGA configuration process execution on the accelerator. Both accelerator FPGAs are configured with a single binary data stream. The verification of configuration completion is performed separately for each FPGA. Each neuromorphic accelerator also has a four-wire JTAG interface controlled by a backplane FPGA.

V. NEUROMORPHIC CROSS CONNECTION BOARD

A. Physical device structure

In order to build scalable modular data processing systems, many manufactures follow a single industry standard in the equipment design. Devices are made as single printed circuit boards, combined in frames, which are then placed in racks. There is a size unification within this industry standard: the equipment height is measured in “units”, U [17].

A simplified image of neuromorphic cross-connection board is shown in Fig. 5. The neuromorphic accelerators are installed in DDR4 DIMM sockets, which are 162 mm long and 6.5 mm wide. Since the backplane has not only sockets, its width was increased to 6U, which corresponds to the generally accepted standard Eurocard 6U [18] or 233.35 mm.

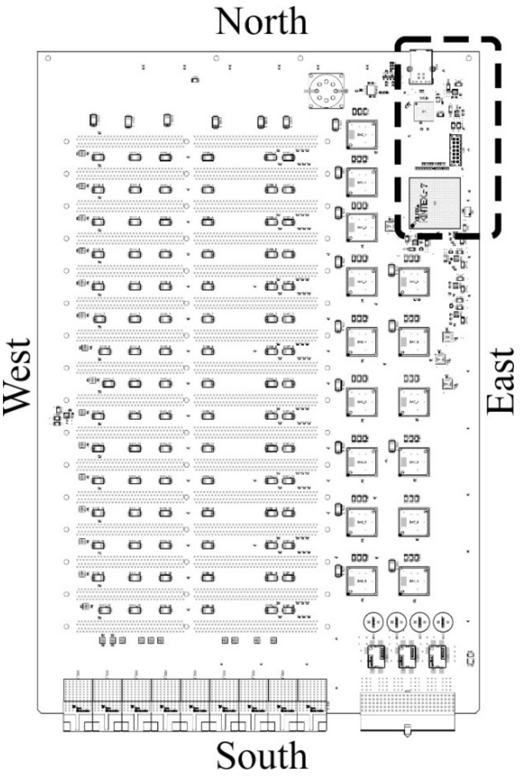

Fig. 5. Neuromorphic cross connection board. This 8-layer backplane is a universal platform for the user-friendly interaction from a PC with accelerator-simulated neural network. Up to 16 neuromorphic accelerators may be installed in DDR4 DIMM sockets. The simulated neural network has four cardinal directions, where “North” is the FPGA (USB bridge with it is marked with a dotted rectangle). An additional increase of the simulated neural network size is possible by means of combining several backplanes by high-speed differential connectors (image bottom).

The developed backplane does not assume any storage of configuration and neural network data on a non-volatile memory. Therefore, an external interface with the highest possible data transfer rate is needed to load configurations. The most popular solutions on the market are USB 3.1 SuperSpeed [19] and Thunderbolt 3 [20]. Despite the fact that Thunderbolt offers the highest possible data transfer rate up to 40 Gbit/s, compared to USB 3.1 (5 Gbit/s), there is a whole set of element base manufacturers in favor of the latter, which greatly simplifies the development process. For a hardware implementation of USB 3.1, a Cypress EZ-USB FX3 microcontroller is used.

The USB interface realization does not solve the problem of interaction with neuromorphic accelerators. Thus, a hardware bridge was made between the USB interface and accelerators. On the one hand, the bridge must be a part of a neural network which will receive incoming spikes. On the other hand, it must provide an interface for a data transfer to the USB standard. The analysis of existing solutions showed that the best design for providing an access to all necessary backplane elements is a set of USB microcontroller and FPGA. The choice of FPGA was the same as for accelerator.

The increase of neural network size is made with high-speed differential connectors. Based on the ability of installing several backplanes in frames and racks, optimal number of output differential pairs and total power consumption, it was decided to install 16 sockets for neuromorphic accelerators. In that case, the number of output differential pairs is 274 mm, and the backplane length is 350 mm.

The neural network configuration is loaded through spike transfer ports by emitting a special configuration packet.

B. Logical device structure

The interaction between backplanes is made with two differential lines using a GTX Transceiver on the FPGA as a separate high-speed data transfer unit.

The I2C transfer lines are routed through the entire backplane. Its operating frequency is 100 kHz, Cypress FX3 operates as a master, and any other device is a slave. That low-speed interface is used to identify the type of module and (in case of the FPGA-based accelerator) to configure FPGAs. Since the installed accelerators may be different, each of them may need its own supply voltage values, and that is controlled by digital potentiometers via I2C.

Therefore, adjustable DC-DC converters are used on the backplane, and the voltage divider based on digital potentiometers is responsible for changing the converters output voltage. Voltage values will be set based on types of installed accelerators. The set value is stored in the non-volatile memory of potentiometers and may be further changed by a user.

The backplane FPGA configuration is performed, using Cypress FX3. After this, FPGA and FX3 act as a bridge between a PC and neuromorphic accelerators. FPGA-based accelerators are also configured using the Cypress FX3 two-wire serial interface.

After the configuration process, FPGA is the "North" of the neural network and is able to receive and transmit spikes on 8 ports. Between FPGA and FX3, the 32-bit data interface, in which the slave is FX3 and the master is FPGA, is realized. Each data line operates at frequency 100 MHz, which provides the total bandwidth of 3.2 Gbps. Since the nature of spikes is sparse in space and time, this bandwidth is more than enough.

Each neuromorphic accelerator has a four-wire JTAG interface controlled by an FPGA on the backplane.

C. Neuromorphic system structure

A neuromorphic system is a network of two-dimensional matrix of cores, that simulates a neural network loaded in it. The neuromorphic system has four directions for scaling: "North", "West", "East" and "West". It is assumed that the backplane with USB is the initial board; any other boards will have no USB.

Increasing of the neuromorphic system sizes for the initial board is available in directions "West" and "South". Direction "North" corresponds to the virtual boundary (USB bridge) between the network and control computer by an FPGA; the "East" direction is terminated on the backplane, thus, forming the second boundary of a neuromorphic system. During the simulated neural network operation, its spike packets can be transmitted outside the backplane and received from the outside.

A series of 16 slots for neuromorphic accelerators extends in direction from "North" to "South". Each accelerator has 4 ports in directions "West" and "East", and 8 ports in directions "North" and "South". In total, the number of "North" and "South" ports remains equal, while each additional accelerator adds 4 ports in directions "East" and "West". Additionally, two differential pairs are used for the communication between the backplanes.

The simulated neural network configuration is loaded through spike transfer ports using special configuration packets. The neuromorphic system operation implies the presence of two main clock signals – a tick and a clock signal. The time step for each neuromorphic accelerator is controlled by issuing a tick. The tick value is adjustable and, in case of FPGA-based accelerators, it can be up to 1 KHz. In terms of the optimal use of resources, Cypress FX3 controls the tick signal issuing. The backplane FPGA generates the 10 MHz system clock frequency for neuromorphic accelerators, which is then multiplied to the required values on each accelerator with the help of a PLL.

VI. Power Consumption

We assume that the power consumption of an FPGA-based neuromorphic accelerator, unlike that of the ASIC-based, can be quite large. For this reason we performed a preliminary analysis of the power consumption using a Xilinx Power Estimator (XPE) [21].

It is worth noting that the main energy consumers on the FPGA-based neuromorphic accelerator are FPGA (70 % of total power) and DDR3 memory chips (20 %). The analysis showed that the single FPGA-based neuromorphic accelerator consumes about 28 W. Therefore, 16 accelerators will consume about 448 W, of which 235 W are supplied to neuromorphic cores, 155 W – to input/output interfaces and 58 W – to peripheral components. The maximum power consumption of neuromorphic cross-connection board only is 16.3 W and it is powered with conventional 12 V.

VII. Conclusion

Within the project for developing the neuromorphic custom application-specific integrated circuit for hardware simulating of spiking neural networks the following issues were developed:

The project of FPGA-based on neuromorphic accelerator module for approbing the main hardware approaches in the development of neuromorphic ASIC. The FPGA-based neuromorphic accelerator is manufactured on the 133.35 mm long and 62.2 mm high 12-layer PCB. The accelerator operates at the clock frequency of 400 MHz, allows implementing 256 neuromorphic cores. In total, the single accelerator can simulate up to 131 thousand of neurons and 67 million of synaptic connections.

The project of the neuromorphic cross connection board, which is a universal platform for working with neuromorphic accelerators. It is manufactured on the 350 mm in long and 233.35 mm wide 8-layer PCB, which corresponds to the industrial standard in the development of scalable modular systems. The backplane allows an installation of up to 16 accelerators, provides the external USB 3.1 interface for the communication with a PC with the maximal bandwidth at 3.2 Gbps. The backplane also has all necessary high-speed interfaces for scaling the simulated neural network sizes by combining several backplanes.

ACKNOWLEDGMENT

This work was supported by FASIE grant No. 249GRNTIS5/36374 and Skolkovo's grant MG23/18.

REFERENCES

[1] I. Hashem et al. The rise of "Big Data" on cloud computing: Review and open research issues. Information Systems, Vol. 47, 2015, pp. 98–115.

[2] C. Snijders, U. Matzat, U.-D. Reips Big Data: Big gaps of knowledge in the field of Internet. International Journal of Internet Science, Vol. 7, No. 1, 2011, pp. 1–5.

[3] S. Herculano-Houzel Scaling of brain metabolism with a fixed energy budget per neuron: implications for neuronal activity, plasticity and evolution. PLoS ONE, Vol. 6, No. 3, 1 March 2011.

[4] C. Mead Analog VLSI and Neural Systems. Boston, MA: Addison-Wesley Longman Publishing Co., Inc., 1989, pp. 63–84.

[5] S. Furber, D. Lester, L. Plana, J. Gaside, E. Painkras, S. Temple, A. Brown Overview of the SpiNNaker System Architecture. IEEE Transactions on Computers, Vol. 62, No. 12, December 2013. pp. 2454 – 2467.

[6] B. Benjamin, P. Gao, E. McQuinn, et al. Neurogrid: A Mixed-Analog-Digital Multichip System for Large-Scale Neural Simulations. Proceedings of the IEEE, Vol. 102, No. 5, 2014. pp. 699 - 716.

[7] BrainScaleS The Participants [Online]. Available: https://brainscales.kip.uni-heidelberg.de/index.html

[8] F. Akopyan 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, Vol. 34, No. 10, 2015. pp. 1537 - 1557.

[9] N. Qiao, H. Mostafa, F. Corradi A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128K synapses. Front. Neurosci., Vol. 9, April 2015.

[10] M. Davies, N. Srinivasa, T. Lin Loihi: A Neuromorphic Manycore Processor with On-Chip Learning. IEEE Micro, Vol. 38, No. 1, February 2018. pp. 82 – 99

[11] Intel Movidius | an Intel Company [Online]. Available: https://www.movidius.com/

[12] Esser S., Merolla P., Arthur J. Convolutional Networks for Fast, Energy-Efficient Neuromorphic Computing. arXiv. 2016. [Online]. Available: https://arxiv.org/pdf/1603.08270.pdf

[13] Physical principles for scalable neural recording. Frontiers in Computational Neuroscience, Vol. 7, Oct 2013. pp. 1 - 34.

[14] Xilinx Inc. 7 Series FPGAs Memory Resources User Guide. Xilinx. 2016. [Online]. Available: https://www.xilinx.com/support/documentation/user_guides/ug473_7Series_Memory_Resources.pdf

[15] DDR3 SDRAM STANDART | JEDEC [Online]. Available: https://www.jedec.org/standards-documents/docs/jesd-79-3d

[16] NXP Semiconductors. UM10204 I2C-bus specification and user manual - UM10204.pdf. [Online]. Available: https://www.nxp.com/docs/en/user-guide/UM10204.pdf

[17] IEC 60297-3-108:2014. Mechanical structures for electronic equipment - Dimensions of mechanical structures of the 482,6 mm (19 in) series - Part 3-108: Dimensions of R-type subracks and plug-in units, International Electrotechnical Commission, 2014.

[18] Eurocard Specifications [Online]. Available: https://www.elma.com/en/services/us-resources/eurocard-specs/

[19] USB.org [Online]. Available: http://www.usb.org/developers/docs/

[20] Thunderbolt™ 3 Technology [Online]. Available: https://www.intel.com/content/www/us/en/io/thunderbolt/thunderbolt-technology-general.html

[21] Xilinx Inc. Xilinx Power Estimator [Online]. Available: https://www.xilinx.com/products/technology/power/xpe.html

Dmitry E. Ipatov was born in Russia in 1995. He received M.S. degree in Nanotechnology from the Novosibirsk State Technical University (NSTU), Russia, in 2018. Now he is a PhD student of Rzhanov Institute of Semiconductor Physics, SB RAS. Since 2017 he has been working in development group of application-specific IC.

Alexey V. Zverev was born in Russia in 1975. He received PhD degree from the Rzhanov Institute of Semiconductor Physics of Siberian Branch of Academy of Sciences in 2007. Now, he is a leader of development group of application-specific IC design.