Evaluation of effectiveness of data processing based on neuroprocessor devices of various models

VA Romanchuk¹

¹Ryazan State University named for S.Yesenin, 46, Svobody ave., Ryazan, 390000, Russia
E-mail: v.romanchuk@365.rsu.edu.ru

Abstract. The paper presents the results of studying the efficiency of information processing based on neuroprocessor devices. We define analytical relationships that characterize information processing in neuroprocessor devices using modern and frequent architectures with sets of functional technical specifications used. The proposed relationships differ in that approaches, and characteristics unique to neuroprocessor devices are offered for evaluation. The obtained relationships can be used for simulating the operation of a neuroprocessor device; analysis of the effectiveness of information processing; optimizing program code to improve performance.

1. Introduction
Neurocomputing is a scientific field dealing with the development of sixth-generation computing systems, neural computers, which consist of a large number of parallel operating simple computing elements (neurons). Elements are interconnected, forming a neural network. They perform uniform computing operations and do not require external control. A large amount of parallel computing elements results in higher performance [1, 2].

Currently, neurocomputers are being developed in most industrialized countries (such manufacturers as Module, Qualcomm, IBM, Toshiba, HumanBrainProject, KnuEdgeInc., Analog Devices, Texas Instruments, Darwin, Google, NVidia, Fujitsu, Eyeriss, Intel). Neurocomputers make it possible to solve many mental problems with high efficiency. These are tasks of pattern recognition, adaptive control, forecasting, diagnostics [3, 4].

However, at present, there are gaps in typical mathematical apparatus for describing and analyzing the functioning of the whole multitude of neurocomputer devices.

The aim of the current study is developing a mathematical apparatus for evaluating some temporal and quantitative characteristics of storage and functioning of standard artificial neural networks in their implementation based on neurocomputer devices of different architectures.

2. Materials and methods
Let \( Z_{INS}^{(j)} \) be somej-th class of neural network problems, which is an artificial neural network when implemented on a microprocessor platform. An artificial neural network can be defined as a tuple of parameters and characteristics [5]:

\[ \text{evaluation of effectiveness of data processing based on neuroprocessor devices of various models} \]
- multiple inputs of the artificial neural network \( \text{Net}_X = \{\text{net}_{x_1}, \ldots, \text{net}_{x_i}, \ldots, \text{net}_{x_n}\} \), where each input is characterized by a type and range of possible values;
- multiple outputs of the artificial neural network \( \text{Net}_Y = \{\text{net}_{y_1}, \ldots, \text{net}_{y_i}, \ldots, \text{net}_{y_m}\} \), where each output \( \text{net}_{y_i} \) is characterized by a type and range of possible values;
- a set of neurons \( \text{Net}_N = \{n_1, \ldots, n_i, \ldots, n_n\} \), each of them requiring emulation, to solve some class of tasks \( Z^{(j)}_{\text{INS}} \);
- a set of weighting coefficients \( \text{Net}_W = \{\text{net}_{w_1}, \ldots, \text{net}_{w_i}, \ldots, \text{net}_{w_n}\} \), characterized by the type and range of possible values;
- topology of the artificial neural network \( \text{Net}_T = \{\text{SLP}, \text{MLP}, \text{RBF}, \text{Hop,Ham}, \text{BAM}, \text{RMLP, Elm, ART, CPN}, \ldots\} \), including the number of layers \( SI \) and the number of neurons in each of the layers;
- the activation function of neurons \( F \);
- method of setting weighting coefficients;
- and other parameters.

The currently used criteria for evaluating such systems as artificial neural networks based on a neurocomputer device reflect two main methodological approaches. In the first case, the system under study is considered as an element of the supersystem, within whose framework its intended use is realized. This approach involves the analysis of the super system's performance, the study of its functional relationships and systems. Efficiency criteria, in this case, are, as a rule, temporary indicators of user interaction with the system, for example, time of access to resources, execution time, downtime.

The other approach considers the parameters of the system under study. Moreover, the characteristics of the system's connections with the supersystem are discussed only in terms of their limitations. Examples of performance criteria of this kind can be speed, availability, load factor. However, they have the same significance and can serve as a measure of comparison only in systems with similar architecture, logic, internal language. Their application is justified in evaluating a system with a primary architecture. These criteria may lose validity, facing any variation of the system parameters.

In the study of more complex structures, such as artificial neural networks based on a neurocomputer device, it is advisable to use criteria selected based on the first approach to assessing a system's performance.

To assess the effectiveness of the organization of data processing in a neuroprocessor device, we take a criterion for the exploitation of equipment, which for the first type of neurocomputer, i.e., a neurocomputer simulating a neuron as a whole; this can be calculated as follows:

\[
K_{\text{NET}} = \sum_{i=1}^{L} \frac{M_i}{H_{\text{NET}}},
\]

where \( M_i \) is the number of neurons emulated on a neurocomputer per \( i \)-th command, \( H_{\text{NET}} \) is the maximum number of neurons emulated on a given neurocomputer per cycle, and \( L \) is the number of neurocommands.

We use two performance criteria in our assessment:
- \( \text{downtime} K_{Tpr} \) – the total sum of times during which any computing elements of a neuroprocessor device stand idle awaiting data;
- \( \text{processing time} K_{To} \) – the total time during which the computing elements of the neuroprocessor device process the data.
3. Neuroprocessors from RTC "Module"

The NeuroMatrix technology combines two modern architectures: VLIW (Very Long Instruction Word) and SIMD (Single Instruction Multiple Data). This combined architecture simultaneously uses two processors on one chip: a vector 64-bit neuroprocessor (DSP) and a scalar 32-bit kernel (RISC) with a single command system. The main operation is that of weighted summation, which requires using hardware nodes such as matrix multipliers with accumulation (Multiply Accumulate, MAC), as this ensures the performance necessary for real-time scale performance. Besides, one of the main ideas embodied in this processor is operating with variable-width operands (from 1 to 64 bits). Due to this data processing speed and calculation, accuracy may vary, and the number of MACs performed per unit of time will depend on the number and lengths of the operands that fit in a 64-bit word. The core of the architecture is a regular structure, similar to a matrix multiplier. The matrix consists of 64x64 cells, each containing a memory element and several logical elements. The matrix can be divided into several sub matrices in two 64-bit programmable registers that define the boundaries of the MAC and the input data.

The $M^{(k)}$ command in the left-hand side of the processor can be represented as a set of separate commands: boot, modification, ftw, and wtw commands $M^{(k)} = <TO^{(k)}_{1tw}, TO^{(k)}_{2tw}, TO^{(k)}_{3tw}, TO^{(k)}_{4tw}>$. All devices will receive data simultaneously [6]; therefore, the structure will remain a vector one.

The command in the right-hand side of the processor can be represented as a set of separate commands: masking, activation, shift, matrix loading, calculation and loading in AFIFO $M^{(k)} = <TO^{(k)}_{5tw}, TO^{(k)}_{6tw}, TO^{(k)}_{7tw}, TO^{(k)}_{8tw}, TO^{(k)}_{9tw}>$.

If $R$ is the bit depth of the data input at the ANN, then for the neuroprocessors of the NM640x family, the number of operations equals:

$$NP = 4 + \left\lceil \frac{64}{R} \right\rceil \times 5.$$ 

The $K_{NET}$ coefficient can be calculated as follows:

$$K_{NET}(NM) = \sum_{i=1}^{k} \frac{M_{Oi}}{4 + \left\lceil \frac{64}{R} \right\rceil \times 5},$$

A vector processor is an independent parallel computing complex of a vector structure [7]. Based on the structure of the processor, $NP \leq 68$, because maximum four operations can be performed on the left side, and maximum 64 operations, on the right side.

Further, we calculate the values of time criteria for a single $k$-th command of the neuroprocessor. The operating time of the neuroprocessor device will be determined, depending on the processor model and type of operation, e.g., for the NM6406 processor $K^{(k)}_{tot}(NM) \approx 1.9 \times 10^{-9}$ c.

The final assessment of performance will be defined as the total of scores for each part of the command. For the left side of the microneurocommand, the downtime is:

$$K_{tp}^{(k)}(left) = 4 \times K_{TP}^{(k)} - M_{O}^{(k)} \times K_{Tw}^{(k)},$$

processing time:

$$K_{T0}^{(k)}(left) = M_{O}^{(k)} \times K_{Tw}^{(k)}.$$

For the right side of the micro-neurocommand:

$$K_{tp}^{(k)}(right) = 5 \times K_{Tw}^{(k)} - M_{O}^{(k)} \times K_{Tw}^{(k)},$$

$$K_{T0}^{(k)}(right) = M_{O}^{(k)} \times K_{Tw}^{(k)}.$$

And then, the final values of the information processing efficiency criteria for the vector command will be determined as follows:
\[ K_{T_p}^{(k)} (NM) = 4 * K_{Tw}^{(k)} - M_{O}^{(k)} * K_{Tw}^{(k)} + \frac{64}{R} (5 * K_{Tw}^{(k)} - M_{O}^{(k)} * K_{Tw}^{(k)}) , \]
\[ K_{T_o}^{(k)} (NM) = \frac{64}{R} (M_{O}^{(k)} * K_{Tw}^{(k)}) + M_{O}^{(k)} * K_{Tw}^{(k)}. \]

4. Neuroprocessors of True North architecture

In 2014, IBM, researching the DARPA SyNAPSE program, introduced a neurosynaptic processor of the new True North architecture. The True North neurosynaptic processor includes about a million programmable electronic neurons and 256M programmable synapses that can transmit signals from one neuron to another. All these elements are grouped in 4,096 neurosynaptic computing cores, which include computing and communication modules, and memory. The cores can work in parallel, and this architecture removes the bottleneck of the traditional processor architecture, which does not allow the simultaneous transfer of instructions and operational data along the same path. Each core allows each block to simulate 256 neurons at a frequency of 1 kHz.

Computations in True North occur according to the spiking type of neural networks, that is, each neuron integrates signals from other neurons that arrive at different times, and only works when integration of incoming signals requires it. Because each neuron has a binary state, it is necessary to use \( R \) neurons to organize the functioning of the ANN with input data of \( R \) capacity [8].

For True North architecture, we get the following coefficient:

\[ K_{NET} (TN) = \sum_{l=1}^{L} \sum_{j=1}^{4096} M_{ij} * R_j \frac{2^8}{2^8}, \]

and the values for criteria are:

\[ K_{T_p}^{(k)} (TN) = \sum_{i=1}^{L} \sum_{j=1}^{4096} \left( \frac{2^8}{R_i} * K_{Tw}^{(k)} - M_{ij}^{(k)} * K_{Tw}^{(k)} \right), \]
\[ K_{T_o}^{(k)} (TN) = \sum_{i=1}^{L} \sum_{j=1}^{4096} (M_{ij}^{(k)} * K_{Tw}^{(k)}). \]

5. Intel’s Loihi Neuroprocessors

The chip has the following features: a fully asynchronous neuromorphic grid supporting a wide range of neural network topologies for each neuron, associated with thousands of others; each neuromorphic core includes a learning mechanism that can be programmed to adapt network parameters during training through various artificial intelligence paradigms; 14-nanometer technological process; 130 K neurons and 130 M synapses; development and testing of many machine learning algorithms with high algorithmic efficiency.

For Loihi architecture, we have the following coefficient:

\[ K_{NET} (L) = \sum_{j=1}^{130*10^6} M_j, \]

and the values for criteria are:

\[ K_{T_p}^{(k)} (L) = 0, \]
\[ K_{T_o}^{(k)} (L) = \sum_{k=1}^{130*10^6} (M_{ij}^{(k)} * K_{Tw}^{(k)}). \]

6. Neuroprocessors of the SpiNNaker architecture

This neurocomputer consists of a regular matrix of 50 chips. Each chip has 20 ARM968 microprocessors, where 19 of them are directly dedicated to the simulation of neurons, and the last one controls the operation of the chip and logs activities. Each chip is a subsystem complete with its
router, internal synchronous dynamic RAM, with direct access mode (32 Kb for storing commands and 64 Kb for storing data), and its messaging system with a throughput of 8 Gbps. Besides, each chip has 1 GB of external memory to store the network topology. According to the developers, the processor allows simulating up to 1,000 neurons in each microprocessor in real-time.

For SpiNNaker architecture, we have the following coefficient:

\[ K_{NET}^{(SNN)} = \sum_{i=1}^{L} \sum_{j=1}^{M} \frac{M_{ij}}{19}, \]

and the values of criteria are:

\[ K_{Tp_i}^{(SNN)} = \sum_{i=1}^{L} \sum_{j=1}^{M} 19 \cdot K_{Tp_i}^{(k)} - M_{ij} \cdot K_{Tp_i}^{(k)}, \]

\[ K_{To_i}^{(SNN)} = \sum_{i=1}^{L} \sum_{j=1}^{M} (M_{ij} \cdot K_{To_i}^{(k)}). \]

7. Neuroprocessors of the KnuEdge architecture

This 256-core processor is designed to perform tasks related to voice identification and a wide range of other tasks related to deep machine learning and self-learning. The KnuPath processor is built under the principles of the Lambda Fabric architecture. The architecture of Lambda Fabric allows creating computing systems that include up to 512 K processors. At the same time, the delay in transmitting data from one rack to another is about 400 nanoseconds, which is quite comparable or higher than the performance of the most modern trunk buses used in supercomputers [9, 10].

On the chip of each KnuPath processor, there are 256 DSP cores, 64 programmable direct memory access modules (DMA), an integrated L1 router, which together provides the processing capacity of one processor at 256 Gflops with a memory bandwidth of 3.702 Gbit/s. The processor has sixteen bidirectional I/O ports, providing data exchange at a speed of 320 Gbit/s.

In this architecture, the processors are combined into computing clusters: on each board, there are eight separate neuroprocessor devices, which are combined into 8 clusters, which in turn are combined into four superclusters.

For KnuPath architecture, we have the following coefficient:

\[ K_{NET}^{(KP)} = \sum_{i=1}^{L} \sum_{j=1}^{M} \sum_{k=1}^{4} \frac{M_{ijk}}{8}, \]

and the values of criteria are:

\[ K_{Tp_i}^{(KP)} = \sum_{i=1}^{L} \sum_{j=1}^{M} \sum_{k=1}^{4} (8 \cdot K_{Tp_i}^{(k)} - M_{ijk} \cdot K_{Tp_i}^{(k)}), \]

\[ K_{To_i}^{(KP)} = \sum_{i=1}^{L} \sum_{j=1}^{M} \sum_{k=1}^{4} (M_{ijk} \cdot K_{To_i}^{(k)}). \]

8. Conclusion

We have developed analytical expressions that allow assessing the values of information processing characteristics when implementing artificial neural networks in neuroprocessor devices: program performance coefficient, downtime and processing time for the most common neuroprocessor architectures: IBM’s TrueNorth architecture, Intel architecture, Loihi architecture, SpiNNaker architecture, processor architecture from KnuEdge, the architecture of neuroprocessors of RTC “Module.” The given analytical relations can be further refined for solving problems of emulation of artificial neural networks of various types and using specific architectural models of modern neuroprocessor devices.
References

[1] Pande S, Morgan F, Cawley S, Bruintjes T, Smit G, McGinley D, Carrillo S, Harkin J and McDaid L 2013 Modular neural tile architecture for compact embedded hardware spiking neural network Neural Process Lett 38 131–153

[2] Izeboudjen N, Larbes C and Farah A 2014 A new classification approach for neural networks hardware: from standards chips to embedded systems on chip ArtifIntell Rev 41 491–534

[3] Kolinummo P, PASI Pulkkinen P, Hämäläinen T and Saarinen J 2000 Parallel implementation of a self-organizing map on the partial tree shape neurocomputer Neural Processing Letters 12 171-82

[4] Romanchuk V 2019 Representation models for parallel processes in parallel, cloud and distributed computing systems based on neuroprocessors Journal of Physics: Conference Series 1333 022014

[5] Romanchuk V 2020 Evaluation of storage and functioning characteristics of artificial neural networks on the basis of a neurocomputer device IOP Conference Series: Materials Science and Engineering 734 012009

[6] Avetisyan A, Grushin D, Ryzhov A 2002 Systems of cluster management Proceedings of the Institute for System Programming of Russian Academy of Sciences 140-147

[7] Ivutin A, Larkin E 2014 Estimation of latency in embedded real-time systems Proceedings of the 3rd mediterranean conference on embedded computing(Montenegro: Budva) 236-239.

[8] Diken E, Jordans R, Corvino R, Jozwiak L, Corporaal H and Chies F 2014 Construction and exploitation of VLIW ASIPs with heterogeneous vector-widths Microprocessors and Microsystems 38 947–959

[9] Jordans R, Jozwiak L, Corporaal H 2014 Instruction set architecture exploration of VLIW ASIPs Using a genetic algorithm Proceedings of the 3rd mediterranean conference on embedded computing(Montenegro: Budva) 32-35

[10] Bekhtin Y, Gurov V, Guryeva M 2014 Algorithmic supply of IR sensors with FPN using texture homogeneity levels Proceedings of the 3rd mediterranean conference on embedded computing(Montenegro: Budva) 252-255.