A. A. Yaroviy, Cand. Sc. (Eng); Yu. S. Bogomolov; K. Yu. Voznesenskiy

APPLIED REALIZATION OF LARGE-SCALE NEURAL AND NEURAL-LIKE PARALLEL-HIERARCHICAL NETWORKS BASED ON GPGPU TECHNOLOGY

Within the frame of research aimed at software-hardware realization of large-scale neural and neural-like parallel-hierarchical networks. The selection of hardware platform for further imitation modeling and practical-applied realization is substantiated. Proceeding from the research carried out and the results obtained programming modules intended for realization on CPU and GPU large-scale neural and neural-like parallel-hierarchical networks of various topologies are suggested.

Key words: *parallel computations, neural networks, digital processing of information, parallel-hierarchical systems, forecasting.*

Introduction

Rapid transition of modern control systems to digital standards resulted in the necessity of rapid processing of super large volumes of information, actuality of such research and the results obtained is typical for systems, where it is necessary to perform complex processing and filtration of signals, for instance, unpacking of compressed audio and video data, routing of data streams, forecasting of dynamic rapidly changing data, all this requires application of rather efficient intelligent computational systems. Similar systems can be realized on various element bases, but nowadays parallel neural-like networking facilities gained wide application.

The main goal of the given research is variance analysis and selection of the most optimal hardware platform intended for modeling of large-scale neural and neural-like parallel-hierarchical networks with further development of the software required for emulation of parallel and parallel-hierarchical computations needed for solution of extremely complex problem of digital processing of information, namely, pattern recognition, dynamic processing of images, forecasting etc.

To solve the above mentioned problems, the paper investigates and analyses principle technologies of hardware realization of artificial neural networks, especially modern specialized neural processors, PLIS, digital signal processors (DSP), multimedia central processors (CPU) and modern alternative hardware facilities (in particular, GPU) to substantiate the selection of basic platform for modeling of various structures of neural and neural-like parallel-hierarchical networks [1 – 5].

Analysis of hardware platforms for modeling of large-scale neural and neural-like parallel-hierarchical networks

Most widely spread schemes – central processors (CPU), video cards and specialized neural processors (neurochips) were considered as the variants of hardware platforms, since their application enables to overlook the level of hardware platform design. Besides the above-mentioned facilities there exist other ways of this problem solution – for instance, manufacture circuits, based on DSP-processors, but one of the drawback of this approach – is the need to use certain topology of the network [2 – 6].

Program emulation on CPU

One of the most widely used methods of neural networks realization is virtual creation of topology – in the form of weight matrices set and activation levels. Topology may be of any dimensionality (dimensions of the network are limited by available main memory). Maximum main memory amount for home computer – 8 GB, that enables to create the array of double type (floating point number of double precision, size – 8bytes) 32768x32768 of size. For servers maximum main

memory amount is32 GB, correspondingly, the dimensions of the given array can be two times larger (65536x65536) [2, 4, 7]. These calculations are rather relative, since we did not take into consideration the memory occupied by operation system and user's program itself.

Table 1

Principle advantage and disadvantages of program emulation on CPU

Advantages	Disadvantages
1) Wide-spread and free access of hardware platform; 2) modeling flexibility – possibility of realization of any topology, using any programming language; 3) high accuracy of the result while performing computations (up to 128 bits); 4) large available main memory (up to 8 GB for home computer and up to 32 GB per processor for server).	1) Lower speed while performing real tasks then in specialized neurochips or videoadaptors; 2) relatively less amount of data loadings from the memory than in neurochips and videoadaptors.

Neurochips

Structure, realized on the crystal, containing numerous cores with intercore connections, which correspond to preset topology or can correspond to multiple topologies. Word-length of the devices is chosen for realization of certain problem – thus surface of chip and correspondingly power supply are used more efficiently (table 2) [2-4].

Table 2

Main advantages and disadvantages of neurochips

Advantages	Disadvantages
1) Specialized devices, concentrated on execution of only one task (relatively higher speed than in CPU); 2) easier realization of connections "all-to-all" for the developer of neural network (user of the device); 3) low energy consumption; 4) relatively low price (approximately 50$); 5) data load rate from the memory is 12.5% greater (relatively data indicated for CPU) for the best neurochip (information available for the year 2005).	1) Considerable structural complexity and low-reliability of systems; 2) high complexity of efficient realization of teaching procedure, self-education, self-organization of integrated circuits on formal neurons for weights of neuron interaction, which constantly change; 3) "Tyranny of interconnections" in neurochips and neuroplates, when connection "all-to-all" is realized, it is a problem at the stage of device design; 4) considerable increase of consumed power and decrease of performance in case of growth of neurochips integration level; 5) strictly set topology (several topologies); 6) outdated technological process of circuits manufacturing from silicon – as compared to manufacturers of CPU, videochips and DSP-processors.

Selection of hardware platform for processing of large-scale neural and neural-like parallel-hierarchical networks

On the whole, application of videoadaptor for a general-purpose computation on graphic processing units (GPGPU) does not differ greatly from CPU emulation. But still there is considerable difference – program, which uses videoadaptor for maximum efficiency (utilization of hardware resources) must be parallel relatively data or tasks (so-called Data Parallelism and Task Parallelism). Principle block of program computation is compiled into DirectX 9 or 10 byte-code, or into corresponding ATI CTM IL byte-code. Such byte-code is translated into special machine-code (so-called device-specific assembler) prior to execution. Let us consider hardware facilities: modern serial videoadaptors by their theoretical operating speed exceed modern processors 1—20 times, number of loadings from memory is considerably greater, this is explained by larger bus width, and higher clock frequency of the memory. Videoadaptors, unlike neurochips, are serial products (what is more — they are products of great demand), that is why they are manufactured in accordance with up-to-date technical requirements and are widely available [6, 8].

Proceeding from the analysis carried out, we can state, that among the competing hardware platforms videoadaptors are the most suitable for practice application. Now we will consider the available solutions (in particular, devices, manufactured by the companies "NVidia" and "ATI"), we will compare the most powerful videocards, manufactured by these companies, in accordance with the following criteria:

Table 3

Criterion	NVidia	ATI
Maximal theoretical performance	1 Tflops	1,2 Tflops
Carrying capacity of the memory	141,7 GB/s	115,2 GB/s
Price	520 USD*	320 USD*
Specific performance	1,92 Gflops/USD	3,75 Gflops/USD
Specific carrying capacity of the memory	0,27 GB/s/USD	0,36 GB/s/USD

*Note: average price available at www.hotline.ua 12.10.2008.

Taking into account specific price of performance, ATI adaptors are the most optimal solution for general-purpose computations.

Analysis of programming platforms for GPGPU

We can distinguish the following programming platforms, intended for realization for large-scale neural and neural-like parallel-hierarchical networks based on GPGPU technologies: assembler (ATI CTM IL), shader languages (GLSL – OpenGL 2.0, HLSL – DirectX 9.0c+), and high-level languages (NVidia CUDA, RapidMind, Brook/Brook+) [8 – 13].

Table 4

Comparative characteristics of GPGPU programming platforms

Possibilities	ATI CTM IL	GLSL/HLSL	NVidia CUDA	RapidMind	Brook/Brook+
Random memory read	+	+	+	+	+
Random memory write	+	–	+	+	– / +
Precision	64 bit	32 bit	64 bit (CUDA 2.0)	32 bit	64 bit
License	Freeware	Freeware	Freeware	Shareware (demo version unavailable)	Open source
Support of videoadaptors	ATI (2XXX+)	Any OpenGL 2.0 – compatible (GLSL); any DirectX 9.0c – compatible (HLSL)	NVidia (8XXX+)	Any DirectX 10-compatible	Any DirectX 9.0c or OpenGL 2.0-compatible (Brook) / ATI (2XXX+ series) (Brook+)
Possibility of low-level optimization	+	–	–	–	– / +
Does not require runtime environment	+	–	+	–	–

Development of programming library for construction and modeling of artificial neural and neural-like network topologies

Developed programming library "NN-Constructor" is intended for construction of artificial neural and neural-like network topologies (namely, parallel-hierarchical and hierarchic-hierarchical) and their simulation modeling. "NN-Constructor" realizes function of loading/storage of corresponding description of network topology in text files of special format, as well as function of teaching and processing (signal carrying) in neural or neural-like network. It should be noted,

that suggested programming library realizes the possibilities of modeling of such classes of neural of neural-like networks as feedforward networks and recurrent networks having the possibility of setting custom structure of the network for by the user. Construction of neural-like network is carried out by means of interconnecting layers of neural elements. The layer contains random number of neural elements; the number of layers is not limited (dynamic list is used).

In programming realization the principle of neural-like processing of data is selected; in accordance with this principle, the pulse is sent from neurons of the layers, belonging to processing step i, to neurons of the layers, belonging to processing step *i+1*. Thus, each value of input signal Ij can be calculated simultaneously, i.e. parallel. Principle of timing processing of neural networks is explained in the following figure:

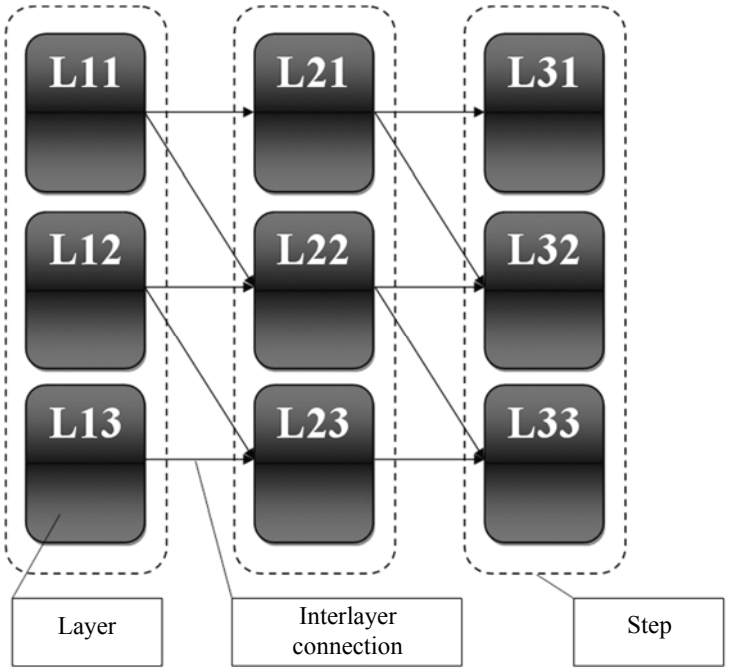

Figure 1. Generalized scheme of timing processing of neural networks in "NN-Constructor"

Language of CPU-version of programming library realization is C# for "MS .NET 2.0" platform. Library functions operate correctly under various operation systems: MS Windows XP (if "MS .NET 2.0" is installed), MS Windows Vista, Linux (if platform "Mono" is installed).

Language of realization of GPU-version of programming library is C++ with programming platform AMD Stream Computing SDK. Functions of the library operate correctly with various operation systems for videoadaptors ATI Radeon HD (series 2000 and higher). Algorithm of pulse propagation processing between steps (Fig. 1), taking into account specific features of parallel device programming – i.e., in order to avoid realization parallel threads synchronization – requires transformation of data format, which is performed in the following way:

1. For each neuron 1D table is constructed, each element of the table is the structure "number of connected neuron in previous layer – weight of interneural connection".
2. Thus, for each layer of the current step the set of tables is obtained according to the number of neurons if the layer, which characterize interneuron connections.
3. Besides, for each layers, additional 1D table is constructed, which contains levels of activation of the neurons of the given layer.

Such transformation allows saving data, needed for pulse propagation between steps, in a single array and load them into the memory of videocard during one step of data transmission. Such realization allows avoiding the usage of operation of random memory writing, which is not supported by videocards of R670 series, and realization of synchronization mechanism between

parallel threads.

Operation with "NN-Constructor" occurs within the limits of such main stages: loading from the file (or creation by user by means of corresponding functions) of a number of layers, connections between them, number of neural elements in the layer, connections between neural elements of different layers of neural or neural-like network; input information processing; network teaching; saving of network topology and modeling results.

Experimental research of simulation modeling of artificial neural and neural-like networks of problems, dealing with forecast of statistic series of currency exchange rates

In research, carried out real statistic series was used; it was obtained from open sources of Forex market, which represents hourly dynamics of euro-dollar exchange rate, entry size being 4137 (12.10.2008). The task of the experiment was to obtain forecast value of exchange rate with forecasting horizon – 1 step [4].

For forecasting of the given problem several structures of neural networks topologies were chosen, for instance, Word network with the structure 100-100-100-1, 100-25-25-1, 9-8-5-1 and multilayer perceptron with different variants of topologies. As test example, neural network – multilayer perceptron having topology 8-3-1 and method of teaching – error backpropagation was chosen. The given forecasting problem was realized by means of the developed library for constructing and modeling of artificial neural and neural-like networks topologies "NN-Constructor" (with possibilities of CPU and GPU processing).

Fig. 2 shows the results of teaching of the given network, realized by means of neuroconstructor "NN-Constructor". As it is seen in the figure, as a result of teaching neural network correctly restores a dynamics of exchange rate values, mean error of forecast if 0.004476721, that is quite acceptable for the given economic problem. Forecasting step in the program was defined in the following way: from the input series 8 elements were chosen, and one element of the output values series (forecast for the 9^{th} element of input series, forecasting horizon was 1).

The data processing rate in neural network was defined, it equals the sum of network teaching speed and testing speed. For the suggested variant of data processing speed in neural network was 14 sec.

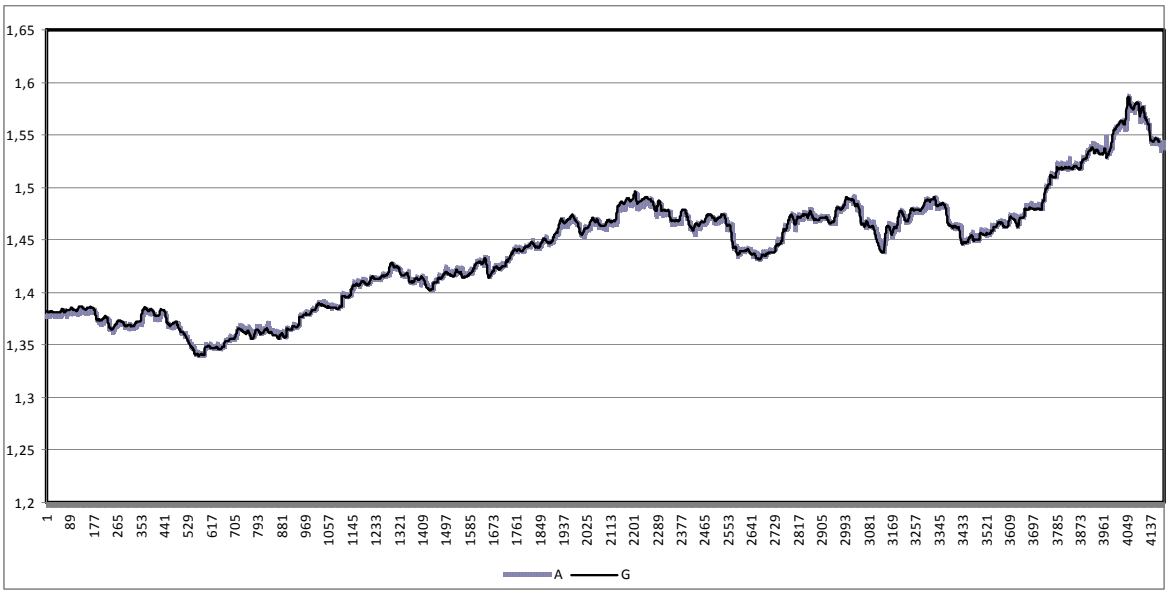

Fig. 2. Results of exchange rate forecast using "NN-Constructor",
series A – original series, series G – forecast series

For proving the adequacy of operation of the given software product and validity of the results obtained, computer modeling was performed in one for professional and known in the sphere

of neural network processing of software products – Statistica Neural Network (SNN), company StatSoft [15].

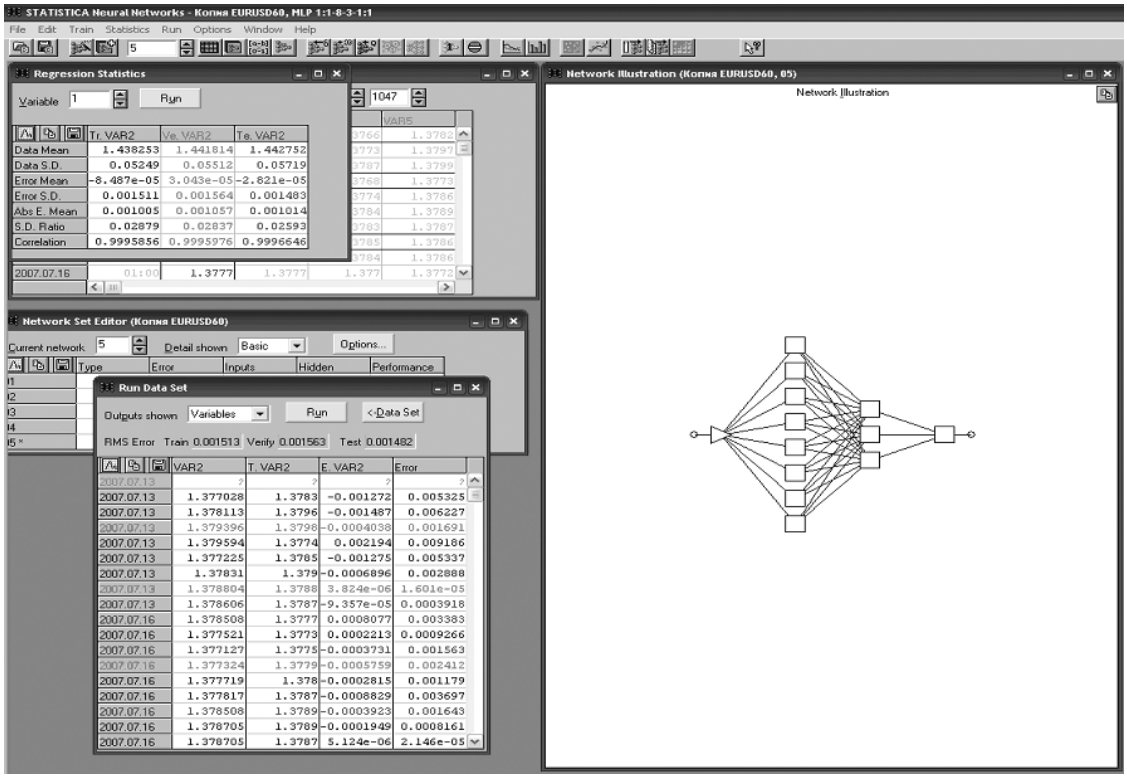

Fig. 3. Screen forms with the results of computer modeling
in software package Statistica Neural Network

In particular, Fig. 3 shows the results of computer modeling of the given neural network (multilayer perceptron with topology 8-3-1 and teaching method – error backpropagation).

As a result of teaching neural network correctly restores dynamics of exchange rate values, mean error of forecast is 0.00127200.

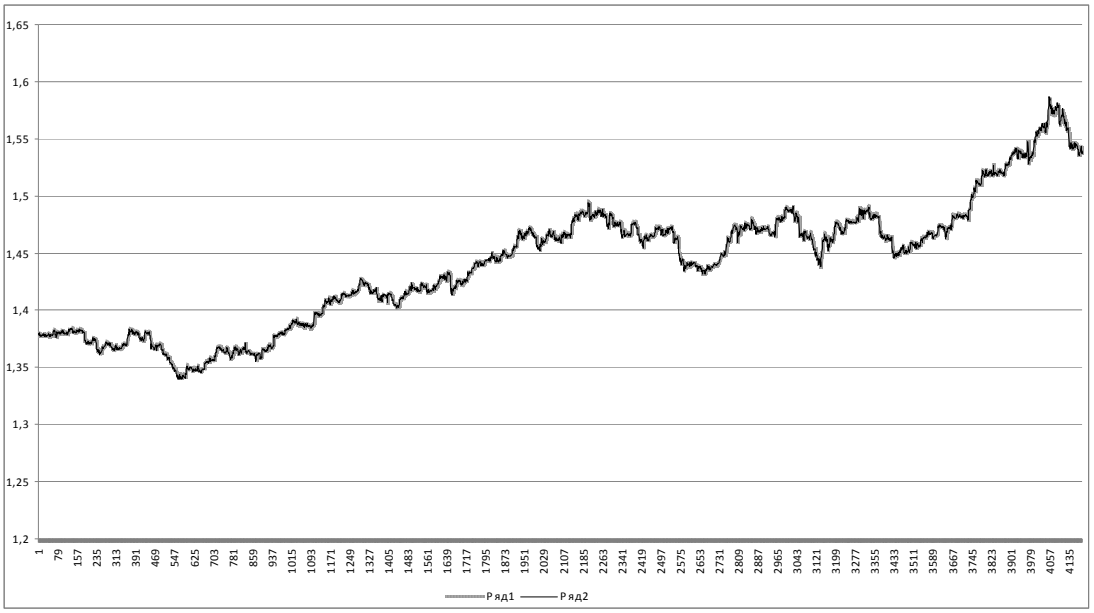

Fig. 4. Results of exchange rate forecast in software package Statistica Neural Network,
series 1 – original series, series 2 – forecast series

Statistica Neural Network for the suggested method also evaluated the speed of data processing in neural network, it was considerably greater (more than 10 min) that in the variant of realization using “NN-Constructor”.

Conclusions

The paper studied and analyzed basic technologies of hardware-software realization of artificial neural networks, in particular, modern specialized neuroprocessors, digital signal processors (DSP), multimedia central processors (CPU) and alternative modern hardware facilities (namely GPU) in order to soubstantiate the choice of basic platform for modeling of various structures of neural and neural-like parallel-hierarchical networks.

Research carried out were performed with the aim of further development of neuroemulator – the system, constructed on the base of series cascade-connected universal SISD, SIMD or MISD processor; the system performs typical neural operations (weighted summation and non-linear transformation) on software level. The paper suggests selecting GPGPU technology as neuroaccelerator as a hardware platform for realization of large-scale neural and neural-like parallel-hierarchical networks. The given technology is based on application of powerful videoadaptor for carrying out specialized, including parallel, computations. Since modern technologies of videoadaptor construction allow using 128-core special processors, as compared with existing quad-core multimedia CPU, their application for neuroemulation of various opologies of large-scale neural and neural-like parallel-hierarchical networks is actual and perspective [6]. Within the frame of software realization, the process of neuropackage intended for realization of different topologies of neural and neural-like parallel-hierarchical networks and their possible computation on GPU is under way. In particular, programming library, which realizes processes of neural network processing and visual editor of neural and neural-like parallel-hierarchical networks is suggested. On the base of solution of testing problem dealing with forecast of economic information the adequacy and efficiency of program product have been checked and proved.

REFERENCES

1. Methodological Principles of Pyramidal and Parallel-Hierarchical Image Processing on the Base of Neural-Like Network Systems / V. Kozhemyako, L. Timchenko, A. Yarovyy // Advances in Electrical and Computer Engineering – Romania, “Stefan cel Mare” University of Suceava. – Volume 8 (15), Number 2 (30). – 2008. – PP. 54 - 60. – ISSN 1582-7445.

2. Воеводин В. В. Параллельные вычисления : учебн. пособие [для студ. высш. учебн. зав.] / В. В. Воеводин, В. В.Воеводин. – СПб.: БХВ-Петербург, 2002. – 608 с. – ISBN 5-94157-160-7.

3. Круг П. Г. Нейронные сети и нейрокомпьютеры : учебн. пособие [для студ. высш. учебн. зав. по курсу «Микропроцессоры»] / Круг П. Г. – М.: Издательство МЭИ, 2002. – 176 с. – ISBN 5-7046-0832-9.

4. Корнеев В., Киселев А. Современные микропроцессоры. – 3 издание : учебн. пособие [для студ. высш. учебн. зав.] / В. Корнеев, А. Киселев. – СПб.: БХВ-Петербург, 2003. – 448 с. – ISBN 5-94157-385-5.

5. Кожем’яко В. П. Паралельно-ієрархічні мережі як структурно-функціональний базис для побудови спеціалізованих моделей образного комп’ютера : [Монографія.] / В. П. Кожем’яко, Л. І. Тимченко, А. А. Яровий. – Вінниця: Універсум-Вінниця, 2005. – 161 с. – ISBN 966-641-142-3.

6. Вибір апаратної платформи для реалізації масштабних нейронних та нейроподібних паралельно-ієрархічних мереж [Електронний ресурс] : IX Міжнародна конференція Контроль і управління в складних системах (КУСС-2008), Вінниця, 21-24 жовтня 2008 року / А. А. Яровий, Ю. С. Богомолов, К. Ю. Вознесенский. – Режим доступу: http://www.vstu.vinnica.ua/ mccs2008/materials/subsection_2.2.pdf.

7. Сравнение производительности графических ускорителей и центрального процессора при вычислениях для больших объемов обрабатываемых данных / Скрибцов П. В., Долгополов А. В. // Нейрокомпьютеры: разработка, применение – М.: Радиотехника, 2007. – № 9. – С. 421 - 425. – ISSN 0869-5350.

8. GPGPU: General Purpose computations on Graphic Processing Unit [Електронний ресурс] – Режим доступу: http://www.gpgpu.org.

9. OpenCL: Open Computing Language – [Електронний ресурс] – Режим доступу: http://en.wikipedia.org/wiki/OpenCL.

10. AMD/ATI StreamComputing SDK – [Електронний ресурс] – Режим доступу: http://ati.amd.com/technology/streamcomputing/index.html.

11. NVidia CUDA – [Електронний ресурс] – Режим доступу: http://www.nvidia.com/object/cuda_home.html.

12. RapidMind – [Електронний ресурс] – Режим доступу: http://www.rapidmind.net.
13. Объектно-ориентированный подход к шейдерам – [Електронний ресурс] – Режим доступу: http://www.dtf.ru/articles/read.php?id=47296 &DTFSESSID=fc58ce864752390b052fd34c3fc1f000.
14. Форекс Украина – [Електронний ресурс] – Режим доступу: www.forexua.com
15. STATISTICA Neural Networks. Техническое описание. – [Електронний ресурс] – Режим доступу: http://www.statsoft.ru/statportal/tabID__32/MId__141/ ModeID__0/PageID__11/DesktopDefault.aspx.

Yarovyy Andriy – Assistant Prof. of the Department of Intelligence Systems.

Bogomolov Yuriy – Student of the Department of Intelligence Systems.

Voznesenskiy Kostiantyn – Student of the Department of Intelligence Systems.
Vinnytsia National Technical University.