



І. Г. Цмоць¹, Б. В. Штогрінець¹, М. В. Терлецький²

¹ Національний університет "Львівська політехніка", м. Львів, Україна

² Львівський національний університет ім. Івана Франка, м. Львів, Україна

СИНТЕЗ НЕЙРОКОМП'ЮТЕРНИХ СИСТЕМ З УЗГОДЖЕНО-ПАРАЛЕЛЬНИМ ОБРОБЛЕННЯМ ІНТЕНСИВНИХ ПОТОКІВ ДАНИХ У РЕАЛЬНОМУ ЧАСІ

Розглянуто особливості синтезу нейрокомп'ютерних систем з узгоджено-паралельним обробленням інтенсивних потоків даних у реальному часі, що дало змогу отримати модульну та регулярну структуру, орієнтовану на реалізацію на сучасній елементній базі з високою ефективністю використання обладнання. Визначено, що початковою інформацією для синтезу нейрокомп'ютерних систем з узгоджено-паралельним обробленням даних у реальному часі є: структура нейромережі; графове відображення нейромережі; алгоритми навчання та функціонування нейромережі; кількість вхідних даних; інтенсивність надходження вхідних даних і вагових коефіцієнтів; вимоги до інтерфейсу; розрядність вхідних даних, вагових коефіцієнтів і точність обчислень; техніко-експлуатаційні вимоги та обмеження. Розроблено метод синтезу нейрокомп'ютерних систем реального часу з узгоджено-паралельним обробленням даних, визначено, що основними етапами такого методу є: оцінювання обчислювальних і структурних характеристик нейромережі та вибір складності функціональних операторів для відображення її структури; просторово-часове відображення структури нейромережі у вигляді конкретизованого узгоджено-го потокового графу; визначення базових компонент для синтезу нейрокомп'ютерних систем реального часу з узгоджено-паралельним обробленням даних; визначення основних характеристик базових компонент; розроблення базових компонент; розроблення зовнішнього інтерфейсу та інтерфейсу систем обміну між шарами нейромережі; розроблення алгоритмів і засобів управління обчислювальним процесом; перехід від конкретизованого узгоджено-го потокового графу нейромережі до його апаратної реалізації. Розглянуто реалізацію кожного із етапів синтезу нейрокомп'ютерних систем реального часу. Розроблено дві структури нейроелементів паралельно-потокового типу: з обчисленням макрочасткового добутку для k розрядів і отриманням макрочасткового добутку шляхом зчитування з таблиці. Вибрано для обміну між шарами нейромережі багато-канальний пристрій обміну даними на базі багатопортової пам'яті. Запропоновано для узгодження тривалості введення даних з тривалістю конвеєрного такту використовувати послідовно-паралельні перетворювачі. Використано для переходу від конкретизованого потокового графу до апаратної реалізації метод адекватного апаратного його відображення. Показано, що узгодження інтенсивності надходження даних з інтенсивністю опрацювання та використання проблемно-орієнтованого підходу забезпечує реалізацію нейрокомп'ютерних систем реального часу з високою ефективністю використання обладнання та високими техніко-експлуатаційними характеристиками.

Ключові слова: нейромережа; синтез; нейрокомп'ютерна система; узгоджено-паралельне оброблення даних; потоковий граф; нейроелемент; конвеєризація.

Вступ / Introduction

Сучасний етап розвитку нейрокомп'ютерних систем реального часу характеризується розширенням галузей застосування та задач, які вони допомагають вирішувати. До основних галузей, де широко використовують нейрокомп'ютерні системи реального часу належать: автомобільна промисловість, робототехніка, охорона здоров'я, енергетика, телекомунікації, безпека, вір-

туальна та доповнена реальність [16]. Основними завданнями, які вирішують нейрокомп'ютерні системи реального часу в цих галузях, є: виявлення пішоходів, розпізнавання дорожніх знаків, контроль смуги руху, автоматичне гальмування, автономне управління рухом мобільної робототехнічної платформи, медичні діагностичні системи [29], моніторинг та управління енергосистемами, інтелектуальне управління мережевими ресурсами, нейромережевий захист даних у мобільних

Інформація про авторів:

Цмоць Іван Григорович, д-р техн. наук, професор, кафедра автоматизованих систем управління.

Email: ivan.h.tsmots@lpnu.ua; <https://orcid.org/0000-0002-4033-8618>

Штогрінець Богдан Володимирович, аспірант, кафедра автоматизованих систем управління.

Email: bohdan.v.shtohrinets@lpnu.ua; <https://orcid.org/0009-0001-4956-3862>

Терлецький Микола Віталійович, аспірант, кафедра інформаційних систем.

Email: mterletskyi@gmail.com; <https://orcid.org/0009-0002-6369-0793>

Цитування за ДСТУ: Цмоць І. Г., Штогрінець Б. В., Терлецький М. В. Синтез нейрокомп'ютерних систем з узгоджено-паралельним обробленням інтенсивних потоків даних у реальному часі. Науковий вісник НЛТУ України. 2024, т. 34, № 6. С. 76–86.

Citation APA: Tsmots, I. H., Shtohrinets, B. V., & Terletskyi, M. V. (2024). Synthesis of neurocomputer systems with coordinated-parallel processing of intensive real-time data flows. *Scientific Bulletin of UNFU*, 34(6), 76–86. <https://doi.org/10.36930/40340611>

системах і взаємодія з віртуальним світом [18, 30]. Нейрокомп'ютерні системи реального часу повинні забезпечувати такі вимоги: передбачуваність і своєчасне виконання завдань, високу швидкість оброблення даних шляхом використання спеціалізованих апаратних пристроїв, зменшення енергоспоживання та апаратних витрат шляхом узгодження інтенсивності надходження даних з інтенсивністю оброблення та видачі даних, мати високу стійкість до відмов, малі габарити та вагу, можливість адаптації до вимог конкретного застосування шляхом використання змінного складу обладнання, інтерфейси для введення/виведення даних мають бути швидкими та надійними. Задовольнити такі вимоги можна шляхом використання сучасної елементної бази (графічних процесорів, нейропроцесорів, FPGA (англ. *Field-Programmable Gate Array*)) та проблемно-орієнтованого підходу, який передбачає використання як програмних, так і спеціалізованих апаратних засобів [10].

Створення нейрокомп'ютерних систем реального часу з високою ефективністю використання обладнання потребує розроблення нових методів синтезу, алгоритмів і структур базових компонент для їх синтезу. Для синтезу нейрокомп'ютерних систем реального часу використовують нейроелементи паралельно-потокowego типу та багатоканальні пристрої безконфліктного обміну. Нейроелементи паралельно-потокowego типу орієнтовані на опрацювання неперервних інтенсивних потоків даних і працюють за конвеєрним принципом. Управління процесом обчислення у таких нейроелементах зводиться до подачі конвеєрних тактових імпульсів, які просувають дані з входу на вихід, записуючи проміжні результати в буферну пам'ять. Частоту надходження конвеєрних тактових імпульсів можна вибирати шляхом вибору складності операції, яка реалізується у сходінці конвеєра. Для ефективної роботи конвеєра всі його сходінки повинні виконувати приблизно однакові операції.

Краще розуміти структуру та функціонування нейронних мереж допомагає відображення нейроалгоритмів у вигляді потоків графів. Потоків графів використовують вузли для представлення операцій або шарів у нейронній мережі, а ребра показують потік даних між цими вузлами. Для аналізу та узгодження потоків даних у нейронних мережах з інтенсивністю роботи апаратури розробляються конкретизовані узгоджені потоків графів. Такі графів забезпечують просторово-часове відображення нейромережі на рівні функціональних операторів заданої складності, аналіз продуктивності мережі, зміну складності функціональних операторів з метою забезпечення узгодженості. У разі переходу від конкретизованого узгодженого графу до його реалізації необхідно передусім здійснити розподіл операцій, які будуть реалізовуватися апаратними засобами, а які програмним шляхом. Розроблення спеціалізованих апаратних компонент використовується здебільшого для синтезу нейрокомп'ютерних систем реального часу.

Тому актуальним завданням є розроблення базових компонент і синтез на їхній основі нейрокомп'ютерних систем реального часу з високими техніко-експлуатаційними характеристиками.

Об'єкт дослідження – синтез нейрокомп'ютерних систем реального часу з узгоджено-паралельним обробленням даних.

Предмет дослідження – методи синтезу комп'ютерних систем реального часу, що містять потоків графів

нейромереж, алгоритми паралельно-потокowego обчислення скалярного добути, структури нейроелементів, пристроїв обміну та нейрокомп'ютерних систем.

Мета роботи – розробити метод синтезу нейрокомп'ютерних систем з узгоджено-паралельним обробленням інтенсивних потоків даних у реальному часі, що дасть змогу отримати модульну та регулярну структуру, орієнтовану на реалізацію на сучасній елементній базі з високою ефективністю використання обладнання.

Для досягнення зазначеної мети визначено такі основні завдання дослідження:

- розробити метод синтезу нейрокомп'ютерних систем з узгоджено-паралельним обробленням інтенсивних потоків даних у реальному часі, що дасть змогу підвищити продуктивність та ефективність оброблення великих обсягів інформації, забезпечити швидке реагування на змінні умови функціонування, а також знизити затримки та енергоспоживання в процесі оброблення даних завдяки інтеграції сучасних алгоритмів машинного навчання та штучного інтелекту;
- розробити структури базових компонент (нейроелементів паралельно-потокowego типу, послідовно-паралельних перетворювачів, засобів багатоканального обміну на базі багатопортової пам'яті), що забезпечить можливість синтезу на їх базі нейрокомп'ютерних систем з узгоджено-паралельним обробленням інтенсивних потоків даних у реальному часі.

Аналіз останніх досліджень та публікацій. Проведений аналіз засобів реалізації штучних нейронних мереж показує, що переважно для реалізації таких мереж використовують різні програмні засоби [22]. Основними недоліками програмних засобів є: невисока швидкість, яка не забезпечує опрацювання інтенсивних потоків даних у реальному часі, великі затрати обладнання та мала ефективність використання обладнання [19]. Забезпечити опрацювання інтенсивних потоків даних у реальному часі можна шляхом використання високопродуктивних комп'ютерних систем, недоліком яких є те, що вони не відповідають обмеженням щодо габаритів і енергоспоживання [28].

Аналіз останніх досліджень щодо створення нейрокомп'ютерних систем реального часу для опрацювання інтенсивних потоків даних [3, 6, 8] показує, що у разі розроблення таких систем використовують таке: інтегрований та проблемно-орієнтований підходи; розпаралелення процедури оброблення даних; просторово-часове відображення нейроалгоритмів; апаратну реалізацію базових штучних компонент; узгодження інтенсивності надходження даних з інтенсивністю їх оброблення та видачі результатів.

Сучасні дослідження [13, 14] зосереджують увагу на використанні сучасної елементної бази для реалізації нейрокомп'ютерних систем реального часу з високою ефективністю використання обладнання. В роботах [17, 20] показано, що для такої реалізації широко використовують графічні процесори, нейропроцесори, процесори цифрового оброблення сигналів і програмовані логічні інтегральні схеми (ПЛІС). Однією з архітектурних різновидів ПЛІС є програмована користувачем вентилярна матриця FPGA (англ. *Field-Programmable Gate Array*). Особливістю FPGA є те, що її програмують шляхом зміни логіки роботи принципової схеми, наприклад, за допомогою початкового коду мовою проєкування (типу VHDL), на якому можна описати цю логіку роботи мікросхеми.

У роботах [4, 21] розглядають використання FPGA для реалізації паралельно-потоківих нейронних елементів. У роботі [1, 5] показано, що розроблення паралельно-потоківих нейронних елементів на базі FPGA доцільно виконувати за такими принципами: використання базису елементарних арифметичних операцій; організація процесу обчислення скалярного добутку як виконання єдиної операції; однотипності сходенок конвеєра; локалізації та спрощення зв'язків між сходінками конвеєра; забезпечення балансу між введенням-виведенням та обчисленнями; використання конвеєризації та просторового розпаралелювання процесу обчислення; мінімізація зовнішнього інтерфейсу зв'язку.

Аналіз робіт [7, 38] показує, що значну увагу під час розроблення базових компонентів нейронних мереж приділяють методам оптимізації структури пристроїв для обчислення скалярних добутків. Існують два основні підходи до апаратної реалізації обчислення скалярного добутку: традиційний метод, що містить операції множення та додавання, та метод, що базується на використанні елементарних арифметичних операцій додавання, інверсії та зсуву [34, 35]. Переваги другого підходу полягають у можливості оптимізації структур за швидкодією та зменшенням апаратних витрат.

У роботах [12, 32] розглянуто інтегрований та проблемно-орієнтований підходи до створення нейрокомп'ютерних систем реального часу. Показано, що використання інтегрованого підходу, який охоплює сучасну елементну базу, методи та алгоритми реалізації базових операцій нейроалгоритмів, архітектури компонент і систем, враховує вимоги конкретних застосувань та інтенсивності надходження даних забезпечує реалізацію нейрокомп'ютерних систем з високими техніко-експлуатаційними характеристиками. У роботі [24, 37] показано, що використання під час створення нейрокомп'ютерних систем реального часу універсального та спеціального підходів, програмних і апаратних засобів дає змогу ефективно реалізовувати нейроалгоритми і розширює діапазон галузей застосування.

Аналіз засобів обміну [33] показує, що використання багатопортової пам'яті для обміну між сходінками нейрокомп'ютерних систем зменшує тривалість обміну та зводить до мінімуму проблеми, що пов'язані з управлінням системою, упорядкуванням, затримкою та комутацією потоків даних.

У роботах [15, 26] розглянуто методи просторово-часового відображення алгоритму розв'язання задачі у вигляді узгодженого поточкового графу. Показано, що таке відображення дає змогу виявити паралелізм і керувати ним, забезпечуючи цим самим розроблення паралельних структур, що узгоджуються з інтенсивністю надходження даних.

З проведеного аналізу випливає, що синтез нейрокомп'ютерних систем реального часу потребує розроблення базових компонент нейронних мереж, орієнтованих на FPGA реалізацію, узгоджених поточкових графів нейронних мереж і їх програмно-апаратної реалізації з високою ефективністю використання обладнання.

Результати дослідження та їх обговорення / Research results and their discussion

Метод синтезу нейрокомп'ютерних систем з узгоджено-паралельним обробленням інтенсивних потоків даних у реальному часі. Метою синтезу нейро-

комп'ютерних систем (НКС) з узгоджено-паралельним обробленням інтенсивних потоків даних у реальному часі є отримання модульної та регулярної структури, орієнтованої на реалізацію на сучасній елементній базі з високою ефективністю використання обладнання. Початковою інформацією для синтезу таких НКС є [25]:

- структура нейромережі;
- графове відображення нейромережі;
- алгоритми навчання та функціонування нейромережі;
- кількість вхідних даних N ;
- інтенсивність надходження вхідних даних і вагових коефіцієнтів;
- вимоги до інтерфейсу;
- розрядність вхідних даних, вагових коефіцієнтів і точність обчислень;
- техніко-економічні вимоги і обмеження.

Процес синтезу НКС реального часу з узгоджено-паралельним обробленням даних можна звести до виконання таких етапів [25, 27]:

- 1) оцінювання обчислювальних і структурних характеристик нейромережі та вибору складності функціональних операторів для відображення її структури;
- 2) просторово-часового відображення структури нейромережі у вигляді конкретизованого узгодженого поточкового графу;
- 3) визначення базових компонент для синтезу НКС реального часу з узгоджено-паралельним обробленням даних;
- 4) визначення основних характеристик базових компонент;
- 5) розроблення базових компонент;
- 6) розроблення зовнішнього інтерфейсу та інтерфейсу систем обміну між шарами нейромережі;
- 7) розроблення алгоритмів і засобів управління обчислювальним процесом;
- 8) перехід від конкретизованого узгодженого поточкового графу нейромережі до його апаратної реалізації.

Розглянемо кожний із етапів детальніше.

Перший етап синтезу НКС реального часу зводиться до оцінювання обчислювальних і структурних характеристик нейромережі та вибору складності функціональних операторів для відображення її структури. Для такої оцінювання необхідно нейромережу подати у вигляді поточкового функціонального графу $F = (\Phi, \Gamma)$, де $\Phi = \{\Phi_j, j = \overline{1, n}\}$ – множина функціональних операторів (базових нейрооперацій), Γ – закон відображення зв'язків між операторами. Графічно такий функціональний граф відображено у вигляді вершин, що відповідають базовим нейроопераціям Φ_i та дуг, які відображають зв'язки між ними (рис. 1).

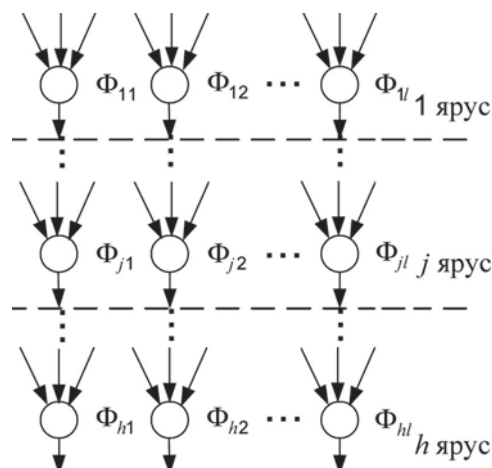


Рис. 1. Поточковий функціональний граф нейромережі / Flow functional graph of a neural network

Для апаратної реалізації алгоритмів є потреба у їх просторово-часовому відображенні на рівні арифметичних або базових нейрооперацій. Таке відображення нейромережі повинно забезпечувати виявлення всіх форм паралелізму та знаходження необхідних просторово-часових рішень для створення НКС реального часу з високою ефективністю використання обладнання.

Забезпечити просторово-часове відображення алгоритмів з виявленням усіх форм паралелізму дає змогу ярусно-паралельна форма (ЯПФ). За такої форми подання алгоритму здійснюють розподіл усіх його функціональних операторів Φ_i за ярусами так, що в j -му ярусі розміщені функціональні оператори, які залежать хоча б від одного функціонального оператора $(j-1)$ -го ярусу і не залежать від операторів наступних ярусів. Усі функціональні оператори одного ярусу виконують незалежно один від одного.

Кожний j -й ярус паралельної форми описується такими параметрами [27]:

- наборами незалежних функціональних операторів Φ_{jk} k -го функціонального оператора в j -му ярусі;
- шириною j -го ярусу l_j , тобто кількістю розміщених у ярусі функціональних операторів Φ_{jk} .

Кількість ярусів у ЯПФ алгоритму є її висотою h , а максимальна ширина ярусів визначає її ширину L . Під час відображення орієнтованого графа алгоритму у ЯПФ j -го ярусу є часовими індексами, а k -го функціонального оператора у ярусі є просторовими індексами. За допомогою просторово-часових індексів у системі координат час-простір задається розміщення функціональних операторів. Таке відображення орієнтованого графа алгоритму будемо називати потоковою паралельною формою або потоковим графом. Параметри потокового графу: складність функціональних операторів Φ_i , ширина L і висота h є взаємно залежними, зміна одного з них веде до зміни інших.

Складність функціональних операторів, які використовують для відображення функціонування нейромережі, залежить від засобів, які використовують для її реалізації. Залежно від засобів реалізації, відображення потокових графів нейромережі може здійснюватись з різним ступенем деталізації. У разі реалізації нейромережі на програмованих логічних інтегральних схемах (ПЛІС) функціональні оператори подають у вигляді виконання арифметичних операцій, а в разі використання для реалізації мікропроцесорів – у вигляді складних базових операцій.

На другому етапі синтезу НКС реального часу здійснюють просторово-часове відображення структури нейромережі у вигляді конкретизованого узгодженого потокового графу. Розроблення конкретизованого узгодженого потокового графу передбачає узгодження інтенсивності надходження вхідних даних, яку визначають за такою формулою [27]:

$$P_d = F_d m_d n_d, \quad (1)$$

де: F_d – частота надходження даних; m_d – кількість каналів надходження даних; n_d – розрядність каналів надходження даних з інтенсивністю обчислень у НКС, яку обчислюють за такою формулою [27]:

$$D_{HKS} = F_k m_o n_o, \quad (2)$$

де: F_k – частота роботи конвеєра НКС; m_o – кількість трактів оброблення даних; n_o – розрядність трактів оброблення даних. Узгодження інтенсивності обчислень

D_{HKS} у НКС досягається за рахунок розпаралелювання та конвеєризації процесів обчислень.

Окрім цього, для забезпечення роботи НКС у реальному часі необхідно виконання такої умови [25]:

$$D_{HC} \geq P_d. \quad (3)$$

Узгодження інтенсивності надходження вхідних даних P_d з інтенсивністю обчислень D_{HC} забезпечує високу ефективність використання обладнання E . Показник ефективності використання обладнання E зв'язує продуктивність НКС з апаратними витратами на їх реалізацію та дає оцінку елементам системи за продуктивністю. Кількісну величину для оцінювання ефективності використання обладнання E визначають за формулою [27]:

$$E = \frac{R_{HM}}{t_{CK} W_{HC}}, \quad (4)$$

де: R_{HM} – складність нейромережевого алгоритму в кількості елементарних арифметичних операцій; t_{CK} – тривалість спрацювання сходинок конвеєра; W_{HC} – витрати обладнання у вентилях на реалізацію НКС реального часу.

Процес просторово-часового відображення структури нейромережі у вигляді конкретизованого узгодженого потокового графу виконують за чотири кроки [25]:

- 1) декомпозиція нейромережевого алгоритму;
- 2) проєктування комунікацій між функціональними операторами;
- 3) укрупнення функціональних операторів;
- 4) планування обчислень.

На першому кроці виконують декомпозицію нейромережевого алгоритму Φ , за якої він розбивається на функціональні оператори Φ_i , між якими установлюють зв'язки, що відповідають даному алгоритму. Чим більший ступінь деталізації алгоритму отримуємо внаслідок декомпозиції, тим гнучкішим буде алгоритм і тим легше можна здійснити адаптацію його до вимог застосування. Декомпозицію можна здійснювати за методом декомпозиції даних або функціональної декомпозиції. Використання методу функціональної декомпозиції дає змогу отримати просторово-часове відображення структури нейромережі на рівні арифметичних операцій Φ_i , які використовують у разі її реалізації на ПЛІС. На практиці для синтезу високоефективних структур НКС реального часу використовують метод функціональної декомпозиції, за якого алгоритм Φ розбивається на операції Φ_i , кожна із яких може бути реалізована відповідною елементною базою. Тривалість і спосіб виконання операції Φ_i є одними із основних параметрів під час визначення частоти F_k роботи конвеєра НКС. Результатом першого кроку розроблення є граф схема нейромережі, де функціональні оператори Φ_i мають приблизно однакову тривалість виконання, а їх складність визначають засобами реалізації.

На другому кроці проєктування комунікацій необхідно визначити структуру каналів обміну даними між функціональними операторами Φ_i . Для цього виконують перехід від граф схеми алгоритму до потокового графу, в якому здійснюють просторово-часове розміщення та закріплення функціональних операторів Φ_i за ярусами. Структура зв'язків у потоковому графі між функціональними операторами Φ_{jk} сусідніх ярусів визначає кількість трактів оброблення даних m_o та їх розрядність n_o .

За результатами перших двох кроків отримуємо поточковий граф нейромережі, який дає змогу оцінити ін-

тенсивність обчислень D_{HC} у НКС. Початковими даними для визначення інтенсивності обчислень D_{HC} є складність функціональних операторів Φ_i та швидкодія елементної бази, які визначають частоту F_k роботи конвеєра НКС. Для оцінювання узгодженості інтенсивності надходження даних P_d із інтенсивністю обчислень D_{HC} вводять коефіцієнт узгодженості, який визначають за формулою [25]:

$$R = \frac{D_{HC}}{P_d}. \quad (5)$$

Узгодження інтенсивності обчислень D_{HC} із інтенсивністю надходження даних P_d досягається за рахунок зміни одного або кількох параметрів: частоти F_k роботи конвеєра НКС, кількості трактів оброблення даних m_o і розрядності n_o трактів оброблення даних. У разі розроблення узгодженого потокового графу нейромережі для узгодження передусім необхідно максимального використати особливості та швидкодію елементної бази, яку використовують для визначення складності функціональних операторів Φ_i і частоти F_k роботи конвеєра НКС. Далі потрібно змінювати кількість m_o і розрядність n_o трактів оброблення даних у потоковому графі. Зміна параметрів потокового графу алгоритму має забезпечити узгодження інтенсивності обчислень D_{HC} із інтенсивністю надходження даних P_d .

У випадку, коли $R = 1$, розроблений узгоджений поточковий граф забезпечує синтез НКС реального часу з високою ефективністю використання обладнання.

Якщо $R < 1$, то для забезпечення процедури оброблення потоків даних у реальному часі необхідно збільшення інтенсивності обчислень D_{HC} , яке може бути досягнуте за рахунок збільшення таких параметрів: кількості трактів m_o оброблення, розрядності n_o трактів оброблення даних і частоти F_k роботи конвеєра. Таке збільшення інтенсивності обчислень D_{HC} забезпечується зменшенням складності функціональних операторів Φ_i та збільшенням їх кількості. У випадку, коли зміною перерахованих параметрів не вдається досягнути необхідної інтенсивності обчислень D_{HC} , використовують паралельне включення НКС, кількість яких визначають виразом [25, 27]:

$$K = \lceil 1/R \rceil, \quad (6)$$

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

У випадку, коли $R > 1$, для забезпечення високої ефективності використання обладнання необхідно переходити до третього кроку розроблення – укрупнення функціональних операторів Φ_{jk} . На цьому етапі здійснюють об'єднання функціональних операторів Φ_{jk} і каналів передачі даних у ярусах потокового графа або об'єднання функціональних операторів сусідніх ярусів. Граф алгоритму, який отримаємо внаслідок такого об'єднання, будемо називати узгодженим поточковим графом. Коефіцієнт об'єднання v визначають за формулою [25]:

$$v \leq \lfloor D_{HC}/P_d \rfloor, \quad (7)$$

де $\lfloor \cdot \rfloor$ – знак округлення до меншого цілого. У разі об'єднання функціональних операторів v сусідніх ярусів утворюють один ярус функціональних макрооператорів, у якому за v ітерацій виконують обчислення таких функціональних макрооператорів V . Обчислення у середині ярусу здійснюють із частотою F_k , яка визначає складність функціональних операторів Φ_{jk} а між ярусами – з макрочастотою $F_{Mq} = vF_k$. Об'єднання функціональних операторів сусідніх ярусів приводить до змен-

шення у v разів кількості ярусів. Таке об'єднання доцільно здійснювати, коли яруси поточкового графа є однотипними. Інше укрупнення здійснюють шляхом об'єднання функціональних операторів і каналів передачі даних у межах ярусу. Для випадку, коли $v \geq L$ укрупнення здійснюють шляхом лінійної проєкції, за якої всі функціональні оператори ярусу поточкового графу відображаються в один функціональний макрооператор, а канали передачі даних – в оператор затримки та перестановки даних. Для забезпечення узгодженості можна використовувати комбіноване укрупнення, яке передбачає об'єднання функціональних операторів і каналів передачі даних як усередині ярусу, так і між ярусами.

На рис. 2 наведено приклад узгодженого потокового графу нейромережі, де OY – оператор управління, $3K$ – оператор збереження та комутації, V – функціональний макрооператор. Опис узгодженого потокового графу алгоритму містить дві частини – функціональну та структурну.

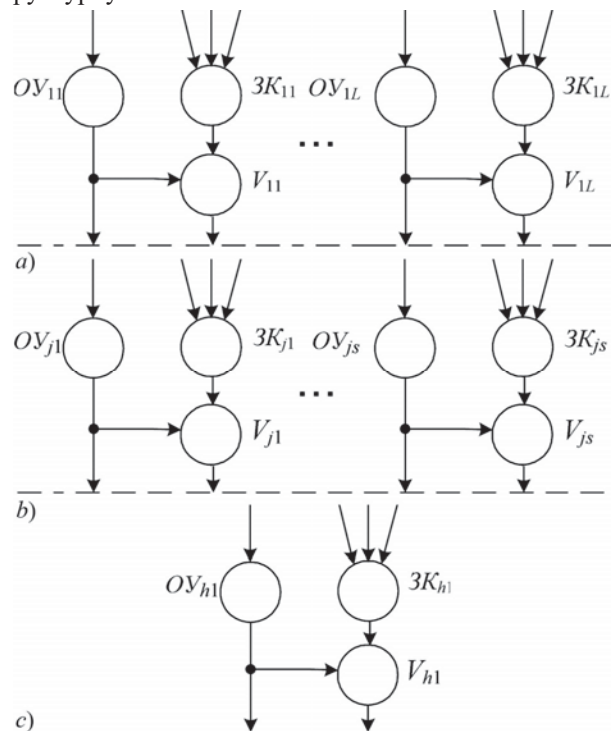


Рис. 2. Узгоджений поточковий граф нейромережі / Consistent flow graph of a neural network

Функціональний опис визначає операційний базис функціональних макрооператорів, а структурний – затримки, комутації даних і зв'язки між макрооператорами. Структуру узгодженого потокового графу нейромережі відображають орієнтованим графом $G = \langle V, E, D(E) \rangle$, де V – функціональні макрооператори ярусу; E – орієнтовані дуги, які моделюють зв'язки між функціональними макрооператорами. Кожна дуга $e \in E$ зв'язує вихід одного функціонального макрооператора з входом другого та володіє вагою, яка дорівнює значенню затримки $D(e)$.

Третій крок розроблення тісно пов'язаний четвертим кроком планування, на якому після об'єднання функціональних операторів для збереження інформації про структуру потокового графу алгоритму здійснюють планування обчислень, визначають величину затримок і перестановки даних. Для відтворення обчислень у кожному ярусу узгодженого графу вводять оператор управління OY і оператор $3K$ затримки та комутації.

На третьому етапі синтезу НКС реального часу визначають базові компоненти, які необхідні для синтезу НКС реального часу з узгоджено-паралельним обробленням даних.

Для реалізації функціональних макрооператорів V використовують нейроелементи, які функціонують згідно з формулою [25]:

$$y = f(Z), \text{ де } Z = \sum_{j=1}^N W_j X_j, \quad (8)$$

де: y – початковий сигнал нейроелемента; X_j – вхідні дані; W_j – вагові коефіцієнти; N – кількість входів даних і вагових коефіцієнтів; f – функція активації. З формули (8) випливає, що нейроелемент реалізується на таких операціях як обчислення скалярного добутку Z та функції активації $f(Z)$, яка реалізується табличним методом. Для забезпечення опрацювання інтенсивних потоків даних у НКС необхідно, щоб нейроелементи працювали за конвеєрним принципом. Конвеєризація у нейроелемента передбачає розділення його на сходинки шляхом введення буферної пам'яті. Для синтезу НКС реального часу здебільшого використовують нейронні елементи паралельно-потокowego типу, особливістю яких є відсутність обернених зав'язків і отримання результату під час проходження даних з входу на вихід.

Обчислення скалярного добутку в нейроелементі паралельно-потокowego типу виконують на базі однотипних операцій, які зводяться до обчислення макрочасткових добутків P_{Mh} і їх додавання до раніше накопичених сум згідно з формулою [25]:

$$Z_h = 2^{-h} Z_{h-1} + P_{Mh}, \forall h \in m, \quad (9)$$

де $Z_0 = 0$. Операцію обчислення макрочасткового добутку P_{Mh} та його додавання за формулою (9) реалізують h -ою сходинкою конвеєра. Кількість сходинок конвеєра визначають розрядністю операндів n і кількістю розрядів k , які аналізують для формування макрочасткових добутків P_{Mh} . Обчислення h -го макрочасткового добутку P_{Mh} ($h = \overline{1, m}$, де $m = \lceil n/k \rceil$, $\lceil \cdot \rceil$ – знак заокруглення до більшого цілого числа) виконують шляхом підсумовування групових часткових добутків P_{jh} , які отримують внаслідок аналізу k розрядів відповідно до формули [27]:

$$P_{Mh} = \sum_{j=1}^N \sum_{k=1}^K 2^{-(k-1)} W_j X_{jkh} = \sum_{j=1}^N P_{jh}, \forall h \in m. \quad (10)$$

Для отримання часткових добутків доцільно використовувати алгоритми, в яких часткові добутки отримуємо завдяки аналізу k молодших розрядів. Від кількості розрядів k , що аналізують для отримання групових часткових добутків P_{jh} залежить структура сходинки конвеєра.

Для реалізації обчислень оператор збереження та комутації $3K$ під час синтезу НКС реального часу доцільно використовувати багатоканальні пристрої обміну на базі багатопортової пам'яті (БПП). Особливістю таких пристроїв обміну є безконфліктний обмін даними з використанням методу часового розподілу ресурсів пам'яті. У багатоканальному пристрої обміну даних між шарами нейронної мережі забезпечення високої ефективності використання обладнання досягається шляхом узгодження інтенсивності надходження даних від нейроелементів з інтенсивністю доступу до БПП. Для реалізації такого узгодження використовують методи і за-

соби послідовно-паралельного та паралельно-послідовного перетворень.

У випадку, коли відомі вагові коефіцієнти W_j , для отримання макрочасткових добутків P_{Mi} доцільно використовувати таблицю з наперед обчисленими макрочастковими добутками. Обчислення таблиці макрочасткових добутків P_{Mi} виконують за такою формулою [25]:

$$P_{Mi} = \begin{cases} 0, & \text{якщо } x_{ji} = 0, j = \overline{1, N}; \\ W_1, & \text{якщо } x_{1i} = 1; x_{ji} = 0, j = \overline{2, N}; \\ W_2, & \text{якщо } x_{1i} = 0, x_{2i} = 1; x_{ji} = 0, j = \overline{3, N}; \\ W_1 + W_2, & \text{якщо } x_{1i} = 1, j = \overline{1, 2}; x_{ji} = 0, j = \overline{3, N}; \\ \vdots & \\ \sum_{j=1}^N W_j, & \text{якщо } x_{1i} = 0; x_{ji} = 1, j = \overline{2, N}; \\ \sum_{j=1}^N W_j, & \text{якщо } x_{ji} = 1, j = \overline{1, N}, \end{cases} \quad i = \overline{1, M}, \quad (11)$$

де $X = \{X_i = \{x_{ji}, j = \overline{1, N}\}, i = \overline{1, M}\}$ – i -ий розрядний зріз j -их вхідних даних.

На четвертому етапі синтезу НКС реального часу визначають основні характеристики базових компонент НКС реального часу.

Основною характеристикою нейроелемента паралельно-потокowego типу є конвеєрний такт T_k роботи пристрою обчислення скалярного добутку. Такий конвеєрний такт визначають за такою формулою [27]:

$$T_k = t_{P_2} + t_{OB}, \quad (12)$$

де t_{P_2} – тривалість звертання до буферної пам'яті (регістра). t_{OB} – тривалість отримання макрочасткового добутку P_{Mh} і його додавання до раніше накопичених сум відповідно до формули (9). Під час обчислення макрочасткового добутку P_{Mh} за формулою (10), тривалість t_{OB} обчислюють за формулою [27]:

$$t_{OB} = t_{\Phi\chi\Delta} + t_{BCM} + t_{CM}, \quad (13)$$

де: $t_{\Phi\chi\Delta}$ – тривалість формування часткових добутків; t_{BCM} – тривалість підсумовування часткових добутків; t_{CM} – тривалість додавання двох чисел. У разі використання для отримання макрочасткового добутку P_{Mi} таблиці обчисленої за формулою (11), тривалість t_{OB} визначають за формулою [25]:

$$t_{OB} = t_{m\Pi} + t_{CM}, \quad (14)$$

де $t_{m\Pi}$ – тривалість зчитування P_{Mi} з пам'яті.

Необхідно відзначити, що тривалість конвеєрного такту роботи T_k в основному залежить від швидкодії елементної бази та часу отримання макрочасткового добутку P_M і його додавання до раніше накопичених сум. Кількість сходинок конвеєра і їх складність залежить від кількості розрядів, які аналізують для отримання макрочасткового добутку P_M . Збільшення кількості сходинок конвеєра веде до зменшення тривалості такту за рахунок зменшення складності операцій, які виконують у сходинці. При зменшенні кількості сходинок збільшується складність операцій, які виконують у сходинці, а відповідно і збільшується тривалість такту. Апаратні витрати на реалізацію компонент зростають із збільшенням кількості сходинок конвеєра.

Тривалість конвеєрного такту T_k при узгоджено-паралельному обробленні даних має відповідати такому виразу [27]:

$$T_k \leq t_d, \quad (15)$$

де t_d – тривалість надходження вхідних даних і вагових коефіцієнтів.

У багатоканальному пристрої обміну на базі БПП для забезпечення узгоджено-паралельного оброблення даних необхідно виконання такої умови [25]:

$$P_d \leq P_{БПП}, \quad (16)$$

де: P_d – інтенсивність надходження даних; $P_{БПП}$ – інтенсивність доступу до БПП.

Інтенсивність надходження даних визначають за такою формулою [25]:

$$P_d = \sum_{i=1}^m n_{di} F_{di}, \quad (17)$$

з інтенсивністю доступу до БПП визначають за формулою [25]:

$$P_{БПП} = \sum_{i=1}^m n_i F_i, \quad (18)$$

де: n_{di} – розрядність каналів надходження даних від i -го нейроелемента; F_{di} – частота надходження даних від i -го нейроелемента; m – кількість нейроелементів, які під'єднують до пристрою обміну, $P_{БПП}$ – інтенсивність доступу до БПП; n_i – розрядність i -го каналу надходження даних до БПП; F_i – частота надходження даних в i -му каналі до БПП. Для реалізації такого узгодження використовують методи і засоби послідовно-паралельного та паралельно-послідовного перетворень.

На н'ятому етапі синтезу НКС реального часу виконують розроблення базових компонент – нейроелементів і багатоканального пристрою обміну на базі БПП.

Розроблено структуру нейроелемента паралельно-потокowego типу, в якій макрочасткові добутки обчислюють за формулою (10). Така структура наведена на рис. 3, де СК – сходи́нка конвеєра, Рг – реєстр, ФЧД – формувач часткових добутків, БФГЧД – блок формування групових часткових добутків, СМ – суматор, $NСМ$

– входовий суматор, W_{jq}, X_{jq} – входи нейроелемента, $\downarrow$ – вихід нейроелемента.

Нейроелемент паралельно-потокowego типу з обчисленням макрочасткового добутку для k розрядів для своєї реалізації потребує: перетворювача форматів даних, СК і табличного обчислювача передавальної функції.

Розроблену структуру нейроелемента паралельно-потокowego типу з табличним отриманням макрочасткового добутку P_{Mi} наведено на рис. 4. Особливістю цієї структури нейроелемента є зчитування в кожній i -й сходи́нці конвеєра з пам'яті за адресою, яка відповідає i -му розрядному зрізу вхідних даних, i -го макрочасткового добутку P_{Mi} . Кількість сходинок конвеєра у пристрої дорівнює розрядності вхідних даних. Зменшити кількість сходинок конвеєра можна шляхом розміщення у сходи́нці конвеєра k блоків для отримання k макрочасткових добутків.

Для реалізації оператор збереження та комутації ЗК розроблено структуру багатоканального пристрою обміну даними на базі БПП, яку наведено на рис. 5, де: БГПІ – блок генерування тактових імпульсів, КБПП – контролер багатопортової пам'яті, ОЗП – оперативний запам'ятовуючий пристрій, ШД – шина даних, ША – шина адреси, $Чм / Зп$ – сигнал зчитування/запису, $\overline{Вк}$ – сигнал вибірки пам'яті.

У багатоканальному пристрої обміну даними між шарами нейромереж інтенсивність доступу до ОЗП з часовим розподілом ресурсів залежить як від тривалості доступу до ОЗП, так і від тривалості формування сигналів управління, адреси та даних на магістраль ОЗП. Забезпечити малу тривалість формування сигналів управління, генерування адреси та надходження даних на входи ОЗП можна шляхом апаратної реалізації КБПП на ПЛІС.

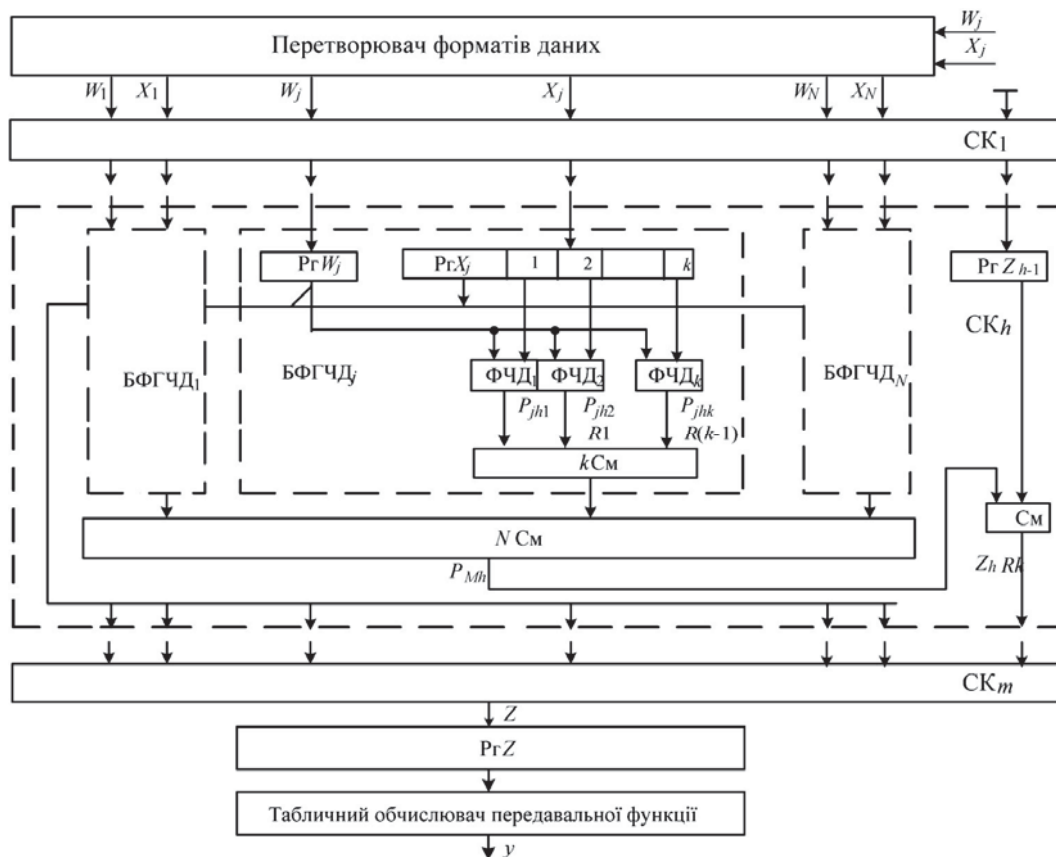


Рис. 3. Структура нейроелемента паралельно-потокowego типу з обчисленням макрочасткового добутку для k розрядів /
Tructure of the neuroelement of the parallel-flow type with the calculation of the macroparticle product for k discharges

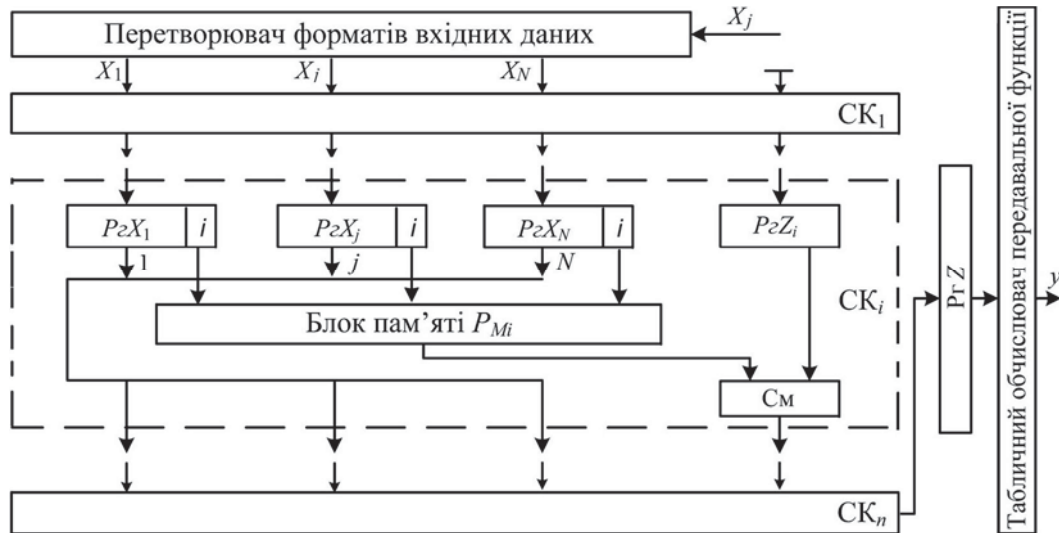


Рис. 4. Структура нейроелемента паралельно-потокowego типу з табличним отриманням макрочасткового добутку / Structure of the neuroelement of the parallel-flow type with tabular obtaining of the macroparticle product

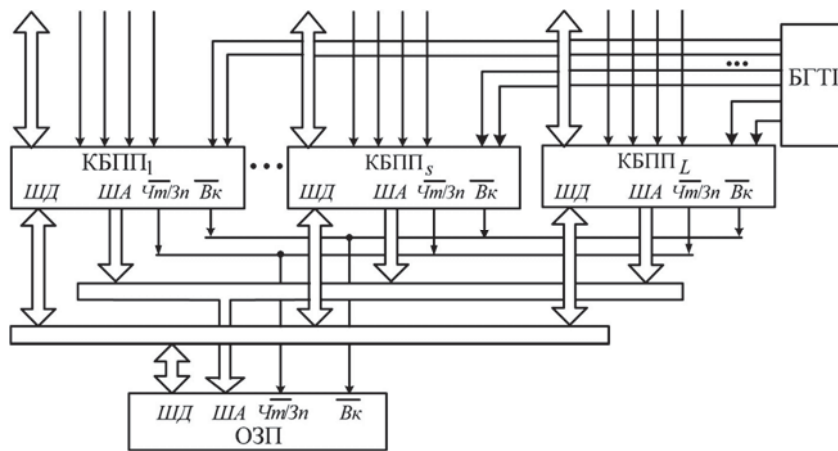


Рис. 5. Структура багатоканального пристрою обміну даними між шарами нейромережі / Structure of a multi-channel data exchange device between neural network layers

На шостому етапі синтезу НКС реального часу розробляють зовнішній інтерфейс та інтерфейс системи обміну між шарами нейромережі. Зовнішній інтерфейс повинен забезпечувати послідовно-паралельне перетворення операндів з мінімальною кількістю виводів. Для нейроелемента паралельно-потокowego типу з обчисленням макрочасткового добутку для k розрядів розроблено структуру послідовно-паралельне перетворювача, яку наведено на рис. 6, де БР₂ – буферний регістр.

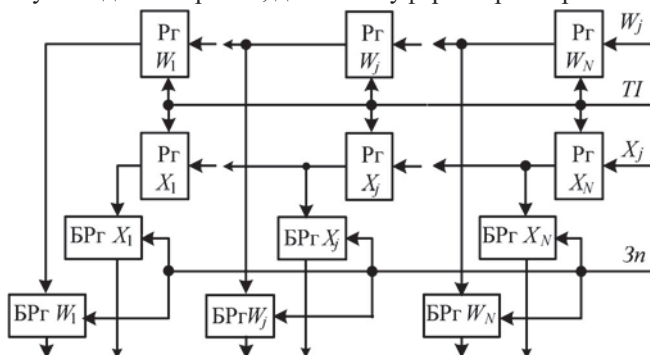


Рис. 6. Структура послідовно-паралельного перетворювача для нейроелемента з обчисленням макрочасткового добутку для k розрядів / Structure of the series-parallel converter for the neuroelement with the calculation of the macroparticle product for k digits

У цьому перетворювачі виконують послідовно-паралельне отримання N -ої кількості вагових коефіцієнтів

$W = \{W_j, j = \overline{1, N}\}$ та N вхідних даних $X = \{X_j, j = \overline{1, N}\}$. Тривалість такого перетворення $t_d = NT_{\Pi}$ для забезпечення роботи у реальному часі повинен відповідати такій умові $t_d \geq T_k$ [27].

Розроблено структуру послідовно-паралельне перетворювача для нейроелемента з табличним отриманням макрочасткового добутку. Структуру такого перетворювача наведено на рис. 7.

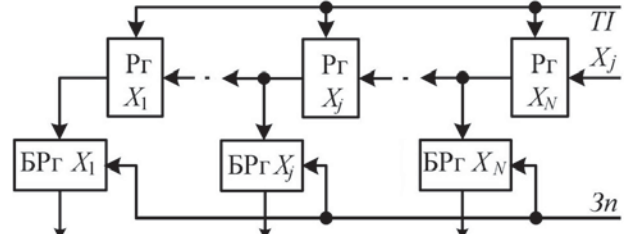


Рис. 7. Структура послідовно-паралельного перетворювача для нейроелемента з табличним отриманням макрочасткового добутку / Structure of the series-parallel converter for the neuroelement with tabular obtaining of the macroparticle product

У цьому перетворювачі виконують послідовно-паралельне отримання N вхідних даних $X = \{X_j, j = \overline{1, N}\}$, які надходять на входи нейроелемента з табличним отриманням макрочасткового добутку. Тривалість такого перетворення визначає конвеєрний такт T_k роботи нейроелемента з табличним отриманням макрочасткового добутку.

Для забезпечення високої інтенсивності доступу до БПП з часовим розподілом ресурсів ОЗП необхідно, щоб кількість КБПП дорівнювала ширині доступу (кількості входів-виходів). Усі КБПП мають бути однотипними та забезпечувати можливості [25]:

- роботи з різними за швидкодією ОЗП і зовнішніми пристроями;
- програмного налаштування на генерацію необхідних послідовностей адреси;
- паралельного включення необхідної кількості КБПП;
- виділення кожному КБПП періодичних фіксованих часових інтервалів доступу ОЗП.

Для обміну з багатоканальним пристроєм на базі БПП використовують інтерфейс КБПП, який має такі сигнали: 16-розрядну шину даних $\overline{ШД}$, сигнали управління $\overline{ЗнОЗП}$, $\overline{ЧтОЗП}$, $\overline{ЗнАД}$ і $\overline{ЧтАД}$, тактові сигнали $П_1$ і $П_2$.

На сьомому етапі синтезу НКС реального часу розробляють алгоритми і засоби управління обчислювальним процесом. Засоби управління обчислювальним процесом під час реалізації нейромережі реалізують оператори управління ОУ, які розгортають обчислювальний процес у часі та просторі.

На восьмому етапі синтезу НКС реального часу здійснюють перехід від конкретизованого узгодженого потокового графу нейромережі до її апаратної реалізації. Для синтезу НКС реального часу з узгоджено-паралельним обробленням використаємо відомий метод адекватного апаратного відображення структури графів алгоритмів і процесів функціонування, за якого кожному функціональному оператору ставиться у відповідність операційний блок, а дугам між функціональними операторами – каналами передачі даних. Синтезовані в такий спосіб НКС є алгоритмічними. У таких НКС нейроалгоритм реалізується під час проходження та оброблення даних від входів до виходів через усі операційні блоки. Для забезпечення високої ефективності використання у таких системах НКС необхідно, щоб оброблення даних здійснювалося у реальному часі за мінімальних апаратних затрат на реалізацію НКС.

У разі розроблення НКС реального часу в більшості випадків потрібне зменшення T_k . Це зменшення досягається такими шляхами [25]:

- зменшення кількості часткових добутків, які обчислюють у конвеєрні сходинок;
- зменшення складності конвеєрних сходінок;
- використання просторового розпаралелення.

Для зменшення апаратних затрат та забезпечення високих техніко-експлуатаційних характеристик реалізацію НКС реального часу доцільно здійснювати на базі проблемно-горієнтованого підходу. Такий підхід до реалізації НКС реального часу потребує у собі використання універсальних і спеціалізованих реалізацій, програмних і апаратних засобів.

Обговорення результатів дослідження. Для зменшення вартості і тривалості створення НКС з узгоджено-паралельним обробленням даних у реальному часі актуальним завданням є розроблення засобів автоматизації процесу синтезу таких систем. Розроблення засобів автоматизації процесу синтезу НКС реального часу вимагає виконання таких етапів: формування вимог конкретного застосування та визначення основних функцій; вибір елементної бази, програмного забезпечення та середовищ розроблення для розроблення засобів автоматизації; адаптація нейронних мереж до розв'язання

заданої задачі; апаратне відображення адаптованих нейронних мереж. Автоматизація цих етапів може бути досягнута за допомогою спеціалізованих інструментів і платформ.

Віднедавня спостерігається значний прогрес у розробленні методів автоматизації процесу синтезу нейронних комп'ютерних систем. Наприклад, у роботах [11, 31] запропонували метод, що використовує оптимізаційні алгоритми для підвищення ефективності синтезу нейронних мереж у реальному часі, що призводить до значного скорочення тривалості розроблення та ресурсів. Додатково представили інструмент для автоматизованого розроблення спеціалізованих апаратних засобів для нейронних мереж, який дає змогу ефективніше використовувати апаратні ресурси та скоротити витрати на розроблення.

Інші дослідження, згадані у роботах [2, 36], стосуються створення адаптивних платформ для автоматизації процесу синтезу нейронних систем, що забезпечує гнучкість та масштабованість у різних застосуваннях. Було розроблено підхід, який використовує машинне навчання для автоматизації процесу адаптації нейронних мереж до конкретних завдань, що значно підвищує точність та швидкість синтезу.

Зокрема, у роботах [9, 23] продемонстрували ефективність використання хмарних обчислень для автоматизації синтезу нейронних мереж, що дає змогу знизити витрати на апаратне забезпечення та збільшити масштабованість систем. Також досліджували можливість використання FPGA для прискорення синтезу нейронних мереж у реальному часі, що дає змогу значно підвищити ефективність використання апаратних ресурсів.

Отже, за результатами виконаної роботи можна сформулювати такі наукову новизну та практичну значущість результатів дослідження.

Наукова новизна отриманих результатів дослідження – розроблено метод синтезу нейрокомп'ютерних систем з узгоджено-паралельним обробленням даних, який за рахунок відображення нейромереж у вигляді узгодженого конкретизованого потокового графу та використання розроблених базових компонент для реалізації його операторів забезпечує синтез на їх основі нейрокомп'ютерних систем реального часу з високою ефективністю використання обладнання.

Практична значущість результатів дослідження – використання розробленого методу забезпечує синтез нейрокомп'ютерних систем реального часу для конкретних застосувань з високою ефективністю використання обладнання.

Висновки / Conclusions

Розроблено структури базових компонент (нейроелементів паралельно-потокowego типу, послідовно-паралельних перетворювачів, засобів багатоканального обміну на базі багатопортової пам'яті) нейрокомп'ютерних систем та метод синтезу нейрокомп'ютерних систем з узгоджено-паралельним обробленням даних у реальному часі з високою ефективністю використання обладнання. За результатами проведених досліджень можна зробити такі основні висновки.

1. Розроблено метод синтезу нейрокомп'ютерних систем з узгоджено-паралельним обробленням даних, який за рахунок відображення нейромереж у вигляді узгодженого конкретизованого потокового графу та використання розроблених базових компонент для реалізації

- його операторів забезпечує синтез на їх основі нейрокомп'ютерних систем реального часу з високою ефективністю використання обладнання.
- Вибрано для відображення нейромережових алгоритмів у ярусно-паралельну форму, яка забезпечує виявлення всіх форм паралелізму та знаходження необхідних просторово-часових рішень для створення нейрокомп'ютерних систем з узгоджено-паралельним обробленням даних у реальному часі.
 - Визначено, що ступень деталізації функціональних операторів під час відображення нейромережових алгоритмів залежить як від засобів реалізації, так і від інтенсивності надходження даних.
 - Показано, що розроблення конкретизованого узгодженого поточкового графу здійснюють узгодженням інтенсивності надходження даних з інтенсивністю обчислень шляхом зміни складності і кількості функціональних операторів, які обчислюють на сходінці конвеєра та визначають такт конвеєра, кількість трактів оброблення даних і їх розрядність.
 - Розроблено два види нейроелементів паралельно-поточкового типу, перший з яких реалізують на базі сходінки конвеєра з обчисленням макрочасткового добутку шляхом підсумовування часткових добутків, які отримують аналізом k розрядних зрізів, а другий – на сходінках конвеєра з отриманням k макрочасткових добутків шляхом зчитування з k таблиць.
 - Визначено, що такт роботи конвеєра нейроелементів паралельно-поточкового типу залежить від кількості зрізів розрядів, які аналізують для отримання часткових добутків на сходінці конвеєра та для зчитування макрочасткових добутків з таблиць на сходінці конвеєра.
 - Розроблено багатоканальний пристрій безконфліктного обміну з часовим розподілом ресурсів пам'яті, який за рахунок узгодження інтенсивності надходження даних із зовнішніх пристроїв з інтенсивністю доступу до пам'яті забезпечує підвищення ефективності використання обладнання.
 - Показано, що узгодження інтенсивності надходження даних з інтенсивністю обчислень у НКС досягається зміною кількості нейроелементів, що паралельно працюють на сходінці конвеєра та зміною тактової частоти роботи конвеєра за рахунок вибору швидкодії елементної бази та складності функціональних операторів.
 - Вибрано метод адекватного апаратного відображення для переходу від конкретизованого узгодженого поточкового графу до його апаратної реалізації, за якого кожному функціональному оператору ставиться у відповідність операційний блок, а дугам між функціональними операторами – канали передачі даних.

References

- Ajay, A., Critch, A., & Agrawal, P. (2023). Dexterous Manipulation from Images: Autonomous Real-World RL via Substep Guidance. *Science Robotics*, 4(32), article ID eaaw1960. <https://doi.org/10.1109/ICRA48891.2023.10161493>
- Chen, M., et al. (2020). Adaptive platforms for automated synthesis of neural systems. *IEEE Transactions on Computers*, 69(11), 1623–1634. <https://doi.org/10.1109/TC.2020.2996531>
- Chen, Y.-H., et al. (2019). Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices. *IEEE Journal of Solid-State Circuits*, 54(1), 294–305. <https://doi.org/10.48550/arXiv.1807.07928>
- Chen, Y.-H., Krishna, T., Emer, J. S., & Sze, V. (2016). Eyeriss: An Energy-Efficient Reconfigurable Accelerator for Deep Convolutional Neural Networks. *IEEE Journal of Solid-State Circuits*, 70(5), 710–721. <https://doi.org/10.1109/ISSCC.2016.7418007>
- Choi, J., et al. (2020). Edge AI: On-demand accelerated AI with adaptive private inference. In: *Proceedings of the 25th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 589–602. <https://doi.org/10.1145/3373376.3378508>
- Deng, L., & Li, X. (2021). Harnessing the power of dynamic reconfiguration in hardware acceleration for deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5), 2048–2059. <https://doi.org/10.54254/2755-2721/47/20241256>
- Han, S., et al. (2018). ESE: Efficient speech recognition engine with sparse LSTM on FPGA. In: *Proceedings of the 2018 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA)*, 75–84. <https://doi.org/10.1145/3174243.3174252>
- Jouppi, N. P., et al. (2021). A domain-specific supercomputer for training deep neural networks. *Communications of the ACM*, 64(5), 67–78. <https://doi.org/10.1145/3360307>
- Kim, D., et al. (2020). FPGA-accelerated synthesis of real-time neural networks. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 28(3), 746–759. <https://doi.org/10.1109/TVLSI.2020.2965943>
- Kozhemiako, V. P., Martynuk, T. B., Kozhemiako, A. V., Vasylykiva, O. S., & Kitaychuk, O. V. (2015). Hardware implementation of the perceptron as a basic neural network node. *Optical-Electronic Information-Energy Technologies*, 28(2), 48–55. URL: <https://oeipt.vntu.edu.ua/index.php/oeipt/article/view/432>
- Lee, J., et al. (2022). Automated hardware design for neural networks: Efficiency and scalability. *ACM Transactions on Design Automation of Electronic Systems*, 27(4), 1–22. <https://doi.org/10.1145/3503181>
- Li, H., et al. (2020). Understanding the synchronization performance of distributed deep learning frameworks on GPU clusters. In: *Proceedings of the 2020 IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 166–179. <https://doi.org/10.1109/HPCA47549.2020.00024>
- Li, Y., et al. (2018). Efficient and scalable graph neural networks via historical embedding. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 705–713. <https://doi.org/10.1145/3219819.3219953>
- Lin, S., Cheng, H.-T., Lu, L., Yang, M.-H., & Wu, B. (2020). Real-Time High-Resolution Background Matting. *ACM Transactions on Graphics (TOG)*, 37(4), 1–10. <https://doi.org/10.48550/arXiv.2012.07810>
- Liu, Z., et al. (2021). Enabling efficient processing of graph neural networks with reconfigurable hardware. In: *Proceedings of the 2021 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA)*, 23–33. <https://doi.org/10.1145/3431920.3431950>
- Markevych, K. (2021). *Smart infrastructure in sustainable urban development: world experience and prospects of Ukraine*. Razumkova Center, Zapovit Publishing House, Kyiv, Ukraine, 400. URL: <https://razumkov.org.ua/uploads/other/2021-SMART-%D0%A1YTI-SITE.pdf>
- Markidis, S., et al. (2018). NVIDIA Tensor Core programmability, performance & precision. In: *Proceedings of the 2018 IEEE/ACM Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems (PMBS)*, 1–9. <https://doi.org/10.1109/PMBS.2018.8658703>
- Meitus, V. Yu., Morozova, H. I., Taran, L. Yu., Kozlova, V. P., & Maidanyuk, N. V. (2019). Cyber-physical systems as a basis for the intellectualization of "Smart" enterprises. *Control Systems and Machines*, 4, 14–26. <https://doi.org/10.15407/csc.2019.04.014>
- Nurvitaldhi, E., Venkatesh, G., Sim, J., Marr, D., Huang, R., Ong Gee Hock, J., Liew, Y. T., Srivatsan, K., Moss, D., & Subhaschandra, S. (2017). Can FPGAs beat GPUs in accelerating next-generation deep neural networks? In: *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays*, 5–14. Monterey, CA, USA: ACM. <https://doi.org/10.1145/3020078.3021740>
- Rabyk, V., Kryvinska, N., Yatsymirsky, M., & Teslyuk, V. (2022). Design of the processors for fast cosine and sine Fourier transforms. *Circuits, Systems, and Signal Processing*, 41(9), 4928–4951. <https://doi.org/10.1007/s00034-022-02012-8>
- Shi, L., et al. (2019). Fast sparse ConvNets. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 39(10), 2128–2139. <https://doi.org/10.48550/arXiv.1911.09723>

22. Subbotin, S. O. (2020). Neural networks: Theory and practice: Textbook. Zhytomyr: O. O. Yevenok Publishing House, 184. URL: <https://eir.zp.edu.ua/server/api/core/bitstreams/2abb401b-9ee6-4afc-a92a-2de5c332d12f/content>
23. Sun, X., et al. (2023). Cloud-based automation of neural network synthesis: Methods and applications. *IEEE Transactions on Cloud Computing*, 11(2), 341–353. <https://doi.org/10.1109/TCC.2022.3148675>
24. Sze, V., Chen, Y.-H., Yang, T.-J., & Emer, J. (2017). Efficient processing of deep neural networks: A tutorial and survey. *Proceedings of the IEEE*, 106(11), 1990–2023. <https://doi.org/10.48550/arXiv.1703.09039>
25. Tsmots, I. G., Tkachenko, R. O., Tesliuk, V. M., Riznyk, O. Y., & Kazymyra, I. Y. (2023). Smart systems: Technologies, architectures, processing, data protection, and encoding. Lviv: SPO-LUM Publishing, 220 p. URL: <https://www.irbis-nbuv.gov.ua/publ/REF-0000817068>
26. Tsmots, I. H., Opotyak, Y. V., Shtohrinets, B. V., Mamchur, T. B., & Holubets, V. M. (2024). Model, structure, and synthesis method of a matrix-type neural element. *Scientific Bulletin of UNFU*, 34(4), 68–77. <https://doi.org/10.36930/40340409>
27. Tsmots, I., Teslyuk, V., Kryvinska, N., Skorokhoda, O., & Kazymyra, I. (2023). Development of a generalized model for parallel-streaming neural element and structures for scalar product calculation devices. *Journal of Supercomputing*, 79(5), 4820–4846. <https://doi.org/10.1007/s11227-022-04838-0>
28. Tsmots, I., Teslyuk, V., Łukaszewicz, A., Lukashchuk, Y., et al. (2023). An approach to the implementation of a neural network for cryptographic protection of data transmission at UAV. *Drones*, 7(507). <https://doi.org/10.3390/drones7080507>
29. Vishnevskiy, V. P., Vyietska, O. V., Harkushenko, et al. (2018). *Smart industry in the era of digital economy: Prospects, directions, and development mechanisms: Monograph* (V. P. Vishnevskiy, Ed.). National Academy of Sciences of Ukraine, Institute of Industrial Economics, Kyiv, 192. URL: <https://iie.org.ua/monografiyi/smart-promislovist-v-epohu>
30. Vishnevskiy, V. P., Vyietska, O. V., Vyietskiy, O. A., et al. (2019). *Smart industry: Directions of formation, problems, and solutions: Monograph* (V. P. Vishnevskiy, Ed.). National Academy of Sciences of Ukraine, Institute of Industrial Economics, Kyiv, 464 p. URL: https://iie.org.ua/wp-content/uploads/2020/04/2019-smart-promislovist-napriamyi-stanovlennia-problemy-i-rishennia_compressed-1.pdf
31. Wang, Y., et al. (2021). Optimizing real-time neural network synthesis using advanced algorithms. *IEEE Transactions on Neural Networks and Learning Systems*, 32(8), 1234–1245. <https://doi.org/10.1109/TNNLS.2021.3067589>
32. Wu, C., Fresse, V., Suffran, B., & Konik, H. (2021). DLAU: A Scalable Deep Learning Accelerator Unit on FPGA. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 70(2), 272–284. <https://doi.org/10.1109/TCAD.2016.2587683>
33. Wu, Y., Wang, H., & Wang, X. (2024). 8T-SRAM-based processing-in-memory (PIM) system with current mirror for accurate MAC operation. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 27(12), 2752–2765. <https://doi.org/10.1109/ACCESS.2024.3358451>
34. Yang, Z., et al. (2020). Interstellar: Using Halide's scheduling language to analyze DNN accelerators. In: *Proceedings of the 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 369–383. <https://doi.org/10.1145/3373376.3378510>
35. Yousefzadeh, A., Kiasari, A., & Radetzki, M. (2020). NeuronFlow: A Neuromorphic Processor Architecture for Live AI Applications. *Transactions on Neural Networks and Learning Systems*, 32(3), 1078–1091. <https://doi.org/10.23919/DAT48585.2020.9116352>
36. Zhou, H., et al. (2019). Machine learning-driven automation for neural network adaptation. *Journal of Parallel and Distributed Computing*, 135, 65–75. <https://doi.org/10.1016/j.jpdc.2019.07.012>
37. Zhou, S., et al. (2018). Adaptive batch sizes for training deep neural networks. In: *Proceedings of the 27th International Conference on Parallel Architectures and Compilation Techniques (PACT)*, 157–169. <https://doi.org/10.1145/3243176.3243204>
38. Zou, Z., Shi, Z., Guo, Y., & Ye, J. (2019). A Survey of Deep Learning-Based Object Detection. *Transactions on Circuits and Systems for Video Technology*, 31(7), 2752–2771. <https://doi.org/10.1109/ACCESS.2019.2939201>

I. H. Tsmots¹, B. V. Shtohrinets¹, M. V. Terletskiy²

¹ Lviv Polytechnic National University, Lviv, Ukraine

² Ivan Franko Lviv National University, Lviv, Ukraine

SYNTHESIS OF NEUROCOMPUTER SYSTEMS WITH COORDINATED-PARALLEL PROCESSING OF INTENSIVE REAL-TIME DATA FLOWS

The peculiarities of the synthesis of neurocomputer systems with coordinated and parallel processing of intensive data streams in real time were considered, which made it possible to obtain a modular and regular structure, oriented to implementation on a modern element base with high efficiency of equipment use. In the course of research, we determined that the initial information for the synthesis of neurocomputer systems with coordinated and parallel data processing in real-time is as follows: the structure of the neural network; graphical representation of a neural network; learning algorithms and neural network functioning; amount of input data; intensity of incoming data and weighting factors; interface requirements; bit rate of input data, weighting factors and accuracy of calculations; technical and operational requirements and restrictions. A method for the synthesis of real-time neurocomputer systems with coordinated-parallel data processing was developed, the main stages of such a method are determined as follows: the assessment of computational and structural characteristics of the neural network and selection of the complexity of functional operators to display its structure; spatio-temporal display of the structure of the neural network in the form of a specified coordinated flow graph; definition of the basic components for the synthesis of real-time neurocomputer systems with coordinated parallel data processing; determination of the main characteristics of the basic components; development of basic components; development of the external interface and the interface of exchange systems between layers of the neural network; development of algorithms and computing process management tools; the transition from the concretized consistent flow graph of the neural network to its hardware implementation. The implementation of each stage of the synthesis of real-time neurocomputer systems is considered as well. Two structures of neuro elements of the parallel-flow type have been developed, in particular, calculating the macroparticle product for k digits and obtaining the macroparticle product by reading from the table. A multi-channel data exchange device based on multi-port memory is selected for exchange between neural network layers. A series-parallel converter is proposed to match the time of data entry with the duration of the conveyor cycle. Moreover, the method of adequate hardware mapping was used to move from a specified flow graph to a hardware implementation. The results obtained show that matching the intensity of data input with the intensity of processing and the use of a problem-oriented approach ensures the implementation of real-time neurocomputer systems with high efficiency of equipment use and high technical and operational characteristics.

Keywords: neural network; synthesis; neurocomputer system; coordinated-parallel data processing; flow graph; neuroelement; pipeline.