УДК 681.325

В. Грицик, Р. Ткаченко, І. Цмоць
Національний університет "Львівська політехніка",
кафедра автоматизованих систем управління,
кафедра інформаційних систем та мереж

ТЕХНОЛОГІЯ НЕЙРОКОМП'ЮТИНГУ РЕАЛЬНОГО ЧАСУ

© Грицик В., Ткаченко Р., Цмоць І., 2010

Проаналізовано особливості апаратної реалізації штучних нейронних мереж, вибрано принципи побудови, визначено шляхи підвищення ефективності використання обладнання, розроблено методи синтезу та базові структури нейрокомп'ютерних систем реального часу.

Ключові слова: нерокомп'ютинг, штучні нейронні мережі, НВІС-технологія, паралельна обробка, конвеєр, реальний час.

Features of hardware representation of artificial neural networks were analyzed, principles of construction were chosen, ways of efficiency increase of equipment use were determined, methods of synthesis and base structures of the neural computing, , real-time systems were developed.

Key words: neurocomputing, artificial neural networks, VLSIC-technology, parallel **processing, conveyer, real time.**

Постановка проблеми

Технологія нейрокомп'ютингу в реальному часі – це створення паралельних, розподілених, адаптивних комп'ютерних систем обробки даних у реальному часі, які здатні "вчитися" опрацьовувати дані, діючи в інформаційному середовищі. Ця технологія не вимагає готових алгоритмів і правил опрацювання, при її використанні комп'ютерні системи повинні "уміти" випрацьовувати правила та модифікувати їх у процесі розв'язання конкретних задач опрацювання даних. Роботи в галузі нерокомп'ютингу ведуться за такими напрямками:

- розроблення нейроалгоритмів;
- створення спеціалізованого програмного забезпечення;
- розроблення спеціалізованих нейрокомп'ютерних систем.

Особливою рисою нейрокомп'ютингу найчастіше є єдиний принцип навчання нейрокомп'ютерних систем – мінімізація емпіричної похибки. Функція похибки, яка оцінює конфігурацію системи, задається іззовні – залежно від того, яку мету передбачає навчання. Поступову модифікацію конфігурації (стан зміни всіх своїх синаптичних ваг) здійснюють так, щоб мінімізувати похибку.

Сучасний етап розвитку технологій нейрокомп'ютингу характеризується розширенням галузей застосування, в значній частині з яких вимагається опрацювання у реальному часі різних за інтенсивністю надходження потоків даних на комп'ютерних системах, що задовольняють обмеження щодо габаритів, енергоспоживання, вартості та часу розроблення. Створення таких нейрокомп'ютерних систем вимагає широкого використання сучасної елементної бази (напівзамовних і замовних НВІС, процесорів цифрової обробки сигналів, мікроконтролерів, трансп'ютерів, нейрочіпів) та розроблення нових методів, алгоритмів і НВІС-структур для реалізації нейроалгоритмів. Режим реального часу накладає обмеження на час розв'язання задачі Тр, який не повинен перевищувати часу обміну повідомленнями $T_{обм}$, тобто:

$$T_p \le T_{обм}.$$

Час обміну залежить як від обсягу N, розрядності n і частоти F_d надходження вхідних даних, так і від кількості k каналів та їх розрядності n_k. Такий час визначається за формулою:

$$T_{обм} = \frac{Nn}{F_d k n_k}.$$

Для забезпечення опрацювання потоків даних у реальному часі за допомогою нейрокомп'ютерних систем їх продуктивність повинна бути:

$$П \geq \frac{BRF_d kn_k}{Nn},$$

де R – складність алгоритмів розв'язання задач; β – коефіцієнт врахування особливостей засобів реалізації алгоритму.

Тому актуальною проблемою є розроблення технології нейрокомп'ютингу в реальному часі, яка використовує сучасні НВІС-технології, просторово-часове розпаралелювання та узгодження інтенсивності надходження потоків даних з обчислювальною інтенсивністю системи.

Аналіз останніх досліджень та публікацій

Аналіз існуючих досліджень у галузі створення нейрокомп'ютерних систем реального часу [1–13] показує, що вони мають такі недоліки:

- опрацювання в режимі реального часу інтенсивних потоків даних вимагає великих затрат обладнання для створення нейрокомп'ютерних систем;
- не враховуються вимоги конкретних застосувань щодо габаритів і споживаної потужності;
- велика вартість і час, які необхідні для розробки нейрокомп'ютерних систем реального часу;
- алгоритми функціонування та структури компонентів нейрокомп'ютерних систем не орієнтовані на НВІС-реалізацію;
- неузгодженість інтенсивності надходження даних з обчислювальною здатністю компонентів нейрокомп'ютерних систем реального часу не забезпечує досягнення високої ефективності використання обладнання.

Результати аналізу свідчать, що для забезпечення високої продуктивності обробки даних у нейрокомп'ютерних системах реального часу необхідно використовувати спеціалізацію, конвеєризацію та просторовий паралелізм. Для орієнтації архітектури нейрокомп'ютерної системи на НВІС-реалізацію необхідно, щоб вона була однорідною та мала локальні зв'язки між елементами системи.

Завдання і мета дослідження

Технологія нейрокомп'ютингу в реальному часі вимагає розпаралелювання і конвеєризації процесів обчислень, широкого використання сучасної елементної бази, розроблення компонентів нейрокомп'ютерних систем реального часу, які просто адаптуються до вимог конкретних застосувань. Компоненти нейрокомп'ютерних систем реального часу повинні мати високу ефективність використання обладнання, якої досягають розробленням узгоджено-паралельних алгоритмів функціонування та методів синтезу та просторово-часового перетворення їх на паралельні структури. Технологія нерокомп'ютингу в реальному часі повинна ґрунтуватися на інтегрованому підході, який охоплює сучасну елементну базу, методи та алгоритми реалізації базових операцій нейроалгоритмів, архітектури компонентів і систем, враховувати вимоги конкретних застосувань і інтенсивності надходження даних.

Мета дослідження полягає у виборі принципів, засобів реалізації нейрокомп'ютерних систем реального часу, визначенні шляхів підвищення ефективності використання обладнання, розробленні методу синтезу та базових структур нейрокомп'ютерних систем.

Виклад основного матеріалу

Аналіз, вибір засобів і варіантів реалізації нейроалгоритмів. Особливістю розвитку нейрокомп'ютерних систем є вдосконалення процесорної компоненти у напрямку збільшення її гнучкості (адаптивності). Ефективність комп'ютерної реалізації нейроалгоритмів безпосередньо пов'язана з вибором засобів реалізації: програмних, мікропрограмних або апаратних [7,13].

Програмна реалізація передбачає використання універсальних і функціонально-орієнтованих мікропроцесорів для синтезу нейрокомп'ютерних систем. При програмній реалізації нейроалгоритмів обчислювальні процеси переважно розгортаються в часі з великим об'ємом пересилання інформації між оперативною пам'яттю і операційними пристроями. Програмні засоби є доступними

для програміста, перед яким виникає задача мінімізації об'єму програм і часу їх реалізації за заданої точності обчислень. Вказані засоби характеризуються низькою швидкодією і гнучкістю з погляду можливості модифікації та заміни алгоритмів.

Головною особливістю мікропрограмних засобів є регулярна структура, до складу, якої входить пам'ять мікропрограм реалізації нейроалгоритмів. Мікропрограмна реалізація обчислень передбачає їхнє розгортання як в часі, так і в просторі. Під час мікропрограмування є доступ до системи мікропрограм процесора, що забезпечується застосуванням постійної пам'яті, програмованих логічних матриць, а також оперативних запам'ятовувальних пристроїв, які використовують як пам'ять мікропрограм. Прикладом нейрокомп'ютерної системи з мікропрограмною реалізацією нейроалгоритмів є комп'ютерні системи на основі однорідних обчислювальних середовищ (ООС). Процесор на базі ООС – це двовимірна регулярна матриця процесорних елементів (ПЕ), кожен з яких фізично зв'язаний входом–виходом з чотирма сусідами: згори, знизу, зліва та справа [2]. Кожний ПЕ може виконувати набір бітових операцій перетворення інформації з вхідних каналів на вихідні. Процесор на базі ООС є універсальною системою, тобто в ньому можливо реалізувати довільну обчислювальну функцію. Бітовий рівень ПЕ та повна система комутації дають змогу розпаралелювати обчислення на найнижчому бітовому рівні. Це є значною перевагою мікропрограмної реалізації нейроалгоритмів. Реалізації повною мірою потенційних можливостей мікропрограмних засобів можна досягти лише при глибокому вивченні як задачі, яка розв'язується, так і внутрішньої мови процесора. Мікропрограмні засоби реалізації нероалгоритмів порівняно з програмними є більш швидкодіючими.

Успіхи у галузі інтегральної технології дають змогу все більше перекладати реалізацію нероалгоритмів на апаратні засоби, які розгортають обчислення не тільки у часі, а і в просторі. Такі обчислення характеризуються введенням додаткового обладнання і відсутністю проміжних пересилок інформації в процесі обчислення, а також спрощенням функції місцевого управління. В основу структурної організації апаратних засобів покладено принцип адекватного апаратного відображення графів алгоритмів функціонування штучних нейронних мереж (ШНМ) на процесорні елементи, які реалізують функції нейронів та з'єднані між собою відповідно з графом [7].

Потрібно зазначити, що всі види реалізації нейроалгоритмів у безпосередньому вигляді зустрічаються доволі рідко. На практиці у більшості випадків для реалізації алгоритмів функціонування ШНМ використовуються комбіновані підходи з перевагою одного з перерахованих засобів. Переважання того чи іншого засобу визначається вимогами, які ставляться до ШНМ за швидкодією, габаритами, потужністю споживання та ціною.

Враховуючи результати аналізу засобів реалізації нероалгоритмів, можливі такі варіанти побудови нейрокомп'ютерних систем:

- на основі універсальних і функціонально-орієнтованих мікропроцесорів шляхом розроблення спеціалізованого програмного забезпечення;
- на основі універсального обчислювального ядра, доповненого базовими апаратно-програмними компонентами, які реалізують часомісткі алгоритми функціонування ШНМ;
- у вигляді спеціалізованої алгоритмічної системи, архітектура та організація обчислювального процесу в якій відображає структуру алгоритму функціонування ШНМ.

Перший варіант є доступним для широкого кола користувачів. Істотною його перевагою є можливість використання раніше розроблених програм. Його недоліками є невисока швидкодія, функціональна і структурна надлишковість комп'ютерних засобів.

Достатньо ефективним шляхом покращання продуктивності нейрокомп'ютерних систем вважають перехід від програмних емуляторів до апаратних реалізацій, які передбачають декілька різних рівнів реалізації так званих нейрокомп'ютерів – від створення плат розширення для прискорення виконання базових операцій (другий варіант) і аж до розроблення повністю спеціалізованої елементної бази, що забезпечує побудову пристроїв у нейромережному елементному базисі (третій варіант).

Другий варіант є перспективним, оскільки він передбачає поєднання універсальних і спеціальних засобів. Таке поєднання забезпечує високу ефективність використання обладнання при

створенні систем для опрацювання у реальному часі потоків даних за алгоритмами, які є нерегулярними з великою кількістю логічних операцій. При цьому розроблення нейрокомп'ютерних систем із заданими технічними параметрами зводиться до доповнення обчислювального ядра додатковими апаратно-програмними компонентами.

Третій варіант орієнтований на обробку у реальному часі інтенсивних потоків даних. При цьому високої ефективності використання обладнання досягають узгодженням обчислювальної здатності апаратних засобів з інтенсивністю надходження потоків даних. Використання для побудови нейрокомп'ютерних систем обчислювальних полів на основі програмованих логічних інтегральних схем (ПЛІС) з динамічним репрограмуванням відкриває новий етап створення реконфігурованих нейрокомп'ютерних систем (РНС) Особливістю РНС є архітектурна гнучкість, яка у рівному степені стосується як апаратних (hardware), так і програмних (software) засобів. У РНС структура може динамічно змінюватися як при підготовці до розв'язання задачі, так і в ході обчислювального процесу. Використання принципу реконфігурованості при побудові нейрокомп'ютерних систем забезпечує високу живучість і нарощування функцій. Застосовуючи такий принцип, можна простим перепрограмуванням структури системи налаштувати її на ефективну реалізацію необхідного нейроалгоритму, що визначає його універсальність.

Особливості апаратної реалізації ШНМ. Розглянемо деякі особливості і обмеження щодо апаратної реалізації однієї з найпоширеніших і апробованих на практиці нейромережних структур – багатошарових перцептронів. Навчають подібні ШНМ, враховуючи напрацьовані модифікації методів, на основі ітеративних алгоритмів багатокритеріальної оптимізації, шляхом поширення вихідних похибок навчання на кожному кроці на вхід перцептрона. Принагідно зауважимо, що сама процедура навчання є по суті завершальним етапом підготовки мережі до виконання своїх функцій – вибір параметрів складності структури, стратегії навчання та тестування може займати надзвичайно багато часу навіть при прискореному виконанні алгоритму. Навчання також може виявитися ненадійним, зупинятися через ефект "паралічу мережі", а принципова відсутність повторюваності результатів значно ускладнює процес налагодження ШНМ. Однак, згадаємо, що в основі методу навчання шляхом зворотного поширення похибок є спосіб прискореного обчислення компонентів градієнта. Для цього функція похибки представляється у вигляді складної функції, для якої послідовно розраховуються часткові похідні. Основними операціями на кожному кроці алгоритму є знаходження комбінованих входів нейронних елементів (НЕ) (скалярний добуток багатокомпонентних векторів), обчислення нелінійних функцій (функції активації і їх похідні) і багаточленні операції підсумовування парних добутків [13]. Отже, виявляється, що алгоритми функціонування нейрокомп'ютерної системи є достатньо складними, містять значну кількість різнотипних операцій.

Оскільки для забезпечення достатньої точності необхідно проводити масштабування, що є проблематичним, в апаратних рішеннях передбачається використання багаторозрядної сітки (наприклад, 64 біти) та арифметики з плаваючою комою. Окрім того, сам алгоритм модифікації вагових коефіцієнтів є достатньо нерегулярним, зокрема процедури прямого і зворотного проходження сигналів суттєво відрізняються. Завдяки цьому до складу нейрокомп'ютерної системи вводять не менше двох процесорів з різними функціями – як для здійснення обчислень, так і для керування алгоритмом. У результаті виявляється, що ефект від спеціалізованої реалізації нейрокомп'ютерної системи помітно знижується, а саме її створення є надто складним і дорогим процесом.

Розглянемо альтернативний варіант побудови нейрокомп'ютерної системи, за структурою і принципами застосування близької до багатошарових перцептронів, що ґрунтується на новій концепції опрацювання даних, яка має загалом значно ширшу основу застосування, окрім варіанта створення нейромереж, зокрема в області нечіткої логіки, компресії і оброблення сигналів, синтезу формул – математичних моделей тощо. Нейромережний варіант геометричних перетворень (ГП), відомий під назвою функціонал на множині табличних функцій, істотно підвищує продуктивність як для програмного, так і апаратного варіантів реалізації нейроподібних структур (НС).

Особливістю НС на базі ГП є однотипність алгоритмів навчання та функціонування. Топологія алгоритму навчання НС на базі ГП представляється в вигляді деякого графу, вершини якого відповідають основним операціям алгоритму – скалярному добутку вектора вхідних сигналів

на вектор вагових коефіцієнтів та нелінійному перетворенню від скалярного добутку. Тобто вершини графу можна розглядати як відповідники нейронних елементів прихованого шару ШНМ, а моделі НС на базі ГП описуються структурами нейроподібного типу. Подібне трактування є продуктивним як при налагодженні моделей НС на базі ГП, так і при побудові архітектури відповідних систем на основі НС на базі ГП для розв'язання багатьох задач, зокрема, прогнозування часових послідовностей, виділення головних компонент, ущільнення даних і т.п. з використанням нейромережних підходів до реалізації.

Одночасно зауважимо принципову відмінність НС на базі ГП від нейромережних засобів – якщо ШНМ навчання здійснюється (як правило, ітераційно) з метою встановлення параметрів для обраної заздалегідь структури, то в НС на базі ГП структура моделі формується за результатами навчання відповідно до заданих його характеристик. Тобто в нейромережному тлумаченні НС на базі ГП являють собою лише графи відповідних алгоритмів, які, однак, містять внутрішній просторовий і часовий паралелізм і можуть бути реалізовані також апаратно. Тобто, концепція НС на базі ГП забезпечує реалізацію принципів функціонального, а не структурного моделювання, оскільки структура моделі НС на базі ГП визначається структурою табличних даних.

НС на базі ГП покликані усунути або зменшити негативні властивості існуючих засобів інформаційного моделювання: регресійних моделей, машин опорних векторів, ШНМ, індуктивних моделей, контролерів нечіткої логіки, статистичних процедур.

Базовими властивостями НС на базі ГП є такі: єдина методологічна основа побудови архітектури для різних завдань та предметних областей застосування; швидке неітераційне навчання за наперед задану кількість кроків обчислень, що відкриває можливість розв'язування завдань великих розмірностей; повна повторюваність результатів навчання; можливість отримання задовільних розв'язків для тренувальних вибірок зменшеного об'єму; можливість розв'язування задач в автоматичному режимі; розширення кібернетичного принципу "чорної скриньки" на користь "сірої скриньки", оскільки НС на базі ГП володіють додатковими можливостями аналізу внутрішньої структури даних; висока точність та покращені генералізуючі властивості.

Незважаючи на помітні переваги НС на базі ГП в плані продуктивності перед традиційними нейромережними засобами, найвищих параметрів швидкодії для них досягаються у випадку апаратної реалізації. В основу отримуваного ефекту покладено особливості відповідних алгоритмів навчання та функціонування.

Проаналізуємо основні співвідношення для перетворення даних, що закладені в основу навчання та застосування (за змістом майже повторюються) НС на базі ГП. Реалізація вказаних та ряду інших функцій покладається на НС на базі ГП, яка попередньо має бути навченою. Для методу ГП навчання забезпечується шляхом послідовних трансформацій тренувальної матриці MT_{ni}, що являє собою підматрицю всіх реалізацій, де n – номер рядка-реалізації матриці, і – номер відповідної ознаки [6,8]. Рядки останньої матриці повинні достатньо повно відображати можливий діапазон реалізацій об'єкта, описувати основні риси його поведінки. Представляємо будь-який елемент матриці реалізацій як обчислене значення функції двох змінних F(ni). Метою навчання є представлення функції двох змінних F(ni) комбінацією функцій однієї змінної f(n), φ(i), які і є шуканими аргументними функціями нейромережі.

Розглядаємо випадок, коли елементи стовпців матриці MT_{ni}, що відповідають вихідним ознакам, є відомими. Це характерне для режиму навчання і завжди – для автоасоціативного варіанта нейромереж. Для представлення функції F(ni) кінцевою сумою добутків функцій однієї змінної послідовно виконуємо ряд перетворень:

1) приймаємо j=0;

2) обираємо базовий рядок $MT^{(j)}_{nb,i}$ серед рядків для $MT^{(j)}_{n,i}$; це може бути, наприклад, рядок, сума квадратів елементів якого є максимальною; можливий ітераційний варіант формування базового рядка на основі заданого критерію оптимальності;

3) від кожного з рядків тренувальної матриці віднімаємо базовий рядок, помножений на коефіцієнт, величина якого визначається з умови мінімуму різниці в сенсі критерію найменших квадратів

$$MT_{n,i}^{(j+1)} = MT_{n,i}^{(j)} - K1_{n}^{(j)} \cdot MT_{nb,i}^{(j)}, \qquad (1)$$

де

$$K1_n^{(j)} = \frac{\sum_{i=1}^{ix+iy} (MT_{n,i}^{(j)} \cdot MT_{nb,i}^{(j)})}{\sum_{i=1}^{ix+iy} (MT_{nb,i}^{(j)})^2}, \quad (2)$$

j – номер кроку реалізацій; n=1,…,n_{max}; ix – число вхідних ознак; iy – число вихідних ознак; n_{max} – число рядків тренувальної матриці.

Якщо j<ix+iy-1, збільшуємо j на одиницю, перехід на пункт 2, інакше – кінець процедури розкладу.

Перетворення за формулами (1) та (2) відповідають процедурі ортогоналізації Грама-Шмітта , отже для j=ix+iy-1

$$MT_{n,i}^{(j+1)} = |0|,$$

тобто отримуємо нульову матрицю. Отже,

$$MT_{n,i}^{(0)} = \sum_{j=0}^{ix+iy-1} K1_n^{(j)} \cdot MT_{nb,i}^{(j)}, \quad (3)$$

де n=1,…,n_{max}, i=1,…,ix+iy. Співвідношення (3) можна подати у такому вигляді:

$$F_{n,i} = \sum_{j=0}^{ix+iy-1} f_j(n) \cdot \varphi_j(i), \quad (4)$$

де f(n), φ(i) – табличні функції, які отримуються шляхом перетворень тренувальної матриці реалізацій. Дані функції представляють результат навчання і використовуються, як параметри нейромережі в режимі її застосування, який реалізується на основі базової формули (4).

Проаналізуємо робочі формули (1)–(4), покладені в основу навчання і застосування НП структури. Можна зробити такі висновки:

1. Базовою операцією, що займає найбільше ресурсів, є векторна операція скалярного добутка. Додатково використовуються нелінійні операції обчислення степеневого полінома, що містить лише декілька членів. Використовуючи схему Горнера для обчислень, приходимо до операції множення з накопиченням, яка потребує у цьому випадку менше ресурсів як скалярний добуток для багатокомпонентних векторів.

2. Враховуючи особливості алгоритму, де операнди представляються в наперед заданому діапазоні чисел від 1 до -1, а при наступних перетвореннях зменшуються за абсолютною величиною до нуля, доходимо висновку про можливість представлення даних в короткій розрядній сітці (до 16 бітів) з фіксованою комою. Останнє відкриває додаткові можливості в плані використання прискорених алгоритмів методів виконання арифметичних операцій.

Компонентно-ієрархічний підхід до розроблення нейрокомп'ютерних систем реального часу. Розробляти нейрокомп'ютерні системи реального часу з високою ефективністю використання обладнання будемо на основі компонентно-ієрархічного підходу, який передбачає поділ процесу розроблення на ієрархічні рівні та види забезпечення (алгоритмічне, апаратне та програмне). Для реалізації такого підходу використовується метод декомпозиції, який передбачає розбиття засобів на окремі компоненти, з яких у міру необхідності комплектують конкретну нейрокомп'ютерну систему реального часу. На кожному рівні ієрархії розв'язуються задачі відповідної складності, які характеризуються як одиницями інформації, так і алгоритмами обробки.

За складністю обробки компоненти нейрокомп'ютерних систем можна розділити на чотири ієрархічні рівні. Збільшенню номера рівня ієрархії відповідає збільшення деталізації алгоритмічних, апаратних і програмних засобів. При цьому на вищих рівнях ієрархії одиниці інформації, алгоритми, програмні та апаратні засоби являють собою впорядковані сукупності одиниць інформації та композиції алгоритмів, програмних і апаратних засобів нижчих рівнів ієрархії (табл. 1). Методологія послідовної декомпозиції, яка використовується для розроблення систем обробки сигналів, відображає процес розроблення "згори донизу".

Таблиця 1

Рівні та види розробок нейрокомп'ютерної системи реального часу

Ієрархічний рівень	Види забезпечення та виконувані розробки		
	Алгоритмічне	**Апаратне**	**Програмне**
1-й	Алгоритми функціонування ШНМ	Структура апаратних засобів нейрокомпютерної системи	Структура програмних засобів нейрокомпютерної системи
2-й	Алгоритми функціонування макрокомпонентів ШНМ (шари ШНМ)	Структури нейропроцесорів, паралельної пам'яті, контролерів пристроїв обміну	Програми реалізації макрокомпонентів ШНМ
3-й	Алгоритми реалізації компонентів ШНМ	Структури нейронних елементів, пристроїв для реалізації макроопераций нейроалгоритмів (сум парних добутків)	Програми реалізацій нейронних елементів, макрооперацій нейроалгоритмів
4-й	Алгоритми обчислення сум парних добутків, множення	Структури операційних пристроїв групового підсумовування, множення і нелінійного перетвореня	Програми реалізацій операцій обчислення сум парних добутків, множення

На першому ієрархічному рівні розроблення алгоритми реалізації ШНМ подаються у вигляді функціонального графу F=(Ф,Г), де Ф={$Ф_1$,$Ф_2$,…,$Ф_n$} – множина функціональних операторів, Г-закон відображення зв'язків між операторами [5].

Другий рівень ієрархії становлять макрокомпоненти ШНМ. До таких макрокомпонентів належать нейропроцесори, які реалізують шари ШНМ, паралельна пам'ять, контролери пристроїв обміну [2].

Третій ієрархічний рівень становлять компоненти, які реалізують базові макрооперації нейроалгоритмів (сум парних добутків) і нейронні елементи. Для реалізації елементів третього рівня розробляються паралельні алгоритми, НВІС-структури та програми.

До четвертого рівня ієрархії належать елементи, які реалізують операції множення, групового підсумовування, нелінійного перетворення. У функціональному і структурному відношеннях елементи четвертого рівня ґрунтуються на елементарних арифметичних операціях.

Компонентно-ієрархічну структуру МІТ можна описати за допомогою такого виразу:

$$C_{MIT}^{1} = \bigcup_{i=1}^{n} C_{MIT}^{2i} \bigcup_{j=1}^{m} C_{MIT}^{3j} \bigcup_{p=1}^{h} C_{MIT}^{4p},$$

де C_{MIT}^{2i}, C_{MIT}^{3j}, C_{MIT}^{4p} – засоби відповідно другого, третього і четвертого ієрархічних рівнів; n – кількість типів макрокомпонентів; m – кількість типів компонентів; h – кількість типів операційних пристроїв.

Принципи побудови нейрокомп'ютерних систем реального часу. Архітектури нейрокомп'ютерних систем реального часу повинні повною мірою використовувати можливості НВІС-технології, враховувати вартість площі кристала, а також кількість вхідних і вихідних виводів. Число зовнішніх виводів НВІС обмежене рівнем технології та розміром кристала. В основу побудови нейрокомп'ютерних систем реального часу пропонується покласти принципи, за якими можна зменшити вартість, терміни і розширити галузі їх застосування. Аналіз показує [5], що забезпечити ці вимоги можна, використовуючи такі принципи:

- модульності, який передбачає розроблення компонентів нейрокомп'ютерних систем у вигляді функціонально завершених пристроїв (модулів);
- узгодженості інтенсивності надходження даних з обчислювальною здатністю нейрокомп'ютерних систем;
- конвеєризації та просторового паралелізму обробки даних;
- однорідності та регулярності архітектури нейрокомп'ютерних систем;
- локалізації та спрощення зв'язків між елементами нейрокомп'ютерних систем;
- спеціалізації та адаптації апаратно-програмних засобів до структури алгоритмів обробки та інтенсивності надходження даних;
- програмованості архітектури шляхом використання репрограмованих ПЛІС.

Синтез нейроподібних систем реального часу на базі моделі геометричних перетворень. Пропонується створювати високоефективні НС реального часу на базі ГП на основі інтегрованого підходу, який ґрунтується на можливостях сучасної елементної бази та охоплює методи, алгоритми НС на базі ГП, архітектури апаратних засобів, враховує вимоги конкретних застосувань та інтенсивності надходження даних.

При цьому задача синтезу НС реального часу на базі ГП зводиться до формування множин вимог **R**={R_1, R_2,…,R_k}, характеристик **H**={H_1, H_2,…,H_m} і обмежень **B**={B_1, B_2,…,B_k} та знаходження такого вектора $\mathbf{H}^*=[H^*_1, H^*_2,\ldots,H^*_m]$, $H^*_i=f_i(R, H, B)$, i=1,…,m, який забезпечить максимальне значення ефективності використання обладнання E=max f(R, H^*,B).

Множину вимог **R** становить: R_1 – кількість каналів надходження даних m_d; R_2 – розрядність каналів надходження даних n_d; R_3 – частота надходження даних F_d; R_4 – швидкодія елементної бази, яка визначається часом затримки вентиля $t_в$; R_5 – кількість елементів (слів) вхідного масиву N; R_6 – розрядність вхідного слова n. Множину характеристик **H** становлять: H_1 – загальна кількість зв'язків Z; H_2 – просторова зв'язкова віддаль Δj ; H_3 – конвеєрний такт $t_к$; H_4 – витрати обладнання W; H_5 – кількість типів функціональних вузлів s; H_6 – кількість каналів введення $m_{вв}$; H_7 – розрядність каналів введення $n_{вв}$; H_8 – кількість виводів інтерфейсу зв'язку Y. Обмеження **B**, які необхідно враховувати при синтезі апаратних засобів реального часу, є наступними: B_1 – точність обчислення, яка визначається розрядністю результату $n_р$; B_2 – час обчислення $T_{обч}$, повинен бути $T_{обч}\le T_{обм}$, де $T_{обч}=\dfrac{t_к Nn}{m_{вв}n_{вв}}$, $T_{обм}$ – час обміну, який визначається так

$$T_{обм}=\frac{Nn}{F_d m_d n_d}.$$

Для вибору варіанта НС реального часу на базі ГП використовується критерій ефективності використання обладнання E, який враховує кількість виводів інтерфейсу, однорідність структури, кількість і локальність зв'язків, зв'язує продуктивність з витратами обладнання та дає оцінку елементам (вентилям) компонента за продуктивністю [2]. Кількісна величина ефективності використання обладнання для такого апаратного засобу визначається так:

$$E=\frac{m_к n_к}{t_к Nn(k_1\sum_{i=1}^{s}W_{ФУ_i}d_i+k_2Q+k_3Y)}$$

де $W_{ФУi}$ – витрати обладнання у вентилях на реалізацію і-го функціонального вузла, d_i – кількість функціональних вузлів і-го типу, k_1 – коефіцієнт врахування однорідності $k_1=f(s)$, k_2 – коефіцієнт врахування регулярності зв'язків $k_2=f(\Delta j)$, k_3 – коефіцієнт врахування кількості виводів інтерфейсу зв'язку $k_3=f(Y)$.

Конвеєрний такт $t_к$ визначається за формулою $t_к=\sum_{j}^{l}\max t_в$, де l – кількість послідовно з'єднаних вентилів у найповільнішій сходинці конвеєра, а Δj – як різниця просторових індексів.

Етапи синтезу апаратних засобів. Синтез НС реального часу на базі ГП складається із таких етапів: вибору та розроблення методів і алгоритмів функціонування ШНМ; визначення основних параметрів НС; переходу від алгоритму до структури нейрокомп'ютерної системи [3].

При виборі та розробленні алгоритмів функціонування НС враховуються вимоги **R** і характеристики **H**, але визначальним є забезпечення обмежень **B**. Для оцінювання розроблених алгоритмів використовуються характеристики: інформаційні, операційні та точності. До інформаційних характеристик належать: кількість констант, вхідних, вихідних і проміжних даних, кількість каналів та їх розрядність, кількість і види операцій. Операційні характеристики дають змогу оцінити час реалізації та обчислювальну здатність. До характеристик точності алгоритму належать: розрядність операційних пристроїв, способи округлення. У НС реального часу на базі ГП необхідно

забезпечити узгодженість інтенсивності надходження даних із обчислювальною здатністю нейрокомп'ютерної системи на всіх етапах обробки.

До основних параметрів оцінювання апаратних засобів реального часу, крім витрат обладнання, швидкодії, ефективності використання обладнання, пропонується використовувати обчислювальну здатність. Для обробки потоків даних у реальному часі доцільно використовувати синхронні структури з конвеєрною реалізацією графів нейроалгоритмів, в яких здійснюється суміщення у часі виконання функціональних операторів нейроалгоритму над різними даними. Конвеєризація нейрокомп'ютерної системи передбачає розділення її на сходинки шляхом введення буферної пам'яті. При цьому кожна сходинка конвеєра складається з двох компонентів: буферної пам'яті та операційних пристроїв, які реалізують функціональні оператори ярусу. Для забезпечення високої швидкодії та ефективності використання обладнання функціональні оператори, які реалізуються у сходинках конвеєра, мають бути простими та мати приблизно однаковий час реалізації. У конвеєрних нейрокомп'ютерних системах обчислювальна здатність визначається так:

$$D_к = \frac{m_к n_к}{t_{БП} + t_{ОБ}},$$

де $t_{БП}$ – час звертання до буферної пам'яті, $t_{ОБ}$ – час обчислення операційним блоком найскладнішого функціонального оператора Φ_{jk}, $m_к$ – кількість каналів надходження даних у сходинках конвеєра, $n_к$ – розрядність каналів надходження даних у сходинках конвеєра.

Шляхи підвищення ефективності використання обладнання. В нейрокомп'ютерних системах реального часу високої ефективності використання обладнання досягають мінімізацією витрат обладнання при забезпечені реального часу. Перехід від алгоритму розв'язання задачі в реальному часі до структури системи формально зводиться до мінімізації витрат обладнання

$$W_{НС} = W_{ІП} + W_{П} + W_{ПУ} + \sum_{i=1}^{k} W_{ПЕ_i} m_i ,$$

де $W_{НС}$, $W_{ПУ}$, $W_{ІП}$, $W_{П}$, $W_{ПЕ}$ – витрати обладнання відповідно на НС, пристрої управління, інтерфейсні пристрої, пам'ять, k – кількість типів процесорних елементів, і-й процесорний елемент, m_i – кількість процесорних елементів і-го типу, при забезпеченні такої умови:

$$\frac{Nn}{F_d m_d n_d} \geq \frac{t_к Nn}{m_{вв} n_{вв}}. \qquad (1)$$

Основними шляхами мінімізації апаратних затрат на реалізацію НС реального часу на базі ГП є:

- врахування величини зміни елементів даних;
- вибір ефективних методів і алгоритмів реалізації функціональних операторів нероалгоритмів;
- зменшення розрядності операційних пристроїв, пам'яті, кількості і розрядності каналів передавання даних;
- узгодження інтенсивності надходження даних із обчислювальною здатністю нейрокомп'ютерної системи.

Метод просторово-часового відображення нейроалгоритмів в узгоджено-паралельні структури. Для переходу від алгоритму до структури НС реального часу на базі ГП необхідно розробити узгоджений потоковий граф [3–8]. Цей процес розроблення можна розбити на такі чотири етапи:

- декомпозиція алгоритму розв'язання задачі;
- проектування комунікацій (обмінів даними) між функціональними операторами (нейроелементами);
- укрупнення функціональних операторів;
- планування обчислень.

На етапі декомпозиції нейроалгоритм **Ф** розбивається на функціональні оператори Φ_i, між якими установлюються зв'язки, що відповідають цьому алгоритму. Чим більшої деталізації алгоритму досягаємо у результаті декомпозиції, тим гнучкішим буде алгоритм і тим легше можна адаптувати його до виконання необхідних умов. Декомпозицію будемо здійснювати за методом

функціональної декомпозиції, при якому нейроалгоритм **Ф** розбивається на операції Φ_i, кожна із яких може бути реалізована операційними блоками певного рівня ієрархії. Використанням методу функціональної декомпозиції можна отримати просторово-часове відображення структури алгоритму на рівні операцій Φ_i, які використовуються при НВІС-реалізаціях. Час і спосіб виконання операції Φ_i операційним блоком є одними із основних параметрів при визначенні конвеєрного такту роботи $T_к$ і розрядності каналів надходження даних $n_к$ для алгоритмічних структур реального часу. Результатом першого етапу розроблення є граф-схема алгоритму, де функціональні оператори Φ_i мають приблизно однаковий час виконання, а їх складність визначається засобами реалізації.

На етапі проектування комунікацій для конвеєрної алгоритмічної реалізації нейроалгоритму необхідно визначити структуру каналів обміну даними між функціональними операторами Φ_i. Для цього виконується перехід від граф-схеми алгоритму до потокового графу, в якому здійснюється просторово-часове розміщення і закріплення функціональних операторів Φ_i за ярусами. Структура зв'язків у потоковому графі між функціональними операторами Φ_{jk} сусідніх ярусів визначає кількість каналів надходження даних m_j і структуру з'єднань між операційними блоками при апаратній реалізації алгоритму.

За результатами перших двох етапів розроблення отримуємо потоковий граф, який дає змогу оцінити обчислювальну здатність $D_к$ апаратних засобів НС. Вихідними даними для визначення обчислювальної здатності $D_к$ є $m_к$ – кількість каналів надходження даних у сходинці конвеєра, $n_к$ – розрядність каналів даних у сходинці конвеєра, складність функціональних операторів Φ_i, особливості і швидкодія елементної бази. Використовувана елементна база та її швидкодія є визначальними для оцінки такту роботи $T_к$ апаратних засобів. Для оцінювання узгодженості інтенсивності надходження даних P_d із обчислювальною здатністю $D_к$ вводиться коефіцієнт узгодженості, який визначається так:

$$R = \frac{D_k}{P_d}.$$

Узгодження обчислювальної здатності $D_к$ апаратних засобів НС із інтенсивністю надходження даних P_d досягається шляхом зміни тривалості такту $T_к$ або кількості $m_к$ і розрядності $n_к$ каналів надходження даних у потоковому графі. Розробляючи узгоджений потоковий граф алгоритму для реалізації апаратних засобів з високою ефективністю використання обладнання, необхідно насамперед максимально використати особливості та швидкодію елементної бази, тобто визначити складність функціональних операторів Φ_i і мінімізувати тривалість такту $T_к$.. Змінювати кількість $m_к$ і розрядність $n_к$ каналів у потоковому графі, яка напряму пов'язана з використовуваною елементною базою. Зміна параметрів потокового графу алгоритму має забезпечити узгодження обчислювальної здатності $D_к$ апаратних засобів НС із інтенсивністю надходження даних P_d.

У випадку, коли R=1, то розроблений узгоджений потоковий граф забезпечує отримання узгоджено-паралельної структури НС реального часу з високою ефективністю використання обладнання.

Якщо R<1, то для забезпечення обробки потоків даних у реальному часі необхідне збільшення обчислювальної здатності $D_к$, якого можна досягти збільшенням кількості каналів $m_к$, їх розрядності $n_к$ або зменшенням такту конвеєра $T_к$., що досягається зменшенням складності функціональних операторів Φ_i. У випадку, коли зміною перерахованих параметрів не вдається досягти необхідної обчислювальної здатності $D_к$, то тоді використовується паралельне включення НС, кількість яких визначається виразом:

$$h = \lceil 1/R \rceil$$

де $\lceil\rceil$ – знак округлення до більшого цілого.

У випадку, коли R>1, для забезпечення високої ефективності використання обладнання необхідно переходити до третього етапу розроблення – укрупнення функціональних операторів Φ_{jk}. На цьому етапі об'єднують функціональні оператори Φ_{jk} і канали передачі даних у ярусах потокового графу, або об'єднання функціональних операторів сусідніх ярусів. Граф алгоритму, який отримаємо у результаті такого об'єднання, будемо називати узгодженим потоковим графом. Коефіцієнт об'єднання v визначається так:

$$v \le \lfloor D_к/P_d \rfloor,$$

де $\lfloor\rfloor$ – знак округлення до меншого цілого.

При об’єднанні функціональних операторів v сусідніх ярусів утворюється один ярус функціональних макрооператорів, у якому за v ітерацій виконується обчислення таких функціональних макрооператорів V. Обчислення у середині ярусу здійснюється із тактом $T_к$, який визначає складність функціональних операторів $Ф_{jk}$, а між ярусами з макротактом $T_{мк} = vT_к$. Об’єднання функціональних операторів сусідніх ярусів приводить до зменшення у v разів кількості ярусів. Таке об’єднання доцільно здійснювати, коли яруси потокового графу є однотипними. Інше укрупнення здійснюють шляхом об’єднання функціональних операторів і каналів передачі даних у межах ярусу. Для випадку, коли $v \geq L$ укрупнення здійснюють шляхом лінійної проекції, при якій всі функціональні оператори ярусу потокового графу відображаються в один функціональний макрооператор, а канали передачі даних – в оператор затримки та перестановки даних. Для забезпечення узгодженості може використовуватися комбіноване укрупнення, яке передбачає об’єднання функціональних операторів і каналів передачі даних як у середині ярусу, так і між ярусами.

Структура узгодженого потокового графу нейроалгоритму відображається орієнтованим графом G=<V,E,D(E)>, де V – функціональні макрооператори ярусу; E – орієнтовані дуги, які моделюють зв’язки між функціональними макрооператорами. Кожна дуга e∈ E зв’язує вихід одного функціонального макрооператора із входом другого та володіє вагою, що дорівнює значенню затримки D(e).

Етап укрупнення тісно пов’язаний з етапом планування, на якому після об'єднання функціональних операторів для збереження інформації про структуру потокового графу нейроалгоритму планують обчислення, визначають величини затримок і перестановки даних. Для відтворення обчислень у кожний ярус узгодженого графу вводяться оператори управління затримки та перестановки даних.

При апаратному відображенні узгодженого потокового графу алгоритму кожному функціональному макрооператору V ставляться у відповідність багатофункціональні операційні пристрої, які забезпечують виконання операцій ярусу, операторам затримки – буферна паралельна пам’ять, яка може забезпечити необхідну затримку та перестановку даних, а операторам управління – пристрої керування, які керують багатофункціональними операційними пристроями і буферною пам’яттю.

Процес розроблення узгодженого потокового графу алгоритму є ітераційним, він тісно пов’язаний з покращенням характеристик нейроалгоритму.

Базові структури нейрокомп’ютерних систем реального часу

Залежно від вимог застосування, частоти зміни задач, алгоритму їх розв’язання, розміру та частоти надходження вхідних даних можуть бути синтезовані різні структури нейрокомп’ютерних систем реального часу, які відрізняються як організацією обчислень, так і технічними параметрами. Задача опису всіх можливих структур є нерозв'язною. Тому доцільн виділити і дослідити узагальнені базові структури, на основі яких можуть бути синтезовані нейрокомп’ютерні системи реального часу для розв’язання конкретних задач з потрібними параметрами. Аналіз нейрокомп’ютерних систем реального часу показує, що їхні структури можуть бути спеціалізованими або проблемно-орієнтованими, містити як спеціалізовані, так і універсальні засоби.

Структури спеціалізованих нейрокомп’ютерних систем реального часу наведені на рис. 1, де БПП – буферна паралельна пам’ять, ПР – пам’ять реконфігурацій. Ці структури апаратно відображають узгоджений потоковий граф нейроалгоритму розв’язання задачі.

Структура спеціалізованої нейрокомп’ютерної системи реального часу на основі реконфігурованого обчислювального поля (рис. 1, а) є гнучкішою від структури на основі алгоритмічного нейропроцесора (рис. 1, б), оскільки вона забезпечує оперативне переналаштування алгоритму розв’язання задачі. Структуру системи на основі алгоритмічного нейропроцесора доцільно використовувати у випадку, коли алгоритми є повністю відпрацьованими і не будуть змінюватися у процесі експлуатації. Перевагою такої структури є висока ефективність використання обладнання.

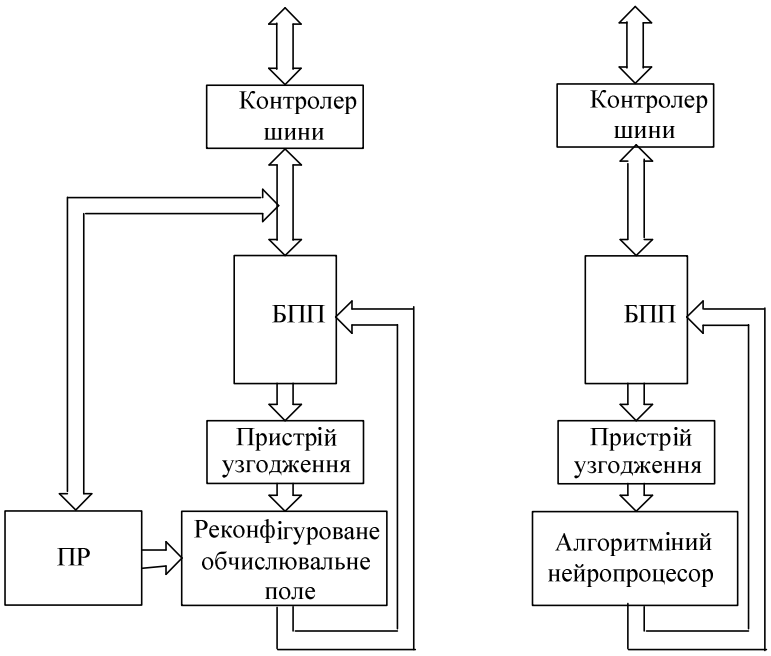

Рис.1. Структури спеціалізованих нейрокомп'ютерних систем реального часу:
а – на основі реконфігурованого обчислювального поля;
б – на основі алгоритмічного нейропроцесора

Структури проблемно-орієнтованих нейрокомп'ютерних систем реального часу з використанням процесора цифрової обробки сигналів (ПЦОС) і апаратного розширювача наведені на рис. 2.

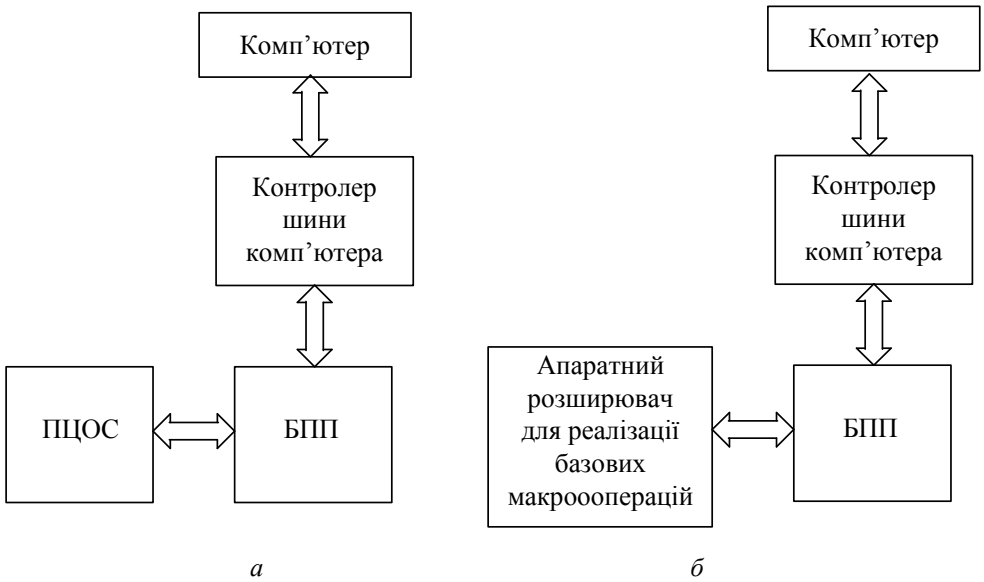

Рис. 2. Структури проблемно-орієнтованих нейрокомп'ютерних систем реального часу:
а – з використанням ПЦОС; б – апаратного розширювача

У проблемно-орієнтованих нейрокомп'ютерних системах реального часу ППОС і апаратний розширювач реалізують найчасоємніші базові макрооперації нейроалгоритмів. Для зменшення часу та синхронізації обміну між комп'ютером і ПЦОС (апаратним розширювачом) в проблемно-орієнтованих нейрокомп'ютерних системах реального часу використовується БПП.

Висновки

1. Розробляти нейрокомп'ютерні системи реального часу доцільно на основі інтегрованого підходу, який охоплює сучасну елементну базу, методи та алгоритми реалізації базових операцій нейроалгоритмів, архітектури компонентів і систем, враховує вимоги конкретних застосувань, інтенсивності надходження даних і ґрунтується на таких принципах побудови: конвеєризації та просторового паралелізму обробки даних; модульності; однорідності та регулярності архітектури; спеціалізації та адаптації апаратно-програмних засобів до структури алгоритмів обробки та інтенсивності надходження даних.

2. Для вибору апаратних компонентів нейрокомп'ютерних систем реального часу запропоновано використовувати критерій ефективності використання обладнання, який враховує кількість виводів інтерфейсу, однорідність структури, кількість і локальність зв'язків, пов'язує продуктивність з витратами обладнання та оцінює елементи системи за продуктивністю.

3. Основними шляхами підвищення ефективності використання обладнання при реалізації НВІС-компонентів нейрокомп'ютерних систем реального часу є: вибір ефективних методів і алгоритмів реалізації для НВІС-реалізації; зменшення розрядності операційних пристроїв, пам'яті, кількості і розрядності каналів передавання даних; узгодження інтенсивності надходження даних із обчислювальною здатністю апаратних засобів на всіх рівнях.

4. Визначено, що узгодити інтенсивність надходження даних із обчислювальною здатністю НВІС-компонентів можна так: зміною тривалості конвеєрного такту, кількості і розрядності каналів надходження даних в операційних пристроях.

5. Основними етапами синтезу НВІС-компонентів нейрокомп'ютерних систем реального часу є: вибір та розроблення методів і алгоритмів узгоджено-паралельної обробки; визначення основних параметрів апаратних засобів; перехід від алгоритму до узгодженої паралельної структури.

1. Круглов В.В., Борисов В.В. Искусственные нейронные сети. Теория и практика. – 2-е изд., стереотип. – М.: Горячая линия-Телеком, 2002. – 382 с. 2. Хайкин С. Нейронные сети: Полный курс, 2-е изд.: Пер. с англ.. –М.: Вильямс, 2006. 3. Галушкин А.И. Нейрокомпьютеры. Кн. 3. – М.: ИПРЖР,2000.-528с. 4. Осовский С. Нейронные сети для обработки информации / Пер. с польск. – М.: Финансы и статистика, 2002. – 344 с. 5. Палагин А.В., Опанасенко В.Н. Реконфигуриремые вычислительные системы. – К.: Просвіта, 2006. – 280 с. 6. Ткаченко Р.О. Нова парадигма штучних нейронних мереж прямого поширення // Вісник Держ. ун-ту "Львівська політехніка": Комп'ютерна інженерія та інформаційні технології. – 1999. – № 386. – С. 43–54. 7. Цмоць І.Г. Інформаційні технології та спеціалізовані засоби обробки сигналів і зображень у реальному часі. – Львів: УАД, 2005. – 227 с. 8. Грицик В.В., Ткаченко Р.О. Нові підходи до навчання штучних нейромереж // Доповіді Національної академії наук України. – 2002. – № 11. – С. 59–65. 9. Грушицкий Р.И., Мурсаев А.Х., Угрюмов Е.П. Проектирование систем на микросхемах программируемой логики. – СПб.: БХВ-Петербург, 2002. – 608 с. 10. Воеводин В.В., Воеводин В.В. Параллельные вычисления. – СПб.: БХВ-Петербург, 2002. – 608 с. 11. Немнюгин С.А., Стесик О.Л. Параллельное программирование для многопроцессорных систем. – СПб.: БХВ – Петербург, 2002. – 400 с. 12. Касьянов В.Н., Евстигнеев В.А. Графы в программировании: обработка, визуализация и применение. – СПб.: БХВ – Петербург, 2003. – 1104 с. 13. Ткаченко Р.О., Ткаченко П.Р., Цмоць І.Г. Апаратна реалізація багатошарових перцептронів з неітераційним навчанням // Збірн. Наук.х пр. – К.: Інститут проблем моделювання в енергетиці НАН України, 2005. –. Вип. 29. – С. 103–113.