УДК 621.398:007

А. А. Яровой, к. т. н., доц.; Ю. С. Богомолов; К. Ю. Вознесенский

ПРИКЛАДНАЯ РЕАЛИЗАЦИЯ МАСШТАБНЫХ НЕЙРОННЫХ И НЕЙРОПОДОБНЫХ ПАРАЛЛЕЛЬНО-ИЕРАРХИЧЕСКИХ СЕТЕЙ НА ОСНОВЕ ТЕХНОЛОГИЙ GPGPU

В контексте исследования проблемы программно-аппаратной реализации масштабных нейронных и нейроподобных параллельно-иерархических сетей обосновано выбор аппаратной платформы для последующего имитационного моделирования и практически-прикладной реализации. На основе проведенных исследований и полученных результатов предложены программные модули для реализации на CPU и GPU масштабных нейронных и нейроподобных параллельно-иерархических сетей различных топологий.

Ключевые слова: *параллельные вычисления, нейронные сети, цифровая обработка информации, параллельно-иерархические системы, прогнозирование.*

Введение

Стремительный переход современных систем управления на цифровые стандарты привел к необходимости обрабатывать с высокой скоростью сверхбольшие объемы информации. Актуальность данных исследований и полученных результатов наиболее характерна для систем, в которых необходимо осуществлять сложную обработку и фильтрацию сигналов, например, распаковку сжатых аудио- и видеоданных, маршрутизацию информационных потоков, прогнозирование динамических быстросменных данных, которое требует применения достаточно производительных интеллектуальных вычислительных систем. Подобные системы могут быть реализованы на разнообразной элементной базе, но наибольшего распространения на современном этапе получили параллельные нейроподобные сетевые устройства.

Основной целью исследований является вариантный анализ и выбор наиболее оптимальной аппаратной платформы моделирования масштабных нейронных и нейроподобных параллельно-иерархических сетей с последующей разработкой программного обеспечения для эмуляции параллельных и параллельно-иерархических вычислений при решении сверхсложных задач цифровой обработки информации, в частности распознавания образов, динамической обработки изображений, прогнозирования и тому подобное.

С целью решения указанных проблем в работе исследуются и анализируются основные технологии аппаратной реализации искусственных нейронных сетей, в частности, современные специализированные нейропроцессоры, ПЛИС, цифровые сигнальные процессоры (DSP), мультимедийные центральные процессоры (CPU) и альтернативные современные аппаратные средства (в частности, GPU), в контексте обоснования выбора базовой платформы для моделирования разнообразных структур масштабных нейронных и нейроподобных параллельно-иерархических сетей [1 – 5].

Анализ аппаратных платформ моделирования масштабных нейронных и нейроподобных параллельно-иерархических сетей

В качестве вариантов аппаратных платформ рассматривались наиболее распространенные схемы – центральные процессоры (CPU), видеоадаптеры и специализированые нейропроцессоры (нейрочипы), поскольку их использование позволяет абстрагироваться от уровня проектирования аппаратной платформы. Кроме отмеченных средств существуют также другие пути решения поставленной задачи – например, собственноручное

изготовление схемы на основе DSP-процессоров, но среди недостатков данного решения есть привязанность к определенной топологии сети [2 – 6].

Программная эмуляция на CPU

Один из распространенных методов коммерческой реализации нейронных сетей заключается в создании топологии виртуально – в виде набора матриц весов и уровней активации. При этом топология может иметь практически любую размерность (размеры сети ограничены имеющимся объемом свободной оперативной памяти). Максимальный объем памяти для домашнего компьютера – 8 Гб, что позволяет создать массив типа double (число с плавающей запятой двойной точности, размер 8 байт) размером 32768×32768. Для серверов максимальный объем оперативной памяти – 32 Гб, соответственно, размеры данного массива могут быть вдвое большими (65536×65536) [2,4,7]. Данные подсчеты немного относительны, поскольку мы не учитывали объем памяти, которую занимает операционная система и, собственно, программа пользователя.

Таблица 1

Основные преимущества и недостатки программной эмуляции на CPU

Преимущества	Недостатки
1) Широкая распространенность и доступность аппаратной платформы; 2) гибкость при моделировании – возможность реализации любой топологии, используя любой язык программирования; 3) высокая точность результата при выполнении вычислений (до 128 бит); 4) высокая пропускная способность памяти (до 12,8 Гб / с в синтетическом тесте при использовании памяти типа DDR3); 5) большой доступный объем памяти (до 8 Гб для домашнего компьютера и до 32 Гб / процессор для сервера).	1) Меньшее быстродействие на реальных задачах, чем у специализированных нейрочипов или видеоадаптеров; 2) относительно меньшее количество загрузок данных из памяти, чем у нейрочипов и видеоадаптеров.

Нейрочипы

Реализованная на кристалле структура из многих ядер со связями между ядрами, которые отвечают определенной заранее заданной топологии или могут отвечать нескольким топологиям. Разрядность устройств подбирается для определенной задачи, таким образом, площадь чипа и, соответственно, электроэнергия для энергопитания используются более эффективно (табл. 2) [2 – 4].

Таблица 2

Основные преимущества и недостатки нейрочипов

Преимущества	Недостатки
1) Специализированные устройства, которые сосредоточены на выполнении лишь одной задачи (относительно большее быстродействие, чем у CPU); 2) облегчена реализация связей «все-со-всеми» для разработчика нейросетей (пользователя устройства); 3) низкое потребление электроэнергии; 4) относительно доступная цена (ориентировочно 50$); 5) на 12,5% больше загрузок данных из памяти (относительно указанной для CPU) для наилучшего нейрочипа (состоянием на 2005 год).	1) Большая структурная сложность и низкая надежность систем; 2) большая сложность эффективной реализации процедуры обучения, самообучения, самоорганизации интегральных схем из формальных нейронов для весов взаимодействия нейронов, которые постоянно изменяются; 3) «тирания межсоединений» в нейрочипах и нейропластинах, когда реализуется связь «все-со-всеми», что является проблемой на этапе проектирования устройства; 4) значительное увеличение потребляемой мощности и потеря быстродействия при увеличении степени интеграции нейрочипов; 5) жестко заданная заранее топология (несколько топологий); 6) отсталый технологический процесс изготовления схем в кремнии – относительно производителей CPU, видеоадаптеров и DSP-процессоров.

Выбор аппаратной платформы для обработки масштабных нейронных и нейроподобных параллельно-иерархических сетей

В целом, использование видеоадаптера для вычислений общего назначения (General-Purpose computation on Graphic Processing Units –GPGPU) мало чем отличаются от эмуляции на CPU. Однако есть существенная разница – программа, которая использует видеоадаптер для максимальной эффективности (утилизации аппаратных ресурсов), должна быть параллельная относительно данных или задач (так называемые Data Parallelism и Task Parallelism). При этом основной блок вычислений программы компилируется в байт-код DIRECTX 9 или 10, или в соответствующий байт-код ATI CTM IL. Такой байт-код транслируется в специальный машинный код (так называемый device-specific assembler) перед выполнением. Рассмотрим аппаратную базу: современные массовые видеоадаптеры по своему теоретическому быстродействию превышают современные процессоры в 10 – 20 раз, количество загрузок из памяти значительно больше, что объясняется большей шириной шины и более высокой тактовой частотой памяти. Видеоадаптеры, в отличие от нейрочипов, являются массовым продуктом (более того – продуктом большого спроса), а потому они изготовляются по актуальному техническому процессу и являются широкодоступными [6, 8].

Из приведенного анализа конкурирующих аппаратных платформ самыми оптимальными для практического использования являются видеоадаптеры. Рассмотрим конкурирующие решения, в частности, продукцию компаний «NVidia» и «ATI», и сравним мощнейшие видеоадаптеры отмеченных компаний по таким критериям:

Таблица 3

Критерий	NVidia	ATI
Максимальное теоретическое быстродействие	1 Tflops	1,2 Tflops
Пропускная способность памяти	141,7 GB/s	115,2 GB/s
Цена*	520 USD*	320 USD*
Удельное быстродействие	1,92 Gflops/USD	3,75 Gflops/USD
Удельная пропускная способность памяти	0,27 GB/s/USD	0,36 GB/s/USD

*Примечание: средняя цена по данным сайта www.hotline.ua на 12.10.2008.

Учитывая удельную стоимость быстродействия, видеоадаптеры ATI являются наиболее оптимальным решением для вычислений общего характера.

Анализ программных платформ для GPGPU

В качестве программных платформ для реализации масштабных нейронных и нейроподобных параллельно-иерархических сетей на основе технологий GPGPU широко применяются и могут быть выделены такие: ассемблер (ATI CTM IL), шейдерные языки (GLSL-OpenGL 2.0, HLSL-DirectX 9.0c+), высокоуровневые языки (NVidia CUDA, RapidMind, Brook/Brook+) [8 – 13].

Таблица 4

Сравнительная характеристика программных платформ GPGPU

Возможности	ATI CTM IL	GLSL/ HLSL	NVidia CUDA	RapidMind	Brook/Brook+
Произвольное считывание из памяти	+	+	+	+	+
Произвольная запись в память	+	–	+	+	– / +
Разрядность	64 bit	32 bit	64 bit (CUDA 2.0)	32 bit	64 bit

Продолжение табл. 4

Лицензия	Freeware	Freeware	Freeware	Shareware (демоверсия отсутствует)	Open source
Поддержка видео-адаптеров	ATI (2XXX+)	Любой OPENGL 2.0 – совместимый (GLSL); любой DIRECTX 9.0c – совместимый (HLSL)	NVidia (8XXX+)	Любой DIRECTX 10 – совместимый	Любой DIRECTX 9.0c или OPENGL 2.0 – совместимый (Brook) / ATI (серия 2XXX+) (Brook+)
Возможность низкоуровневой оптимизации	+	–	–	–	– / +
Не требует среды исполнения	+	–	+	–	–

Разработка программной библиотеки для конструирования и моделирования топологии искусственных нейронных и нейроподобных сетей

Разработана программная библиотека «NN-Constructor», предназначенная для конструирования топологий искусственных нейронных и нейроподобных сетей (в частности, параллельно-иерархических и иерарх-иерархических) и их имитационного моделирования. «NN-Constructor» реализует функции загрузки / сохранения соответствующего описания топологии сети в текстовых файлах специального формата, а также функции для обучения и обработки (распространение сигнала) в нейронной или нейроподобной сети. Необходимо отметить, что в предложенной программной библиотеке реализованы возможности моделирования таких классов топологий нейронных сетей, как сети прямого распространения (Feedforward) и рекуррентные сети с возможностью задания пользователем произвольной структуры сети.

В программной реализации избран принцип нейроподобной обработки данных, который заключается в передаче импульса от нейронов слоев, которые принадлежат к такту обработки i, к нейронам слоев, которые принадлежат к такту обработки $i+1$. Таким образом, каждое значение входного сигнала I_j может быть рассчитано одновременно, то есть параллельно. Принцип разделенной на такты обработки нейронных сетей объясняет следующий рисунок:

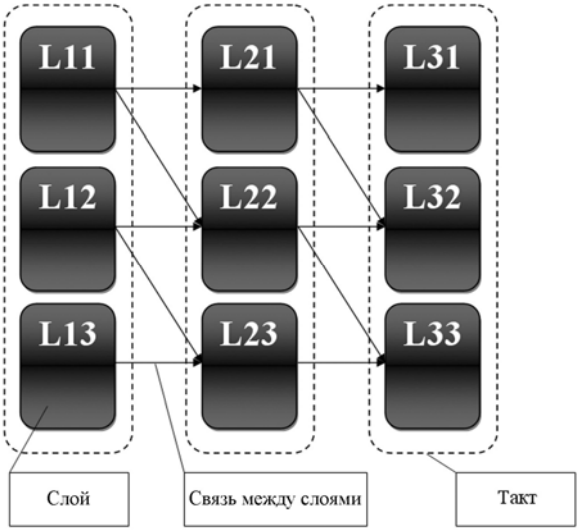

Рис. 1. Обобщена схема процесса разделенной на такты обработки нейронных сетей в «NN-Constructor»

Язык реализации CPU-версии программной библиотеки – C# для платформы «MS .NET 2.0». Функции библиотеки корректно работают с разными операционными системами: MS Windows XP (при условии наличия установленного «MS .NET 2.0»), MS Windows Vista, Linux (при условии установленной платформы «Mono»).

Язык реализации GPU-версии программной библиотеки – C++ с использованием программной платформы AMD StreamComputing SDK. Функции библиотеки корректно работают с разными операционными системами для видеоадаптеров ATI Radeon HD (серии 2000 и выше). Алгоритм обработки передачи импульса между тактами (рис. 1) с учетом специфики программирования параллельных устройств, то есть с целью избежания реализации механизма синхронизации параллельных потоков, требует преобразования формата данных, которое происходит таким образом:

1. Для каждого нейрона строится одномерная таблица, каждый элемент которой является структурой «номер связанного нейрона в предыдущем слое – вес межнейронной связи».
2. Таким образом, для каждого слоя текущего такта получается набор таблиц по количеству нейронов в слое, которые характеризуют межнейронные связи.
3. Кроме того, для каждого слоя строится дополнительная одномерная таблица, которая содержит в себе уровни активации нейронов данного слоя.

Такое преобразование позволяет хранить данные, необходимые для передачи импульса между тактами, в единственном массиве и загружать их в память видеоадаптера за один цикл передачи данных. Такая реализация позволяет избежать использования операции произвольной записи в память, которая не поддерживается видеоадаптерами младше серии R670, и реализации механизма синхронизации между параллельными потоками.

Работа с «NN-Constructor» происходит в рамках таких основных этапов: загрузка из файла (или создание пользователем с помощью соответствующих функций) количества слоев, связей между ними, количества нейронных элементов в слое, связей между нейронными элементами разных слоев нейронной или нейроподобной сети; обработка входной информации; обучение сети; сохранение топологии сети и результатов моделирования.

Экспериментальные результаты имитационного моделирования искусственных нейронных и нейроподобных сетей для задач прогнозирования статистических рядов курсов валют

В проведенных исследованиях использовался реальный статистический ряд, полученный из открытых источников рынка Forex, который отображает почасовую динамику изменения курса евро-доллар, размерностью 4137 записей (12. 10. 2008 г.). Задачей эксперимента было получение прогнозируемого значения изменения курса с горизонтом прогнозирования 1 шаг [14].

Для прогнозирования указанной задачи было избрано несколько структур топологий нейронных сетей, в частности, сеть Ворда со структурой 100-100-100-1, 100-25-25-1, 9-8-5-1, а также многослойный персептрон с разнообразными вариантами топологии. В качестве тестового примера экспериментально была избрана нейронная сеть – многослойный персептрон с топологией 8-3-1 и методом обучения обратного распространения ошибки. Поставленная задача прогнозирования была реализована с помощью разработанной программной библиотеки для конструирования и моделирования топологии искусственных нейронных и нейроподобных сетей «NN-Constructor» (с возможностями обработки на CPU и GPU).

В частности, на рис. 2 представлены результаты обучения указанной нейронной сети, которая реализована с использованием нейроконструктора «NN-Constructor». Как видно из рисунка, в результате обучения нейронная сеть корректно воспроизводит динамику изменения значений курса, в частности, средняя погрешность прогнозирования составляет 0,004476721, что является приемлемым для поставленной экономической задачи. Шаг

прогнозирования в программе определялся таким образом: из входного ряда было избрано 8 элементов и один элемент ряда исходных значений (прогноз на 9 элемент входного ряда, поскольку горизонт прогнозирования избран 1).

Также была определена скорость обработки данных в нейронной сети, которая равна сумме скорости обучения сети и скорости тестирования. Для предложенного варианта скорость обработки данных в нейронной сети составила 14 секунд.

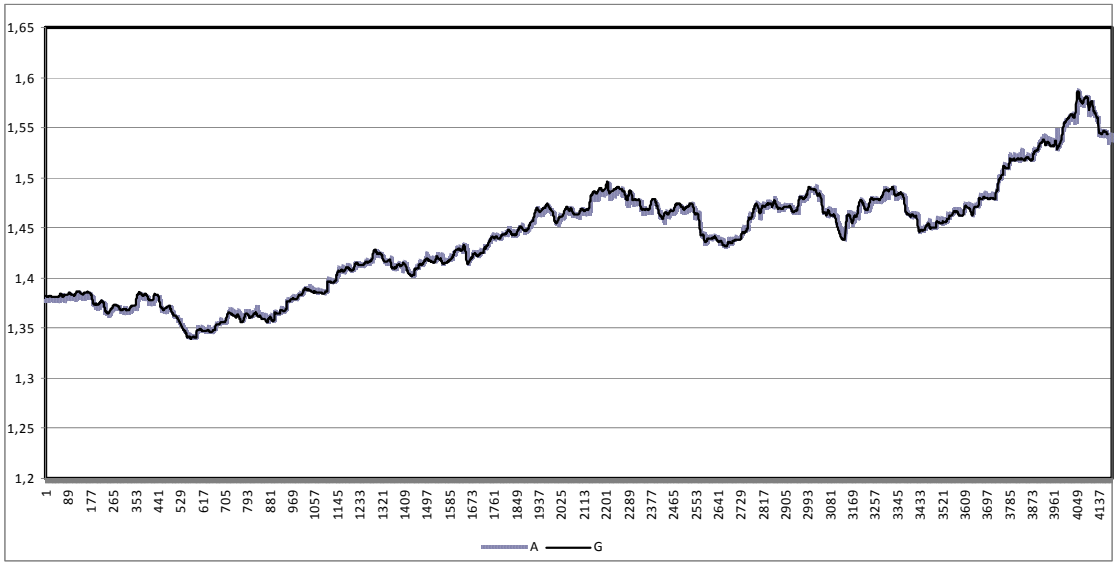

Рис. 2. Результаты прогнозирования курсов валют с использованием «NN-Constructor», где ряд A – оригинальный ряд, ряд G – ряд прогноза

Для подтверждения адекватности работы предложенного программного продукта и корректности полученных результатов также было осуществлено компьютерное моделирование в одном из профессиональных и признанных в отрасли нейросетевой обработки программных пакетов – Statistica Neural Network (SNN), компании StatSoft [15].

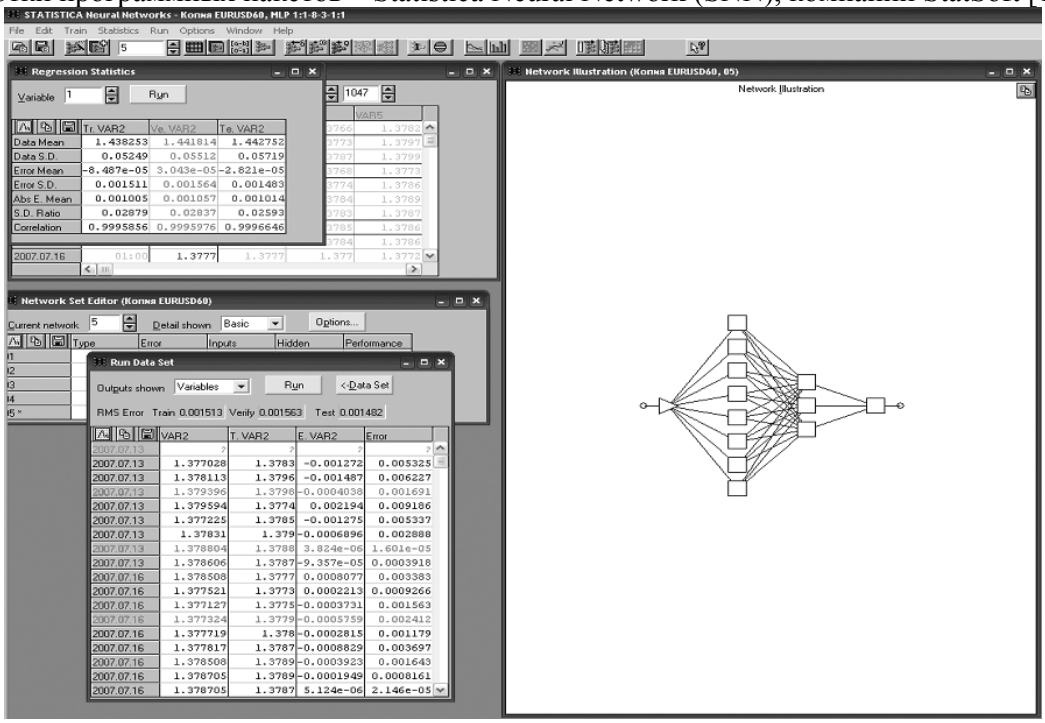

Рис. 3. Экранные формы с результатами компьютерного моделирования в программном пакете Statistica Neural Network

В частности, на рис. 3 представлены результаты компьютерного моделирования указанной нейронной сети (многослойный персептрон с топологией 8-3-1 и методом обучения обратным распространением ошибки) в программном пакете Statistica Neural Network.

В результате обучения нейронная сеть корректно воспроизводит динамику изменения значений курса, в частности, средняя погрешность прогнозирования составляет 0,00127200.

Рис. 4. Результаты прогнозирования курсов валют
в программном пакете Statistica Neural Network
где ряд 1 – оригинальный ряд, ряд 2 – ряд прогноза

Для Statistica Neural Network для предложенного варианта также была оценена скорость обработки данных в нейронной сети, которая была значительно больше (больше 10 минут), чем в варианте реализации с использованием «NN-Constructor».

Выводы

В работе было исследовано и проанализировано основные технологии аппаратно-программной реализации искусственных нейронных сетей, в частности современные специализированые нейропроцессоры, цифровые сигнальные процессоры (DSP), мультимедийные центральные процессоры (CPU) и альтернативные современные аппаратные средства (в частности, GPU), в контексте обоснования выбора базовой платформы для моделирования разнообразных структур масштабных нейронных и нейроподобных параллельно-иерархических сетей.

Проведенные научные исследования осуществлялись в контексте последующей разработки нейроэмулятора – системы, построенной на базе каскадного соединения универсальных SISD-, SIMD- или MISD-процессоров, которая реализует типичные нейрооперации (взвешенное суммирование и нелинейное преобразование) на программном уровне. В работе предложено, в качестве нейроускорителя, как аппаратной платформы для реализации масштабных нейронных и нейроподобных параллельно-иерархических сетей, избрать технологию GPGPU, которая базируется на использовании мощного видеоадаптера для выполнения специализированных, в том числе параллельных, вычислений. Поскольку современные технологии построения видеоадаптеров позволяют использовать 128-ядерные спецпроцессоры, в сравнении с современными 4-ядерными мультимедийными CPU, то

применение их для нейроэмуляции различных топологий масштабных нейронных и нейроподобных параллельно-иерархических сетей является актуальным и перспективным [6]. В контексте программной реализации проведена работа по созданию нейропакета для реализации различных топологий масштабных нейронных и нейроподобных параллельно-иерархических сетей и возможности просчета их на GPU. В частности, предложена программная библиотека, которая непосредственно выполняет процессы обработки нейросети, а также визуальный редактор топологии нейронной и нейроподобных параллельно-иерархических сетей. На основе решения тестовой задачи прогнозирования экономической информации экспериментально проверено и доказано адекватность и эффективность программной разработки.

СПИСОК ЛИТЕРАТУРЫ

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.

Яровой Андрей Анатолиевич – к.т.н., доцент, доцент кафедры интеллектуальных систем.

Богомолов Юрий Сергеевич – студент кафедры интеллектуальных систем.

Вознесенский Константин Юрьевич – студент кафедры интеллектуальных систем.
Винницкий национальный технический университет.