УДК 621.398:007

А. А. Яровий, к. т. н., доц.; Ю. С. Богомолов; К. Ю. Вознесенський

ПРИКЛАДНА РЕАЛІЗАЦІЯ МАСШТАБНИХ НЕЙРОННИХ ТА НЕЙРОПОДІБНИХ ПАРАЛЕЛЬНО-ІЄРАРХІЧНИХ МЕРЕЖ НА ОСНОВІ ТЕХНОЛОГІЙ GPGPU

У контексті дослідження проблеми програмно-апаратної реалізації масштабних нейронних та нейроподібних паралельно-ієрархічних мереж обґрунтовано вибір апаратної платформи для подальшого імітаційного моделювання та практично-прикладної реалізації. На основі проведених досліджень та отриманих результатів запропоновано програмні модулі для реалізації на CPU та GPU масштабних нейронних та нейроподібних паралельно-ієрархічних мереж різноманітної топології.

Ключові слова: *паралельні обчислення, нейронні мережі, цифрова обробка інформації, паралельно-ієрархічні системи, прогнозування.*

Вступ

Стрімкий перехід сучасних систем управління на цифрові стандарти привів до необхідності обробляти з високою швидкістю надвеликі обсяги інформації. Актуальність цих досліджень та отриманих результатів найбільш характерна для систем, у яких необхідно здійснювати складну обробку й фільтрацію сигналів, наприклад, розпакування стислих аудіо- та відеоданих, маршрутизацію інформаційних потоків, прогнозування динамічних швидкозмінних даних, яке вимагає застосування досить продуктивних інтелектуальних обчислювальних систем. Подібні системи можуть бути реалізовані на різноманітній елементній базі, але найбільшого поширення на сучасному етапі одержали паралельні нейроподібні мережні пристрої.

Основною метою досліджень є варіантний аналіз та вибір найбільш оптимальної апаратної платформи моделювання масштабних нейронних та нейропобідних паралельно-ієрархічних мереж з подальшою розробкою програмного забезпечення для емуляції паралельних та паралельно-ієрархічних обчислень під час розв'язання надскладних задач цифрової обробки інформації, зокрема, розпізнавання образів, динамічної обробки зображень, прогнозування тощо.

З метою вирішення вказаних проблем у роботі досліджуються та аналізуються основні технології апаратної реалізації штучних нейронних мереж, зокрема, сучасні спеціалізовані нейропроцесори, ПЛІС, цифрові сигнальні процесори (DSP), мультимедійні центральні процесори (CPU) та альтернативні сучасні апаратні засоби (зокрема, GPU), у контексті обґрунтування вибору базової платформи для моделювання різноманітних структур масштабних нейронних та нейропобідних паралельно-ієрархічних мереж [1 – 5].

Аналіз апаратних платформ моделювання масштабних нейронних та нейропобідних паралельно-ієрархічних мереж

Як варіанти апаратних платформ розглядались найбільш поширені схеми – центральні процесори (CPU), відеокарти та спеціалізовані нейропроцесори (нейрочіпи), оскільки їх використання дозволяє абстрагуватися від рівня проектування апаратної платформи. Окрім зазначених засобів є також інші шляхи розв'язання поставленої задачі – наприклад, власноручне виготовлення схеми на основі DSP-процесорів, але серед недоліків цього рішення є прив'язаність до певної топології мережі [2 – 6].

Програмна емуляція на CPU

Один із розповсюджених методів комерційної реалізації нейронних мереж полягає у створенні топології віртуально – у вигляді набору матриць ваг та рівнів активації. При цьому топологія може мати практично будь-яку розмірність (розміри мережі обмежені наявним обсягом вільної оперативної пам'яті). Максимальний обсяг пам'яті для домашнього комп'ютера – 8 Гб, що дозволяє створити масив типу double (число з плаваючою комою подвійної точності, розмір 8 байт) розміром 32768×32768. Для серверів максимальний обсяг оперативної пам'яті – 32 Гб, відповідно, розміри цього масиву можуть бути вдвічі більшими (65536×65536) [2, 4, 7]. Ці підрахунки є дещо відносними, оскільки ми не враховували обсяг пам'яті, яку займає операційна система та, власне, програма користувача.

Таблиця 1. Основні переваги та недоліки програмної емуляції на CPU

Переваги	Недоліки
1) Широка розповсюдженість та доступність апаратної платформи; 2) гнучкість під час моделювання – можливість реалізації будь-якої топології, використовуючи будь-яку мову програмування; 3) висока точність результату під час виконання обчислень (до 128 біт); 4) висока пропускна спроможність пам'яті (до 12,8 Гб / с у синтетичному тесті під час використання пам'яті типу DDR3); 5) великий доступний обсяг пам'яті (до 8 Гб для домашнього комп'ютера та до 32 Гб / процесор для сервера).	1) Нижча швидкодія на реальних задачах, ніж у спеціалізованих нейрочіпів чи відеоадаптерів; 2) відносно менша кількість завантажень даних з пам'яті, ніж у нейрочіпів та відеокарт.

Нейрочіпи

Реалізована на кристалі структура з багатьох ядер з між'ядерним зв'язками, які відповідають певній наперед заданій топології або можуть відповідати декільком. Розрядність пристроїв підбирається для певної задачі, таким чином, площа чіпа і, відповідно, електроенергія для енергоживлення використовуються більш ефективно (табл. 2) [2 – 4].

Таблиця 2

Основні переваги та недоліки нейрочіпів

Переваги	Недоліки
1) Спеціалізовані пристрої, які зосереджені на виконанні лише однієї задачі (відносно більша швидкодія, ніж в CPU); 2) полегшена реалізація зв'язків «всі-зі-всіма» для розробника нейромереж (користувача пристрою); 3) низьке споживання електроенергії; 4) відносно доступна ціна (орієнтовно 50$); 5) на 12,5% більше завантажень даних з пам'яті (відносно вказаної для CPU) для найкращого нейрочіпа (станом на 2005 рік).	1) Велика структурна складність і низька надійність систем; 2) велика складність ефективної реалізації процедури навчання, самонавчання, самоорганізації інтегральних схем із формальних нейронів для ваг взаємодії нейронів, які постійно змінюються; 3) «Тиранія міжз'єднань» у нейрочіпах та нейропластинах, коли реалізується зв'язок «всі-зі-всіма», що є проблемою на етапі проектування пристрою; 4) значне збільшення споживаної потужності та втрата швидкодії під час збільшення ступеню інтеграції нейрочіпів; 5) жорстко задана наперед топологія (декілька топологій); 6) відсталий технічний процес виготовлення схем у кремнії – відносно виробників CPU, відеочіпів та DSP-процесорів.

Вибір апаратної платформи для обробки масштабних нейронних та нейропобідних паралельно-ієрархічних мереж

У цілому, використання відеоадаптера для обчислень загального призначення (General-Purpose computation on Graphic Processing Units – GPGPU) мало чим відрізняються від емуляції на CPU. Проте є суттєва різниця – програма, яка використовує відеоадаптер для максимальної ефективності (утилізації апаратних ресурсів), повинна бути паралельна

відносно даних або задач (так звані Data Parallelism та Task Parallelism). За цих умов основний блок обчислень програми компілюється у байт-код DirectX 9 чи 10, або у відповідний байт-код ATI CTM IL. Такий байт-код транслюється у спеціальний машинний код (так званий device-specific assembler) перед виконанням. Розглянемо апаратну базу: сучасні масові відеоадаптери за своєю теоретичною швидкодією перевищують сучасні процесори у 10 – 20 разів, кількість завантажень із пам'яті значно більша, що пояснюється більшою шириною шини та більш високою тактовою частотою пам'яті. Відеоадаптери, на відміну від нейрочіпів, є масовим продуктом (більше того – продуктом великого попиту), а тому вони виготовляються за актуальним технічним процесом та є широкодоступними [6, 8].

З наведеного аналізу конкуруючих апаратних платформ найоптимальнішою для практичного використання є відеоадаптери. Розглянемо конкуруючі рішення, зокрема, продукцію компаній „NVidia" та „ATI" та порівняємо найпотужніші відеокарти означених компаній за такими критеріями:

Таблиця 3

Критерій	NVidia	ATI
Максимальна теоретична швидкодія	1 Tflops	1,2 Tflops
Пропускна спроможність пам'яті	141,7 GB/s	115,2 GB/s
Ціна*	520 USD*	320 USD*
Питома швидкодія	1,92 Gflops/USD	3,75 Gflops/USD
Питома пропускна спроможність пам'яті	0,27 GB/s/USD	0,36 GB/s/USD

*Примітка: середня ціна за даними сайту www.hotline.ua на 12.10.2008.

З огляду на питому вартість швидкодії відеоадаптери ATI є найбільш оптимальним рішенням для обчислень загального характеру.

Аналіз програмних платформ для GPGPU

Як програмні платформи для реалізації масштабних нейронних та нейроподібних паралельно-ієрархічних мереж на основі технологій GPGPU широко застосовуються та можуть бути виділені такі: асемблер (ATI CTM IL), шейдерні мови (GLSL-OpenGL 2.0, HLSL-DirectX 9.0c+), високорівневі мови (NVidia CUDA, RapidMind, Brook/Brook+) [8 – 13].

Таблиця 4

Порівняльна характеристика програмних платформ GPGPU:

Можливості	ATI CTM IL	GLSL/HLSL	NVidia CUDA	RapidMind	Brook/Brook+
Довільне зчитування з пам'яті	+	+	+	+	+
Довільний запис у пам'ять	+	–	+	+	– / +
Розрядність	64 bit	32 bit	64 bit (CUDA 2.0)	32 bit	64 bit
Ліцензія	Freeware	Freeware	Freeware	Shareware (демоверсія відсутня)	Open source
Підтримка відеоадаптерів	ATI (2XXX+)	Будь-який OpenGL 2.0 – сумісний (GLSL); будь-який DirectX 9.0c – сумісний (HLSL)	NVidia (8XXX+)	Будь-який DirectX 10-сумісний	Будь-який DirectX 9.0c чи OpenGL 2.0-сумісний (Brook) / ATI (серія 2XXX+) (Brook+)

Продовження табл. 4

Можливості	ATI CTM IL	GLSL/HLSL	NVidia CUDA	RapidMind	Brook/Brook+
Можливість низькорівневої оптимізації	+	–	–	–	– / +
Не потребує середовища виконання	+	–	+	–	–

Розробка програмної бібліотеки для конструювання та моделювання топологій штучних нейронних та нейроподібних мереж

Розроблена програмна бібліотека „NN-Constructor" призначена для конструювання топологій штучних нейронних і нейроподібних мереж (зокрема, паралельно-ієрархічних та ієрарх-ієрархічних) та їхнього імітаційного моделювання. „NN-Constructor" реалізує функції завантаження / збереження відповідного опису топології мережі у текстових файлах спеціального формату, а також функції для навчання та обробки (проведення сигналу) в нейронній або нейроподібній мережі. Необхідно відзначити, що в запропонованій програмній бібліотеці реалізовано можливості моделювання таких класів топологій нейронних мереж як мережі прямого поширення (Feedforward) та рекурентні мережі з можливістю задання користувачем довільної структури мережі. Конструювання нейроподібної мережі виконується шляхом об'єднання між собою шарів нейронних елементів. Шар може містити довільну кількість нейронних елементів; кількість шарів не обмежується (використовується динамічний список).

У програмній реалізації обрано принцип нейроподібної обробки даних, який полягає у передачі імпульсу від нейронів шарів, які належать до такту обробки i, до нейронів шарів, що належать до такту обробки $i+1$. Отже, кожне значення вхідного сигналу I_j може бути обчислено одночасно, тобто паралельно. Принцип тактування обробки нейронних мереж пояснює наступний рисунок:

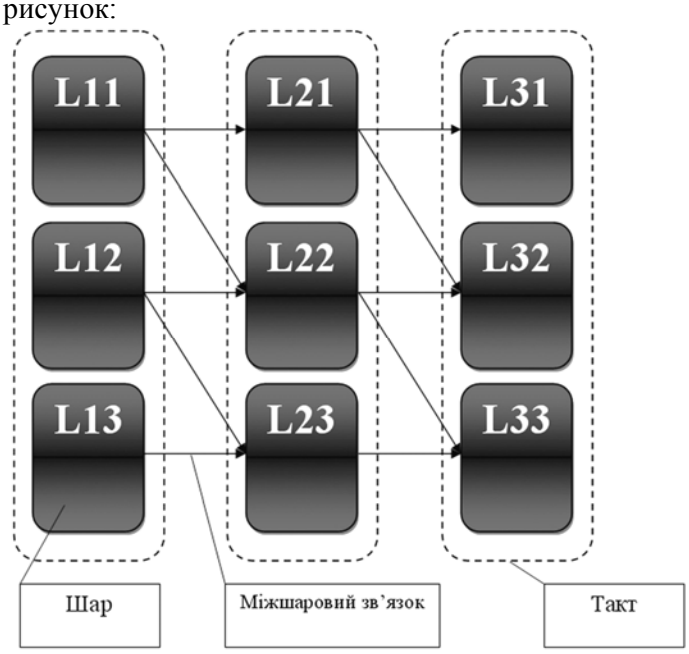

Рис. 1. Узагальнена схема процесу тактування обробки нейронних мереж у „NN-Constructor"

Мова реалізації CPU-версії програмної бібліотеки – C# для платформи „MS .NET 2.0". Функції бібліотеки коректно працюють з різними операційними системами: MS Windows XP (за умови наявності встановленого „MS .NET 2.0"), MS Windows Vista, Linux (за умови

встановленої платформи „Mono").

Мова реалізації GPU-версії програмної бібліотеки – C++ з використанням програмної платформи AMD StreamComputing SDK. Функції бібліотеки коректно працюють з різними операційними системами для відеоадаптерів ATI Radeon HD (серії 2000 і вище). Алгоритм обробки передачі імпульсу між тактами (рис. 1) з врахуванням специфіки програмування паралельних пристроїв, тобто із метою уникнення реалізації механізму синхронізації паралельних потоків, потребує перетворення формату даних, яке відбувається таким чином:

1. Для кожного нейрона будується одновимірна таблиця, кожний елемент якої являється структурою „номер зв'язаного нейрона у попередньому шарі – вага міжнейронного зв'язку".
2. Отже, для кожного шару поточного такту отримується набір таблиць за кількістю нейронів у шарі, які характеризують міжнейронні зв'язки.
3. Крім того, для кожного шару будується додаткова одновимірна таблиця, яка містить у собі рівні активації нейронів даного шару.

Таке перетворення дозволяє зберігати дані, необхідні для передачі імпульсу між тактами, у єдиному масиві та завантажувати їх у пам'ять відеокарти за один цикл передачі даних. Така реалізація дозволяє уникнути використання операції довільного запису в пам'ять, яка не підтримується відеокартами молодше серії R670, та реалізації механізму синхронізації між паралельними потоками.

Робота з „NN-Constructor" відбувається в межах таких основних етапів: завантаження із файлу (або створення користувачем за допомогою відповідних функцій) кількості шарів, зв'язків між ними, кількості нейронних елементів у шарі, зв'язків між нейронними елементами різних шарів нейронної або нейроподібної мережі; обробка вхідної інформації; навчання мережі; збереження топології мережі та результатів моделювання.

Експериментальні результати імітаційного моделювання штучних нейронних та нейроподібних мереж для задач прогнозування статистичних рядів курсів валют

У проведених дослідженнях використовувався реальний статистичний ряд, отриманий з відкритих джерел ринку Forex, який відображає погодинну динаміку зміни курсу євро–долар розмірністю 4137 записів (12. 10. 2008 р.). Завданням експерименту було отримання прогнозованого значення зміни курсу з горизонтом прогнозування – 1 крок [14].

Для прогнозування вказаного завдання було обрано декілька структур топологій нейронних мереж, зокрема, мережа Ворда із структурою 100-100-100-1, 100-25-25-1, 9-8-5-1, а також багатошаровий персептрон із різноманітними варіантами топологій. Як тестовий приклад експериментально було обрано нейронну мережу – багатошаровий персептрон з топологією 8-3-1 та методом навчання зворотного поширення помилки. Поставлену задачу прогнозування було реалізовано за допомогою розробленої програмної бібліотеки для конструювання та моделювання топологій штучних нейронних та нейроподібних мереж „NN-Constructor" (з можливостями обробки на CPU та GPU).

Зокрема, на рис. 2 представлено результати навчання вказаної нейронної мережі, яка реалізована з використанням нейроконструктора „NN-Constructor". Як видно з рисунку, в результаті навчання нейронна мережа коректно відтворює динаміку зміни значень курсу, зокрема, середня похибка прогнозування складає 0,004476721, що є прийнятним для поставленої економічної задачі. Крок прогнозування у програмі визначався таким чином: з вхідного ряду було обрано 8 елементів та однин елемент ряду вихідних значень (прогноз на 9 елемент вхідного ряду, оскільки горизонт прогнозу обрано 1).

Також було визначено швидкість обробки даних у нейронній мережі, яка дорівнює сумі швидкості навчання мережі та швидкості тестування. Для запропонованого варіанту швидкість обробки даних у нейронній мережі склала 14 секунд.

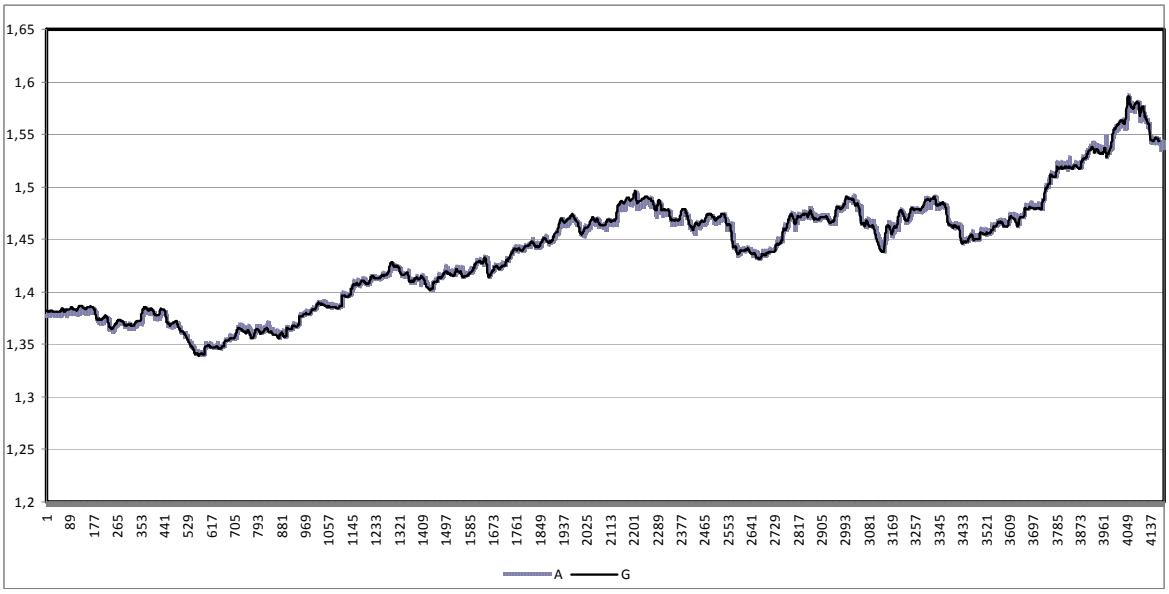

Рис. 2. Результати прогнозування курсів валют з використанням „NN-Constructor", де ряд A – оригінальний ряд, ряд G – ряд прогнозу

Для підтвердження адекватності роботи запропонованого програмного продукту та коректності отриманих результатів, також було здійснено комп'ютерне моделювання в одному із професійних та визнаних у галузі нейромережевої обробки програмних пакетів – Statistica Neural Network (SNN), компанії StatSoft [15].

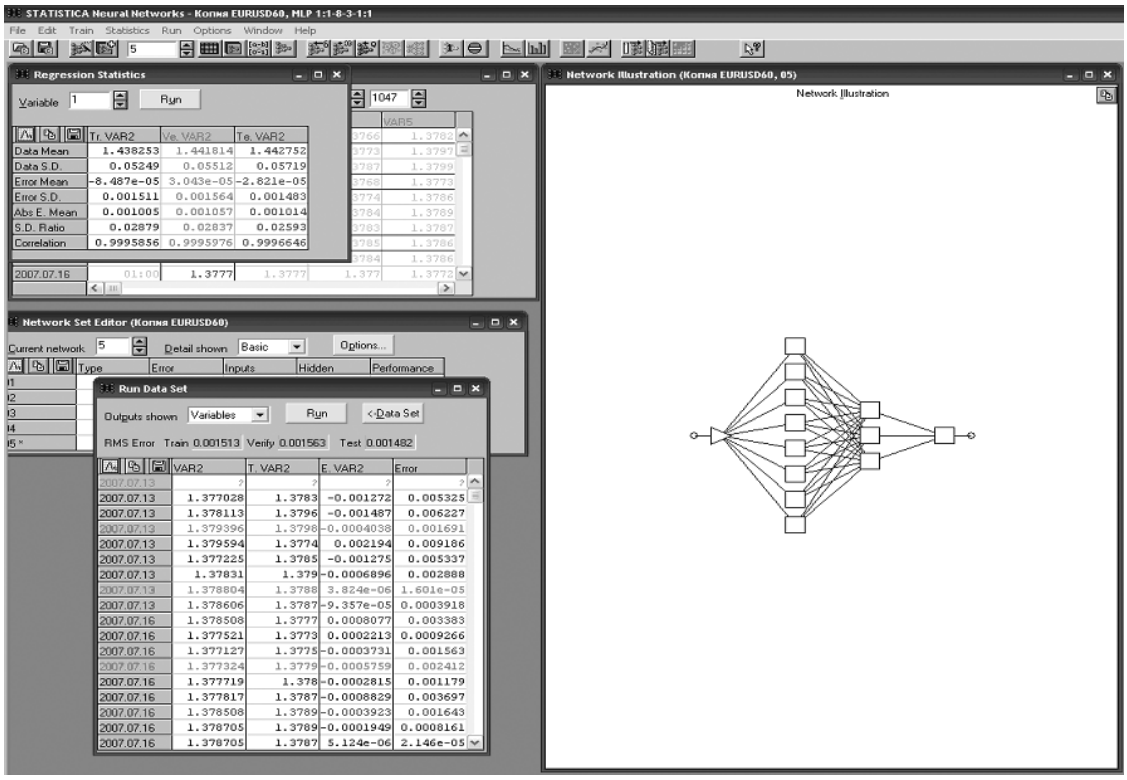

Рис. 3. Екранні форми з результатами комп'ютерного моделювання у програмному пакеті Statistica Neural Network

Зокрема, на рис. 3 представлено результати комп'ютерного моделювання вказаної нейронної мережі (багатошаровий персептрон з топологією 8-3-1 та методом навчання

зворотного поширення помилки) у програмному пакеті Statistica Neural Network.

У результаті навчання нейронна мережа коректно відтворює динаміку зміни значень курсу, зокрема, середня похибка прогнозування складає 0,00127200.

Рис. 4. Результати прогнозування курсів валют
у програмному пакеті Statistica Neural Network,
де ряд 1 – оригінальний ряд, ряд 2 – ряд прогнозу

Для Statistica Neural Network для запропонованого варіанту також було оцінено швидкість обробки даних у нейронній мережі, яка була значно більшою (більше 10 хвилин), ніж у варіанті реалізації з використанням „NN-Constructor".

Висновки

У роботі було досліджено та проаналізовано основні технології апаратно-програмної реалізації штучних нейронних мереж, зокрема сучасні спеціалізовані нейропроцесори, цифрові сигнальні процесори (DSP), мультимедійні центральні процесори (CPU) та альтернативні сучасні апаратні засоби, зокрема GPU, в контексті обґрунтування вибору базової платформи для моделювання різноманітних структур масштабних нейронних та нейропобідних паралельно-ієрархічних мереж.

Проведені наукові дослідження здійснювались у контексті подальшої розробки нейроемулятора – системи, побудованої на базі каскадного з'єднання універсальних SISD-, SIMD- або MISD-процесорів, яка реалізує типові нейрооперації (зважене підсумовування й нелінійне перетворення) на програмному рівні. У роботі запропоновано, як нейроприскорювач в контексті апаратної платформи для реалізації масштабних нейронних та нейропобідних паралельно-ієрархічних мереж, обрати технологію GPGPU, яка базується на використанні потужного відеоадаптера для виконання спеціалізованих, у тому числі паралельних, обчислень. Оскільки сучасні технології побудови відеоадаптерів дозволяють використання 128-ядерних спецпроцесорів, у порівнянні із сучасними 4-ядерними мультимедійними CPU, то застосування їх для нейроемуляції різних топологій масштабних нейронних та нейропобідних паралельно-ієрархічних мереж є актуальним та перспективним [6]. У контексті програмної реалізації проведена робота із створення нейропакета для реалізації різних топологій масштабних нейронних та нейропобідних паралельно-ієрархічних мереж і можливості прорахунку їх на GPU. Зокрема, запропоновано програмну бібліотеку, яка безпосередньо виконує процеси обробки нейромережі, а також візуальний

редактор топологій нейронних та нейропобідних паралельно-ієрархічних мереж. На основі вирішення тестової задачі прогнозування економічної інформації експериментально перевірено та доведено адекватність та ефективність програмної розробки.

СПИСОК ЛІТЕРАТУРИ

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.

Яровий Андрій Анатолійович – к. т. н., доцент, доцент кафедри інтелектуальних систем.

Богомолов Юрій Сергійович – студент кафедри інтелектуальних систем.

Вознесенський Костянтин Юрійович – студент кафедри інтелектуальних систем.
Вінницький національний технічний університет.