Міністерство освіти і науки України
Тернопільський національний економічний університет
Факультет комп'ютерних інформаційних технологій
Кафедра комп'ютерної інженерії

Кураш Ярина Ярославівна

Апаратна реалізація нейроелемента вертикально-групового типу / Hardware implementation of vertically group type neural element

Спеціальність 8.091501 – Комп'ютерні системи та мережі

Дипломна робота за освітньо-кваліфікаційним рівнем «магістр»

Науковий керівник
к.т.н., професор Цмоць І. Г.
_____

Дипломну роботу допущено до захисту
«___»_____ 20 __ р.
Зав. кафедри КІ
Березький О.М. _____

Тернопіль – 2017

РЕЗЮМЕ

Дипломна робота на тему «Апаратна реалізація нейроелементів вертикально-групового типу» на здобуття освітньо-кваліфікаційного рівня «Магістр» зі спеціальності «Комп'ютерні системи та мережі» написана обсягом 92 сторінка і містить 30 ілюстрацій, 2 таблиць, 4 додатків та 42 джерел за переліком посилань.

Метою роботи є розроблення апаратної реалізації нейроелементів вертикально-групового типу та створення високопродуктивних апаратних нейромереж реального часу синтезованих на основі швидкодіючих нейроелементів.

Методи досліджень. Для розв'язання поставлених задач у дипломній роботі використано: методи: групового опрацювання даних у нейромережах; моделі: нейромережі вертикально-групового типу для структури нейроелемента з можливістю зміни розрядності каналів надходження і кількості розрядів множника у групі, що одночасно аналізуються для формування часткових добутків.

Результати дослідження: швидкодія обчислення оператора групового підсумовування, для досягнення комплексного використання вертикального та багатооперандного підходів, при якому обчислення ґрунтується на вертикальних операціях і розглядається як єдиний процес підсумовування.

Орієнтовні напрямки розвитку досліджень: застосовуються в таких галузях як інформатика, економіка, медицина, зв'язок, інтернет, автоматизації виробництва, робототехніка, , введенні і обробці інформації.

КЛЮЧОВІ СЛОВА: НЕЙРОЕЛЕМЕНТ, ПАРАЛЕЛЬНИЙ МЕТОД ВЕРТИКАЛЬНО-ГРУПОВОГО ОПРАЦЮВАННЯ ДАНИХ, МОДЕЛЬ ФОРМАЛЬНОГО НЕЙРОНА, РЕАЛЬНИЙ ЧАС, ГРУПОВЕ ПІДСУМОВУВАННЯ, БАГАТОРОЗРЯДНИЙ СУМАТОР, НВІС-СТРУКТУРА.

RESUME

Diploma work: " Hardware implementation of vertically group type neural element " to education and qualification of "Master" specialty "Computer systems and networks" written 92 page volume and contains 30 illustrations, 2 tables, 4 applications and 42 sources for references.

The aim is to develop hardware implementation neuro element vertical-type group and the creation of high-performance hardware neural network real-time synthesized based on speed neuro element.

Research Methods. To solve the tasks of the thesis work used: Methods: group data processing in neural networks; Model: neural network group vertical type structure for neural elements bit with the ability to change channels revenues and the number of bits of the multiplier in the group simultaneously analyzed to form the partial products.

Results: speed calculation summation operator group to achieve integrated use of vertical and many operant approach, in which the calculation is based on vertical transactions and is regarded as a single process summation.

The estimated directions of research: applied in such fields as computer science, economics, medicine, communication, Internet, automation, robotics, input and processing of information.

KEY WORDS: NEURO ELEMENT, PARALLEL METHOD OF VERTICALLY-GROUP DATA PROCESSING. FORMAIL NEURON MODEL, REAL TIME, GROUP SUMMAATION, MULTI ADDER, VLSI-STRUCTURE.

ЗМІСТ

Вступ…… 8
1 Аналіз галузей застосування, архітектура та апаратні засоби реалізації штучних нейронних мереж……. 11
1.1 Галузі застосування апаратних нейронних мереж…….. 11
1.2 Структури штучних нейронних мереж і нейроелементів…….. 13
1.3 Апаратні засоби реалізації штучних нейронних мереж та нейроелементів……... 34
2 Розробка алгоритмів реалізації паралельного нейрона вертикально-групового типу……. 42
2.1 Формування вимог до апаратної реалізації нейроелемента……... 42
2.2 Алгоритм реалізації паралельного нейроелемента вертикально-групового типу…… 47
2.3 Вибір принципів побудови та варіантів реалізації паралельного нейроелемента вертикально-групового типу……... 50
2.4 Модель формального нейрона вертикально-групового типу з мультиплексуванням шин вагових коефіцієнтів і даних…… 56
3 Розробка, моделювання компонентів і синтез паралельного нейрона вертикально-групового типу…….. 64
3.1 Вибір апаратних засобів моделювання нейроелемента…….. 64
3.2 Структура нейроелемента вертикально-груповго типу…….. 71
3.3 Реалізація нейроелемента вертикально-групового типу з використанням багатовходового суматора……. 75
Висновки…..…… 85
Список використаних джерел…… 87
Додаток А Комбінаційна схема 7-входового однорозрядного суматора ….. 94
Додаток Б Лістинг коду програми……..…… 95
Додаток В Довідка про впровадження…..……..
Додаток Г Апробація…….……..

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

ШНМ – штучні нейронні мережі

ДАП – двонаправлена асоціативна пам'ять

ПЛІС – програмовані логічні інтегральні схеми

НВІС – надвеликі інтегральні схеми

ПЕ – процесорні елементи

ФЧД – формувач часткових добутків

ТФА – таблиця функції активності

САПР – система автоматизованого проектування і розрахунку

HDL - Hardware Description Language

RTL - Register TransferLlevel

VHDL - Very high speed integrated circuits Hardware Description Language

FPGA - Field-Programmable Gate Array

ВСТУП

Актуальність теми. «Апаратні засоби нейроелементів вертикально-групового типу» є використання високопродуктивних апаратних нейромереж реального часу синтезованих на основі швидкодіючих нейроелементів, в яких висока швидкодія досягається за рахунок розпаралелювання обчислень як в часі, так і в просторі. Також створення високоефективних нейромережевих засобів реального часу, яке потребує широкого використання сучасної елементної бази, розроблення нових моделей нейрона, орієнтованих на реалізацію у надвеликих інтегральних схемах (НВІС), методів, алгоритмів і спеціалізованих НВІС-структур для реалізації нейроелементів та нейромереж. З огляду на це особливої актуальності набуває завдання синтезу нейроелементів і нейромереж, орієнтованих на опрацювання даних у реальному часі та НВІС-реалізацію з високою ефективністю використання обладнання.

Метою дослідження є розроблення паралельного методу вертикально-групового опрацювання даних у нейромережах, моделі та структури паралельного нейроелемента вертикально-групового типу з можливістю зміни розрядності каналів надходження і кількості розрядів множника у групі, що одночасно аналізуються для формування часткових добутків.

Завдання дослідження. Для досягнення мети необхідно розвязати наступні завдання:

1) проаналізувати галузі застосування, архітектури та апаратні засоби реалізації штучних нейронних мереж;

2) розробити алгоритми реалізації паралельного нейрона вертикально-групового типу;

3) здійснити, моделювання компонентів і синтез паралельного нейрона вертикально-групового типу.

Об’єктом дослідження є процеси опрацювання даних в нейромережах.

Предметом дослідження є методи, алгоритми, та засоби реалізації

нейроелемента нейронних мереж

Методи досліджень. При розв’язанні завдань застосовується метод групового підсумовування та синтез семивходового однорозрядного суматора.

Наукова новизна одержаних результатів. Науковою новизною роботи є:

1) розроблений вертикально-груповий алгоритм обчислення скалярного добутку, який за рахунок вибору кількості розрядів для аналізу множників забезпечує високу ефективність використання обладнання;

2) розроблений нейрон забезпечує адаптацію структури до інтенсивності надходження даних, що забезпечує високу ефективність використання обладнання;

3) розроблений багатовходовий суматор забезпечує зменшення часу підсумовування.

Практичне значення отриманих результатів. Розроблені алгоритми та структура нейроелемента може бути використана при синтезі нейромереж реального часу з високою ефектиністю використання обладнання. Більшість відомих моделей штучного нейрона є аналогами або модифікаціями розглянутих вище чотирьох моделей нейрона. Аналіз відомих моделей нейронів показав, що вони не орієнтовані на НВІС-реалізацію, оскільки не ґрунтуються на елементарних арифметичних операціях і вимагають значної кількості виводів.

Для розроблення паралельного методу вертикально-групового опрацювання даних у нейромережах, який у порівнянні з відомими забезпечує підвищення швидкодії шляхом збільшення розрядів каналів надходження множників і кількості часткових добутків, які формуються у результаті їхнього аналізу вибрано принципи побудови, розроблено модель та структуру формального нейрона вертикально-групового типу з мультиплексуванням шин вагових коефіцієнтів і даних. Модель забезпечує узгодження інтенсивності надходження даних із обчислювальною здатністю

нейроемента шляхом зміни розрядності каналів надходження і кількості розрядів множників у групі, які одночасно аналізуються для формування часткових добутків.

Публікації та апробація ДР. Результати наукового дослідження опубліковано в матеріалах VI Всеукраїнської школи-семінару молодих вчених і студентів ACIT’2016 [1].

Впровадження результатів ДР. Впровадження наукових результатів здійснено в тернопільській компанії «Олімп».

1 АНАЛІЗ ГАЛУЗЕЙ ЗАСТОСУВАННЯ, АРХІТЕКТУРА ТА АПАРАТНІ ЗАСОБИ РЕАЛІЗАЦІЇ ШТУЧНИХ НЕЙРОННИХ МЕРЕЖ

1.1 Галузі застосування апаратних нейронних мереж

"Штучна нейронна мережа" виникла в галузі штучного інтелекту в процесі певного моделювання структури мозку, що надає змогу живим істотам навчатися, виправляючи власні помилки. Цим терміном позначають певний клас математичних моделей та їх програмних або апаратних реалізацій, які побудовані за принципами організації й функціонування біологічних нейронних мереж, тобто мереж нервових клітин всіх живих організмів.

Особливість нейронних мереж полягає в тому, що їх застосування, завдяки так званому навчанню на наявних даних, дає можливість зпрогнозувати, які значення прийматимуть досліджуванні змінні у нових спостереженнях, посилаючись на дані попередніх спостережень. При грамотному застосуванні нейронних мереж точність таких прогнозів значно перевищує точність прогнозів, здійснених за допомогою класичних статистичних методів. Завдання прогнозування вирішується нейронними мережами аналогічно завданню розпізнавання образів, а загальною умовою застосування нейронних мереж в прогнозуванні є наявність "історичних даних", за допомогою яких мережа може «навчитись». Крім того, певні типи нейронних мереж (наприклад, мережі Кохонена) можуть застосовуватись й без навчаючих вибірок для вирішення завдань класифікації та кластеризації, що розширює можливості типологічного аналізу, оскільки з’являється новий інструмент перевіркиякості типологій, які побудовані завдяки застосуванню інших методів [35].

Нейромережі - це гнучкий і потужний набір інструментів для вирішення різноманітних завдань обробки та аналізу даних. У теперішній час нейронні мережі застосовуються у різних галузях: інформатиці, економіці, медицині, зв’язку, інтернеті, автоматизації виробництва, робототехніці,

політичній і соціальній технології, безпеці і охоронній системі, введенні і обробці інформації та геологорозвідці. Застосовують їх і в соціології, але поки ще досить рідко.

Застосовуються нейронні мережі в економіці і бізнесі для прогнозування часових рядів (курсів валют, цін на сировину, обсягів продажів), автоматичний трейдинг (торгівля на валютній, фондовій або товарній біржі), оціннюють ризик неповернених кредатів, пророкують банкрутство, оцінка вартості нерухомості, виявлення переоцінених і недооцінених компаній, оптимізація товарних і грошових потоків, зчитування і розпізнавання чеків і документів.

В медицині це постановка діагнозу, обробка медичних зображень, моніторинг стану пацієнта, аналіз ефективності лікування, очищення показань приладів від шумів.

В таких галузях як «зв'язок» нейронні мережі застосовуються для стиснення відеоінформації, швидкого кодування-декодування, оптимізації стільникових мереж і схем маршрутизації пакетів, а в інтернеті - асоціативний пошук інформації, електронні секретарі та автономні агенти в інтернеті, фільтрація і блокування спаму, автоматична рубрикація повідомлень з стрічок новин, адресні реклама і маркетинг для електронної торгівлі.

Автоматизація виробництва слугує для оптимізації режимів виробничого процесу, контроль якості продукції, моніторинг і візуалізація багатовимірної диспетчерської інформації, попередження аварійних ситуацій.

Робототехніка дає можливість розпізнавати сцени, об'єкти і перешкоди перед роботом, прокладку маршруту руху, керування маніпуляторами, підтримання рівноваги.

Політологічні та соціологічні технології дозволяють передбачувати результат виборів, аналіз опитувань, пророкування динаміки рейтингів, виявлення значущих чинників, дослідження і візуалізація соціальної

динаміки населення.

Завдяки безпеці та охороні системи: розпізнавання осіб, ідентифікація особи за відбитками пальців, голосу, підпису або особі, розпізнавання автомобільних номерів, аналіз аерокосмічних знімків, моніторинг інформаційних потоків в комп'ютерній мережі і виявлення вторгнень, виявлення підробок, аналіз даних з відеодатчиків і різноманітних сенсорів.

Введення і обробка інформації це розпізнавання рукописних текстів, відсканованих поштових, платіжних, фінансових та бухгалтерських документів.

1.2 Структури штучних нейронних мереж і нейроелементів

Штучна нейронна мережа є структурою, яка складається з великої кількості процесорних елементів, кожен з яких має локальну пам’ять і може взаємодіяти з іншими процесорними елементами за допомогою комунікаційних каналів з метою передачі даних, що можуть бути інтерпретовані довільним чином. Процесорні елементи незалежно в часі обробляють локальні дані, що поступають до них через вхідні канали. Зміна параметрів алгоритмів такої обробки залежить тільки від характеристик даних [52].

В загальному випадку штучні нейронні мережі – це обчислювальні парадигми, які реалізують спрощені моделі біологічних нейронних мереж (БНМ). Під БНМ будемо розуміти локальні ансамблі нейронів, які об’єднані синаптичними зв’язками. Сукупність таких ансамблів формує мозок із його різноманітними функціональних можливостями.

Сьогодні відома велика кількість нейронних структур та їх модифікацій, що орієнтовані на вирішення конкретного типу задач. Найбільш відомі типи таких структур показані на рисунку 1.1 [35].

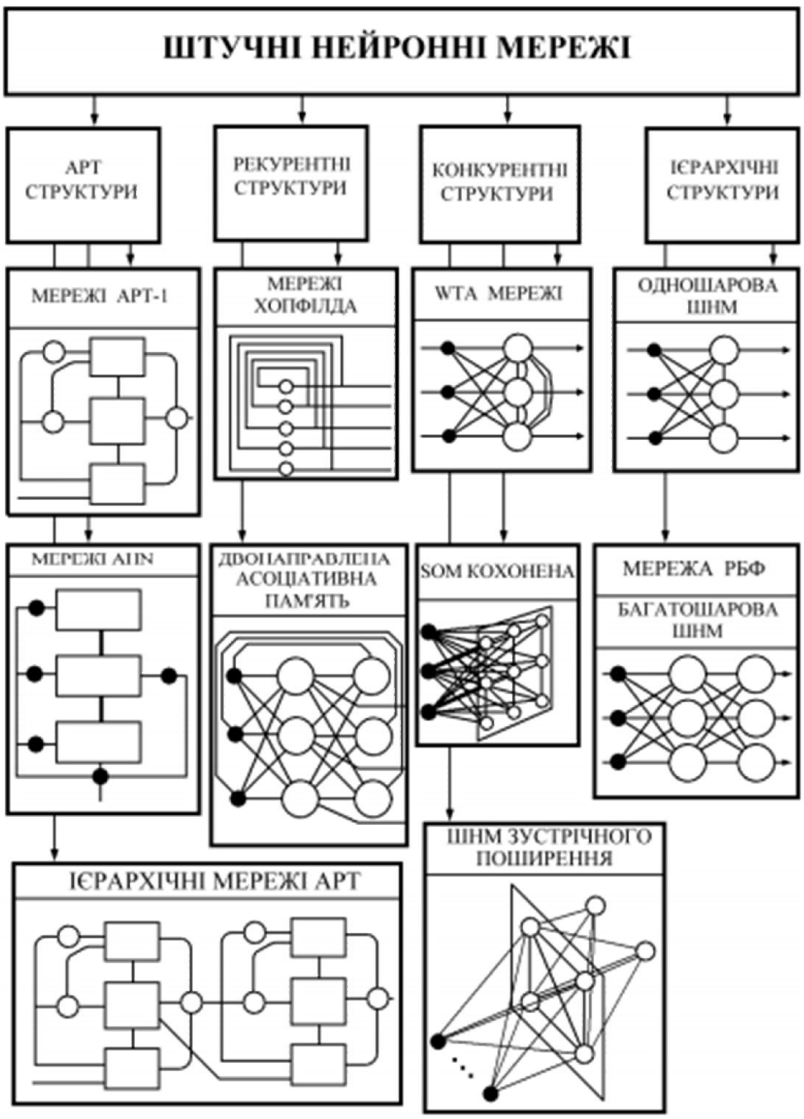


Рисунок 1.1 – Структура штучних нейронних мереж

ШНМ мають такі властивості:

1) локальна обробка інформації в штучному нейроні, який є базовою структурною одиницею мережі;

2) паралелізм, результатом якого є вирішення глобальної задачі шляхом представлення її у вигляді множини локальних задач, що тісно взаємодіють між собою;

3) здатність до навчання, яке підвищує ефективність роботи мережі;

4) здатність до розподіленого зберігання знань, які були одержані в ході навчання.

ШНМ задають у вигляді направлених графів, вершинами яких є

нейрони, а ребрами даних мереж є міжнейронні зв’язки.

Архітектури сучасних нейронних мереж найчастіше поділяють на три категорії:

1) мережі з повним набором міжнейронних зв’язків;
2) мережі з фіксованим індексом оточення;
3) мережі з пошаровою структурою.

У ШНМ із повним набором міжнейронних зв’язків забезпечується можливість взаємодії кожного нейрона мережі з будь-яким іншим. На рисунку 1.2 наведений приклад повного з’єднання чотирьох нейронів.

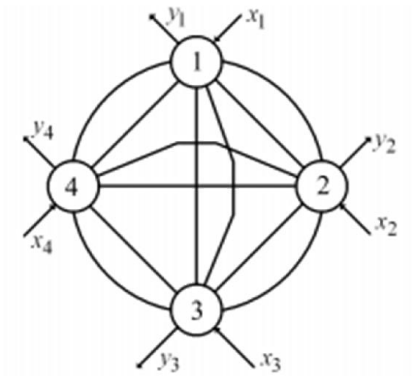


Рисунок 1.2 – З’єднання чотирьох нейронів

Структура з повним з’єднанням є узагальненою структурою, оскільки всі інші довільні об’єднання нейронів можуть розглядатися як підмножини даної структури. Тому ШНМ із повним з’єднанням є універсальним середовищем для реалізації мережних алгоритмів. Широке використання таких структур обмежується недоліком, який полягає в значному зростанні кількості міжнейронних зв’язків при збільшенні кількості нейронів [51].

У випадку, коли необхідно використовувати структури з великою кількістю нейронів, застосовують кліткові структури з фіксованим індексом оточення. На рисунку 1.3 наведений приклад структури такого типу з індексом оточення чотири.

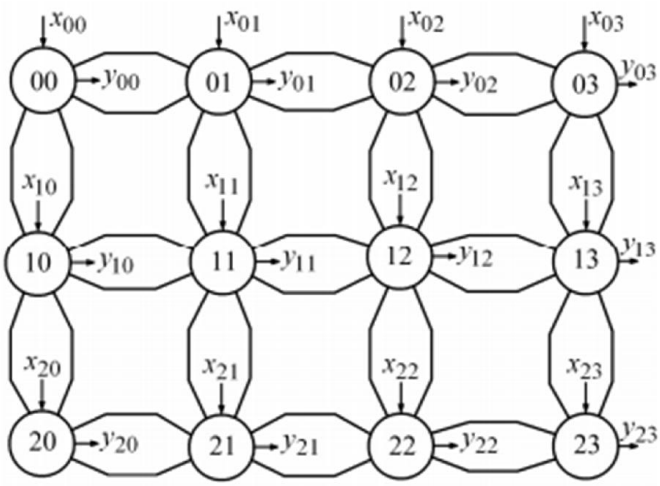


Рисунок 1.3 – Кліткова нейронна мережа з індексом чотири

Ще однією важливою характеристикою нейронних мереж із постійним індексом оточення є модульність. Потужність структури у цьому випадку може нарощуватись простим додаванням елементів без зміни ідеології алгоритму, що на ній працює.

Нейронні структури з повним з'єднанням можуть бути як одношаровими, так і багатошаровими. У одношаровій структурі з повним з'єднанням всі вхідні сигнали можуть поступати на всі нейрони. Класичною структурою даного типу є персептрон Розенблатта.

Багатошарові мережі з повним з'єднанням забезпечують можливість передачі інформації з кожного нейрона попереднього шару на будь-який нейрон наступного. Найчастіше це — багатошарові персептрони.

Найбільш поширеними є структуровані за шарами нейронні мережі, які, в залежності від свого функціонального призначення, можуть містити однотипні або різнотипні нейрони. Виходячи з шарової структури ШНМ, характер міжнейронних зв'язків має свої міжшарові та внутрішньошарові особливості. На рисунку 1.4 показана класифікація видів міжнейронних зв'язків [9].

Рисунок 1.4 – Основні вид міжнейронних зв’язків

У випадку прямого міжшарового поширення інформація передається однонаправлено у напрямку зростання номера шару. Пряме поширення в межах одного шару використовують у випадку, коли група нейронів даного шару з’єднана з попереднім шаром опосередковано через виділені нейрони. Двонаправлене поширення допускає також зворотну передачу, що дозволяє створювати алгоритми, за якими враховувався б взаємний між нейронний обмін.

Таким чином, однонаправленість зв’язків призводить до побудови виключно ієрархічних структур, у яких обробка інформації розподіляється по рівнях. За кожний рівень ієрархічної обробки інформації відповідає свій шар нейронів. Вихідна інформація більш високого рівня попереднього шару є вхідною для нейронів наступного шару, який забезпечує глибший рівень обробки.

Двонаправлені міжшарові зв’язки необхідні для реалізації рекурентних структур, які дають можливість застосування ітераційних алгоритмів. Загальною рисою таких структур є те, що подальша передача інформації відбувається тільки у випадку завершення ітераційного процесу.

Двонаправлені зв’язки у межах одного шару використовуються для створення конкуруючих груп нейронів. При активації сигналом з попереднього шару кожен з нейронів передає сигнал активації нейронам своєї групи та сигнал гальмування всім іншим нейронам. В результаті конкурентоздатною стає та група нейронів, що одержала найбільше збудження [57].

1.2.1 Ієрархічна структура одношарової нейронної мережі прямого поширення

Першою штучною нейронною мережею, яку за сучасною термінологією відносять до одношарових, був персептрон Розенблатта, який показано на рисунку 1.5.

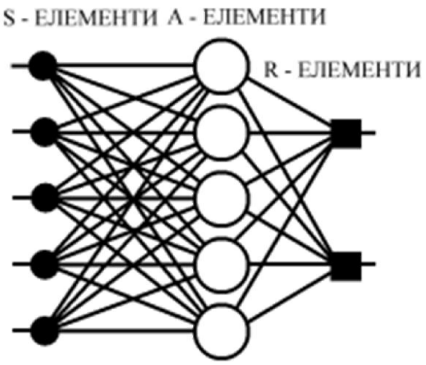


Рисунок 1.5 – Структура персептрона Розенблата

Оскільки персептрон розглядався автором як модель роботи мозку, то елементи його структури відповідали елементам простої рефлекторної мережі. S-елементи моделюють роботу сенсорних клітин, які збирають інформацію про навколишнє середовище та передають її до А-елементів, які насправді і є формальними нейронами з пороговою активаційною функцією.

Для формування реакції персептрона призначені R-елементи, що мають фіксовані вхідні ваги для визначення внеску кожного А-елемента.

Сучасним прототипом персептрона Розенблатта є одношарова

нейронна мережа прямого поширення, яку показано на рисунку 1.6.

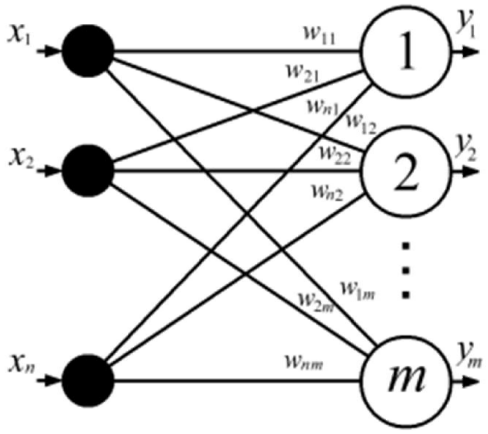


Рисунок 1.6 – Структура нейронної мережі прямого поширення

Вона складається з m нейронів, які одночасно здатні прийняти вхідний вектор сигналів X = $(x_1, \dots, x_i, \dots, x_n)$. Для розмноження елементів x_i цього вектора використовують спеціальні пристрої, які показані зліва від нейронів.

Ці пристрої не виконують обробки інформації, тому не вважаються шаром нейронної мережі. Згідно з моделлю формального нейрона, кожен з його вхідних сигналів множать на ваговий коефіцієнт w_{ij}, де i — поточний номер елемента вектора X, а j — поточний номер нейрона.

Всі вагові коефіцієнти одношарової нейронної мережі утворюють матрицю вагових коефіцієнтів:

$$W = \begin{Bmatrix} w_{11} & \dots & w_{1j} & \dots & w_{1m} \\ w_{i1} & \dots & w_{ij} & \dots & w_{im} \\ w_{n1} & \dots & w_{nj} & \dots & w_{nm} \end{Bmatrix} \quad (1.1)$$

Тоді вектор вектор вихідних сигналів є вектором значень активаційних функцій:

$$Y = F(V)\begin{cases} f_1(v_1), \\ \dots \\ f_j(v_j), \\ \dots \\ f_m(v_m). \end{cases} \quad (1.2)$$

Згадані функції можуть бути однаковими для всіх нейронів, що складають мережу. У цьому випадку її називають гомогенною. Мережу, в якій вигляд активаційної функції залежить від номера нейрона, називають гетерогенною.

На рисунку 1.6 показаний загальний випадок одношарової нейронної мережі, у якій кожен нейрон приймає участь в обробці всіх елементів вхідного вектора даних. Такий підхід не завжди є економічно та технічно виправданим. Тому для вирішення конкретних задач можуть використовуватись архітектури зі структурою зв'язків, яка є підмножиною повної зв'язності [40].

1.2.2 Ієрархічна структура багатошарової нейронної мережі прямого поширення

Спроби застосувати одношарові нейронні мережі для розв'язування широкого кола задач наштовхнулися на ряд труднощів, пов'язаних з проблемою лінійної роздільності. Природним вирішенням цієї проблеми стало застосування багатошарових ШНМ, що нагадують багатошарові структури мозку.

Розглянемо ієрархічну структуру, в якій нейрони структуровані за шарами. Вона складається з m нейронів першого (прихованого) шару, які одночасно здатні прийняти вхідний вектор сигналів $X = (x_1, \dots, x_i, \dots, x_n)$, та k нейронів другого (вихідного) шару. Така структура зображена на рисунку 1.7 і є узагальненою структурою багатошарової нейронної мережі прямого поширення.

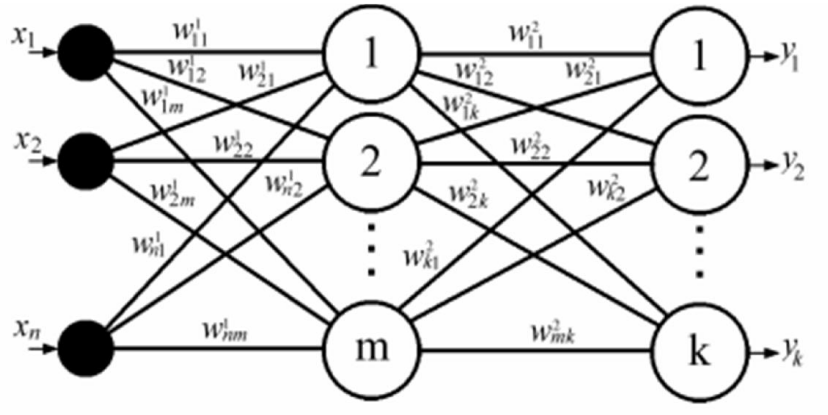


Рисунок 1.7 – Структура багатошарової нейронної мережі прямого поширення

У нейронних мережах прямого поширення синаптичні зв'язки організовані таким чином, що кожний нейрон даного рівня ієрархії сприймає інформацію тільки від деякої непустої множини нейронів, які розташовані на більш низькому рівні. Назва мереж вказує на те, що у них існує виділений напрям поширення сигналів, які рухаються, починаючи з входу, через один або декілька прихованих шарів до вихідного шару [41].

Легко помітити, що багатошарова нейронна мережа може бути одержана шляхом каскадного об'єднання одношарових мереж з матрицями вагових коефіцієнтів $W^1, W^2, \ldots, W^p$, де p — кількість шарів нейронної мережі.

У випадку лінійності активаційних функцій багатошарова нейронна мережа може бути зведена до еквівалентної одношарової з матрицею вагових коефіцієнтів $W = W^1 \cdot W^2 \cdot \ldots \cdot W^p$, тому формування подібних структур має сенс тільки у випадку застосування у нейронах нелінійних активаційних функцій.

1.2.3 Ієрархічна структура РБФ-мережі

Штучні нейронні мережі, які використовують радіальні базисні функції, називають РБФ-мережами. Вони є окремим випадком двошарових нейронних мереж прямого поширення, в яких прихований шар нейронів використовує радіальні базисні функції типу гаусової як

активаційні.

В просторі вхідних векторів вибирають вектор, який називають центром, і відповідно до нього задають вагові коефіцієнти прихованого шару. Аргумент активаційної функції v_j для нейрона j прихованого шару визначатиметься відстанню між вхідним вектором та вектором прихованого шару:

$$v_j = \|X - W_j\| = \sqrt{\sum_{i=1}^{n}(x_i - w_{ij})^2} \quad (1.3)$$

Тоді значення вихідного сигналу цього нейрона дорівнюватиме:

$$y_i = exp\left(-\left(\frac{v_j}{\sigma_j}\right)^2\right) = exp\left(-\left(\frac{\|X-W_j\|}{\sigma_j}\right)^2\right) \quad (1.4)$$

Процес настройки мережі, крім вибору кількості центральних векторів та їх координат, також включає вибір параметра σ_j, який є індивідуальним для кожного нейрона і задає крутизну активаційної функції.

Отже, прихований шар формує сукупність функцій, що утворюють базисну систему, а на вихідному шарі формується лінійна комбінація цих функцій [43].

1.2.4 Конкурентна структура WTA структури Ліпмана–Хеммінга

Для розв'язання задач класифікації застосовують велику кількість структур, елементами яких є зірки Гроссберга. Вхідною зіркою Гроссберга називають нейрон з вхідним вектором $X = (x_1, \dots, x_i, \dots, x_n)$, та активаційною функцією $Y = f(\sum_{i=1}^{n} w_i x_i)$, для якого існує вектор вагових коефіцієнтів $W = (x_1, \dots, x_i, \dots, x_n)$, який забезпечує збудження нейрона тільки у випадку, коли на його вхід поступає потрібний вектор вхідних сигналів.

Існує також вихідна зірка Гроссберга з одним входом X та вектором вихідних сигналів $Y = (y_1, y_2, \dots, y_n)$, що моделює функції командного нейрона. При надходженні сигналу на вхід, на його виході формується

заданий вектор вихідних сигналів. Елементи нейронної мережі Ліпмана–Хеммінга мають властивості вхідної та вихідної зірок Гроссберга.

Структура, показана на рис, реалізує принцип “Winner Take All” (WTA). Вона містить один шар з m нейронів, кожен з яких відповідає за свій клас об’єктів, які потрібно класифікувати. У загальному випадку кожний елемент вхідного вектора зв’язаний з кожним класифікуючим нейроном. Вагові коефіцієнти утворюють матрицю:

$$W = \begin{Bmatrix} w_{01} & \cdots & w_{0m} \\ \cdot & \cdots & \cdot \\ w_{n1} & \cdots & w_{nm} \end{Bmatrix}, \tag{1.5}$$

кожний елемент якої визначається за формулою:

$$w_{ij} = \frac{x_i^j}{\sum_{k=1}^{n} x_k^j}, \tag{1.6}$$

де x_i^j - і-й елемент вхідного вектора, що через відповідний зв’язок подається на j-й нейрон;

w_{ij} - ваговий коефіцієнт елемента x_i^j вхідного вектора.

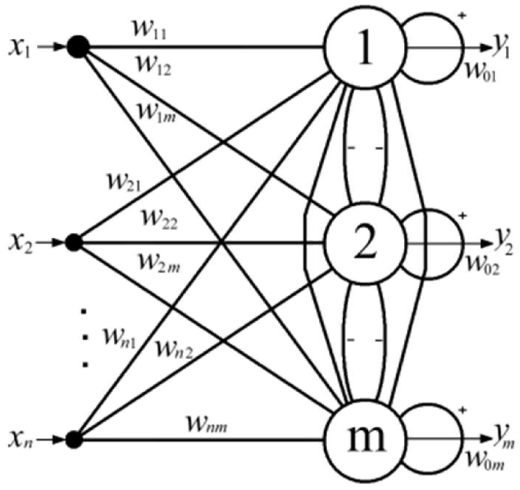


Рисунок 1.8 - WTA структура Ліпмана–Хеммінга

Після подачі на вхід нейронної мережі вектора вхідного сигналу, вихідний сигнал кожного з нейронів буде відповідати його початковій активності, що визначається за формулою:

$$y_j(0) = f(v_j) = f\left(\frac{\sum_{i=1}^{n} x_i w_{ij}}{2} - \theta\right) \quad (1.7)$$

Функція $v(f)$ є пороговою функцією:

$$y = f(v) = \begin{cases} v & \text{при } v \geq 0, \\ 0 & \text{при } v < 0. \end{cases} \quad (1.8)$$

Тому θ визначає зсув порога, який зазвичай дорівнює нулю. Отже, на першому етапі одержуємо множину активних елементів мережі, сформовану з тих нейронів, аргумент активаційної функції яких перевищив поріг. Наступні етапи розв'язання задачі класифікації на даній мережі полягають у виборі з даної множини елементів нейрона з максимальною активністю. Такий вибір реалізують шляхом введення внутрішньошарових зв'язків, які виконують функції «літерального гальмуванн». Кожен нейрон пригнічує активність інших нейронів і пропорційно підвищує свою активність за формулою:

$$y_j(t+1) = f\left(y_1(t) - \frac{1}{m-1}\sum_{\substack{i=1 \\ i \neq j}}^{m} y_i(t)\right) \quad (1.9)$$

Після кількох ітерацій такого типу активним залишається лише один нейрон, який відповідає за клас, до якого належить даний вхідний вектор сигналів.

Принцип «літерального гальмування» має свої аналоги у живій природі, тому широко використовується при реалізації ШНМ. Загальним недоліком подібного типу структур є те, що кількість класів простору класифікації дорівнює кількості елементів нейронної мережі. Тому вихід

з ладу довільного нейрона призводить до втрати інформації про клас, за який відповідав даний нейрон. Така структура пам'яті називається прямою [57].

1.2.5 Конкурентна структура нейронної мережі Кохонена

На відміну від мережі Ліпмана–Хеммінга модель нейронної мережі Кохонена характеризується структурою розподіленої пам'яті. Така структура дозволяє уникнути катастрофічної деградації у випадку відмови одного з нейронів. Ефект підвищеної живучості досягається саме завдяки розподіленій пам'яті, дія якої проявляється за рахунок того, що за класифікацію вхідного вектора відповідає не один нейрон, а кластер нейронів. На рисунку 1.9 показана структура нейронної мережі Кохонена.

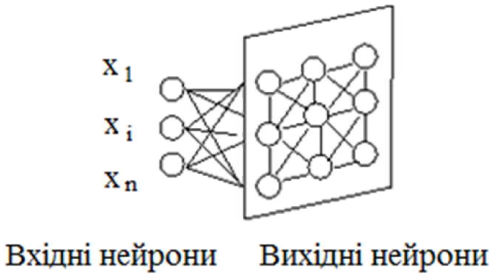


Рисунок 1.9 – Структура нейронної мережі Кохонена

У цій структурі кожний елемент вхідного вектора сигналів $X=(x_1, x_2, \dots, x_n)$, поступає на вхід кожного елемента двовимірної матриці нейронів. Тому множина вагових коефіцієнтів структурована у вигляді матриці:

$$W = \begin{Bmatrix} w^{11} & w^{12} & \dots & w^{1m} \\ w^{21} & w^{22} & \dots & w^{2m} \\ w^{k1} & w^{k2} & \dots & w^{km} \end{Bmatrix}, \quad (1.10)$$

елементами якої є вектори вагових коефіцієнтів $w^{ij} = (w_1^{ij}, w_2^{ij}, \dots, w_n^{ij})$, що

масштабують відповідні елементи вхідного вектора.

Всім ваговим коефіцієнтам на початку роботи мережі задають випадкові значення. Наступний етап полягає в обчисленні відстані $d_{ij} = \sum_{p=1}^{n}(x_p(t) - w_p^{ij}(t))^2$ між вектором вхідного сигналу та кожним з нейронів мережі.

Нейрон із координатами (i^*, j^*), для якого ця відстань є мінімальною, вибирається центром кластера. Відносно нього проводиться підстроювання вагових коефіцієнтів всіх нейронів, які входять до початкового кластера, за формулою:

$$w^{ij}(t+1) = w^{ij}(t) + \eta r\left(l^{ij}\right)\left[X(t) - w^{ij}(t)\right], \qquad (1.11)$$

де $0<\eta<1$ - коефіцієнт навчання, який змінюється обернено пропорційно до часу навчання; $r(l^{ij}) = \exp\left(\frac{-(l^{ij})^2}{\sigma^2}\right)$ - одна з можливих функцій сусідства, що задає зміну вагових коефіцієнтів у залежності від відстані $l^{ij} = \left|\sqrt{(j^* - j)^2 + (i^* - i)^2}\right|$ нейрона (j,i) до центра кластера.

1.2.6 Конкурентна структура нейронної мережі зустрічного поширення

Нейронні мережі зустрічного поширення — це ще одна відома парадигма з використанням механізму WTA. Узагальнена структура такої мережі показана на рисунку 1.8. Вона складається з двох шарів нейронів різного типу. Перший шар є двовимірною матрицею Кохонена, а другий шар називають шаром Гроссберга. Об'єднання цих двох шарів надає структурі додаткові властивості, які відсутні у кожного з них окремо. Вхідний вектор сигналів $X=(x_1, x_2, \ldots, x_n)$, поступає на кожний з нейронів шару Кохонена. Масштабування відбувається за допомогою матриці вагових коефіцієнтів, що містить вектори вагових коефіцієнтів. Відомі два методи функціонування шару Кохонена, що отримали назви «метод акредитації» та «метод інтерполяції».

1.2.6.1Метод акредитації

У випадку застосування методу акредитації шар Кохонена реалізує механізм WTA, який забезпечує активацію лише одного нейрона-переможця. Аргумент активаційної функції V_{IJ} для нейрона (j,i) шару Кохонена обчислюється за формулою:

$$v_{ij} = \sum_{k=1}^{n} w_k^{ij} x_k \tag{1.12}$$

Сукупність всіх аргументів утворює матрицю:

$$V = \begin{pmatrix} v_{11} & v_{12} & \cdots & v_{1m} \\ v_{21} & v_{22} & \cdots & v_{2m} \\ . & . & \cdots & . \\ v_{k1} & v_{k2} & \cdots & v_{km} \end{pmatrix} \tag{1.13}$$

На рисунку 1.10 зображено нейронну мережу зустрічного поширення.

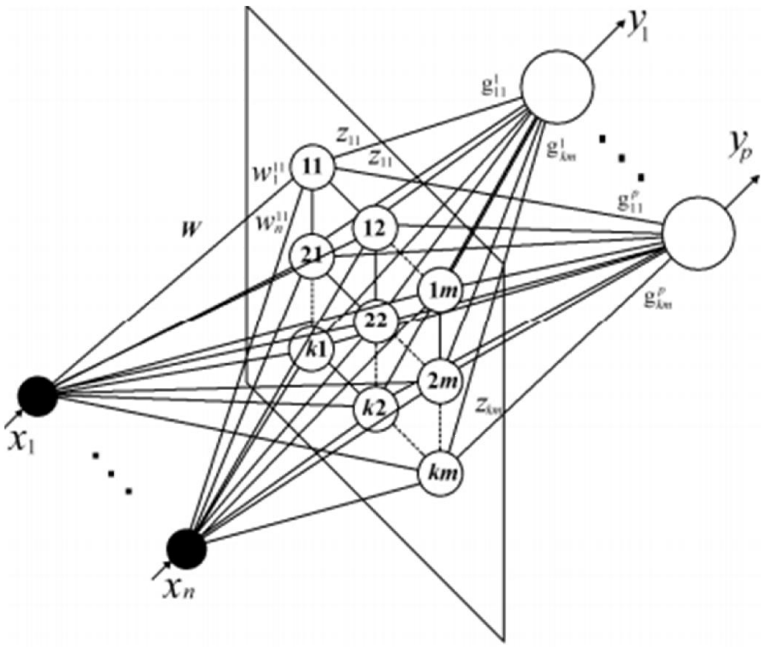


Рисунок 1.10 – Нейронна мережа зустрічного поширення

Вихідний сигнал z_{IJ} нейрона (j,i) заданий пороговою активаційною функцією:

$$z_{ij} = f_{ij}(v_{ij}) = \begin{cases} 1 & \text{при } v_{ij} > r, \\ 0 & \text{при } v_{ij} \le r, \end{cases} \quad (1.14)$$

де r — поріг чутливості нейрона.

Механізм типу WTA у даному випадку реалізується за допомогою ітераційного процесу:

$$w^{ij}(t+1) = w^{ij}(t) + \eta\left(X(t) - w^{ij}(t)\right), \quad (1.15)$$

де $0<\eta<1$ - коефіцієнт навчання.

Результатом його є пропорційне зменшення вагових коефіцієнтів до того моменту, поки не залишиться лише один нейрон, для якого справджується умова $v_{IJ} > r$.

Кожний нейрон шару Гроссберга формує свій аргумент активаційної функції за формулою:

$$b_n = \sum_{i=1}^{m} \sum_{j=1}^{k} g_{ij}^{h} z_{ij}\,, \quad (1.16)$$

Вихідний сигнал нейрона h шару Гроссберга формується активаційною функцією

$$y_h = f_h(b_n) = b_n \quad (1.17)$$

Оскільки в результаті роботи шару Кохонена залишається активним тільки один нейрон, то тільки один елемент z_{cd} матриці вихідних сигналів Z має значення $z_{cd}=1$. У цьому випадку шар Гроссберга видає вихідний вектор сигналів:

$$Y = (g_{cd}^{1}, g_{cd}^{2}, \dots, g_{cd}^{h}, \dots, g_{cd}^{p},) \quad (1.18)$$

1.2.6.2 Метод інтерполяції

Даний метод функціонування нейронної мережі зустрічного поширення полягає в тому, що в результаті роботи механізму WTA в шарі

Кохонена залишається переможцем кластер нейронів, який характеризувався максимальною активністю. Вихідні сигнали кластера нормалізують. Якщо в такий кластер входить підмножина нейронів, що задана матрицею розмірності (c,d), то довільний вихідний сигнал нейрона $z_{\alpha\beta}$, який належить до даного кластера, нормалізуємо за формулою:

$$\overline{z_{\alpha\beta}} = \frac{z_{\alpha\beta}}{\sqrt{\sum_{i=1}^{c}\sum_{j=1}^{d} z_{ij}^2}}\ , \qquad (1.19)$$

тоді активаційна функція шару Гроссберга має вигляд:

$$y_h = \sum_{i=1}^{m}\sum_{j=1}^{k} g_{ij}^h \overline{z_{ij}} \qquad (1.20)$$

Оскільки всі нейрони шару Кохонена, крім тих, що входять до кластера-переможця, видають нульові значення вихідного сигналу, результатом роботи шару Гроссберга є значення сигналів активних нейронів, промасштабовані відповідними ваговими коефіцієнтами. Завдяки розподіленій пам'яті шару Кохонена метод інтерполяції дає можливість реалізувати більш складні алгоритми класифікації та категоризації, ніж метод акредитації.

1.2.7 Рекурентна структура

Структури мереж прямого поширення, що уже розглядались, не мають зворотних зв'язків, тобто сигнали в них поширюються від шару до шару тільки в одному напрямку, формуючи статичний стан кожного нейрона мережі. Дещо складнішим є принцип функціонування конкурентних структур. В мережах такого типу затухаючий ітераційний процес відбувається в межах одного шару нейронів, а ітераційна формула завжди має властивість зниження рівня вихідного сигналу, що викликає затухання слабких вихідних сигналів до рівня, нижчого від порога чутливості. Таким чином реалізується стратегія «переможець забирає все», що зупиняє ітераційний процес у випадку перемоги одного нейрона

або кластера нейронів. Принципово новою є ситуація, коли структура нейронної мережі допускає зворотні зв'язки, тобто, коли обчислення в нейроні даного шару відбуваються з урахуванням попереднього стану цього ж шару. Мережі такого типу називають рекурентними та говорять про динамічний характер їх функціонування. Динаміка зміни станів сукупності взаємопов'язаних об'єктів традиційно вимагає визначення правила взаємної координації дій цих об'єктів у часі та просторі.

Коли встановлюється послідовність спрацьовування нейронів одного шару, то таку нейронну мережу називають синхронною. У випадку, коли час спрацьовування кожного з нейронів не регламентується, мережу називають асинхронною.

Синхронні мережі можуть бути паралельними, послідовними та паралельно-послідовними. Для паралельних мереж визначальною є властивість одночасного спрацьовування всіх нейронів одного шару, для послідовних мереж задають послідовність спрацьовування нейронів одного шару, а для паралельно-послідовних мереж визначається послідовність спрацьовування кластерів нейронів. Про асинхронні мережі говорять, що вони є паралельними, маючи на увазі те, що такі мережі допускають одночасне функціонування всіх нейронів [35].

1.2.7.1 Нейронна мережа Хопфілда

Структура рекурентної нейронної мережі, яка складається з одного шару нейронів, виходи яких через спеціальні пристрої з'єднані зі входами всіх нейронів цього ж шару, крім тих зв'язків, що з'єднують вихід нейрона з його власним входом. Формування аргументу активаційної функції v_J довільного нейрона j відбувається за формулою:

$$v_j = \sum_{\substack{i=1 \\ 1 \neq j}}^{n} w_{ij} y_i + x_j \qquad (1.21)$$

На рисунку 1.11 зображена структура нейронної мережі Хопфілда. В залежності від вигляду активаційної функції розрізняють дискретну та

аналогову моделі мереж Хопфілда.

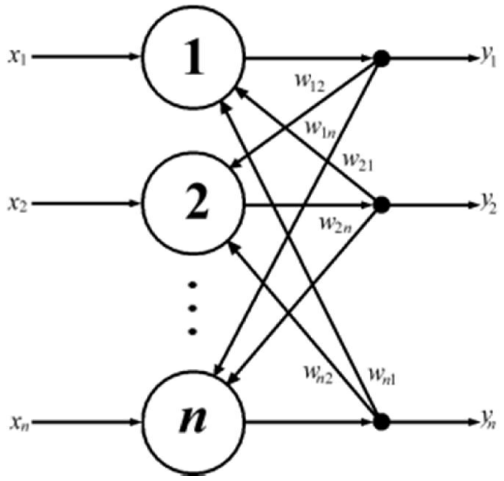


Рисунок 1.11 – Структура нейронної мережі Хопфілда

1.2.7.2 Двонаправлена асоціативна пам’ять

Основною властивістю асоціативної пам’яті, що реалізована на мережі Хопфілда, є здатність відтворювати образи шляхом їх асоціації з частково помилковим запитом. Таку мережу ще називають 34ро масштабований. Було запропоновано більш досконалу версію асоціативної пам’яті — двонаправлену асоціативну пам’ять (ДАП). Ця структура дозволяє додатково встановлювати асоціативні зв’язки між різними образами. На рис. Показана структурна схема такої мережі.

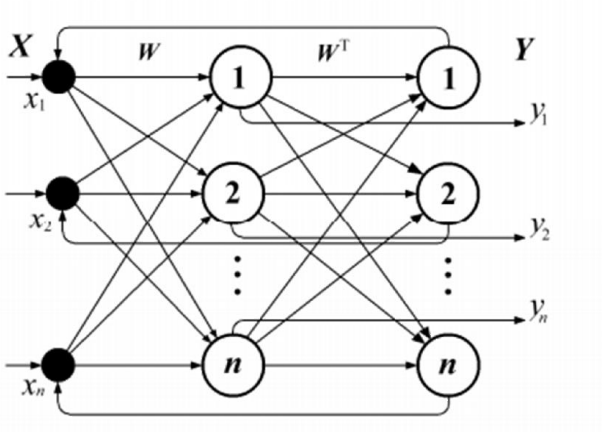


Рисунок 1.12 – Структура двонаправленої асоціативної пам’яті

Вектор вхідних сигналів X, відповідно масштабований елементами матриці вагових коефіцієнтів W, надходить на вхід першого шару нейронів. Результатом спрацьовування цього шару є вихідний вектор сигналів Y, який надходить на вхід другого шару з транспонованою матрицею вагових коефіцієнтів W^T.

До моменту його спрацьовування зовнішній вхідний вектор X знімається і заміщується тим, який є результатом роботи другого шару. Далі починається новий цикл рекурентних обчислень, і він продовжується до того часу, поки мережа не досягне стійкого стану, в якому вектори X та Y залишаються незмінними.

Активаційна функція першого шару нейронів визначається за формулою :

$$y_i = f(v_j) = f\left(\sum_{i=1}^{n} w_{ij} x_i\right) \qquad (1.22)$$

Нейрони другого шару також працюють в традиційному режимі, формуючи вхідний вектор сигналів за формулою:

$$x_i = f(v_i^t) = f\left(\sum_{j=1}^{n} w_{ji}^t y_i\right), \qquad (1.23)$$

де w_{ji}^t - ваговий коефіцієнт, що є елементом транспонованої матриці W^T.

Найчастіше в ролі активаційної функції використовують традиційну сигмоїдальну функцію з великими значеннями коефіцієнта α, що наближає її до порогової функції.

За принципами роботи розрізняють синхронні та асинхронні ДАП. У синхронному режимі відбувається одночасне спрацьовування всіх нейронів одного шару під дією зовнішнього синхронізуючого сигналу. Найчастіше у цьому випадку використовують порогову активаційну функцію.

Для досягнення більшої подібності з біологічним прототипом

застосовують асинхронний режим функціонування ДАП з сигмоїдальною активаційною функцією. В цьому режимі кожний нейрон обробляє неперервні сигнали і може бути змодельований на операційному підсилювачі з нелінійним зворотним зв'язком. ДАП — більш стійкі структури, ніж мережі Хопфілда. Для них притаманним є прагнення до пошуку локального енергетичного мінімуму у випадку, коли матриця вагових коефіцієнтів W має довільну форму.

Існує ряд підходів, які дозволяють покращити розпізнавальні властивості ДАП і, відповідно, збільшити кількість образів, що можуть викликати адекватні асоціації. До таких підходів слід віднести застосування нейронів з різними порогами спрацьовування активаційних функцій. Мережі, що мають такі властивості, називають адаптивними ДАП. Активаційні функції нейронів адаптивної ДАП визначаються за формулою у випадку синхронного режиму функціонування. При цьому для кожного нейрона j встановлюється свій поріг спрацьовування T_j.

Асинхронний режим функціонування використовує неперервну сигмоїдальну активаційну функцію, амплітуда якої визначається параметром B_J окремо для кожного нейрона j. Активаційна функція у цьому випадку:

$$y_i = f_i(v_j) = \frac{\beta_j}{1+e^{-avj}} \qquad (1.24)$$

Алгоритм настройки порогів підвищує чутливість мережі у вибраному діапазоні вхідних сигналів. Ще одним підходом такого типу є застосування конкурентних зв'язків на кожному шарі нейронів, що приводить до формування нової матриці вагових коефіцієнтів. Як і в попередньому випадку, підвищення чутливості відбувається за рахунок зниження допустимого діапазону вхідних образів [9].

1.3 Апаратні засоби реалізації штучних нейронних мереж і нейроелементів

Забезпечення режиму реального часу потребує великих витрат обладнання, вартості та термінів розроблення. Відомі програмні та мікропрограмні засоби не враховують вимог конкретних застосувань щодо габаритів і споживаної потужності, а апаратні засоби мають низьку ефективність використання обладнання. Встановлено, що всі види реалізації нейроалгоритмів у безпосередньому вигляді зустрічаються доволі рідко. Для створення нейромереж реального часу доцільно використовувати комбіновані підходи з переважанням апаратних засобів, які реалізують розпаралелювання обчислення не тільки у часі, а і в просторі. Для таких обчислень характерне введення додаткового обладнання, відсутність проміжних пересилань інформації, а також апаратне управління.

Сучасна елементна база для реалізації нейромереж реального часу визначає, що найчастіше для цих цілей використовуються нейрочіпи, процесори цифрової обробки сигналів та ПЛІС. Реалізація нейроелементів і нейромереж реального часу здійснюється на основі ПЛІС, які характеризуються високою швидкодією, можливістю динамічного репрограмування та наявністю доступних засобів розроблення.

DSP (Digital Signal Processor — цифровий сигнальний процесор), маючи потужну обчислювальну структуру, дозволяє реалізувати різні алгоритми обробки інформаційних потоків. Порівняно невисока ціна, а також розвинені засоби розробки програмного забезпечення дозволяють легко застосовувати їх при побудові обчислювальних систем з масовим паралелізмом.

Стрімкий перехід сучасних систем управління на цифрові стандарти, привів до необхідності обробляти з високою швидкістю достатньо великі об’єми інформації. Складна обробка і фільтрація сигналів, наприклад,

розпаковування стислих аудіо- і відеоданих, маршрутизація інформаційних потоків і тому подібне, вимагає застосування достатньо продуктивних обчислювальних систем. Подібні системи можуть бути реалізовані на різній елементній базі, але найбільшого поширення набули пристрої із застосуванням цифрових сигнальних процесорів і ПЛІС [56].

Програмована логіка здатна працювати на вищих частотах, але оскільки управління реалізовано апаратно, та зміна алгоритмів роботи вимагає перепрограмування ІС. Низька тактова частота DSP поки що обмежує максимальну частоту оброблюваного аналогового сигналу до рівня в 10-20 Мгц, але програмне управління дозволяє достатньо легко змінювати не лише режими обробки, але і функції, виконувані DSP. Окрім обробки і фільтрації даних DSP можуть здійснювати маршрутизацію цифрових потоків, вироблення сигналів урпавління і навіть формування сигналів системних шин ISA, PCI і ін.

Оцінювати швидкодію тих або інших пристроїв на основі DSP і ПЛІСУ прийнято за часом виконання типових операцій цифрової обробки сигналів (Фільтр Собеля, БПФ, перетворення Уолша-Адамара і ін.). Проте оцінки продуктивності нейрообчислювачів використовують інші показники:

1. CUPS (connections update per second) — число змінених значень вагів в секунду (оцінює швидкість навчання).
2. CPS (connections per second) — число з’єднань (множень з накопиченням) в секунду (оцінює продуктивність).
3. CPSPW = CPS/Nw, де Nw — число синапсів в нейроні.
4. CPPS — число з’єднань в секунду, CPPS=CPS·Bw·Bs, де Bw, Bs — розрядність вагів і синапсів.
5. MMAC — мільйонів множень з накопиченням в секунду.

Особливістю використання DSP і ПЛІС як елементна база нейрообчислювачів є те, що орієнтація у виконанні нейромережевих операцій обмежує з одного боку підвищення швидкостей обміну між пам’яттю і паралельними арифметичними пристроями, а з іншого боку

зменшення часу вагового підсумовування (множення і накопичення) за рахунок застосування фіксованого набору команд типу регістр-регістр.

Цифрові сигнальні процесори (DSP) ось вже впродовж декількох десятиліть є елементною базою для побудови як нейроприскорювачів, так і контура логіки загальносистемного управління нейрокомп'ютерів.

При створенні нейрообчислювальних систем на базі сигнальних процесорів необхідно пам'ятати, що DSP мають високий ступінь спеціалізації. У них широко використовуються методи скорочення тривалості командного циклу, характерної для універсальних процесорів RISC, такі як конвеєризація на рівні окремих мікроінструкцій і інструкцій, розміщення операндів більшості команд в регістрах, використання тіньових регістрів для збереження стану обчислень при перемиканні контексту, розділення шин команд і даних (Гарвардська архітектура). В той же час для сигнальних процесорів характерною є наявність апаратного помножувача, що дозволяє виконувати множення як мінімум двох чисел за один командний такт. Іншою особливістю сигнальних процесорів є включення в систему команд таких операцій, як множення з накопиченням MAC (C=A×B+C) з вказаним в команді числом виконань в циклі і з правилом зміни індексів використовуваних елементів масивів A і B, тобто вже реалізовані прообрази базових нейрооперацій — зважене підсумовування з накопиченням), інверсія біт адреси, різноманітні бітові операції. У сигнальних процесорах реалізується апаратна підтримка програмних циклів, кільцевих буферів. Один або декілька операндів витягуються з пам'яті в циклі виконання команди.

Велика продуктивність, потрібна при обробці сигналів в реальному часі, спонукала Texas Instruments і Analog Devices випустити трансп'ютероподібні сімейства мікропроцесорів TMS320C4x і ADSP2106x, орієнтовані на використання в мультипроцесорних системах. На цьому фоні перший російський сигнальний процесор (нейросигнальний процесор) фірми Модуль — «Neuro Matrix», виглядає вельми гідно серед DSP з фіксованою

точкою. При тактовій частоті 50 Мгц «Neuro Matrix» практично не поступається за продуктивністю виробам світових лідерів, а за деякими показниками навіть перевершує їх. Представимо порівняльні тести Cisc процесорів, DSP TI і нейросигнального процесора NM6403 (таблиця 2.1)

Таблиця 2.1 – Порівняльні тести Cisc процесорів, DSP TI і нейросигнального процесора NM6403

Найменування тесту	Intel Pentium II 300 Мгц	Intel PENTIUMMMX 200 Мгц	TI TMS320C40 50 Мгц	Нтц»модуль» NM6403 40 Мгц
ФільтрСобеля (розмір кадру 384X288 байт) кадрів/с.	-	21	6,8	68
Швидке перетворення Фур'є (256 точок, 32 розряди), мкс (тактів)	200	-	464 (11588)	102 (4070)
Перетворення	2,58	2,80	-	0,45

Реалізація однотактного множення і команд, що використовують як операнди вміст елементів пам'яті, обумовлює порівняно низькі тактові частоти роботи сигнальних процесорів. Спеціалізація не дозволяє піднімати продуктивність за рахунок швидкого виконання коротких команд типу R,R->R, як це робиться в універсальних процесорах. Цих команд просто немає в програмах цифрової обробки сигналів [54].

Сигнальні процесори різних компаній-виробників утворюють два класи, що істотно розрізняються за ціною: дешевші мікропроцесори для обробки даних у форматі з фіксованою точкою і дорожчі мікропроцесори, апаратні, які підтримують операції над даними у форматі з плаваючою точкою.

Типові операції DSP вимагають виконання безлічі простих додавань і множень.

Додавання і множення вимагають:

1) провести вибірку двох операндів;
2) виконати додавання або множення (зазвичай і те і інше);
3) зберегти результат або затримувати його до повторення.

Для вибірки двох операндів за один командний цикл необхідно здійснити два доступи до пам’яті одночасно. Але насправді окрім вибірки двох операндів необхідно ще зберегти результат і прочитати саму інструкцію. Тому число доступів в пам’ять за один командний цикл буде більше двох і, отже, DSP процесори підтримують великий доступ до пам’яті за один і той же командний цикл. Але неможливо здійснити доступ до двох різних адрес в пам’яті одночасно, використовуючи для цього одну шину пам’яті.

Існує два види архітектури DSP процесорів що дозволяють реалізувати механізм великого доступу до пам’яті:

1. Гарвардська архітектура.
2. Модифікована архітектура фон Неймана.

1.3.1 Гарвардська архітектура

Гарвардська архітектура має дві фізично розділені шини даних. Це дозволяє здійснити два доступи до пам’яті одночасно: справжня Гарвардська архітектура виділяє одну шину для вибірки інструкцій (шина адреси), а іншу для вибірки операндів (шина даних). Але для виконання DSP операцій цього недостатньо, оскільки в основному всі вони використовують по два операнди. Тому Гарвардська архітектура використовує шину адреси для

цифрової обробки сигналів і для доступу до даних. Поважно відзначити, що часто необхідно провести вибірку трьох компонентів — інструкції з двома операндами, на що власне Гарвардська архітектура нездібна. У такому разі дана архітектура включає кеш-пам’ять. Вона може бути використана для зберігання тих інструкцій, які використовуватимуться знов. При використанні кеш-пам’яті шина адреси і шина даних залишаються вільними, що робить можливою вибірку двох операндів. Таке розширення — Гарвардська архітектура плюс кеш — називають розширеною Гарвардською архітектурою або SHARC (Super Harvard ARChitecture) [35].

Гарвардська архітектура потребує наявність двох шин пам’яті. Це значно підвищує вартість виробництва чипа. Так, наприклад, DSP процесор працює з 32-бітовими словами і в 32-бітовому адресному просторі потребує наявність принаймні 64 виводів для кожної шини пам’яті, а в сумі виходить 128 виводів. Це призводить до збільшення розмірів чипа і до труднощів при проектуванні схеми.

1.3.2 Архітектура фон Неймана

Архітектура фон Неймана використовує тільки одну шину пам’яті. Вона має ряд позитивних характеристик:

1) дешевша;
2) вимагає меншої кількості виводів шини;
3) є більш простішою у використанні, оскільки програміст може розміщувати і команди і дані в будь-якому місці вільної пам’яті.

З погляду реалізації нейроприскорювачів ми зупинимося тільки на деяких найбільш яскравих представниках DSP, що в основному відносяться до класу трансп’ютероподібних DSP з плаваючою комою.

Окремо слід розгледіти можливість створення паралельних обчислювачів (у тому числі і нейро) на базі ПЛІС (програмованих логічних інтегральних схем). На ПЛІС можна реалізовувати системи як другого, так і третього типу, також останнім часом широко поширені гібридні нейрообчислювачі, коли блок обробки даних реалізується на DSP, а логіка

управління на ПЛІС. В наш час безліч фірм в світі займається розробкою і випуском різних ПЛІС, проте лідерство ділять дві фірми Xilinx і ALTERA. Виділити продукцію будь-якої однієї з цих фірм неможливо, оскільки за технічними характеристиками вони розрізняються дуже мало.

Елементною базою перспективних нейрообчислювачів є нейрочипи. Їх виробництво ведеться в багатьох країнах світу, причому більшість з них на сьогодні орієнтовані на закрите використання (тобто створювалися для конкретних спеціалізованих систем управління).

Перш ніж перейти до розгляду найбільш цікавих нейрочипів зупинимося на їх класифікації:

1. За типом логіки їх можна розділити на цифрові, аналогові і гібридні.

2. За типом реалізації нейроалгоритмів: з повністю апаратною реалізацією і з програмно-апаратною реалізацією (коли нейроалгоритми зберігаються в ПЗП).

3. За характером реалізації нелінійних перетворень: на нейрочипи з жорсткою структурою нейронів (апаратних реалізованих) і нейрочипи із структурою нейронів, що настроюється (перепрограмовані).

4. За можливостями побудови нейромереж: нейрочипи із жорсткою і змінною нейромережевою структурою (тобто нейрочипи, в яких топологія нейромереж реалізована жорстко або гнучко).

У окремий клас слід виділити так звані нейросигнальні процесори, ядром яких є типовий сигнальний процесор, а реалізована на кристалі додаткова логіка забезпечує виконання нейромережевих операцій (наприклад, додатковий векторний процесор і тому подібне).

Орієнтація у виконанні нейромережевих операцій обумовлює з одного боку підвищення швидкостей обміну між пам’яттю і паралельними арифметичними пристроями, а з іншого боку зменшення часу вагового підсумовування (множення і накопичення) за рахунок застосування

фіксованого набору команд типу реєстр-регістр. Приклад узагальненої класифікації нейрочіпів наведено на рисунку 1.13 [35].

Рисунок 1.13 – Узагальнена класифікація нейрочіпів

Постановкою задачі створення високоефективних нейромережевих засобів реального часу потребує широкого використання сучасної елементної бази, розроблення нових моделей нейрона, методів і алгоритмів, орієнтованих на реалізацію у вигляді надвеликих інтегральних схем (НВІС). Режим реального часу та НВІС-реалізація нейромереж з високою ефективністю використання обладнання забезпечується апаратним відображенням структури нейромережевих алгоритмів у архітектуру, яка адаптована до інтенсивності надходження потоків даних. Орієнтація структур нейроелементів і нейромереж на НВІС-реалізацію вимагає зменшення кількості виводів інтерфейсу, міжнейронних зв’язків і апаратних затрат. Забезпечити ці вимоги можна за допомогою використання паралельних методів і структур нейроелементів і нейромереж, які адаптуються до інтенсивності надходження даних.

2 РОЗРОБКА АЛГОРИТМІВ РЕАЛІЗАЦІЇ ПАРАЛЕЛЬНОГО НЕЙРОНА ВЕРТИКАЛЬНО-ГРУПОВОГО ТИПУ

2.1 Формування вимог до апаратної реалізації нейроелемента

Для апаратної реалізації нейромереж запропоновано їхнє відображення здійснювати за допомогою напрямлених графів на рівні одно-, дво- і багатооперандних нейроопераций. Графове відображення нейромереж забезпечує виявлення всіх форм паралелізму та знаходження необхідних просторово-часових рішень у разі апаратної реалізації. Для синтезу апаратних нейромереж реального часу з високою ефективністю використання обладнання графове відображення нейромереж необхідно представити у вигляді узгодженого графа. Для отримання такого графа потрібно узгодити інтенсивність надходження даних $\mathrm{P_d = F_d kn_k}$ з обчислювальною здатністю графа нейромережі (паралельно-конвеєрної структури) $\mathrm{D_{НМ} = L_{НМ} n_{НМ} F_t}$, де $\mathrm{L_{НМ}}$– кількість нейронів у шарі нейромережі, $n_{НМ}$ – розрядність міжнейронних зв’язків, F_t – тактова частота опрацювання даних.

Показано, що узгодження інтенсивності надходження даних з обчислювальною здатністю нейромережі може здійснюватися зміною тактової частоти опрацювання, або кількості нейроелементів у шарі нейромережі. Визначено, що насамперед необхідно мінімізувати тривалість конвеєрного такту, яка визначається швидкодією елементної бази та складністю функціональних операторів, які реалізуються сходинкою конвеєра.

Запропоновано синтез нейромереж реального часу з високою ефективністю використання обладнання здійснювати за допомогою апаратного відображення функціональних операторів узгодженого графа нейромережі відповідними апаратними засобами. Перехід від узгодженого

графа до апаратної структури нейромережі з високою ефективністю використання обладнання зводиться до мінімізації апаратних витрат при забезпеченні роботи в реальному часі [39].

Для розроблення апаратних нейромереж реального часу з високою ефективністю використання обладнання вибрано компонентно-ієрархічний підхід. Цей підхід передбачає поділ процесу розроблення на ієрархічні рівні та види забезпечення (алгоритмічне, апаратне та програмне).

За складністю виконуваних робіт процес розроблення нейромереж розділено на чотири ієрархічні рівні:

1. На першому ієрархічному рівні розробляють архітектуру (визначають тип нейромережі, кількість її шарів та нейронів у кожному шарі, способи зв'язків між нейронами), методи, алгоритми навчання і функціонування та структуру програмних засобів нейромережі для розв'язання конкретної задачі.

2. На другому рівні розробляють методи й алгоритми функціонування, структури пристроїв та програми реалізації шарів нейромереж, нейроелементів, паралельної пам'яті та перетворювачів даних.

3. Третій ієрархічний рівень складають пристрої для реалізації багатооперандних нейроопераций попереднього та процесорного опрацювання даних.

4. Четвертий – пристрої для обчислення одно- і двооперандних нейроопераций.

Процес розроблення на основі компонентно-ієрархічного підходу описують таким виразом:

$$C_{HM}^{1} = \bigcup_{i=1}^{n} C_{HM}^{2i} \bigcup_{j=1}^{m} C_{HM}^{3j} \bigcup_{p=1}^{h} C_{HM}^{4p}, \qquad (2.1)$$

де $C_{НМ}^{2i}$, $C_{НМ}^{3j}$, $C_{НМ}^{4p}$ – засоби відповідно другого, третього і четвертого ієрархічних рівнів;

n – кількість типів компонентів другого рівня;

m – кількість типів компонентів третього рівня;

h – кількість типів компонентів четвертого рівня.

Основними компонентами, на базі яких синтезуються апаратні нейромережі є нейроелемети. При вертикально-груповвй реалізації даних у нейроелементі вхідні дані X_j та вагові коефіцієнти W_j (j=1,…,N, де N – кількість входів даних і вагових коефіцієнтів) подаються у порозрядному вигляді згідно з формулою:

$$W_j = \sum_{i=1}^{n} 2^{-1} W_{ji}\,, X_j = \sum_{i=1}^{n} 2^{-1} W_{ji} \qquad (2.2)$$

де W_{ji}, X_{ji} – значення і-х розрядів множників W_j і X_j,

n – розрядність множників.

У загальному випадку нейроелемент здійснює перетворення у відповідності з формулою:

$$y = f\left(\sum_{j=1}^{n} W_j X_j\right) \qquad (2.3)$$

де y – вихідний сигнал нейроелемента,

f – функція активації.

З формули (2.3) випливає, що опрацювання даних у нейроелементах зводиться до виконання таких етапів:

1) обчислення скалярного добутку $Z = \sum_{j=1}^{n} W_j X_j$;

2) обчислення функції активації *f*.

Обчислення скалярного добутку в нейроелементі з використанням паралельно-вертикального опрацювання даних записується так:

$$Z = \sum_{j=1}^{N} W_j X_j = \sum_{i=1}^{n} 2^{-i} \sum_{j=1}^{N} W_j X_{ji} = \sum_{i=1}^{n} 2^{-i} \sum_{j=1}^{N} P_{ji} = \sum_{i=1}^{n} 2^{-i} P_{Mi}\,, \quad (2.4)$$

де P_{ji} – ji-й частковий добуток,

P_{Mi} – і-й макрочастковий добуток, який формується додаванням N часткових добутків P_{ji}, тобто $P_{Mi} \sum_{j=1}^{N} P_{ji}$.

З формули (2.4) випливає, що паралельно-вертикальне обчислення

скалярного добутку виконується за *n* тактів, в кожному *i*-у такті виконуються такі операції:

1) формування для кожного і-о розрядного зрізу X_{ji} і-х часткових добутків P_{ji};

2) обчислення і-о макрочасткового добутку P_{Mi} шляхом підсумовуванням і-х часткових добутків P_{ji};

3) підсумовування макрочасткових добутків P_{Mi} у відповідності з виразом $Z_h = 2^{-h}Z_{h-1} + P_{Mi}$, де $Z_0 = 0$.

Основним шляхом підвищення швидкодії обчислення скалярного добутку є зменшення кількості тактів роботи, яке можна забезпечити за допомогою вертикально-групового опрацювання даних. При такому опрацюванні даних у нейроелементах вхідні дані X_j та вагові коефіцієнти W_j надходять групами із *k* розрядів. Для реалізації такого опрацювання вхідні дані X_j та вагові коефіцієнти W_j записуються так:

$$W_j = \sum_{h-1}^{m} 2^{-(h-1)k} W_{jh1} W_{jh2} \dots W_{jhg}, \quad (2.5)$$

$$X_j = \sum_{n-1}^{m} 2^{-(h-1)k} X_{jh1} X_{jh2} \dots X_{jhg}, \quad (2.6)$$

Паралельний метод вертикально-групового обчислення скалярного добутку в нейроелементі ґрунтується на формулах:

$$Z = \sum_{j=1}^{N} W_j X_j, \quad (2.7)$$

$$Z = \sum_{n-1}^{m} 2^{-(h-1)k} \sum_{j=1}^{N} (W_j X_{jh1} + 2^{-1} W_j X_{jh2} + \dots + 2^{-1(k-1)} W_j X_{jhk}, \quad (2.8)$$

$$Z = \sum_{n-1}^{n} 2^{-(h-1)} \sum_{j=1}^{N} P_{jh} = \sum_{h=1}^{m} 2^{-(h-1)} P_{Mh}. \quad (2.9)$$

де P_{jh} – jh-й груповий частковий добуток,

P_{Mh} – h-й груповий макрочастковий добуток, який формується додаванням N групових часткових добутків P_{jh}, тобто $P_{Mh} = \sum_{j=1}^{N} P_{jh}$.

З формули (2.9) випливає, що паралельне вертикально-групове

обчислення скалярного добутку виконується за *m* тактів, у кожному *h*-у такті виконуються такі операції:

1) формування для кожної j-ї пари операндів k часткових добутків у відповідності з формулою $P_{jh} = W_j X_{jhr}$, де r=1,…,k;

2) обчислення для j-ї пари операндів групового часткового добутку P_{jh} у відповідності з формулою $P_{jh} = \sum_{r=1}^{k} 2^{-(r-1)} W_j X_{jhr}$;

3) обчислення h-о макрочасткового добутку P_{Mh} шляхом підсумовуванням h-х часткових добутків P_{jh} у відповідності з формулою $P_{Mh} = \sum_{j=1}^{N} P_{jh}$;

4) підсумовування макрочасткових добутків P_{Mh} у відповідності з виразом $Z_h = 2^{-h} Z_{h-1} + P_{Mh}$, де $Z_0 = 0$.

Аналіз формул (2.6) і (2.9) показує, що основою опрацювання даних у нейроелементі є операція групового підсумовування:

$$Z = \sum_{j=1}^{M} C_j \,, \tag{2.10}$$

де *M* – кількість часткових результатів*;* C_j – j-й частковий результат.

Нехай доданки C_j є двійковими n-розрядними додатними числами меншими за одиницю, які записуються так:

$$C_j = \sum_{i=1}^{n} 2^{-i} C_{ji}. \tag{2.11}$$

Підставивши значення формули (2.10) у формулу (2.11), отримаємо:

$$C = \sum_{j=1}^{M} \sum_{i=1}^{n} 2^{-1} C_{ji}. \tag{2.12}$$

Формула (2.12) відображає горизонтальну модель обчислення оператора групового підсумовування. Замінивши у формулі (2.12) порядок підсумовування переходимо до вертикальної моделі обчислення оператора групового підсумовування, яка записується так:

$$C = \sum_{i=1}^{n} 2^{-1} \sum_{j=1}^{M_i} C_{ji} \,, \tag{2.13}$$

де M_i – кількість доданків у і-у розрядному зрізі.

У цій моделі групового підсумовування процес підсумовування зводиться до перетворення багаторядного коду в однорядний [11].

2.2 Алгоритм реалізації паралельного нейроелемента вертикально-групового типу

Алгоритм реалізації паралельного нейроелемента вертикально-групового типу зводиться до формування множин вимог R={R1, R2,…,Rk}, характеристик H={H1, H2,…,Hm} і обмежень B={B1, B2,…,Bk} та знаходження такого вектора H*=[H*1, H*2,…,H*m], H * i=fi(R, H, B), i=1,…,m, який забезпечить максимальне значення ефективності використання обладнання E= max f(R, H* ,B).

Множина вимог R складається з:

1) R1 – кількості каналів надходження даних m_d;
2) R2 – розрядності каналів надходження даних n_d;
3) R3 – частоти надходження даних F_d;
4) R4 – швидкодії елементної бази, яка визначається часом затримки вентиля t_B;
5) R5 – кількості вхідних даних X_j і вагових коефіцієнтів W_j;
6) R6 – розрядності вхідного слова n.

Множину характеристик H становлять:

1) H1 – загальна кількість зв’язків Z;
2) H2 – просторова зв’язкова віддаль Dj ;
3) H3 – конвеєрний такт t_k;
4) H4 – витрати обладнання W;
5) H5 – кількість видів функціональних вузлів s;

6) Н6 – кількість каналів введення $m_{ВВ}$;

7) Н7 – розрядність каналів введення $n_{ВВ}$;

8) Н8 – кількість виводів інтерфейсу зв'язку Y.

Обмеження В, які необхідно враховувати при синтезі паралельного нейроелемента вертикально- групового типу, є такими:

1) В1 – точність обчислення, яка визначається розрядністю np;

2) В2 – час обчислення $T_{обч}$ повинен бути $T_{обч} \leq T_{обм}$, де $T_{обм}$ – час обміну, який визначається так: $T_{обм} = \frac{N_n}{F_d k n_k}$.

Для вибору варіанта паралельного нейроелемента вертикально-групового типу використовується критерій ефективності використання обладнання Е, 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 . Кількісну величину ефективності використання обладнання для такого компонента визначають так:

$$E = \frac{R}{t_o}\left(\sum_{j=1}^{M} W_{ЕПj} + \sum_{i=1}^{n} W_{НЕi} + k_1 Y + k_2 P\right), \qquad (2.14)$$

де R – складність алгоритмів навчання та функціонування нейромережі, яка визначається кількістю елементарних арифметичних операцій, необхідних для їхньої реалізації;

t_o – час реалізації алгоритмів навчання та функціонування нейромережі;

$W_{ЕПj}$ – витрати обладнання на реалізацію j-го елемента попереднього опрацювання;

M – кількість елементів попереднього опрацювання;

$W_{НЕi}$ – витрати обладнання на реалізацію і-го нейроелемента;

N – кількість нейроелементів;

Y–кількість виводів інтерфейсу;

k_1 – коефіцієнт врахування кількості виводів інтерфейсу,

$k_1=f(Y)$;

P – кількість міжнейронних зв’язків;

k_2 – коефіцієнт врахування кількості міжнейронних зв’язків, $k_2=f(P)$.

Синтез паралельного нейроелемента вертикально-групового типу складається із таких етапів: вибору та розроблення методів і алгоритмів обчислення скалярного добутку та функції активації, визначення основних параметрів та переходу від алгоритму до структури [12].

Під час вибору та розроблення методів і алгоритмів реалізації скалярного добутку та функції активації враховуються вимоги R і характеристик H, але визначальним є забезпечення обмежень B. Для оцінювання розроблених алгоритмів використовуються інформаційні, операційні та точнісні характеристики. До інформаційних характеристик належать кількість констант, вхідних, вихідних і проміжних даних, кількість каналів та їхня розрядність, кількість і види операцій. Операційні характеристики дають змогу оцінити час реалізації та обчислювальну здатність. До точнісних характеристик алгоритму належать розрядність операційних пристроїв, способи округлення. У паралельних нейромережах реального часу одним із найважливіших параметрів є забезпечення балансу часу при проходженні даних через усі компоненти системи.

При синтезі паралельного нейроелемента вертикально-групового типу необхідно забезпечити обчислення скалярного добутку та функції активації в реальному часі при мінімальних апаратних затратах. Перехід від алгоритму розв’язання задачі в реальному часі до структури паралельного нейроелемента вертикально-групового типу формально зводиться до мінімізації апаратних затрат із забезпеченям режиму реального часу [13].

Для забезпечення високої швидкодії нейроелементів і нейромереж при НВІС-реалізації та зменшення кількості виводів інтерфейсу опрацювання даних здійснювати паралельно розрядними зрізами (вертикально) на основі багатооперандного підходу, тобто паралельно-вертикально. Розроблено метод паралельно-вертикального опрацювання

даних, за яким вагові коефіцієнти W_j та вхідні дані X_j (j=1,…,k, де k – кількість входів даних і вагових коефіцієнтів) подають у порозрядному вигляді згідно з формулою:

$$W_j = \sum_{i=1}^{n} 2^{-i} w_{ji}, X_j = \sum_{i=1}^{n} 2^{-i} x_{ji} \quad (2.15)$$

де w_{ji}, x_{ji} – значення і-х розрядів вагових коефіцієнтів та вхідних даних;

n– розрядність вагових коефіцієнтів та вхідних даних.

Основними етапами розроблення методу паралельно-вертикального опрацювання даних у нейроелементах і нейромережах є:

1) формування для кожного розрядного зрізу часткових результатів P_{ji}; підсумовування часткових результатів та отримання макрочасткового результату P_{Mi};

2) підсумовування макрочасткових результатів;

3) обчислення функції активації f.

Опрацювання даних за розробленим паралельно-вертикальним методом у р-му нейроелементі можна записати так:

$$Y_P = f\left(\sum_{j=1}^{k} W_j X_j\right) = f\left(\sum_{i=1}^{n} 2^{-i} \sum_{j=1}^{k} W_j x_{ji}\right), \quad (2.16)$$

$$Y_P = f\left(\sum_{i=1}^{n} 2^{-i} \sum_{j=1}^{k} P_{ji}\right) = f\left(\sum_{i=1}^{n} 2^{-i} P_{Mi}\right), \quad (2.17)$$

де P_{ji} – ji-й частковий результат;

P_{Mi}– і-й макрочастковий результат, який формується додаванням k часткових результатів.

2.3 Вибір принципів побудови та варіантів реалізації паралельного нейроелемента вертикально-групового типу

Для розробки широкого спектру нейромережевих засобів, потрібно щоб в основу їхньої побудови був покладений принцип змінного складу обладнання, який передбачає реалізацію нейромережевих засобів у вигляді ядра (компютера), яке доповняється відповідними компонентами.

Концептуальна модель нейромережевих засобів, які побудовані за принципом змінного складу обладнання наведена на рисунку 2.1.

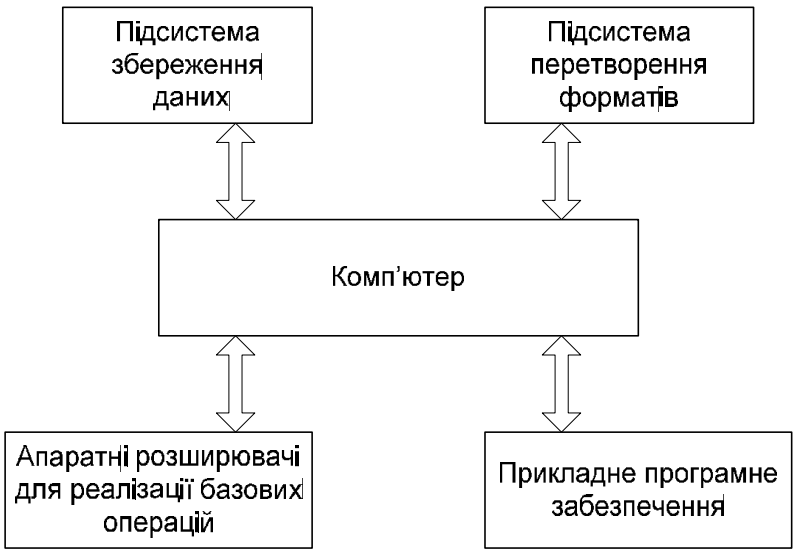


Рисунок 2.1 - Концептуальна модель нейромережевих технологій з вертиально-паралельною обробкою даних

Ця модель передбачає поєднання універсальних і спеціальних підходів та програмних і апаратних засобів. Таке поєднання забезпечує побудову високоефективних нейромережевих засобів для розвязання конкретних задач. При даному підході компютер використовується для реалізації алгоритмів з великою кількістю нерегулярних і логічних операцій, а часомісткі обчислювальні операції реалізуються за допомогою апаратних засобів.

Основними компонентами, які використовуються для синтезу нейромережевих технологій є підсистема збереження даних, підсистема перетворення форматів, апаратні розширювачі для реалізації базових операцій і прикладне програмне забезпечення.

Підсистема збереження даних забезпечує збереження вхідних потоків

даних, проміжних та кінцевих результатів обробки. Основними вимогами, що висуваються до даної системи, є забезпечення високої швидкодії та доступу до множини даних за одне звернення, тобто, пам’ять, що використовується в системі збереження даних, повинна бути паралельною [24].

Підсистема перетворення форматів повинна забезпечувати паралельно-послідовне та послідовно-паралельне перетворення. Паралельно-послідовне перетворення здійснюється шляхом послівного приймання N чисел у кожному такті та видачу в кожному такті розрядних зрізів всіх чисел. Таке перетворення використовується для завантаження вертикальних операційних пристроїв, у яких здійснюється вертикальна обробка, яка передбачає у кожному такті послідовну обробку i-х розрядів всіх чисел, тобто, здійснюється послідовна обробка розрядів при паралельній обробці чисел. Послідовно-паралельне перетворення використовується для перетворення виходів результатів обробки вертикальних операційних пристроїв. При послідовно-паралельному перетворенні на вхід результати надходять розрядними зрізами, а на вихід видаються послівно.

Апаратні розширювачі реалізують часомісткі базові обчислювальні операції: множення, обчислення сум парних добутків, множення матриці на вектор та передатні функції. Реалізація даних базових операцій здійснюється на основі вертикальних обчислень, що забезпечує мінімізацію інтерфейсу та орієнтує структури на НВІС-реалізацію. Крім операційних пристроїв до складу апаратних розширювачів повинні входити інтерфейсні пристрої, які забезпечують підключення апараиних розширювачів до комп’ютера. Для управління обміном між апаратним розширювачем та компютером потрібно розробити програму обміну.

Прикладне програмне забезпечення повинне забезпечувати навчання нейромережевих засобів, попередні обчислення при реалізації перцептрону з неітераційним навчанням та управління процесом функціонування нейромережі.

Архітектури ШНМ реального часу повинні в повній мірі

використовувати можливості НВІС-технології, враховувати вартість площі кристала, а також кількість вхідних і вихідних виводів. Число зовнішніх виводів НВІС обмежене рівнем технології та розміром кристалу. В основу побудови ШНМ реального часу пропонується покласти принципи, які дозволять зменшити вартість, терміни і розширити галузі їх застосування. Аналіз показує , що забезпечити дані вимоги можна при використанні таких принципів:

1) модульності, який передбачає розробку компонентів АЗ у вигляді функціонально завершених пристроїв (модулів), що мають вихід на стандартний інтерфейс;

2) узгодженості інтенсивності надходження даних з обчислювальною здатністю АЗ;

3) конвеєризації та просторового паралелізму обробки даних;

4) однорідності та регулярності архітектури ШНМ;

5) локалізації та спрощення зв’язків між елементами АЗ;

6) відкритості програмного забезпечення, що передбачає можливості нарощування та його вдосконалення, максимального використанням стандартних драйверів та програмних засобів;

1) спеціалізації та адаптації апаратно-програмних засобів до структури алгоритмів обробки та інтенсивності надходження даних;

2) програмованості архітектури шляхом використання репрограмованих логічних інтегральних мікросхем.

Переважна кількість нейромережевих реалізацій на даний час здійснюється на базі процесорів загального призначення. Цьому сприяє широке розповсюдження персональних комп’ютерів і вражаюче зростання їхньої продуктивності. На даний момент створена велика кількість програм і програмних бібліотек. Але дані нейромережеві реалізації здійснюються винятково на програмному рівні [25].

Принцип побудови нейроелемента поєднує в собі цифровий і аналоговий підходи до апаратного моделювання нейронних мереж. Переваги

та недоліки подібних архітектур пов’язані з особливостями конкретних апаратних реалізацій.

Незважаючи на те, що в даний час більшість нейромережевих додатків випускають виключно в програмному виконанні з використанням комерційних програмних емуляторів нейронних мереж, існує досить широка ринкова ніша для нейрочіпів або нейромережних апаратно-програмних комплексів.

Спеціалізоване апаратне забезпечення варто використовувати у таких випадках:

1) якщо складність задачі, що вирішується, досить велика і виключно програмні нейромережні реалізації не забезпечують потрібну швидкість виконання або продуктивність (наприклад, експерименти в області фізики високих енергій);

2) якщо вартість проектування спеціалізованої апаратної нейромережевої системи на готових нейромережних модулях значно нижча, ніж у її програмного аналога;

3) якщо потрібно забезпечити вирішення задачі, алгоритм якої наперед допускає масове розпаралелювання, і система буде експлуатуватися в польових умовах, що вимагає підвищених вимог до її надійності (наприклад, система збору і обробки даних погодних датчиків на аеродромі);

4) якщо є суттєві обмеження на габарити системи або її вагу;

5) якщо є підвищені вимоги до безпеки системи, оскільки апаратні рішення більш захищені від несанкціонованого доступу і забезпечують більший захист авторських прав на схемотехнічні та архітектурні рішення.

Потрібно зауважити, що при проектуванні апаратної реалізації нейросистем враховуються, перш за все, такі характеристики, як можливість масштабування, вартість розробки, сумісність з попередніми і майбутніми версіями.

Стосовно розглянутих категорій чіпів це означає, що у випадку, коли для розробника критичні терміни розробки і вартість, то потрібно звернути

увагу на ПЛІС, сигнальні процесори і процесори для каскадованих архітектур. ПЛІС здатні добре масштабуватися і достатньо дешеві, але затрати апаратних ресурсів ПЛІС при проектуванні достатньо великі і швидко зростають з підвищенням складності мережі. Сигнальні процесори масштабуються гірше, ніж ПЛІС, але розробка нейронних мереж на них завдяки розвинутій технічній підтримці зі сторони виробників достатньо проста. При цьому прив'язка до апаратної частини DSP не дозволить легко переносити нейронну мережу з одної елементної бази на другу. Крім того, сигнальні процесори не мають достатньо доброї масшабованості, що ускладнює побудову великих багатопроцесорних систем. Систолічні процесори і процесори з каскадної архітектурою мають на порядок кращу масштабованість, ніж сигнальні, але для них потрібно досить багато периферійних модулів. Зі збільшенням кількості систолічних процесорів збільшуються затримки з метою проходження сигналу [26].

Для нейросигнальних процесорів в цілому характерне все те ж саме, що й для сигнальних процесорів. Відмінність полягає в більш великій продуктивності за рахунок наявності вбудованого векторного співпроцесора і більш вузької спеціалізації, в більшості випадків орієнтованої на конкретний вид нейронних мереж. Таким чином, загальноприйнятих рекомендацій з вибору певної елементної бази на даний час немає. Все це досить сильно визначається особистим досвідом розробника, його уподобанням, а також доступністю на даній території тих або інших електронних компонентів, і, звичайно, багато в чому залежить від вимог самого проекту.

Як вже було написано вище, переважна кількість нейромережних реалізацій на даних час здійснюється на базі процесорів загального призначення. При цьому відповідні реалізації можна робити на основі універсальних систем розробки програмного забезпечення, або використовувати спеціальні програмні пакети. Перший спосіб є більш гнучким, оскільки не накладає жодних обмежень на вимоги замовника. Але розробка нейромереж "з нуля" є досить трудомістким процесом, який

потребує значних затрат часу на розробку.

Тому значне поширення здобули готові програмні рішення реалізації нейронних мереж. Цей клас програм дає повний доступ до створення нейромережі і її навчання, а також дослідження результатів. Вони дозволяють як навчатися теорії роботи нейромереж, так і проводити розрахунки, прогнози і дослідження за допомогою них, а також здійснювати ряд прикладних задач, що складно реалізовуються за допомогою стандартних засобів. Такі програми називаються нейросимуляторами. Далі описано найбільш розповсюджені нейросимулятори.

2.4 Модель формального нейрона вертикально-групового типу з мультиплексуванням шин вагових коефіцієнтів і даних

Розроблено три моделі формального нейрона вертикально-групового типу, особливістю яких є порозрядне надходження та опрацювання вхідних даних і вагових коефіцієнтів, таблична реалізація функції активації і порозрядне формування результату [28].

Перша модель використовує мультиплексування шин, за яким вхідні дані X_j та вагові коефіцієнти W_j надходять почергово порозрядними зрізами (W_{ji}, x_{ji}). Аналітично цю модель записують так:

$$y = f_{(p-s)}\left(f_a\left(f_z\left(f_{P_{Mi}}\left(f_{P_{ji}}\left(f_{(s-p)}(W_{ji}), x_{ji}\right)\right)\right)\right)\right), \qquad (2.19)$$

де y – вихід результату;

$f_{(p-s)}: R^n \to R^1$ – функціонал паралельно-послідовного перетворення;

f_a – функція активації;

f_Z – підсумовування макрочасткових результатів відповідно до

формули $\mathrm{Z_i} = 2^{-1}\mathrm{Z_{i-1}} + \mathrm{P_{Mi}}, \mathrm{Z_0} = 0$;

$\mathrm{f_{P_{Mi}}}$ – формування макрочасткового результату за формулою $\mathrm{f_{P_{Mi}}} = \sum_{\mathrm{j=1}}^{\mathrm{N}} \mathrm{P_{ji}}$;

$\mathrm{f_{P_{ji}}}$ – формування часткових результатів згідно з формулою $\mathrm{f_{P_{ji}}} = \mathrm{W_j x_{ji}}$;

$\mathrm{f_{(s-p)}}: \mathrm{R^1} \to \mathrm{R^n}$ – оператор послідовно-паралельного перетворення.

Друга модель формального нейрона характеризується одночасним порозрядним надходженням і опрацюванням вхідних даних і вагових коефіцієнтів старшими розрядами вперед. Аналітично цю модель записують так:

$$y = f_{(p-s)}(f_a\left(f_z\left(f_{P_{Mi}}\left(f_{P_{ji}}\left(f_{P_{ji}}f_{(s-p)}(W_{ji}), f_{(s-p)}(x_{ji})\right)\right)\right)\right)) \quad (2.20)$$

Особливістю такої моделі є суміщення у часі процесів надходження розрядних зрізів вхідних даних і вагових коефіцієнтів, їхніх послідовно-паралельних перетворень і формування часткових результатів P_{ji}, що забезпечує збільшення швидкодії порівняно з першою моделлю нейрона. Формування часткових результатів у цій моделі відбувається так:

$$P_{ji}\begin{cases} 0, \text{ якщо } w_{ji} = x_{j1} = 0 \\ x_{j0}x_{j1} \dots x_{j(i-1),} \text{ якщо } w_{ji} = 1, x_{j1} = 0 \\ w_{j0}w_{j1} \dots w_{j(i-1),} \text{ якщо } w_{ji} = 0, x_{j1} = 1 \\ w_{j0}w_{j1} \dots w_{j(i-1)} + w_{j0}w_{j1} \dots w_{j(i-1),} \text{ якщо } w_{ji} = x_{j1} = 1 \end{cases} \quad (2.21)$$

Третя модель формального нейрона орієнтована на використання у нейромережах, у яких вагові коефіцієнти є постійними або змінюються дуже рідко. Аналітично третю модель формального нейрона записують так:

$$y = f_{(p-g)}f_a(f_Z\left(f_{P_{Mi}}(x_{ji})\right)). \quad (2.22)$$

Особливістю цієї моделі є табличний спосіб формування

макрочасткових результатів для кожного розрядного зрізу вхідних даних. Макрочастковий результат формується згідно з формулою:

$$P_{Mi}\begin{cases} 0,\ \text{якщо } x_{1i}=x_{2i}=x_{3i}=\cdots=x_{Ni}=0 \\ W_1,\text{якщо } x_{1i}=1; x_{2i}=x_{3i}=\cdots=x_{Ni}=0 \\ W_2,\text{якщо } x_{1i}=0; x_{2i}=1;\ x_{3i}=\cdots=x_{Ni}=0 \\ W_1+W_2,\text{якщо } x_{1i}=1; x_{2i}=1;\ x_{3i}=\cdots=x_{Ni}=0 \\ \quad\cdots \qquad\qquad \cdots \qquad\qquad \cdots \\ W_2+W_3+\cdots+W_N,\text{якщо } x_{1i}=0; x_{2i}=x_{3i}=\cdots=x_{Ni}=1 \\ W_1+W_2+\cdots+W_N,\text{якщо } x_{1i}=x_{2i}=x_{3i}=\cdots=x_{Ni}=1 \end{cases} \quad (2.23)$$

Модель формального нейрона вертикально-групового типу з мультиплексуванням шин вагових коефіцієнтів і даних аналітично записується так:

$$y = f_{(p-g)}(f_a(f_z(f_{P_{Mk}}(f_{P_{jk}}(f_{P_{jk}}, f_{(g-p)}(W_{jkk})x_{jkr} \quad (2.24)$$

де y – вихід результату;

$f_{(p-g)}: R^n \to R^g$ – функціонал паралельно-групового перетворення результату Y;

$f_a(Z)$– функція активації;

f_Z– обчислення суми макрочасткових результатів у відповідності до формули $Z_h = 2^{-h}Z_{i-1} + P_{Mi}, Z_0 = 0$;

f_{PMi}– обчислення h-о макрочасткового добутку P_{Mh}шляхом підсумовуванням h-х часткових добутків P_{jh} у відповідності з формулою $P_{Mh} = \sum_{j=1}^{N} P_{jh}$;

f_{Pjh}– обчислення для j-ї пари операндів групового часткового добутку P_{jh} у відповідності з формулою $P_{jh} = \sum_{r=1}^{k} 2^{-(r-1)} W_j X_{jhr}$;

f_{Pjhr}– формування для кожної j-ї пари операндів k часткових добутків у відповідності з формулою $f_{jhr} = W_j X_{jhr}$, де r = 1, … ,k; $f_{(g-p)^{w_{jhk}}}: R^g \to R^n$ – оператор групового паралельного перетворення вагових коефіцієнтів W_j.

Структуру моделі формального нейрона паралельно-вертикального типу з мультиплексуванням шин вагових коефіцієнтів і даних, яка реалізує

вираз (2.24), подано на рисунку 2.2.

Основними компонентами даної моделі є: групово-паралельні перетворювачі $f(g-p)_j$, формувачі часткових добутків P_{jhr}, K-входові та N-входовий суматори, підсумовувач макрочасткових результатів $Z_h = 2^{-k}Z^{h-1} + P_{Mh}$, обчислювач функції активації $f_a(Z)$ і паралельно-груповий перетворювач $f(p-s)$.

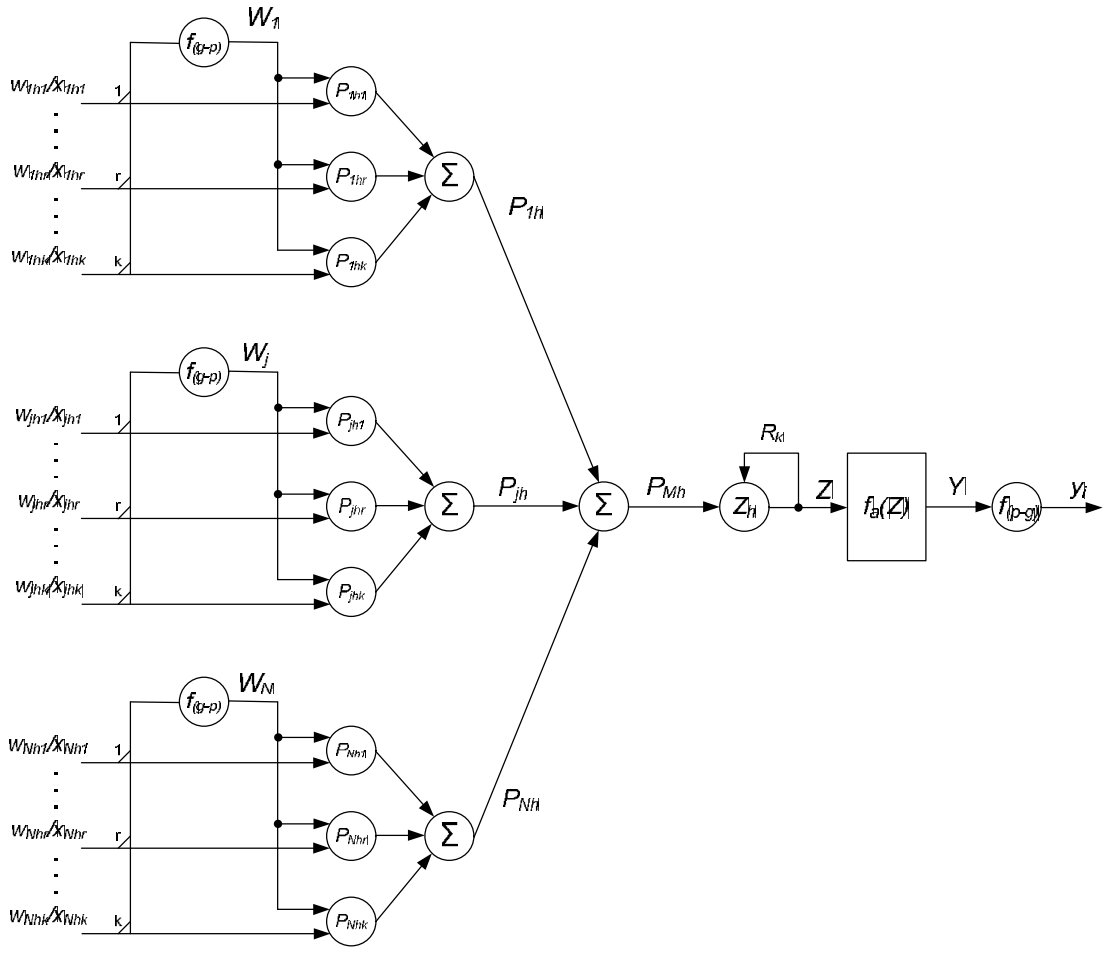


Рисунок 2.2 - Модель формального нейрона вертикально-групового типу з мультиплексуванням шин вагових коефіцієнтів і даних

Для зменшення розрядності підсумовувача макрочасткових результатів надходження вагових коефіцієнтів, вхідних даних і формування часткових добутків P_{jhr} доцільно здійснювати починаючи з молодших розрядів. Особливістю даної моделі є надходження вагових коефіцієнтів і

даних групами з k розрядів та формування для кожної j-ї пари операндів k часткових добутків P_{jhrr}. Така особливість, забезпечує зменшення в k-разів кількості тактів роботи. Одним із критеріїв вибору кількості розрядів у групі є узгодження інтенсивності надходження даних із обчислювальною здатністю нейроелемента.

Для синтезу формального паралельного нейрона вертикально-групового типу з мультиплексуванням шин вагових коефіцієнтів і даних розроблено структуру, яку подано на рисунку 2.2, де:

1) $x_{11}/w_{11},\ldots, x_{Nk}/w_{Nk}$ – Nk мультиплексованих однорозрядних інформаційних входів;

2) TI_1, TI_2 – перший та другий тактові входи;

3) ПУ – вхід початкової установки;

4) $БФЧД_j$ – j-й блок формування часткових добутків;

5) $Pr\,x_{jh}$, $Pr\,P_{jh}$, $Pr\,P_{Mh}$ та $Pr\,Z_h$ – регістри відповідно: множника,

де h-о групового часткового добутку h=1,…,m, $m = \left[\frac{n}{k}\right]$,

n – розрядність множника W_j,

k – кількість розрядів у групі),

h-о макрочасткового добутку та h-ї суми макрочасткових результатів;

6) ФЧД – формувач часткових добутків;

7) См – суматор;

8) ТФА – таблична функція активації.

Перед початком роботи імпульсом початкової установки, який надходить із входу ПУ, регістри $Pr\,x_{jh}$, $Pr\,P_{jh}$, $Pr\,P_{Mh}$ та $Pr\,Z_h$ у всіх блоках БФЧД встановлюються в нуль.

Функціонування формального паралельного нейрона вертикально-групового типу з мультиплексуванням шин вагових коефіцієнтів і даних можна розбити на два етапи, кожний з яких виконується за m тактів.

На рисунку 2.3 подано структуру формального паралельного нейрона вертикально-групового типу з мультиплексуванням шин вагових коефіцієнтів і даних.

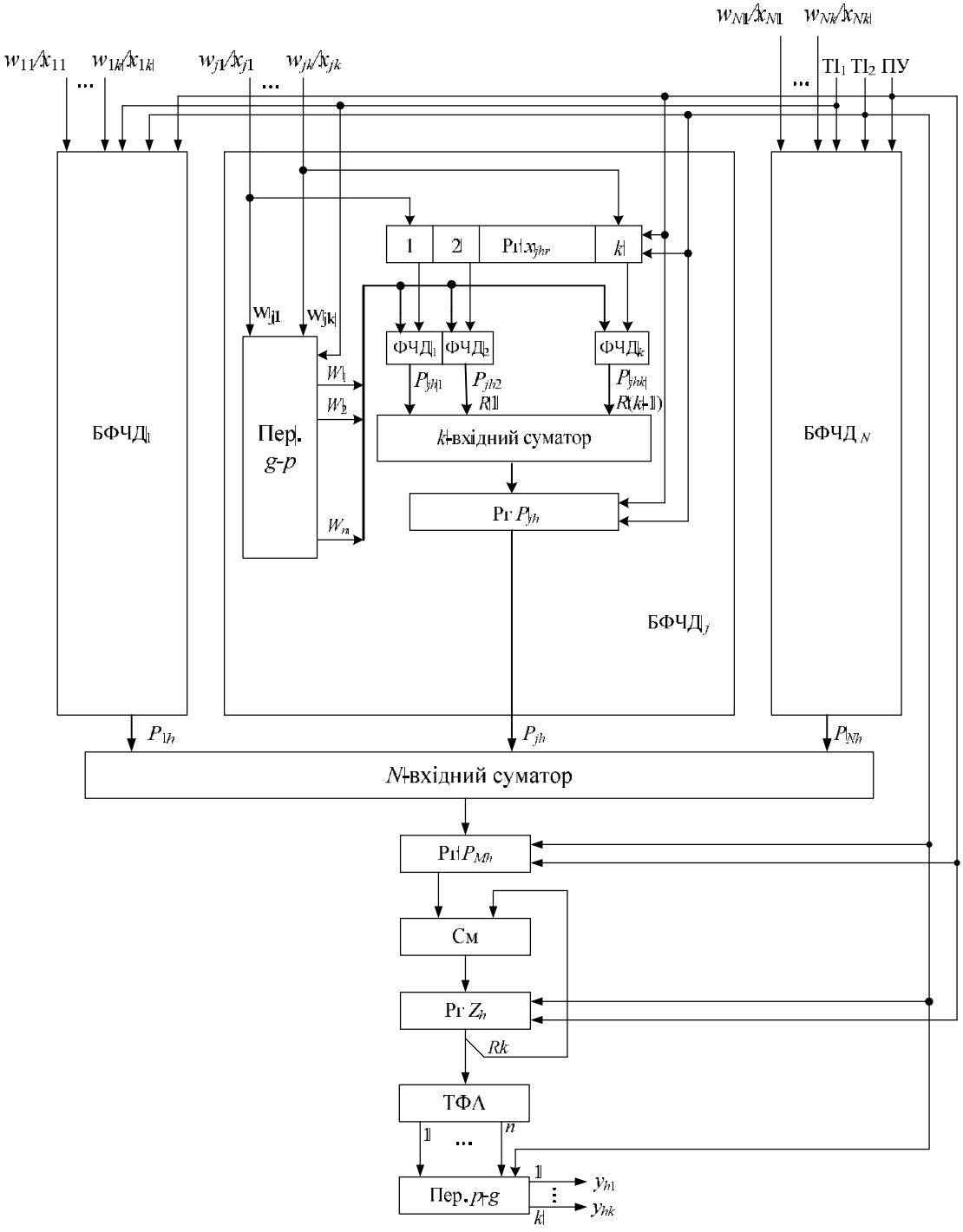


Рисунок 2.3 - Структура формального паралельного нейрона вертикально-групового типу з мультиплексуванням шин вагових коефіцієнтів і даних

На першому етапі в кожному h-у такті, починаючи з молодших розрядів, на j-й інформаційний вхід надходять k розрядів множника W_j. В перетворювачі g-р виконується накопичення та паралельне перетворення множеного W_j.

На другому етапі у кожному h-у такті роботи в блоці $БФЧД_j$ для h-ї групи розрядів множника $X_{jk1}, X_{jk2}, \dots, X_{jkk}$ на виходах $ФЧД_1$,…, $ФЧД_k$ формується k часткових добутків у відповідності з формулою $P_{jhr} = W_j X_{jhr}$. Сформовані часткові добутки надходять на вхід k-входового суматора, причому r-й (r=1,…,k) частковий добуток $W_j X_{jhr}$ зсунутий відносно (r-1)-о часткового добутку $W_j X_{jh(r-1)}$ на один розряд вправо. Шляхом додавання часткових добутків на виході k-входового суматора отримуємо груповий частковий добуток P_{jh} у відповідності з формулою:

$$P_{jh} = \sum_{r=1}^{k} 2^{-(r-1)} W_j X_{jhr} \qquad (2.25)$$

Сформований груповий частковий добуток P_{jh} записується у регістр $РгP_{jh}$. Групові часткові добутки P_{jh} з виходів регістрів $РгP_{jh}$ надходять на входи m-входового суматора, де вони додаються. На виході m-входового суматора отримуємо h-й макрочастковий добуток P_{Mh} у відповідності з формулою:

$$P_{Mh} = \sum_{j=1}^{N} P_{jh} \qquad (2.26)$$

Макрочастковий добутку P_{Mh} з виходів m-входового суматора записується в регістр $РгP_{Mh}$. На суматорі См здійснюється підсумовування макрочасткових добутків P_{Mh} у відповідності з виразом:

$$Z_h = 2^{-h} Z_{h-1} + P_{Mh}, \text{ де } Z_0 = 0 \qquad (2.27)$$

Накопичена сума Z з виходу суматора СМ тактовим імпульсом з входу записується в регістр результату $РгZ_h$ В нейроелементі обчислення функції активації $f_a(Z)$ здійснюється табличним шляхом.

Обчислення результату *y* в даному нейроелементі виконується за час:

$$t_{HE} = (2m + 3)t_{TI} = (2m + 3)(t_{P_z} + t_{mC_м}) \qquad (2.28)$$

де n – розрядність множників; k≥3;

k – кількість розрядів множників, які одночасно аналізуються для отримання групового часткового добутку P_{jh};

t_{TI} – тривалість такту;

$t_{Pг}$ – час запису в регістр;

$t_{mCм}$ – час додавання m чисел.

Отже, в даному розділі було розроблено алгоритм реалізації паралельного нейроелемента вертикально-групового типу. Також розроблено такі нейроелементи та нейромережі найдоцільніше на основі інтегрального підходу, який охоплює сучасну елементну базу, моделі та НВІС-структури нейроелементів, архітектури нейромереж, паралельні методи обчислень і враховує вимоги конкретних застосувань. Для синтезу нейроелементів і нейромереж реального часу з високою ефективністю використання обладнання було розроблено нові моделі, паралельні методи опрацювання даних і НВІС-структури нейронів, які забезпечують узгодження інтенсивності надходження даних з їхньою обчислювальною здатністю. Тому розроблено НВІС-нейроелементи, які орієнтовані на синтезі апаратних нейромереж реального часу.

3 РОЗРОБКА І СИНТЕЗ КОМПОНЕНТІВ ПАРАЛЕЛЕЛЬНОГО НЕЙРОНА ВЕРТИКАЛЬНО-ГРУПОВОГО ТИПУ

3.1 Вибір апаратних засобів моделювання нейроелемента

VHDL – мова опису апаратних засобів яка широко використовується для проектування програмованих логічних інтегральних схем (ПЛІС) та надвеликих інтегральних схем (НВІС).

Розробка сучасних інтегральних мікросхем є складною інженерно-технічною проблемою, вирішення якої неможливе без широкого застосування систем автоматизованого проектування. Створення подібних систем потребує засобів формального опису структур і функцій об’єктів проектування.

Мова VHDL (Very high speed integrated circuits Hardware Description Language) була розроблена у 1980 році в результаті реалізації в США проекту по створенню надшвидкісних інтегральних схем. У 1987 році Інститутом Інженерів з Електрики та Електроніки (IEEE) ця мова була визнана в якості стандарту США. Зараз VHDL є найбільш поширеною у світі мовою такого призначення. Вона застосовується у багатьох системах автоматизованого проектування, кількість користувачів яких стрімко зростає [59].

VHDL служить для опису моделі цифрового пристрою (приладу, системи). Опис на мові VHDL визначає зовнішні зв'язки пристрої ("вид ззовні" або інтерфейс) і один або декілька "видів зсередини" (рисунок 3.1).

Вид зовні задає інтерфейс пристрою, набір сигналів, якими обмінюється з зовнішнім світом. Цей вид описує абстрактне уявлення пристрої "в цілому" і позначається англійським терміном entity, що в дослівному перекладі означає «сутність» і найбільш точно відображає зміст вистави.

Однак у літературі термін «сутність» не знайшов широкого розповсюдження, для позначення зовнішнього опису об'єкта використовуються терміни «інтерфейс об'єкта», «декларативна частина» та

інші. У цьому посібнику буде використовуватися термін «інтерфейс об'єкта» або просто «інтерфейс».

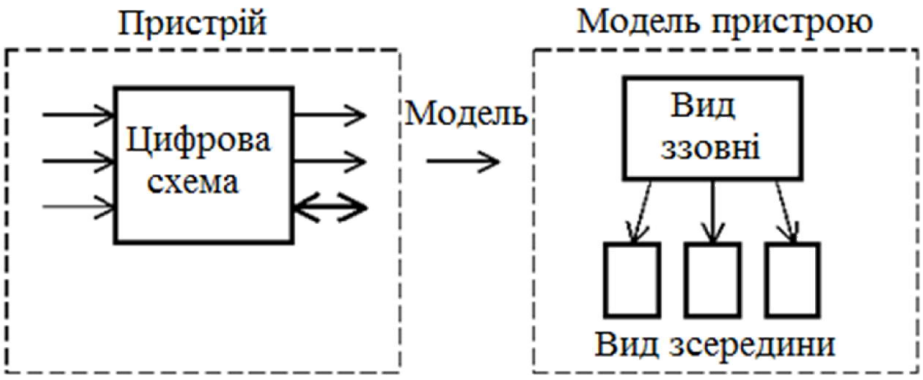


Рисунок 3.1 - Цифровий пристрій і його модель

Вид зсередини визначає функціональні можливості пристрою або його структуру. Внутрішня будова об'єкта визначає архітектура (architecture body).

Як і в мовах програмування, мова VHDL має свої правила, в тому числі правила опису імен змінних, об'єктів, типів даних та інших параметрів. Основні правила мови VHDL описані в наступних розділах. Кожен об'єкт в мові VHDL може зберігати значення, відносяться до певного набору. Це безліч значень декларується за допомогою оголошення типу (type declaration). Тип - це ім'я, яке зв'язується з певним набором значень і набором операцій. Деякі типи зумовлені мовою VHDL. Наприклад, BOOLEAN має набір значень FALSE, TRUE і набір операторів: and, or, nor, nand, not.

У мові VHDL є можливість створювати нові типи з використанням декларацій і завдання набору операцій. Всі можливі типи в VHDL розпадаються на чотири великі категорії:

1) scalar (скалярні);

2) composite (композитні) - вони складаються з елементів одного типу (масиви) або різного типу (записи);

3) access type (типи доступу) - забезпечують доступ до даного типу через покажчики;

4) file types (тип - файл) - забезпечує доступ до об'єктів, що містить послідовності значень даного типу.

У свою чергу скалярні типи поділяються на чотири види:

1) Enumeration (перераховувані тип)
2) Integer (цілий тип)
3) Physical (фізичний тип)
4) Floating point (тип "з плаваючою комою").

У мові VHDL є операції наступних категорій:

1) Логічні операції
2) Операції відносин
3) Операції додавання / віднімання
4) Операції множення / ділення
5) Інші

Verilog HDL (Verilog Hardware Description Language) — мова опису апаратури (HDL), що використовується для опису та моделювання електронних систем. Verilog HDL не слід плутати з VHDL (конкуруюча мова), найбільш часто використовується у проектуванні, верифікації і реалізації (наприклад, у вигляді HBIC) аналогових, цифрових та змішаних електронних систем на різних рівнях абстракції [60].

Розробники Verilog зробили його синтаксис дуже схожим на синтаксис мови C, що спрощує його освоєння. Verilog має препроцесор, дуже схожий на препроцесор мови C і основні керуючі конструкції if, while також подібні однойменним конструкціям мови C.

Слід зазначити, що опис апаратури, написаний мовою Verilog (як і іншими HDL-мовами) прийнято називати програмами, але, на відміну від загальноприйнятого поняття програми, як послідовності інструкцій, тут програма представляє множину операторів, які виконуються паралельно і циклічно під керуванням об'єктів, названих сигналами. Кожен такий оператор є моделлю певного елемента реальної функціональної схеми апаратури, а сигнал — аналогом реального логічного сигналу. Так само для

мови Verilog не застосовується термін «виконання програми». Фактично, виконання Verilog-програми є моделюванням функціональної схеми, яку вона описує, що виконується спеціальною програмою — Verilog-симулятором [59].

Існує підмножина інструкцій мови Verilog, придатна для синтезу. Модулі, які написані в межах цієї підмножини, називають RTL (англ. register transfer level) — рівень регістрових передач). Вони можуть бути фізично реалізовані з використанням САПР синтезу.

САПР за певними алгоритмами перетворить абстрактний вихідний Verilog-код на перелік зв'язків — логічно еквівалентний опис, що складається з елементарних логічних примітивів (наприклад, елементи AND, OR , NOT татригери), які доступні у вибраній технології виробництва НВІС або програмування ПЛІС.

ModelSim - система HDL-моделювання цифрових пристроїв. Одна з характерних тенденцій сучасного етапу розвитку технології проектування цифрових систем - застосування мов опису апаратури HDL (Hardware Description Language) високого рівня, серед яких найбільшого поширення набули VHDL і Verilog.

При цьому не тільки зростає роль засобів моделювання в процесі розробки пристроїв, але і змінюються методи і вимоги, що пред'являються до процедур верифікації. Для своєчасного виявлення можливих помилок кошти моделювання повинні забезпечувати можливість контролю результатів кожного етапу процесу проектування: створення вихідних HDL-описів, синтезу, розміщення і трасування в кристал. Такий підхід забезпечує мінімальний час розробки пристрою і скорочує вартість цього процесу, так як ціна помилки зростає з кожним наступним кроком проектування.

Active-HDL — середовище розробки, моделювання та верифікації проектів для програмованих логічних інтегральних схем, розроблене фірмою Aldec. Перша версія програмного забезпечення вийшла в 1997 році.

Програма дозволяє створювати описи пристроїв за допомогою мов опису апаратури, а також за допомогою структурних схем. Спочатку програма підтримувала тільки мову VHDL, але з часом додалася підтримка мови Verilog [59].

Програми також дозволяє створювати графічні моделі кінцевих автоматів та конвертувати HDL опис в графічні структурні схеми і назад. Також вона забезпечена потужним ядром моделювання. Підтримується спільна робота з програмами MATLAB і Simulink.

Таблиця 3.1 - Компоненти Active-VHDL

Компонент	Опис
Console	Вікно консолі - це інтерактивне вікно, призначене для введення виведення тексту забезпечує введення команд Active-VHDL. Висновок повідомлень згенерованих Active-VHDL інструментальними засобами.
Design Browser	Вікно перегляду Проекту показує поточне утримання проекту, включаючи: файли ресурсів, приєднаних до проекту, зміст файлів робочих бібліотек визначаються за умолчанію, структуру проектованих пристроїв обраних для моделірованія, сигнали і змінні, оголошені всередині обраної області поточного проекту.
HDL Editor	HDL редактор - текстовий редактор, розроблений для створення вихідних файлів VHDL. Він показує певні категорії синтаксису в різних кольорах. Редактор глибоко інтегрований з моделюючим пристроєм, щоб дозволяє просто налагоджувати вихідний текст.
Language Assistant	Мовний Помічник є допоміжним інструментом який забезпечує ряд типових VHDL шаблонів і іхлогіческое перетворення в функціональні блоки.

Продовження таблиці 3.1 - Компоненти Active-VHDL

State Machine Editor	Редактор автоматів з кінцевими станами - графічний інструмент, розроблений для редагування діаграммконечного автомата. Редактор здійснює автоматичний переклад графічних примітивів в коди VHDL.
Waveform Viewer	Просмотршік Форми сигналу показує результати моделювання, під дією тестових сигналів. Це дозволяє нам графічно редагувати форму сигналу.
List	Вікно списку показує результати моделювання, виконані в зведеному в таблиці текстовому форматі. Це дозволяє нампрослеідіть результати моделювання з точністю до дельти циклу.
Watch	Вікно середовища спостереження показує поточні значення вибраних сигналів і змінних протягом моделювання.
Processes	Вікно Процесів показує поточний стан одночасних процесів в розробленому проекті протягом моделювання.
Library Manager	Бібліотечний менеджер розроблений, щоб управляти VHDL бібліотеками та їх змістом.
Design Explorer	Дослідник Проекту полегшує управління Active - VHDL проектами. Він дозволяє не запам'ятовувати фізичну локалізаціюфайлов проекту.
Script Editor	Редактор сценарію - текстовий редактор з вмонтованим отладчиком. Він розроблений, щоб редагувати VBA Basic сценарії і виконувати команди Active-VHDL.

Xilinx ISE являє собою програмний інструмент виробництва для синтезу та аналізу HDL конструкцій, що дозволяє розробнику синтезувати ("компілювати") свої проекти, виконувати тимчасової аналіз, вивчити схеми RTL і налаштувати пристрій програмно [61].

Xilinx ISE це середовище розробки для FPGA продуктів від Xilinx, щільно з'єднані з архітектурою чіпів, і не можуть бути використані з FPGA продуктом від інших постачальників [62]. Xilinx ISE в основному використовується для синтезу схем і дизайну, в той час як ISIM або логіка симулятор ModelSim використовується для тестування на системному рівні.

Основне призначення для користувача інтерфейсу ISE є Project Navigator, який включає в себе ієрархію дизайну (джерела), редактор вихідного коду (на робочому місці), вихід консолі (стенограма), і дерево процесів.

Ієрархічний дизайн складається з проектних файлів (модулів), чиї залежності інтерпретуються в ISE і відображається у вигляді дерева. Для конструкцій може бути один основний модуль, з іншими модулями, включених в основний модуль, схожий на основний підпрограми в програмах на C ++. Обмеження проектування вказані в модулях, які включають конфігурацію роз'єму і відображення.

Ієрархія процесів описує операції, які ISE буде виконувати на поточному активному модулі. Ієрархія включає в себе функції компіляції, їх функції залежності та інших послуг. Вікно також позначає проблеми або помилки, які виникають при кожній функції [63].

Тестування на системному рівні може бути виконана з ISIM або логічний симулятор ModelSim, і такі тестові програми також повинні бути написані на мовах HDL. Випробувальний стенд програми можуть включати в себе імітацію осцилограм вхідного сигналу, або монітори, які спостереження та контроль вихідних сигналів тестованого пристрою.

ModelSim або ISIM можуть бути використані для виконання наступних видів моделювання:

1) логічна перевірка, щоб забезпечити модуль очікування результату;

2) поведінкова перевірка, щоб перевірити логічні і тимчасові проблеми;

3) пост-місце і моделювання маршруту, щоб перевірити поведінку після розміщення модуля в межах перебудовується логіки FPGA.

3.2 Структура нейроелемента вертикально-групового типу

Структура паралельного нейроелемента вертикально-групового типу визначається за такими ознаками:

1) методи обчислення;
2) режими роботи;
3) спосіб організації зв’язків між елементами .

За режимами роботи нейроелементи можна поділити на синхронні та асинхронні. В останньому випадку такі нейроелементи називають однотактними, оскільки опрацьовують вхідні дані без проміжних запам’ятовувань. Швидкодія однотактного нейроелемента визначається часом спрацювання елементів, що лежать на найдовшому шляху проходження даних. Однотактні нейро- елементи є послідовними з погляду реалізації ярусів потокового графу алгоритму. У однотактних нейроелементах процесорні елементи (ПЕ), які реалізують функціональні оператори, починаючи з верхнього ярусу, поступово виключаються з процесу обчислення і не використовуються до завершення обчислення. Це є причиною обмеженої швидкодії і неефективного використання обладнання при обробці потоків даних.

Для опрацювання інтенсивних потоків даних доцільно використовувати синхронні паралельні нейроелементи вертикально-групового типу, опрацювання в яких здійснюється за конвеєрним

принципом. Конвеєрні нейроелементи розбиваються на сходинки буферними регістрами. Для забезпечення високої швидкодії та ефективності використання обладнання ПЕ сходинок кон- веєрних нейроелементів повинні реалізовувати якомога простіші функціональні оператори з приблизно рівним часом виконання. У конвеєрі при подачі тактових імпульсів здійснюються обчислення функціональних операторів і запис результатів у буферні регістри. Обчислювальну здатність конвеєрних нейроелементів визначають так:

$$D_{\text{ОП}} = \frac{m_0 n_0}{t_{p_2} + t_{\text{ПЕ}}}, \tag{3.1}$$

де t_{p_2}- час запису в регістр;

$t_{\text{ПЕ}}$- час працювання ПЕ

Варіанти реалізації нейроелемента при паралельному вертикально-груповому опрацюванні даних залежать від:

1. Способу надходження даних:

1) паралельно-групове надходження вхідних даних X_{jh} і вагових коефіцієнтів W_{jh};

2) почергове паралельно-групове надходження вхідних даних X_{jh} і вагових коефіцієнтів W_{jh};

1) суміщення процесу паралельно-групового надходження вхідних даних X_{jh} і табличного формування і підсумовування макрочасткових добутків P_{Mi}.

2. Формування для кожного групового зрізу добутків $W_j X_{jkk}$:

1) з прямим формуванням;

2) на базі попередніх обчислень.

3. Обчислення групового часткового добутку P_{jh}:

1) послідовне;

2) паралельне;

3) послідовно-паралельне;

4) табличне.

4. Обчислення макрочасткового добутку P_{Mh}:

1) послідовне;

2) паралельне;

3) послідовно-паралельне;

4) табличне.

5. Обчислення функції активації f:

1) послідовне;

2) паралельне;

3) табличне та таблично-алгоритмічне.

Підвищення швидкодії паралельного вертикально-групового опрацювання даних у нейроелементі можна досягнути такими шляхами:

1) зменшенням часу формування часткових добутків $W_j X_{jkk}$;

2) збільшенням розрядності *g* груп надходження та опрацювання вхідних даних X_j та вагових коефіцієнти $W_{j;}$

3) зменшенням часу обчислення групового часткового добутку P_{jh};

4) зменшенням часу обчислення макрочасткового добутку P_{jh};

5) зменшенням часу підсумовування макрочасткових добутків P_{Mh}.

Для підвищення швидкодії паралельно- вертикального опрацювання даних у нейроелементах і нейромережах потрібно використовувати поряд з просторовим розпаралелюванням часове (конвеєр). За допомогою зменшення складності операцій, які реалі- зуються сходинкою конвеєра, підвищується тактова частота роботи конвеєра, за допомогою чого досягається підвищення швидкодії. Використання паралельно-вертикального опрацювання даних у нероелементах та нейромережах забезпечує зменшення кількості виводів нейроелементів, розрядності міжнейронних зв’язків та зменшує витрати обладнання.

Високопродуктивні апаратні нейромережі реального часу синтезуються

на основі швидкодіючих нейроелементів, в яких висока швидкодія досягається за рахунок розпаралелювання обчислень як в часі, так і в просторі. Швидкодія нейроелемента в значні мірі визначається часом реалізації базової операції – обчислення скалярного добутку:

$$Z = \sum_{j=1}^{k} w_j X_j, \tag{3.2}$$

де k – кількість входів нейроелемента,

w_j – j -й ваговий коефіцієнт,

X_j– значення j-го входу.

Наявні методи обчислення скалярного добутку розглядають процес обчислення як виконання сукупності операцій множення та додавання, а не як виконання єдиної операції. Тому для розробки НВІС-орієнтованих алгоритмів обчислення скалярного добутку необхідно записати їх у базисі елементарних арифметичних операцій. У загальному випадку обчислення скалярного добутку в базисі елементарних арифметичних операцій зводиться до макрооперації групового підсумовування:

$$Z = \sum_{j=1}^{M} C_j, \tag{3.3}$$

де M – кількість доданків;

C_j – j-й доданок.

Нехай доданки C_j є двійковими n-розрядними додатними числами меншими за одиницю, які записуються так:

$$C_j = \sum_{i=1}^{n} 2^{-1} C_{ji}, \tag{3.4}$$

Підставивши значення у формулу , отримаємо:

$$Z = \sum_{j=1}^{M} \sum_{i=1}^{n} 2^{-i} C_{ji} \tag{3.5}$$

Дана формула відображає вертикальну модель обчислення оператора групового підсумовування. Найшвидшим варіантом реалізації моделі

групового підсумовування є вертикально-груповий метод обчислення, граф алгоритму якого поданий на рисунку 3.2.

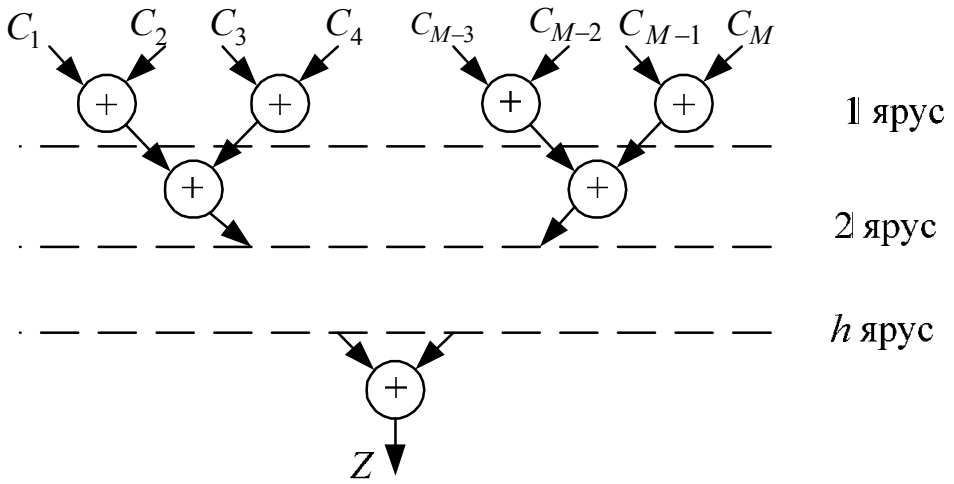


Рисунок 3.2 - Граф алгоритму вертикально-групового методу обчислення

Алгоритм вертикально-групового підсумовування є каскадним. Час обчислення суми макрооперації групового підсумовування за таким алгоритмом залежить від висоти графа (кількості ярусів), яка обчислюється так:

$$h = \lceil log_2 M \rceil, \tag{3.6}$$

де $\lceil\ \rceil$ – операція округлення до більшого цілого числа.

У кожному ярусі операнди розбиваються на пари, для кожної з яких обчислюється сума. Загальна кількість операцій додавання для обчислення суми макрооперації групового підсумовування рівна:

$$U = \frac{M}{2} + \frac{M}{4} + \frac{M}{8} + \cdots + 1 = M - 1\,, \tag{3.7}$$

Підвищити швидкодію обчислення, ефективність використання та орієнтувати структуру макрооперації групового підсумовування на НВІС-реалізацію можна шляхом використання вертикального та багатоопераднього додавання.

3.3 Реалізація нейроелемента вертикально-групового типу з використанням багатовходового суматора

У нейромережах для ефективного обчислення в реальному часі оператора вертикально-групового підсумовування доцільно використовувати вертикальний та багатоопперандний підходи для його реалізації. При використанні даних підходів процес обчислення оператора групового підсумовування розглядається як виконання єдиної операції, що ґрунтується на базовій операції додавання значень бітів розрядного зрізу, тобто зводиться до вертикальної моделі обчислення. Замінивши у останній формулі порядок підсумовування переходимо до вертикальної моделі обчислення оператора групового підсумовування, яка записується так:

$$Z = \sum_{i=1}^{n} 2^{-i} \sum_{i=1}^{M_i} C_{ji} \tag{3.8}$$

де M_i – кількість доданків у і-у розрядному зрізі.

Існуючі вертикальні методи обчислення операції групового підсумовування зводять процес обчислення до перетворення багаторядного коду в однорядний. Таке перетворення ґрунтується на базовій операції перетворення трирядного коду в дворядний:

$$\begin{cases} C_{(j-1)} \dots C_{(j-1)(n-1)} C_{(j-1)n} \\ + \\ C_{j1} \dots C_{j(n-1)} C_{jn} \\ + \\ C_{(j-1)1} \dots C_{(j-1)(n-1)} C_{(j-1)n} \end{cases} = \begin{cases} 0 S_1 \dots S_{n-1} S_n \\ + \\ P_0 P_1 \dots P_{n-1} 0 \end{cases} \tag{3.9}$$

Перетворення трирядного коду в дворядний здійснюється за допомогою шару однорозрядних суматорів, що не мають зв’язків між собою. Для зменшення часу перетворення багаторядного коду в однорядний шари

однорозрядних суматорів необхідно об’єднати за принципом дерева Уоллеса. Кількість шарів однорозрядних суматорів для обчислення оператора групового підсумовування визначається за формулою:

$$K = \lceil log_{1,5} 0{,}5M \rceil, \tag{3.10}$$

Обчислення оператора групового підсумування за таким методом розглядається як виконання єдиної операції, де одиниці переносів враховуються тільки один раз при заключному етапі перетворення дворядного коду в однорядний.

Пришвидшення процесу перетворення багаторядного коду в однорядний пропонується здійснити шляхом використання для перетворення наступних однорозрядних операцій:

$$E_{3-2} = \begin{cases} C_{ji} \\ + \\ C_{(j+1)i} \\ + \\ C_{(j+2)i} \end{cases} = \begin{cases} P_{i-1} \\ + \\ S_i \end{cases}, \quad E_{7-3} = \begin{cases} C_{ji} \\ + \\ C_{(j+1)i} \\ + \\ C_{(j+2)i} \\ + \\ C_{(j+3)i} \\ + \\ C_{(j+4)i} \\ + \\ C_{(j+5)i} \\ + \\ C_{(j+6)i} \end{cases} = \begin{cases} P_{i-1} \\ + \\ S_{i=1} \\ + \\ S_i \end{cases} \tag{3.11}$$

де $E_{3\text{-}2}$, $E_{7\text{-}3}$ – результати однорозрядних операцій додавання відповідно трьох, семи і п’ятнадцяти операндів.

Для реалізації таких операцій використовуються тривходові та семи входові однорозрядні суматори. Розробку аналітичного виразу для синтезу семивходового однорозрядного суматора будемо здійснювати поетапно.

На першому етапі розробки розбиваємо вхідні дані на дві групи таким чином: $L=C_1\ C_2\ C_3\ C_4$, $Y=C_5\ C_6\ C_7$,

На другому етапі для кожної групи формуємо аналітичні вирази для визначення кількості одиниць у групі:

1) група L

$L_0 = \overline{c_4 c_3 c_2 c_1}$; $L_1 = \overline{c_4 c_3 c_2} c_1 \vee \overline{c_4 c_3} c_2 \overline{c_1} \vee \overline{c_4} c_3 \overline{c_2 c_1} \vee c_4 \overline{c_3 c_2 c_1}$;

$L_0 = \overline{c_4 c_3 c_2 c_1}$; $L_1 = \overline{c_4 c_3 c_2} c_1 \vee \overline{c_4 c_3} c_2 \overline{c_1} \vee \overline{c_4} c_3 \overline{c_2 c_1} \vee c_4 \overline{c_3 c_2 c_1}$;

$L_2 = \overline{c_4 c_3} c_2 c_1 \vee \overline{c_4} c_3 \overline{c_2} c_1 \vee c_4 \overline{c_3 c_2} c_1 \vee \overline{c_4} c_3 c_2 \overline{c_1} \vee c_4 \overline{c_3} c_2 \overline{c_1} \vee c_4 c_3 \overline{c_2 c_1}$;

$L_3 = \overline{c_4} c_3 c_2 c_1 \vee c_4 \overline{c_3} c_2 c_1 \vee c_4 c_3 \overline{c_2} c_1 \vee c_4 c_3 c_2 \overline{c_1}$; $L_4 = c_4 c_3 c_2 c_1$.

2) група Y

$Y_0 = \overline{c_7 c_6 c_5}$; $Y_1 = \overline{c_7 c_6} c_5 \vee \overline{c_7} c_6 \overline{c_5} \vee c_7 \overline{c_6 c_5}$;

$Y_2 = \overline{c_7} c_6 c_5 \vee c_7 \overline{c_6} c_5 \vee c_7 c_6 \overline{c_5}$;

$Y_3 = c_7 c_6 c_5$.

На останньому етапі на основі попередніх аналітичних виразів запишемо аналітичні вирази для синтезу 7-ми входового однорозрядного суматора:

$S_2(2^0) = Y_0 L_1 \vee Y_1 L_0 \vee Y_0 L_3 \vee Y_1 L_2 \vee Y_2 L_1 \vee Y_3 L_0 \vee Y_1 L_4 \vee Y_2 L_3 \vee$

$Y_3 L_2 \vee Y_3 L_4$;

$S_1(2^1) = Y_0 L_2 \vee Y_1 L_1 \vee Y_2 L_0 \vee Y_2 L_1 \vee Y_3 L_0 \vee Y_2 L_4 \vee Y_3 L_3 \vee Y_3 L_4 \vee$

$Y_0 L_3 \vee Y_1 L_2$;

$P(2^2) = Y_0 L_4 \vee Y_1 L_3 \vee Y_2 L_2 \vee Y_3 L_1 \vee Y_1 L_4 \vee Y_2 L_3 \vee Y_3 L_2 \vee Y_2 L_4 \vee$

$Y_3 L_3 \vee Y_3 L_4$;

де

$Y_0 = \overline{c_7 c_6 c_5}$; $Y_1 = \overline{c_7 c_6} c_5 \vee \overline{c_7} c_6 \overline{c_5} \vee c_7 \overline{c_6 c_5}$;

$Y_2 = \overline{c_7} c_6 c_5 \vee c_7 \overline{c_6} c_5 \vee c_7 c_6 \overline{c_5}$;

$Y_3 = c_7 c_6 c_5$;

$L_0 = \overline{c_4 c_3 c_2 c_1}$;

$L_1 = \overline{c_4 c_3 c_2} c_1 \vee \overline{c_4 c_3} c_2 \overline{c_1} \vee \overline{c_4} c_3 \overline{c_2 c_1} \vee c_4 \overline{c_3 c_2 c_1}$;

$L_2 = \overline{c_4 c_3} c_2 c_1 \vee \overline{c_4} c_3 \overline{c_2} c_1 \vee c_4 \overline{c_3 c_2} c_1 \vee \overline{c_4} c_3 c_2 \overline{c_1} \vee c_4 \overline{c_3} c_2 \overline{c_1} \vee c_4 c_3 \overline{c_2 c_1}$;

$L_3 = \overline{c_4} c_3 c_2 c_1 \vee c_4 \overline{c_3} c_2 c_1 \vee c_4 c_3 \overline{c_2} c_1 \vee c_4 c_3 c_2 \overline{c_1}$; $L_0 = c_4 c_3 c_2 c_1$;

$L_4 = c_4 c_3 c_2 c_1$.

На основі аналітичних виразів синтезуємо комбінаційний семивходовий однорозрядний суматор, схема якого подана на рисунку 3.3.

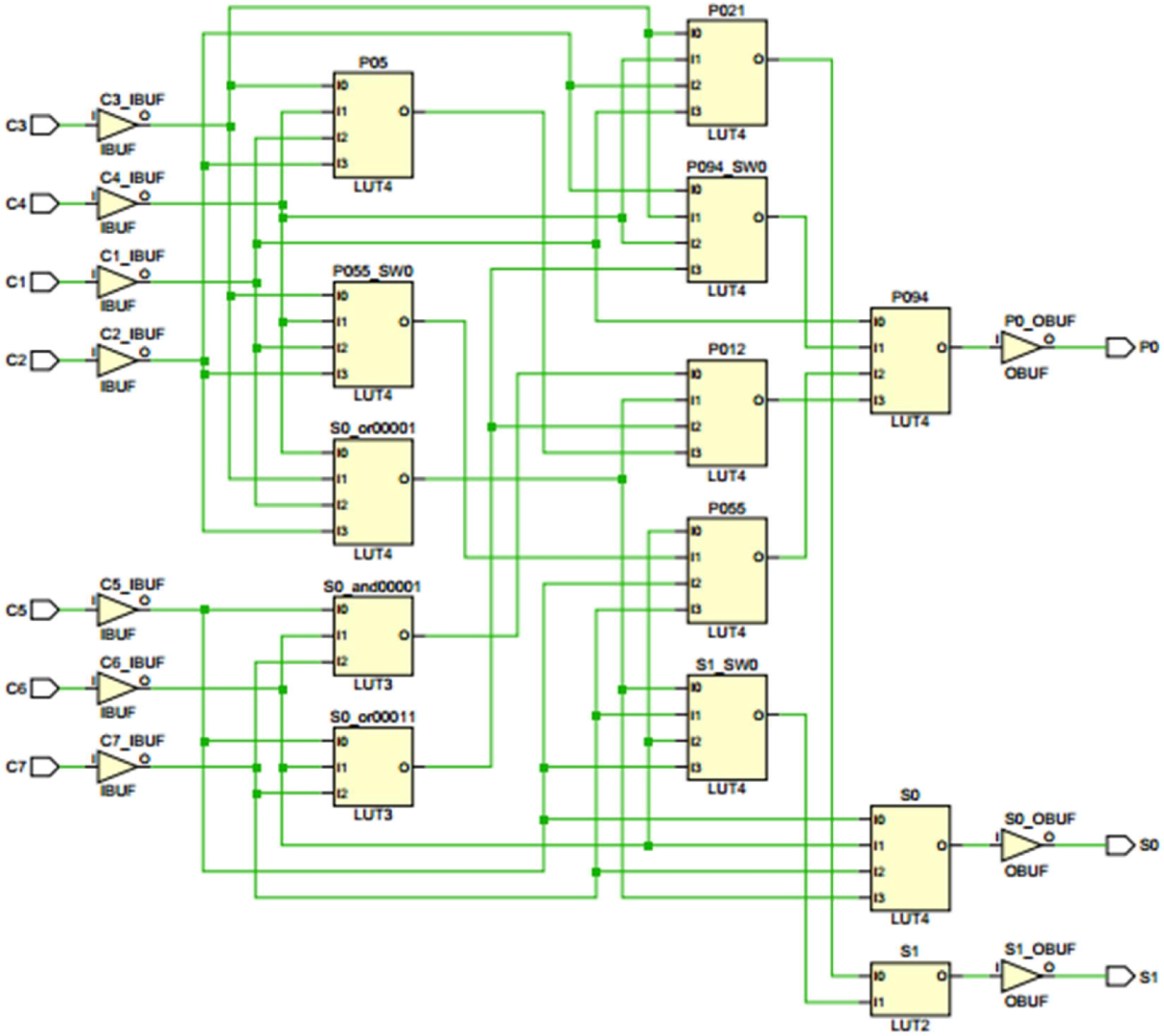


Рисунок 3.3 - Схема комбінаційного семивходового однорозрядного сумматора

Швидкодія комбінаційного семивходового однорозрядного суматора визначається часом затримки проходження даних з входу на вихід і обчислюється так:

$$t_{C_{M_{7-3}}} = 5t_{\text{лог}}I, \tag{3.12}$$

де $t_{\text{лог}}I$ – час спрацювання логічного елемента І.

Аналогічно можна розробляти аналітичні вирази для синтезу багатовходових однорозрядних суматорів з більшою кількістю входів. На рисунку 3.4 представлено схема пристрою семивходового суматора.

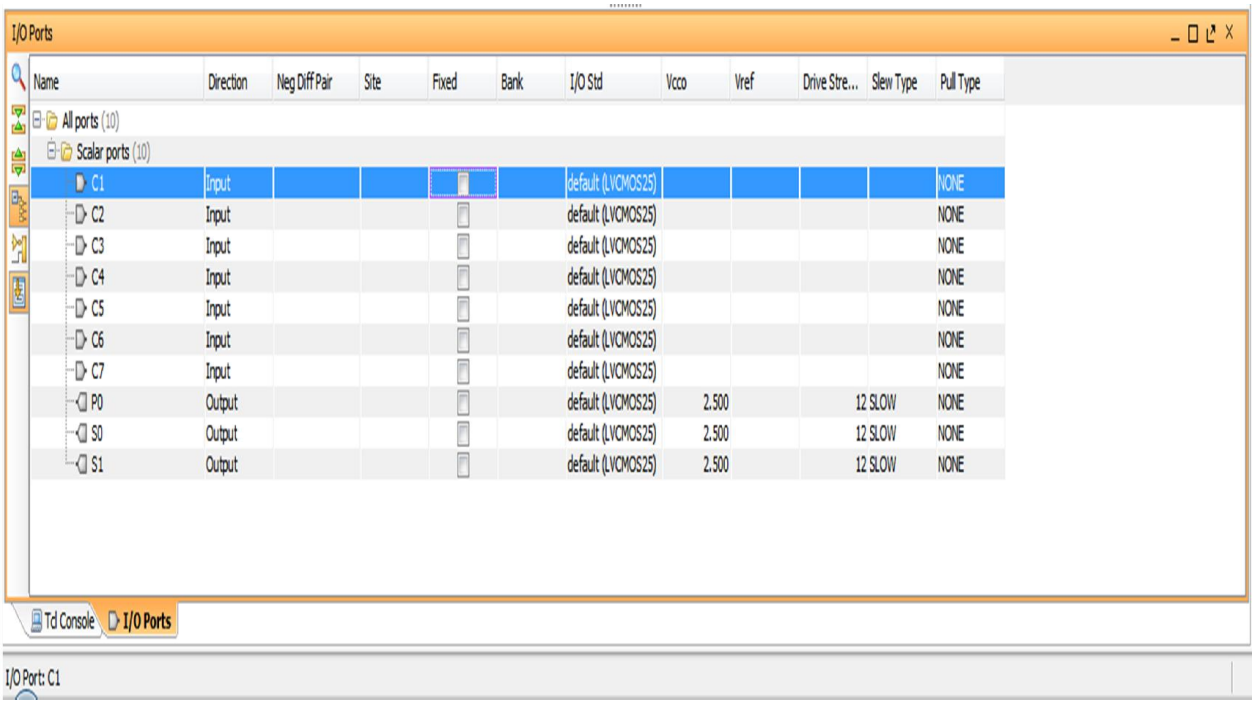


Рисунок 3.4 – Вхідні і вихідні порти багатовходового суматора

Для групового підсумовування багаторозрядних чисел використовуються багатовходові однорозрядні суматори, які синтезуються за вище розробленими аналітичними виразами.

Об’єднання таких суматорів забезпечує перетворення багаторядного коду в дворядний, який перетворюється в однорядний за допомогою паралельного суматора. На рисунку 3.5 предаствлено ядро пристрою суматора.

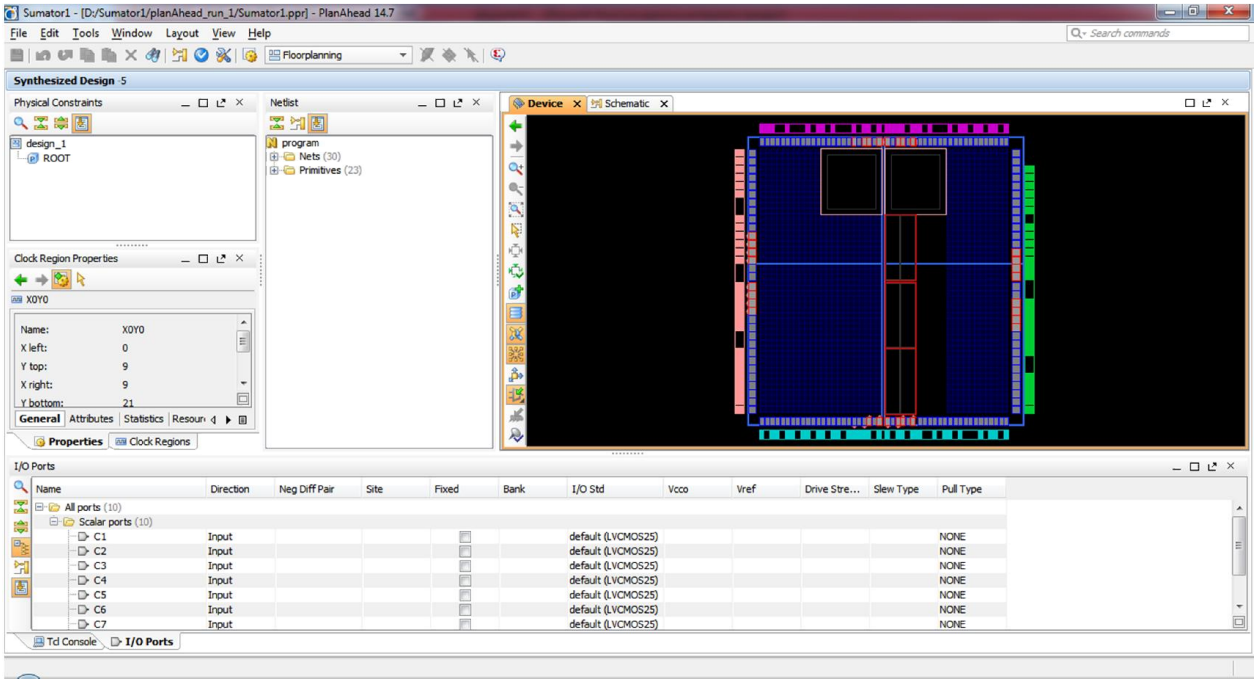


Рисунок 3.5 – Ядро пристрою

На рисунку 3.6 представлена згенерована RTL-схема пристрою для групового підсумовування багаторозрядних чисел з використанням багатовходових вхідних і вихідних портів.

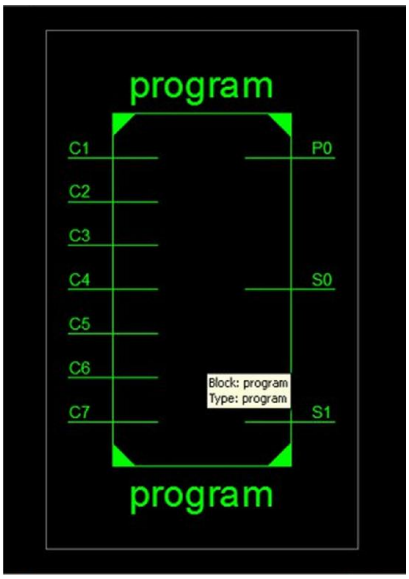


Рисунок 3.6 – RTL схема пристрою (C1 – C6 – вхідні порти, P0,S0,S1 – вихідні порти)

Внутрішня будова даного пристрою на логічних елементах подана на рисунку 3.7.

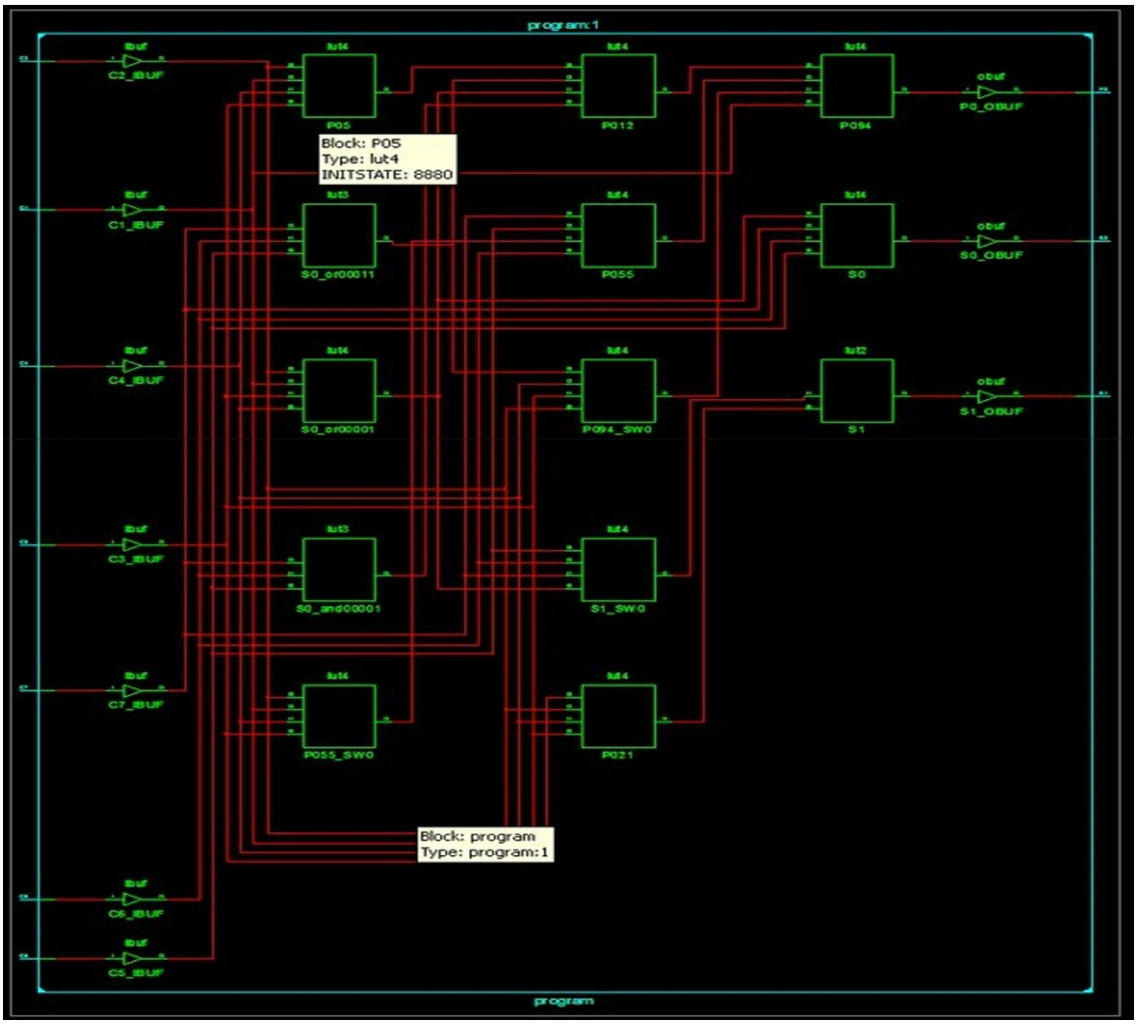


Рисунок 3.7 – Внутрішня будова пристрою

На рисунку 3.8 представлена часова діаграма для перевірки коректності роботи пристрою. Дана діаграма послідовно відображає взаємодію об’єктів впорядкованих за часом та дає змогу побачити послідовність відправлених повідомлень даного пристрою суматора.

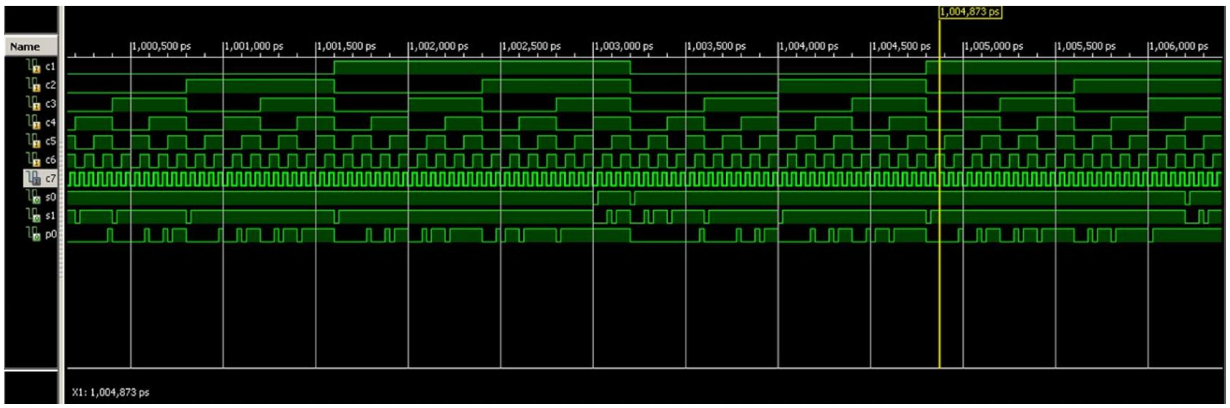


Рисунок 3.8 – Часова діаграма пристрою

В програмному продукті Xilinx було розроблено схему розміщення вхідних і вихідних портів на кристалі ПЛІС, дану схему показано на рисунку 3.9 та рисунку 3.10.

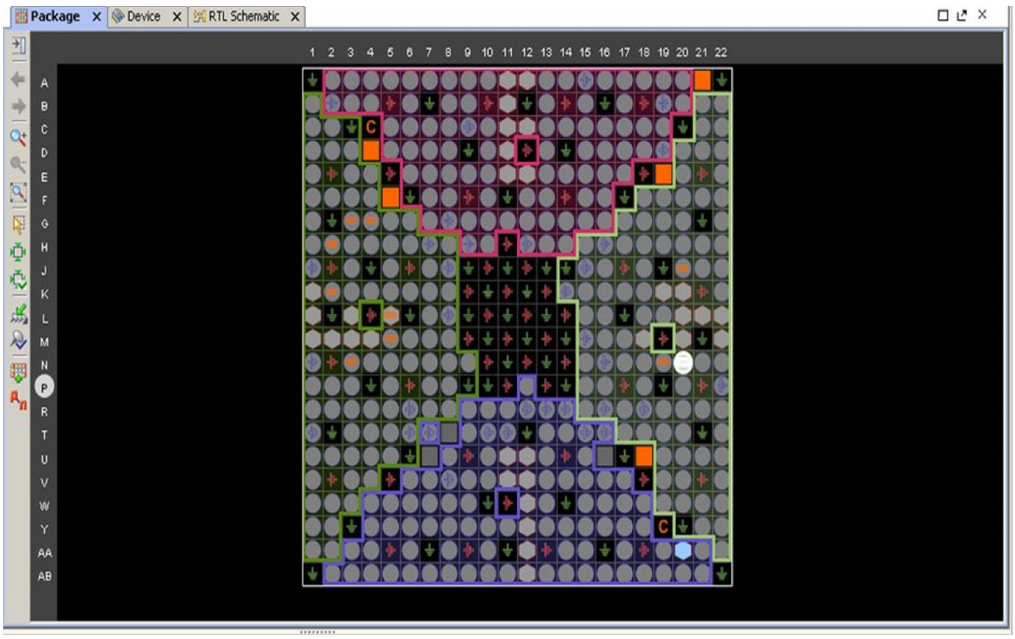
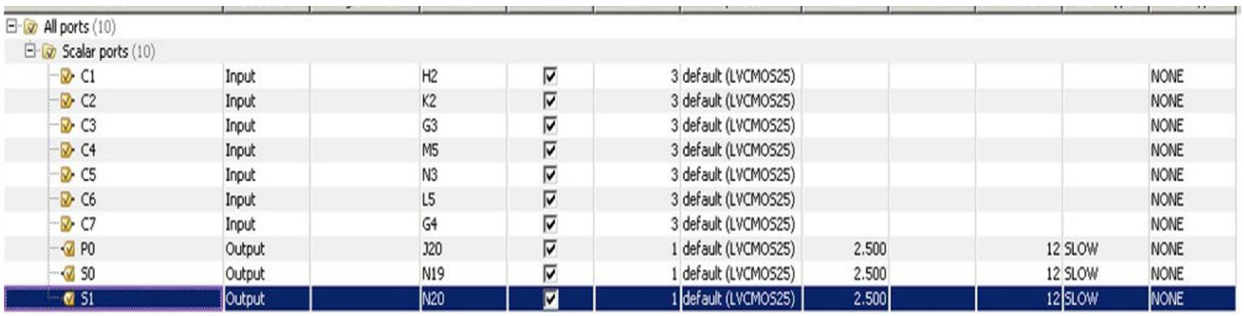


Рисунок 3.9 – Розміщення вхідних і вихідних портів на кристалі



All ports (10)										
Scalar ports (10)										
C1	Input		H2	☑	3	default (LVCMOS25)				NONE
C2	Input		K2	☑	3	default (LVCMOS25)				NONE
C3	Input		G3	☑	3	default (LVCMOS25)				NONE
C4	Input		M5	☑	3	default (LVCMOS25)				NONE
C5	Input		N3	☑	3	default (LVCMOS25)				NONE
C6	Input		L5	☑	3	default (LVCMOS25)				NONE
C7	Input		G4	☑	3	default (LVCMOS25)				NONE
P0	Output		J20	☑	1	default (LVCMOS25)	2.500		12	SLOW NONE
S0	Output		N19	☑	1	default (LVCMOS25)	2.500		12	SLOW NONE
S1	Output		N20	☑	1	default (LVCMOS25)	2.500		12	SLOW NONE

Рисунок 3.10 – Ініціалізація портів

Отже, у даному розділі дипломної роботи було модифіковано метод обчислення оператора групового підсумовування на основі вертикального і багатоопérандного підходів, при якому обчислення ґрунтується на вертикальних операціях і розглядається як єдиний процес підсумовування. Використали для синтезу пристроїв обчислення оператора групового підсумовування 7-входовий однорозрядний суматор, який забезпечує зменшення кількості перетворень і, відповідно, часу обчислення. В програмному продукті Xilinx було розроблено схему розміщення вхідних і вихідних портів на кристалі ПЛІС, внутрішню будову пристрою, часову діаграму та схему комбінаційного семивходового однорозрядного суматора.

ВИСНОВКИ

Розроблено паралельний метод вертикально-групового опрацювання даних у нейромережах, який у порівнянні з відомими забезпечує підвищення швидкодії шляхом збільшення розрядів каналів надходження множників і кількості часткових добутків, які формуються в результаті їх аналізу. Для вибору структур нейроелементів нейросистем реального часу доцільно використовувати критерій ефективності використання обладнання, який враховує кількість виводів інтерфейсу, однорідність структури, кількість і локальність зв’язків, зв'язує продуктивність з витратами обладнання та дає оцінку елементам пристрою за продуктивністю. Визначено, що узгодження інтенсивності надходження даних із обчислювальною здатністю нейроелемента у нейросистемах реального часу може здійснюватися шляхом зміни тривалості конвеєрного такту, кількості і розрядності каналів надходження даних.

Запропоновано розроблення НВІС-структур паралельного нейроелемента вертикально-групового типу для синтезу нейромереж реального часу з високою ефективністю використання обладнання здійснювати на основі інтегрованого підходу, який ґрунтується на можливостях сучасної елементної бази, охоплює методи, алгоритми і НВІС-структури, враховує вимоги конкретних застосувань і інтенсивності надходження даних. Розроблено модель та структуру формального паралельного нейрона вертикально-групового типу з мультиплексуванням шин вагових коефіцієнтів і даних, яка забезпечує узгодження інтенсивності надходження даних із обчислювальною здатністю нейроемента шляхом зміни розрядності каналів надходження і кількості розрядів множників у групі, які одночасно аналізуються для формування часткових добутків.

Основними етапами синтезу паралельного нейроелемента вертикально-групового типу є: вибір та розроблення методів і алгоритмів

обчислення скалярного добутку та функції активації; визначення основних параметрів апаратних засобів; перехід від алгоритму до узгодженої паралельної структури. Показано, що перехід від алгоритму функціонування нейрона в реальному часі до структури паралельного нейроелемента вертикально-групового типу формально зводиться до мінімізації апаратних затрат при забезпечені режиму реального часу.

Підвищення швидкодії обчислення оператора групового підсумовування досягається комплексним використанням вертикального та багатооперандного підходів, при якому обчислення ґрунтується на вертикальних операціях і розглядається як єдиний процес підсумовування. Використання для синтезу пристроїв обчислення оператора групового підсумовування 15-входових і 7-входових однорозрядних суматорів забезпечує зменшення кількості перетворень і, відповідно, часу обчислення.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Цмоць І.Г., Кураш Я.Я. Апаратна реалізація нейроелемента вертикально-групового типу. ACIT'2016, Тернопіль, 20-21 травня 2016. - С. 59-60.

2. Ежов А.А, Шумский С.А. Нейрокомпьютинг и его применение в экономике и бизнесе. М.: МИФИ, 1998. - 224с.

3. Haykin S. Neural networks and learning machines. Third Edition. / S. Haykin. – New York: Prentice Hall, 2009. – 936 p.

4. Руденко О.Г., Бодянський Є.В. Штучні нейронні мережі / О.Г. Руденко, Є.В. Бодянський. – Харків: ТОВ «Компанія СМІТ», 2006. – 404 с.

5. McCulloch W.S. A logical calculus of the ideas immanent in nervous activity / W.S. McCulloch, W. Pitts // The Bulletin of Mathematical Biophysics. – Vol. 5, Issue 4. – pp. 115–133. 4. ADALINE (Adaptive linear) [Електронний ресурс]. – Режим доступу: http://www.cs.utsa.edu /~bylander/cs4793/learnsc32.pdf

6. Fukushima K. Cognitron: A self-organizing multilayered neural network / K Fukushima // Biological cybernetics. – 1975. – Vol. 20, Issue 3-4. – pp. 121–136.

7. Hopfield, J.J. Neural networks and physical systems with emergent collective computational abilities / J.J. Hopfield // Proceedings of the national academy of sciences. – 1982. – Vol. 79, Issue 8. – pp. 2554–2558. 7. Cao J. Boundedness and stability for Cohen–Grossberg neural network with time-varying delays / J. Cao, J. Liang // Journal of Mathematical Analysis and Applications. – 2004. – Vol 296, Issue 2. – pp. 665–685.

8. Грибачев В. П. Элементная база аппаратных реализаций нейронных сетей // Компоненты и технологии. 2006. № 8.

9. Нейроподібні методи, алгоритми та структури обробки сигналів і зображень у реальному часі: монографія / Ю.М. Рашкевич, Р.О. Ткаченко, І.Г

Цмоць, Д.Д. Пелешко. – Львів: Видавництво Львівської політехніки, 2014. - 256 с. 1

10. Палагин А.В. Реконфигуриремые вычислительные системы./ А.В. Палагин, В.Н. Опанасенко. – К.: Просвіта, 2006.- 280с.

11. Цмоць І.Г. Модель та НВІС-структури формального нейрона паралельно-вертикального типу з використанням мультиплексування шин / І.Г. Цмоць, О.В. Скорохода, Б.І. Балич // Збірник наукових праць Інституту проблем моделювання в енергетиці ім. Г.Є. Пухова – Львів. – 2013. – Випуск № 67. – С. 160-166.

12. Цмоць І.Г. Моделі та НВІС-структури формального нейрона паралельно-вертикального типу з суміщенням процесів надходження та опрацювання даних / І.Г. Цмоць, О.В. Скорохода, В.Б. Красовський // Збірник Наукових праць Інституту проблем моделювання в енергетиці ім. Г.Є. Пухова – Львів. – 2013. – Випуск №70. – С. 137-145.

13. Цмоць І.Г. Модель та НВІС-структура формального нейрона паралельно-вертикального типу з табличним формуванням макрочасткових результатів / І.Г. Цмоць, О.В. Скорохода, Б.І. Балич // Збірник Наукових праць «Моделювання та інформаційні технології» Інституту проблем моделювання в енергетиці ім. Г.Є. Пухова – Львів. – 2014. – Випуск №73. – С. 133-138.

14. Патент №101922 Україна, G06F 7/38. Пристрій для обчислення скалярного добутку/ Цмоць І.Г., Скорохода О.В., Теслюк В.М. Бюл. №9, 2013.

15. Егупов Н. Д. Методы робастного , нейро-нечеткого и адаптивного управления . – М.: Изд-во МГТУ им. Н. Э. Баумана, 2001. – 744 с.

16. Терехов В.А. Нейросетевые системы Управления: Учеб .пособие для вузов. – М: Высшая школа. 2002. -183с.

17. Соловьев В. Проектирование цифровых систем на основе ПЛИС . – М.: Радио и связь, 2003. – 376 с.

18. Гильгурт С. Я. Анализ применения реконфигурируемых вычислителей на базе ПЛИС для реализации нейронных сетей // Моделювання та інформаційні технології. Зб. наук . пр. ІПМЕ НАН України. – Вип . 37. – Київ : 2006. – С. 168-174.

19. Логовский А. Технология ПЛИС и ее применение для создания нейрочипов. // Открытые системы . – 2000. – № 4. – С. 100-102.

20. Barren A. R. Universal approximation bounds for super position of a sigmoidal function // IEEE Trans. In-form. Theory. 1993. Vol. 39. P. 930 – 945.

21. Саймон Хайкин. Нейронные сети . Полный курс . – М.: Вильямс , 2006. – 1104 с.

22. Сигеру Омату . Нейроуправление и его приложения. – М.: ИПРЖР, 2000. – 272 с.

23. Рутковская Д., Пилиньский М., Рутковский Л. Нейронные сети , генетические алгоритмы и нечеткие системы : Пер. с польск . И. Д. Рудинского . – М.: Горячая линия – Телеком, 2007. – 452 с.

24. Цмоць І.Г. Реалізація нейронного елемента на основі попередніх обчислень / І.Г. Цмоць, О.В. Скорохода, Б.І. Балич // Вісник Нац. ун-ту. «Львівська політехніка»: Комп'ютерні науки та інформаційні технології. – Львів, 2011. – № 710. – С. 11–18.

25. Цмоць І.Г. Синтез високоефективних багатошарових перцептронів з неітераційним навчанням / І.Г. Цмоць, Р.О. Ткаченко, О.В. Скорохода // Вісник Нац. ун-ту. «Львівська політехніка»: Комп'ютерні науки та інформаційні технології. – Львів, 2009. – № 650. – С. 45–56.

26. Цмоць І.Г. Принципи побудови та способи НВІС-реалізації нейромереж реального часу / І.Г. Цмоць, О.В. Скорохода, І.Є. Ваврук // Наук. вісник НЛТУ України: зб. наук.-техн. праць. – Львів: РВВ НЛТУ України, 2012. – Вип. 22.6. – С. 292–300.

27. Скорохода О.В. Особливості реалізації нейромереж реального часу / О.В. Скорохода, І.Г. Цмоць, Б.І. Сенах // Науково-публіцистичний часопис «Технічні вісті». – Львів, 2011. – № 1(33)–2(34). – С. 28–30.

28. Цмоць І.Г. Методи та НВІС-структури пристроїв паралельно-вертикального обчислення сум парних добутків / І.Г. Цмоць, Б.І. Балич, О.В. Скорохода // Відбір і обробка інформації. – Львів, 2011. – № 33 (109). – С. 109–116.

29. Скорохода О.В. Засоби оцінювання параметрів динамічних об’єктів на основі нейромережевого сингулярного спектрального аналізу / О.В. Скорохода, Б.Р. Андрієцький, І.Г. Цмоць, Р.О. Ткаченко // Наук. вісник НЛТУ України: зб. наук.-техн. праць. – Львів: РВВ НЛТУ України, 2012. – Вип. 22.11. – С. 362–369.

30. Цмоць І.Г. Модифікований метод та НВІС-структура пристрою групового підсумовування для нейроелемента / І.Г. Цмоць, О.В. Скорохода, Б.І. Балич // Вісник Нац. ун-ту «Львівська політехніка»: Комп’ютерні науки та інформаційні технології. – Львів, 2012. – № 732. – С. 51–57.

31. Ткаченко Р.О. Програмно-апаратна реалізація багатошарового перцептрона з неітераційним навчанням на базі різницевого вертикального таблично-алгоритмічного методу / Р.О. Ткаченко, І.Г. Цмоць, О.В. Скорохода, Б.І. Балич // Вісник Нац. ун-ту «Львівська політехніка»: Комп’ютерні науки та інформаційні технології. – Львів, 2010. – № 686. – С. 65–71.

32. Цмоць І.Г. Інтегрований підхід до синтезу високоефективних апаратних засобів нейромережевих технологій реального часу / І.Г. Цмоць, Я.П. Кісь, О.В. Скорохода // Наук. вісник НЛТУ України: зб. наук.-техн. праць. – Львів: РВВ НЛТУ України, 2009. – Вип. 19.9. – С. 269–279.

33. Пат. № 66138, Україна, МПК G06F 7/38. Пристрій для обчислення сум парних добутків: Патент на корисну модель / І.Г. Цмоць, О.В. Скорохода; заявник і патентовласник Національний університет «Львівська політехніка». – № u201106811; заявл. 30.05.2011; опубл. 26.12.2011, Бюл. № 24. – 8 с.

34. Tsmots I. Methods and VLSI-structures for neural element implementation / I. Tsmots, O. Skorokhoda // Proc. of the VI International

Scientific and Technical Conference «MEMSTECH'2010», Polyana, 20–23 April 2010. – Lviv, 2010. – P. 135.

35. Скорохода О.В. Таблично-алгоритмічна реалізація штучних нейромереж / О.В. Скорохода, П.В. Романюк, О.Р. Якимів // Матеріали міжнародної наукової конференції «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту ISDMCI'2010», Євпаторія, 17–21 травня 2010 р. – Євпаторія, 2010. – Т. 1. – С. 398.

36. Грицик В.В. Особливості технології нейрокомп'ютингу реального часу / В.В. Грицик, Р.О. Ткаченко, І.Г. Цмоць, О.В. Скорохода // Матеріали науково-технічної конференції «Обчислювальні методи і системи перетворення інформації», Львів, 7–8 жовтня 2010 р. – Львів, 2010. – С. 229–232.

37. Skorokhoda O. Vertical-tabular implementation of neural element for the synthesis of multilayer perceptron with non-iterative learning / O. Skorokhoda, B. Balych, R. Tkachenko, I. Tsmots // Proc. of the V International Scientific and Technical Conference «CSIT'2010», Lviv, 14–16 October 2010. – Lviv, 2010. – P. 41.

38. Tsmots I. Hardware implementation of the real time neural network components / I. Tsmots, O. Skorokhoda // Proc. of the VII International Scientific and Technical Conf. «MEMSTECH'2011», Polyana, 11–14 May 2011. – Lviv, 2011. – P. 124–126.

39. Скорохода О.В. Вертикально-паралельний метод та структури для реалізації базових компонентів нейроелемента з використанням попередніх обчислень / О.В. Скорохода, І.Г. Цмоць, Я.П. Кісь // Матеріали міжнародної наукової конференції «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту ISDMCI'2011», Євпаторія, 16–20 травня 2011 р. – Євпаторія, 2011. – Т. 1. – С. 311–313.

40. Skorokhoda O. Stages of FPGA-Based Neural Networks Construction / O. Skorokhoda, I. Tsmots // Proc. of the V International Scientific and Technical Conference «CSIT'2011», Lviv, 16–19 November 2011. – Lviv, 2011. – P. 28.

41. Скорохода О.В. Методологія розробки апаратних нейромереж реального часу / О.В. Скорохода, І.Г. Цмоць, Р.О. Ткаченко // Матеріали міжнародної наукової конференції «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту ISDMCI'2012», Євпаторія, 27–31 травня 2012 р. – Євпаторія, 2012. – С. 415–416.

42. Андрієцький Б.Р. Використання нейромережевого сингулярного спектрального аналізу в задачах оцінювання та прогнозування параметрів динамічних об'єктів / Б.Р. Андрієцький, О.В. Скорохода, І.Г. Цмоць // Збірник тез доповідей шостої міжнародної науково-технічної конференції «Фотоніка ОДС-2012», Вінниця, 1–4 жовтня 2012 р. – Вінниця, 2012. – С. 48.

43. Цмоць І.Г. Інформаційні технології та спеціалізовані засоби обробки сигналів і зображень у реальному часі. – Львів: УАД, 2005.- 227с.

44. Справочник по цифровой вычислительной технике:(Электрон. вычисл. машины системы)/Б.Н. Малиновський, В.Я. Александров. В.п. Боюн и др. Под ред. Б.Н. Малиновського. – К. : «Техніка», 1980. 320 с.

45. Цифровая обработка информации на основе быстродействующих БИС. С.А. Гамкрелидзе, А.В. Завьялов, П.П. Мальцев, В.Г. Соколов; Под ред. В.Г. Домрачева.- М.: Энероатомиздат, 1988.- 136 с.

46. Коуги П.М. Архитектура конвейерных ЭВМ.: Пер. с англ. – М.: Радио и связь, 1985. – 360 с.

47. Гамаюн В.П. О развитии многооперандных вычислительных структур / В.П. Гамаюн // Управляющие системы и машины. – 1990. - №4. – С. 31-33.

48. Ромм Я.Е. Методы обработки потока целочисленных групповых данных. Групповые арифметические операции / Я.Е. Ромм // Кибернетика и системный анализ. – 1998. – №3. – С.123-151.

49. Паулин О.Н. Модель и метод проектирования многооперандных сумматоров на базе симметрических функций / О.Н. Паулин, А.М. Ляховецкий // Тези доповідей на міжнар. конф. з індуктив. моделювання МКІМ-2002. – Львів: ДНДІ. – 2002. – С. 208-213.

50. Хайкин С. Нейронные сети / Пер. с английского – М.: Вильямс, 2006. – 1104 с.

51. Рассел С., Норвиг П. Искусственный интеллект: современный подход / Пер. с английского – М.: Вильямс, 2007. – 1408 с.

52. Каллан Р. Основные концепции нейронных сетей / Пер. с английского – М.: Вильямс, 2001. – 288 с.

53. Рутковская Д., Пилиньский Л., Рутковский Л. Нейронные сети, генетические алгоритмы и нечеткие системы / Пер. с польского – М.: Горячая линия-Телеком, 2007. – 452 с.

54. Головко В.А. Нейронные сети: обучение, организация и применение – М.: ИПРЖР, 2002. – 256 с.

55. Круглов В.В., Борисов В.В. Искусственные нейронные сети. Теория и практика – М.: Горячая Линия-Телеком, 2002 – 382 с.

56. Круглов В.В., Борисов В.В. Искусственные нейронные сети. Теория и практика. – 2-е изд., стереотип. – М.: Горячая линия-Телеком, 2002. – 382 с.

57. Галушкин А.И. Нейрокомпьютеры. Кн.3.-.М; ИПРЖР,2000.-528с.

58. Осовский С. Нейронные сети для обработки информации / Пер. с польского. – М.: Финансы и статистика, 2002. – 344 с.

59. Поляков А. К. Языки VHDL и Verilog в проектировании цифровой аппаратуры. — М.: СОЛОН-Пресс, 2003. — 320 с.: ил. — (Серия «Системы проектирования»). — ISBN 5-98003-016-6.

60. Каршенбойм И., Косткин М. Шпаргалка для перехода от AHDL к VHDL. — Компоненты и технологии № 1, 2003.

61. Xilinx Company Overview. Xilinx. Архів оригіналу за 2013-07-16. Процитовано 2011-10-25.

62. Xilinx, Inc. Funding Universe. Архів оригіналу за 2013-07-16. Процитовано 2011-10-25.

63. Xilinx confirms: Samsung, TSMC in, UMC out at 28-nm. Архів оригіналу за 2013-07-16. Процитовано 2011-10-25.