

СИНТЕЗ ПАРАЛЕЛЬНО-ПОТОКОВИХ ПРИСТРОЇВ ВЕРТИКАЛЬНОГО ОБЧИСЛЕННЯ БАЗОВИХ БАГАТООПЕРАНДНИХ НЕЙРООПЕРАЦІЙ

Іван Цмоць¹, Богдан Штогрінець²

^{1, 2} Національний університет “Львівська політехніка”,
кафедра автоматизованих систем управління, Львів, Україна.

¹ E-mail: ivan.h.tsmots@lpnu.ua, ORCID: 0000-0002-4033-8618

² E-mail: bohdan.v.shtohrincts@lpnu.ua, ORCID: 0009-0001-4956-3862

© Іван Цмоць, Богдан Штогрінець, 2025

У статті представлено дослідження апаратної реалізації базових багатооперандних нейрооперацій для штучних нейронних мереж. Виділено операційний базис нейронних мереж, що включає групи операцій попередньої обробки, процесорних операцій та обчислення передатних функцій. Обґрунтовано вибір базових багатооперандних нейрооперацій: пошуку екстремальних значень у одновимірних масивах, обчислення суми квадратів різниць, скалярного добутку та групового підсумовування. Вдосконалено методи вертикального обчислення зазначених операцій за рахунок одночасного опрацювання розрядного зрізу всіх операндів та адаптивної зміни складності операцій у конвеєрній схемі. Це забезпечує узгодження часу надходження даних з часом обчислення та високу ефективність використання обладнання. Запропоновано інтегрований підхід до розробки паралельно-потоківих пристроїв, який базується на можливостях сучасної елементної бази та враховує вимоги конкретних застосувань. Визначено принципи розробки паралельно-потоківих пристроїв вертикально-групового обчислення: використання базису елементарних арифметичних операцій, модульність, конвеєризація, просторовий паралелізм, однорідність структури, узгодження часових параметрів та спеціалізація під конкретні завдання. Розроблено послідовно-паралельний перетворювач форматів даних для перетворення потоку послідовних вхідних даних у паралельний одновимірний масив. Створено базові паралельно-потоківі структури, які апаратно реалізують вертикальні алгоритми обчислень. Удосконалено метод синтезу паралельно-потоківих пристроїв із використанням механізмів узгодження тривалості конвеєрного такту з часом надходження даних. Показано, що застосування розроблених методів та структур забезпечує обробку даних у реальному часі з високою ефективністю використання обладнання.

Ключові слова: багатооперандні нейрооперації, паралельно-потоківі пристрої, вертикальна обробка даних, реальний час, ефективність використання обладнання, штучні нейронні мережі, конвеєрна обробка, апаратна реалізація.

Постановка проблеми

Сучасні інформаційні технології активно впроваджують нейромережеві рішення реального часу у промисловості, енергетиці, транспорті, медицині та військовій сфері. Вони застосовуються для управління процесами, оптимізації навантаження, технічного зору, руху мобільних роботів,

криптозахисту, діагностики та розпізнавання образів. Проте ефективне використання штучних нейронних мереж (ШНМ) вимагає апаратних засобів, здатних опрацьовувати інтенсивні потоки даних за умов жорстких обмежень щодо габаритів, маси й енергоспоживання. Це досягається лише спеціалізованою апаратною реалізацією базових операцій (Panda et al., 2020; Xian et al., 2015). Базові операції ШНМ поділяються на одно-, дво- та багатооперандні, при цьому продуктивність системи визначається насамперед ефективністю реалізації багатооперандних операцій: пошуку екстремумів, скалярного добутку, групового підсумовування та суми квадратів різниць (Davies et al., 2018; Guo et al., 2018). Найперспективнішим підходом вважається побудова паралельно-потоків пристроїв, орієнтованих на НВІС та адаптованих до конкретних галузей застосування (Hager & Wellein, 2010; Sze et al., 2017). У таких структурах дані обробляються конвеєрно: сходинок складаються з операційних блоків і буферної пам'яті, а швидкодія визначається часом їх взаємодії (Das & Serrano-Gotarredona, 2024; Zhang et al., 2019). Ключову роль відіграє вертикальне обчислення – метод організації апаратної обробки, при якому багатооперандні нейрооперації реалізуються у вигляді багаторівневих потікових структур. На кожному рівні виконується однакова елементарна операція над частиною даних, а результат передається далі по вертикалі. Такий підхід забезпечує регулярність, простоту масштабування та можливість синхронної обробки великих масивів інформації. Ефективність вертикального обчислення досягається завдяки рекурсивним алгоритмам з локальними залежностями, що гарантують однаковий набір операцій на кожній сходинці (Albert & Ming, 2021; Liu et al., 2017). Це дозволяє виконувати збалансовану декомпозицію алгоритмів та підтримувати рівномірний час обробки незалежно від розміру задачі. У результаті досягається висока продуктивність, робота в реальному часі та повне завантаження обладнання (Das & Serrano-Gotarredona, 2024; Zhang et al., 2019). Таким чином, актуальність дослідження визначається потребою у вдосконаленні методів і структур вертикального обчислення для синтезу паралельно-потікових пристроїв багатооперандних нейрооперацій. Це безпосередньо впливає на ефективність апаратної реалізації ШНМ, що є критично важливим для сучасних застосувань у сферах із жорсткими вимогами до швидкодії, енергоефективності та надійності (Nouacer et al., 2020; Olaoye, 2025).

Аналіз останніх досліджень та публікацій

Аналіз засобів реалізації ШНМ показує (Saini et al., 2021; Muhammad Sarg et al., 2021), що переважна кількість засобів, які використовуються для реалізації нейромереж є програмними. Основними недоліками програмних засобів є невисока швидкодія, яка не забезпечує опрацювання інтенсивних потоків даних. Підвищення швидкодії програмних засобів може бути досягнуте при використанні високопродуктивних комп'ютерних систем, недоліком яких є те, що вони не відповідають обмеженням щодо габаритів і енергоспоживання. Тому для нейромережевого опрацювання інтенсивних потоків даних у реальному часі доцільно використовувати паралельно-потікову апаратну НВІС-реалізацію, яка одночасно забезпечує малі габарити та високу швидкодію. Апаратні нейронні мережі синтезуються на базі паралельно-потікових пристроїв, які реалізують найскладніші операції та орієнтовані на узгодження часу обчислення з часом надходження вхідних даних. У роботах (Baranovsky et al., 2023; Tsmots et al., 2024a) розглянуті методи обчислень та структури пристроїв для вертикального обчислення базових нейрооперацій. Недоліком розглянутих методів і структур є те, що вони не орієнтовані на паралельно-потікову реалізацію у вигляді НВІС. Характеристики ШНМ в значній мірі залежать від варіантів апаратної реалізації базових багатооперандних нейрооперацій таких як: пошук максимальних і мінімальних чисел у масивах, обчислення сум квадратів різниць, обчислення скалярного добутку та групового підсумовування (Tsmots et al., 2023; Tsmots & Antoniv, 2022). Проведений аналіз структур для реалізації базових багатооперандних нейрооперацій (Tsmots et al., 2021; Tsmots et al., 2024b) показав, що для апаратної реалізації використовуються два типи структур: рекурсивні та нерекурсивні. Структурною особливістю рекурсивних пристроїв є присутність обернених зв'язків. У таких пристроях обчислення операцій здійснюється за декілька ітерацій, кількість яких в основному залежить від розрядності операндів, які аналізуються. Недоліком рекурсивних пристроїв для реалізації базових багатооперандних нейрооперацій є відносно невисока швидкодія (Tsmots et al., 2017). У нерекурсивних пристроях обчислення операцій здійснюється за декілька ітерацій, кількість яких в основному залежить від розрядності операндів, які аналізуються. Недоліком нерекурсивних пристроїв для реалізації базових багатооперандних нейрооперацій є відносно невисока швидкодія (Tsmots et al., 2017). У нерекурсивних пристроях обчислення операцій здійснюється за декілька ітерацій, кількість яких в основному залежить від розрядності операндів, які аналізуються. Недоліком нерекурсивних пристроїв для реалізації базових багатооперандних нейрооперацій є відносно невисока швидкодія (Tsmots et al., 2017). У нерекурсивних пристроях обчислення операцій здійснюється за декілька ітерацій, кількість яких в основному залежить від розрядності операндів, які аналізуються. Недоліком нерекурсивних пристроїв для реалізації базових багатооперандних нейрооперацій є відносно невисока швидкодія (Tsmots et al., 2017).

курсивних пристроях відсутні обернені зв'язки, вони мають більшу швидкість і їх структури орієнтовані на обробку інтенсивних потоків даних у реальному часі. У роботах (Usha & Kanthimathi, 2024; Wang & Fu, 2024) розглядаються два підходи до апаратної реалізації базових багатооперандних нейрооперацій (обчислення суми квадратів різниць і скалярного добутку), перший з яких ґрунтується на операціях множення, додавання, а другий – на елементарних арифметичних операціях додавання, інверсії та зсуву. Однак у цих підходах невирішеним залишається питання оптимізації структур пристроїв та розрахунку його часових параметрів та затрат обладнання. У роботах (Wu et al., 2022; Preethi, 2021) розглядається багатооперандний підхід, при якому процес обчислення базових багатооперандних нейрооперацій розглядається як виконання єдиної операції, що ґрунтується на елементарних арифметичних операціях. Використання такого підходу для апаратної реалізації базових багатооперандних нейрооперацій забезпечить розроблення конвеєрних паралельно-потоківих пристроїв з високою ефективністю використання обладнання. З аналізу розглянутих публікацій випливає, що апаратна реалізація базових багатооперандних нейрооперацій пошуку максимальних (мінімальних) чисел у масивах, обчислення сум квадратів різниць, обчислення скалярного добутку та групового підсумовування вимагає розробки та вдосконалення вертикальних методів, алгоритмів і синтезу паралельно-потоківих структур реального часу з високою ефективністю використання обладнання.

Формулювання цілі статті

Аналіз сучасних нейромережових технологій показує, що наявна науково-технічна база не повністю забезпечує ефективну апаратну реалізацію базових багатооперандних операцій ШНМ у реальному часі. Ключова проблема полягає в узгодженні часу надходження даних із часом їх вертикального опрацювання при мінімізації апаратних витрат. Актуальним завданням є розроблення та вдосконалення вертикальних методів і паралельно-потоківих структур, які забезпечують високу ефективність використання обладнання та адаптацію до різних застосувань.

Об'єктом дослідження є процеси узгодженого часу надходження даних з часом вертикального опрацювання даних та мінімізації витрат обладнання, які забезпечують розроблення паралельно-потоківих пристроїв обчислення базових багатооперандних операцій ШНМ у реальному часі з високою ефективністю використання обладнання.

Предметом дослідження є вертикальні методи, паралельно-потоківі структури пристроїв обчислення базових багатооперандних операцій ШНМ у реального часу з високою ефективністю використання обладнання.

Метою дослідження є розроблення методів і паралельно-потоківих структур пристроїв вертикального обчислення базових багатооперандних операцій ШНМ у реального часу з високою ефективністю використання обладнання.

Для досягнення зазначеної мети визначено такі основні **завдання дослідження**:

- виділити операційний базис ШНМ та вибрати найскладніші багатооперандні операції (пошук екстремумів, сума квадратів різниць, скалярний добуток, групове підсумовування);
- удосконалити методи їх вертикального обчислення з орієнтацією на паралельно-потоківі структури;
- розробити базові структури пристроїв вертикально-групового обчислення для синтезу систем реального часу;
- удосконалити метод синтезу паралельно-потоківих пристроїв, що забезпечує адаптацію до конкретних застосувань та високу ефективність використання обладнання.

Виклад основного матеріалу

Сучасні штучні нейронні мережі (ШНМ) є одним з найбільш перспективних напрямків розвитку інформаційних технологій, що знаходять широке застосування у задачах розпізнавання образів, обробки природних мов, машинного навчання та інших галузях. Однак ефективна апаратна реалізація

нейронних мереж потребує глибокого аналізу їх операційної структури та розробки спеціалізованих обчислювальних архітектур. На основі аналізу алгоритмів роботи ШНМ (Saini et al., 2021) виділений операційний базис, який складається із таких груп операцій: перша – нейрооперації попередньої обробки даних; друга – процесорні нейрооперації; третя – функції активації.

Перша група нейрооперацій забезпечує перетворення вхідних даних до вигляду, який дасть найкращі результати. Навчальний вектор містить по одному значенню на кожен вхід нейромережі і по одному значенню для кожного виходу мережі, у залежності від типу навчання (з вчителем або без вчителя). Навчання мережі на «сирому» наборі, як правило, не дає якісних результатів. Для підвищення якості роботи нейронної мережі здійснюється попередня обробка вхідних даних, яка зводиться до виконання таких операцій: нормалізації, квантування і фільтрації. Нормалізація – це процедура попередньої обробки вхідних даних (навчальних, тестових і робочих вибірок), при якій значення ознак, які формують вхідний вектор, приводиться до деякого заданого діапазону. Після нормалізації всі значення вхідних ознак будуть приведені до деякого вузького діапазону $[0, 1]$ або $[-1, 1]$. Нормалізації вхідних даних до діапазону $[0, 1]$ виконується так:

$$x_i^{\times} = \frac{x_i - x_{min}}{x_{max} - x_{min}}, \quad (1)$$

де x_i – вхідні дані; x_{max} – максимальне значення вхідних даних; x_{min} – мінімальне значення вхідних даних. Нормалізації вхідних даних до діапазону $[-1, 1]$ виконується так:

$$x_i^{\times} = \frac{x_i}{|x|_{max}}. \quad (2)$$

Такі види нормалізації не вимагають виконання складних обчислень і широко використовуються для вхідних даних x_i , які щільно заповнюють певний інтервал. Після нормалізації вхідних даних у RBF- та GRNN-мережах потрібно здійснити обчислення евклідової відстані від кожного вхідного вектору до всіх інших. Таке обчислення евклідової відстані виконується з використанням нейрооперації:

$$y = ||x_i^e - x_i^b||^2 = (x_1^e - x_1^b)^2 + (x_2^e - x_2^b)^2 + \dots + (x_N^e - x_N^b)^2. \quad (3)$$

Для інших типів нейронних мереж можуть використовуватися фільтрація, яка виконується над зашумленими вхідними даними і зводиться до відкидання значень, які є некоректними. Крім того, над неперервними величинами виконується квантування, яке передбачає виділення скінченного набору дискретних значень. Процесорні операції над вхідними даними та ваговими коефіцієнтами виконуються безпосередньо у самій нейронній мережі у процесі навчання та функціонування та зводяться до обчислення зваженої суми. При обробці даних у самій нейронній мережі можуть використовуватися такі операції: додавання, множення, групове підсумовування та обчислення скалярного добутку. Значення зваженої суми перетворюється у вихідний сигнал через алгоритмічний процес, відомий під назвою функція активації або передатна функція. У нейронних мережах можуть використовуватися різні функції активації, які вибираються у залежності від задач, що розв'язуються та типу нейронної мережі. Найчастіше у нейронних мережах використовуються такі функції активації: лінійна, порогова, сигмоїдальна та гіперболічна. Із аналізу операційного базису ШНМ видно, що найбільше швидкодія апаратних нейронних мереж залежить від таких операцій: пошуку у максимальних і мінімальних значень у одновимірному масиві даних, обчислення суми квадратів різниць, обчислення скалярного добутку та групового підсумовування чисел. Особливістю даних нейрооперацій є те, що вони є багатооперандними і виконуються над множиною операндів.

Результатом виконання багатооперандної нейрооперації є одне число. Багатооперандні нейрооперації пропонується виконувати на основі багатооперандного підходу, при якому процес обчислення нейрооперації розглядається як виконання єдиної операції, що ґрунтується на елементарних арифметичних операціях.

Методи вертикального обчислення базових багатооперандних нейрооперацій

При вертикальній обробці даних вхідні дані X_j та вагові коефіцієнти W_j ($j=1, \dots, N$, де N – кількість даних і кількість вагових коефіцієнтів) представляються в порозрядному вигляді згідно з формулою:

$$W_j = \sum_{i=1}^n 2^{-(i-1)} w_{ji}, \quad W_j = \sum_{i=1}^n 2^{-(i-1)} w_{ji} \quad (4)$$

де w_{ji} , x_{ji} – значення i -х розрядів j -х вагових коефіцієнтів та j -х вхідних даних; $i=1, \dots, n$, де n – розрядність вагових коефіцієнтів та вхідних даних. Вертикальне обчислення базових багатооперандних нейрооперацій передбачає надходження та опрацювання даних паралельно розрядними зрізами (вертикально) відповідно з виразом:

$$w_{1i}, \dots, w_{ji}, \dots, w_{Ni}; \quad x_{1i}, \dots, x_{ji}, \dots, x_{Ni} \quad (5)$$

Використання методів вертикального опрацювання даних для реалізації пристроїв паралельно-потокowego обчислення базових багатооперандних операцій ШНМ забезпечить одночасне опрацювання розрядних зрізів всіх операндів, масштабування за кількістю операндів, ефективну реалізацію на FPGA, зменшення апаратної складності та енергоспоживання.

Вертикальний метод пошуку максимального X_{max} і мінімального X_{min} значень у одновимірному $\{X_j\}_{j=1}^N$ масиві даних передбачає паралельне надходження у кожному i -у такті i -о розрядного зрізу всіх N чисел одновимірного масиву. Опрацювання розрядних зрізів при пошуку максимального X_{max} і мінімального X_{min} чисел у одновимірному масиві даних починається із старших розрядів. Пошук максимального X_{max} і мінімального X_{min} чисел у одновимірному масиві $\{X_k\}_{k=1}^N$ за вертикальним методом ґрунтується виконанні макрооперацій: пошуку i -о розряду максимального числа x_{maxi} та пошуку i -о розряду мінімального числа x_{mini} . Макрооперація пошуку i -о розряду максимального числа x_{maxi} у одновимірному масиві $\{X_j\}_{j=1}^N$ виконується так:

- 1) визначається значення i -о розрядного зрізу P_i за формулою:

$$P_i = \cap_{j=1}^N, \quad (6)$$

де x_{ij} – значення i -о розряду j -о числа масиву, y_{ji} – значення j -о розряду i -о слова управління, значення 1-о слова управління дорівнює $y_{11}=y_{12}=\dots=y_{1N}=1$;

- 2) визначається значення i -о розряду максимального числа x_{maxi} за виразом:

$$x_{maxi} = \begin{cases} 0, & \text{коли } P_i=1 \\ 1, & \text{коли } P_i=0 \end{cases}, \quad (7)$$

- 3) визначаються j розряди $(i+1)$ -о слова управління за формулою:

$$y_{(i+1)j} = \begin{cases} 0, & \text{коли } P_i=1, x_{ji} \neq y_{ji} \\ 1, & \text{коли } P_i=0, x_{ji} = y_{ji} = 1 \end{cases} \quad (8)$$

Макрооперація пошуку i -о розряду мінімального числа x_{mini} у одновимірному масиві $\{X_j\}_{j=1}^N$ виконується так:

- 1) визначається значення i -о розрядного зрізу P_i за формулою:

$$P_i = \cup_{j=1}^N x_{ij} \wedge y_{ij}, \quad (9)$$

де x_{ij} – інверсне значення i -о розряду j -о числа масиву; y_{ij} – значення j -о розряду i -о слова управління, значення 1-о слова управління дорівнює $y_{11}=y_{12}=\dots=y_{1N}=1$;

- 2) визначення i -о розряду мінімального числа x_{mini} за виразом:

$$x_{mini} = \begin{cases} 0, & \text{коли } P_i=1 \\ 1, & \text{коли } P_i=0 \end{cases}, \quad (10)$$

3) формування j розрядів $(i+1)$ -о слова управління, яке здійснюється за виразом:

$$y_{(i+1)j} = \begin{cases} 0, & \text{коли } P_i=1, x_{ji} \neq y_{ji} \\ 1, & \text{коли } P_i=0, x_{ji} = y_{ji} = 1 \end{cases} \quad (11)$$

Особливість методу вертикально пошуку максимального (мінімального) числа є те, що у кожному i -у такті роботи визначаються i -і розряди максимального $X_{\max}$ та мінімального $X_{\min}$ чисел. Особливістю вертикально пошуку максимальних і мінімальних чисел у одновимірному масиві є:

- використання однотипних базових макрооперації: пошуку максимального та пошуку мінімального числа;
- можливість використання розпаралелення та конвеєризації пошуку;
- можливість одночасного опрацювання i -о розрядного зрізу всіх N чисел;
- час пошуку в основному визначається розрядністю чисел n , а не їх кількістю N .

Метод вертикального обчислення суми квадратів різниць

Для обчислення суми різниць квадратів використаємо багатооперандний вертикальний підхід, який передбачає одночасне послідовно-порозрядне надходження операндів, формування та підсумовування в кожному такті макрочасткових результатів. Вертикальний метод обчислення суми квадратів різниць вимагає, щоб операнди були представлені в порозрядному вигляді згідно з формулою (1). Вертикальне обчислення суми квадратів різниць ґрунтується на алгоритмі вертикального піднесення до квадрату числа X :

$$X^2 = (0.01) \wedge x_1 + 2^{-1}(0x_101) \wedge x_2 + 2^{-2}(0x_1x_201) \wedge x_3 + \dots + 2^{-(n-1)}(0x_1x_2, \dots, x_{n-1}01) \wedge x_n \quad (12)$$

У формулі (12) частковий результат піднесення до квадрату $(0x_1x_2 \dots x_{i-1}01) \wedge x_i$ замінюємо на Q_i і тоді алгоритм піднесення до квадрату запишеться так:

$$X^2 = \sum_{i=1}^n 2^{-(i-1)} Q_i. \quad (13)$$

Обчислення N сум квадратів різниць будемо здійснювати на основі багатооперандного підходу, який полягає у одночасному опрацюванню всіх операндів і формуванні для них часткових результатів суми квадратів різниць. Обчислення N сум квадратів різниць за вертикальним методом виконується за такою формулою:

$$\begin{aligned} Y &= (X_1^e - X_1^b)^2 + (X_2^e - X_2^b)^2 + \dots + (X_N^e - X_N^b)^2 = \Delta X_1^2 + \Delta X_2^2 + \dots + \Delta X_N^2 = \\ &= \sum_{i=1}^n 2^{-(i-1)} Q_{1i} + \dots + \sum_{i=1}^n 2^{-(i-1)} Q_{Ni} = \sum_{j=1}^n \sum_{i=1}^n 2^{-(i-1)} Q_{ji} = \sum_{i=1}^n 2^{-(i-1)} Q_{ji} = \\ &= \sum_{i=1}^n 2^{-(i-1)} Q_{Mi} \end{aligned} \quad (14)$$

де Q_{Mi} – i -й макрочастковий результат суми квадратів різниць.

Основними етапами вертикального методу обчислення суми квадратів різниць Y є:

- одночасне послівне надходження операндів X_j^e, X_j^b і обчислення N модулів різниці ΔX_j ;
- формування починаючи із старших розрядів для кожного j -о модуля ΔX_j у i -у такті i -о часткового результатів піднесення до квадрату Q_{ji} ;
- обчислення i -о макрочасткового результату суми квадратів різниць Q_{Mi} шляхом підсумовування всіх i -х часткових результатів піднесення до квадрату Q_{ji} ;
- обчислення i -о результату суми квадратів різниць Y_i шляхом підсумовування $(i-1)$ -о результату суми квадратів різниць Y_{i-1} зсунутого на один розряд вліво із i -м макрочастковим результатом піднесення до квадрату Q_{Mi} у відповідності з виразом $Y_i = 2Y_{i-1} + Q_{Mi}$, де $Y_0=0$.

Метод вертикального обчислення скалярного добутку

Метод вертикально обчислення скалярного добутку реалізується на базі елементарних арифметичних операцій та орієнтований на паралельно-потоківу реалізацію. Використання даного методу забезпечує опрацювання інтенсивних потоків даних у реальному часі.

Обчислення скалярного добутку за вертикальним методом зводиться до формування та підсумовування часткових добутків у відповідності з наступною формулою:

$$Z = \sum_{j=1}^N W_j X_j = \sum_{j=1}^N \sum_{i=1}^n 2^{-(i-1)} W_j X_{ji} \quad (15)$$

Зробивши необхідні зміни у формулі (15) обчислення скалярного добутку можна записати так:

$$Z = \sum_{i=1}^n 2^{-(i-1)} \sum_{j=1}^N W_j X_{ji} = \sum_{i=1}^n 2^{-(i-1)} \sum_{j=1}^N P_i = \sum_{i=1}^n 2^{-(i-1)} P_{iM}, \quad (16)$$

де P_i – i -й частковий добуток обчислення скалярного добутку, P_{iM} – i -й макрочастковий результат обчислення скалярного добутку.

З формули (16) видно, що для обчислення скалярного добутку вимагає виконання n тактів, у кожному i -у такті виконуються такі операції:

- формування для кожної j -ї пари операндів i -о часткового добутку у відповідності з формулою $P_{ji} = W_j x_j$;
- обчислення i -о макрочасткового результату обчислення скалярного добутку P_{iM} шляхом підсумовування всіх i -х часткових добутків P_{ji} у відповідності з формулою $P_{iM} = \sum_{j=1}^N P_{ji}$;
- додавання i -о макрочасткового результату обчислення скалярного добутку P_{iM} до результату підсумовування, який зсунутий вправо на один розряд, у відповідності з виразом $Z_i = 2^{-1} Z_{i-1} + P_{iM}$, де $Z_0=0$.

Метод вертикального групового підсумовування чисел

Групове підсумовування M двійкових n -розрядних чисел записуються так:

$$Z_i = \sum_{j=1}^M \sum_{i=1}^n 2^{-i} C_{ji} \quad (17)$$

Формула (17) відображає горизонтальну модель групового підсумовування M двійкових n -розрядних чисел. Для вдосконалення групового підсумовування використаємо вертикальний та багатооперандний підходи. При використанні таких підходів процес 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 формулі (17) порядок підсумовування переходимо до вертикальної моделі групового підсумовування, яка записується так:

$$Z = \sum_{j=1}^n 2^{-i} \sum_{i=1}^{M_i} C_{ji}, \quad (18)$$

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

Вертикальні методи групового підсумовування зводять процес підсумовування до перетворення багаторядного коду в однорядний. Таке перетворення ґрунтується на однорозрядних операціях у формулі (19), де E_{3-2} , E_{7-3} і E_{15-4} – результати однорозрядних операцій додавання відповідно трьох, семи і п'ятнадцяти операндів.

Розробку паралельно-потоківих пристроїв вертикального обчислення базових багатооперандних нейрооперацій запропоновано здійснювати на основі інтегрованого підходу, що враховує можливості сучасної елементної бази, поєднує вертикальні методи, алгоритми та структури й орієнтується на вимоги конкретних застосувань. Основні принципи: використання елементарних арифметичних операцій і багатооперандного підходу, модульність та локалізація зв'язків між сходинками, конвеєризація й просторовий паралелізм, однорідність і регулярність структур, узгодженість часу надходження даних з часом обчислення, а також спеціалізація й адаптація під алгоритм. Особливістю паралельно-потоківих конвеєрних пристроїв вертикального обчислення базових багатооперандних нейрооперацій є те, що вони найбільше відповідають умовам роботи у реальному часі. У таких пристроях суміщається одночасне опрацювання n масивів даних. Основними перевагами таких структур є те, що вони реалізуються на базі однотипних сходинок конвеєра та орієнтовані на НВІС-реалізацію.

$$\begin{aligned}
 E_{3-2} &= \begin{cases} C_{ji} \\ + \\ C_{(j+1)i} \\ + \\ C_{(j+2)i} \end{cases} = \begin{cases} P_{i-1} \\ + \\ S_i \end{cases}, \\
 E_{7-3} &= \begin{cases} C_{ji} \\ + \\ C_{(j+1)i} \\ + \\ C_{(j+2)i} \\ + \\ C_{(j+3)i} \\ + \\ C_{(j+4)i} \\ + \\ C_{(j+5)i} \\ + \\ C_{(j+6)i} \end{cases} = \begin{cases} P_{i-2} \\ + \\ S_{i-1} \\ + \\ S_i \end{cases}, \\
 E_{15-4} &= \begin{cases} C_{ji} \\ + \\ C_{(j+1)i} \\ + \\ C_{(j+2)i} \\ + \\ C_{(j+3)i} \\ + \\ C_{(j+4)i} \\ + \\ C_{(j+5)i} \\ + \\ C_{(j+6)i} \\ + \\ C_{(j+7)i} \\ + \\ C_{(j+8)i} \\ + \\ C_{(j+9)i} \\ + \\ C_{(j+10)i} \\ + \\ C_{(j+11)i} \\ + \\ C_{(j+12)i} \\ + \\ C_{(j+13)i} \\ + \\ C_{(j+14)i} \end{cases} = \begin{cases} P_{i-3} \\ + \\ S_{i-2} \\ + \\ S_{i-1} \\ + \\ S_i \end{cases} \quad (19)
 \end{aligned}$$

**Структура паралельно-потоків пристрою пошуку
максимальних і мінімальних чисел у одновимірному масиві даних**

Паралельно-потоків пристрій пошуку максимальних і мінімальних значень реалізується на базі послідовно-паралельного перетворювача форматів даних ПППФ та n однотипних сходенок конвеєра (СК). Кожна СК _{i} апаратно реалізує i -у базову макрооперацію обчислення максимального та мінімального чисел у одновимірному масиві даних. Структура паралельно-потоків пристрою пошуку максимальних і мінімальних чисел у одновимірному масиві $\{X_j\}_{j=1}^N$ за паралельним вертикальним методом наведена на рис.1, де X_j – вхід чисел, TI – вхід перших тактових імпульсів, TK – вхід конвеєрних тактових імпульсів, Tg – тригер, Rg – регістр, $X_1, \dots, X_N$ – вхідний одновимірний масив чисел, $x_{1min}, \dots, x_{nmin}$ – виходи розрядів мінімального числа, $x_{1max}, \dots, x_{nmax}$ – виходи розрядів максимального числа, ПЕ – процесорний елемент, ПППФ – послідовно-паралельний перетворювач форматів, СК – сходинка конвеєра. Пошук максимального і мінімального чисел у одновимірному масиві відбувається у два етапи. На першому етапі у кожному такті на вхід X_j надходять послідовно вхідні дані, які тактовим імпульсом TI записуються у ПППФ. Після N тактового імпульсу TI всі числа першого одновимірного масиву будуть записані у регістри ПППФ. На другому етапі роботи числа X_j з виходів ПППФ надходять на входи регістрів ПЕ _{$j1$} першої СК₁ і переднім фронтом першого конвеєрного тактового імпульсу TK записуються у дані регістри. Кількість ПЕ _{ji} у кожній СК _{i} рівна кількості чисел в одновимірному масиві $\{X_j\}_{j=1}^N$. Всі ПЕ _{ji} однієї i -ї СК _{i} підключаються до виходу i -х розрядів мінімального числа та максимального чисел. Використання спільних шин розрядів результатів забезпечує розпаралелення процесу опрацювання розрядного зрізу, час опрацювання якого визначає такт роботи пристрою.

Пошук максимальних і мінімальних чисел у розробленому паралельно-потоківому пристрої виконується з конвеєрним тактом, який визначається за наступною формулою:

$$T_{\text{ПММЧ}} = t_{\text{Рг}} + 3t_{\text{логI}} \quad (20)$$

де $T_{\text{ПММЧ}}$ – тривалість конвеєрного такту пристрою пошуку максимального та мінімального чисел; $t_{\text{Рг}}$ та $t_{\text{логI}}$ час спрацювання відповідно регістра та логічних елементів типу АБО, І, І-НЕ.

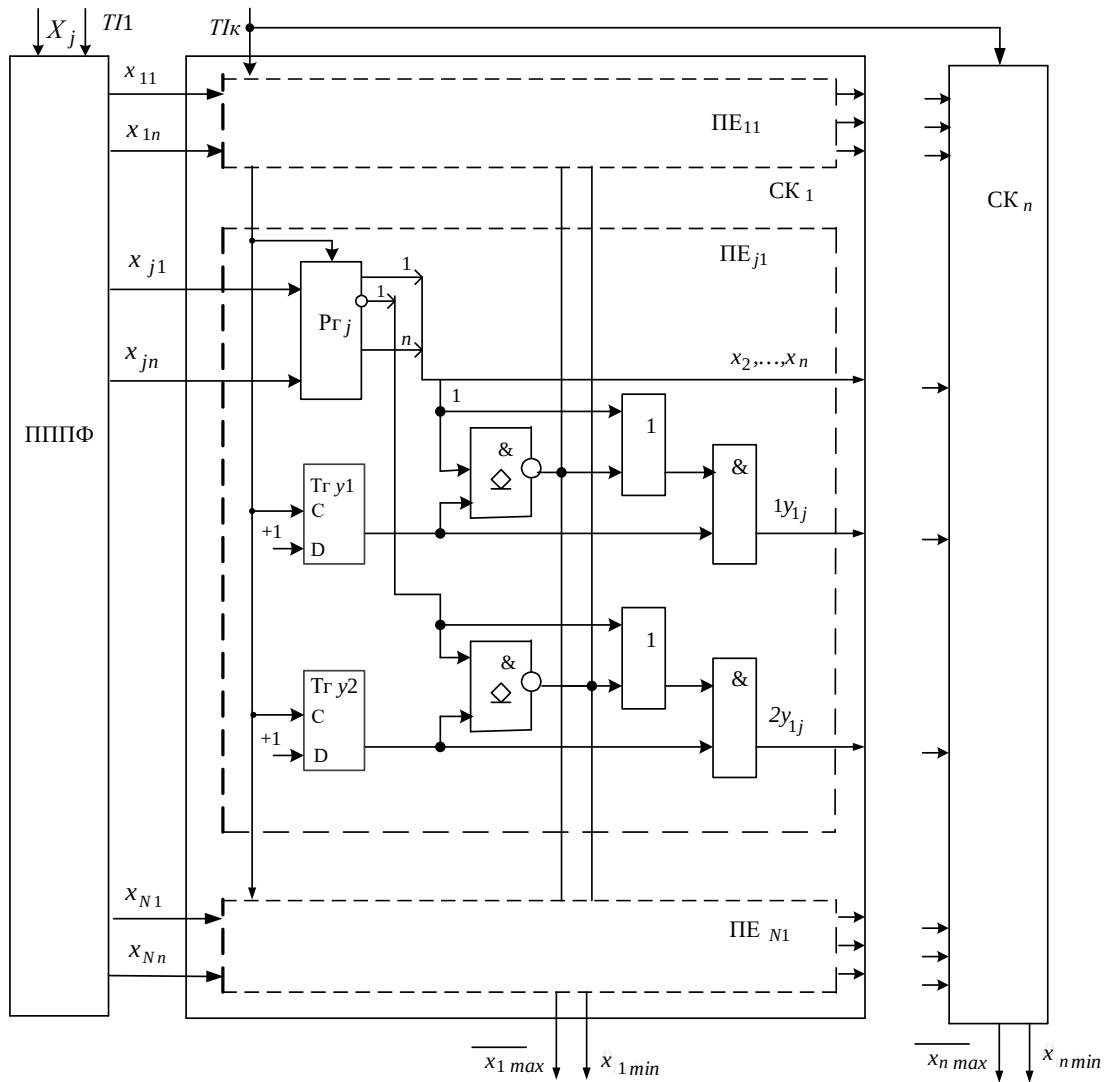


Рис.1. Структура паралельно-потоківому пристрою пошуку максимальних і мінімальних чисел у одновимірному масиві

Структура паралельно-потоківому пристрою вертикального обчислення суми квадратів різниць

Паралельно-потоківий пристрій вертикального обчислення суми квадратів різниць реалізується на базі n однотипних сходинок конвеєра (СК). Кожна i -а сходинка конвеєра $СК_i$ апаратно реалізує i -у ітерацію алгоритму вертикального обчислення суми квадратів різниць. Структура паралельно-потоківому пристрою вертикального обчислення суми квадратів різниць наведена на рис.2, де $ТП1$ – вхід перших тактових імпульсів, X_j^e і X_j^b – входи даних, $ТІК$ – вхід конвеєрних тактових імпульсів, ВД – віднімач, Рг – регістр, БСм – багатовходовий суматор, См – суматор, ПППФР – послідовно-паралельний перетворювач формату різниць, Y – вихід суми квадратів різниць.

Перший результат обчислення суми квадратів різниць отримаємо на виході Y після N -о конвеєрних тактових імпульсу TI_k . У кожному наступному конвеєрному такті на виході Y будемо отримувати наступні суми квадратів різниць.

Час конвеєрного такту TI_k роботи пристрою вертикального обчислення суми квадратів різниць визначається за наступною формулою:

$$t_k = t_{p_r} + t_{бсм} + t_{см}, \quad (21)$$

де t_{p_r} – час запису/читання даних у регістр, $t_{бсм}$ – час підсумовування N чисел, $t_{см}$ – час підсумовування двох чисел.

Для забезпечення режиму реального часу необхідно, щоб виконувалася наступна умова:

$$t_k \leq N \times T_{TI1}, \quad (22)$$

де T_{TI1} – час періоду тактового імпульсу $TI1$.

Структура паралельно-потокowego пристрою вертикального обчислення скалярного добутку

Паралельно-потокowy пристрій вертикального обчислення скалярного добутку реалізується на базі n однотипних сходинок конвеєра (СК). Кожна i -а сходина конвеєра $СК_i$ апаратно реалізує i -у ітерацію алгоритму обчислення скалярного добутку з прямим формуванням часткових добутків на основі аналізу одного розряду множників.

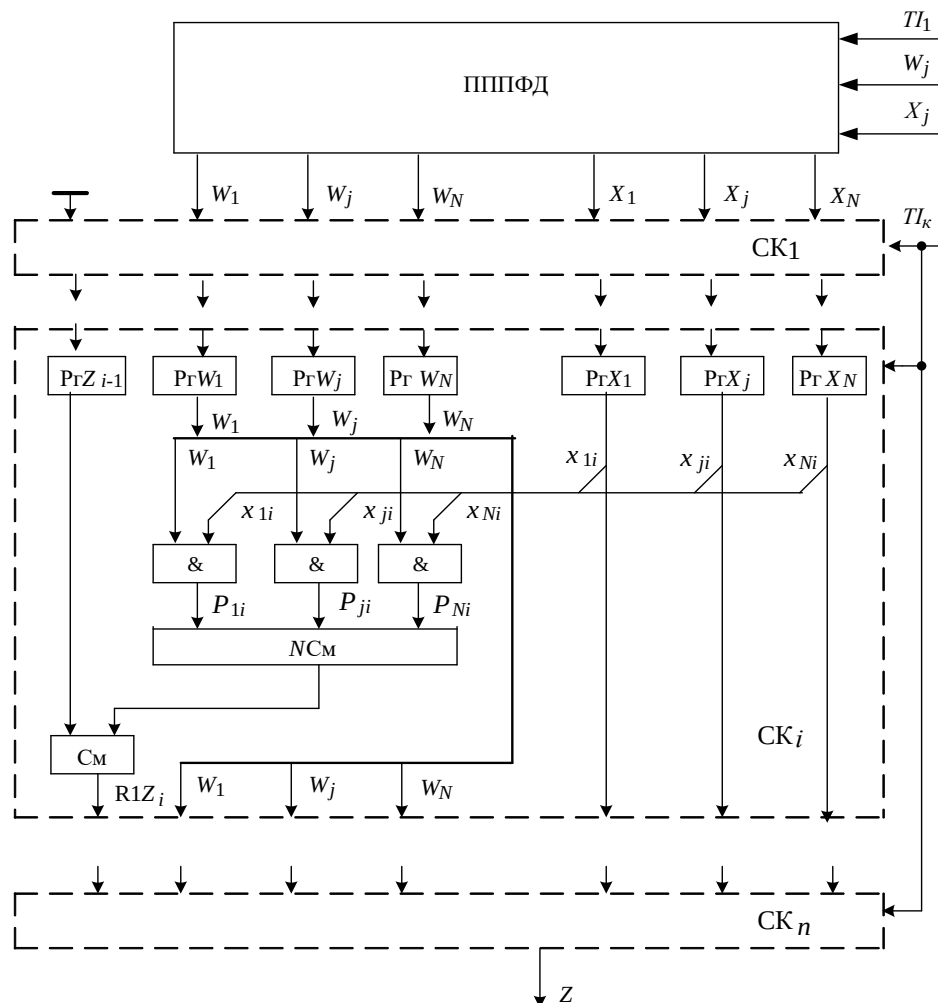


Рис.3. Паралельно-потокова структура пристрою
ветикального обчислення скалярного добутку

Структура паралельно-потоків пристрою вертикального обчислення скалярного добутку, яка реалізовує алгоритм обчислення скалярного добутку з прямим формуванням часткових добутків на основі аналізу одного розряду множників, наведена на рис.3, де TI_1 – вхід перших тактових імпульсів, TI_k – вхід конвеєрних тактових імпульсів, X_j – вхід даних, W_j – вхід вагових коефіцієнтів, ПППФД – послідовно-паралельний перетворювач форматів даних, СК – сходінка конвеєра, Рг – регістр, См – суматор, $NC_{см}$ – N -вхідний суматор, Z – вихід результату обчислення скалярного добутку. Основними компонентами паралельно-потоків пристрою вертикального обчислення скалярного добутку є ПППФД та n сходінок конвеєра $СК_i$. Обчислення скалярного добутку у даному пристрої відбувається у два етапи.

На першому етапі у кожному j -у такті роботи на входи X_j та W_j надходять j -і вхідні дані j -і вагові коефіцієнти, які переднім фронтом тактового імпульсу TI_1 записуються в регістри ПППФД. Після N -о тактового імпульсу TI_1 на виходах ПППФД отримуємо N вагових коефіцієнтів $W_1, \dots, W_N$ і N вхідних даних $X_1, \dots, X_N$.

На другому етапі роботи N вагових коефіцієнтів $W_1, \dots, W_N$ і N вхідних даних $X_1, \dots, X_N$ переднім фронтом першого конвеєрного тактового імпульсу TI_k записуються в регістри $РгW_1, \dots, РгW_N$ та регістри $РгX_1, \dots, РгX_N$ першої сходінки конвеєра $СК_1$. На виходах груп елементів I отримуємо часткових добутків, які надходять на входи N -входного суматора $NC_{см}$. Шляхом підсумовування N часткових добутків на виході N -входного суматора $NC_{см}$ отримуємо перший макрочастковий результату обчислення скалярного добутку P_{1M} . На суматорі См перший макрочастковий результат обчислення скалярного добутку P_{1M} додається до нульового результату скалярного добутку Z_0 , який зсунутий на один розряд вправо.

В наступному конвеєрному такті дані з виходів регістрів $РгW_1, \dots, РгW_N$ та регістрів $РгX_1, \dots, РгX_N$ та перший частковий результат скалярного добутку Z_1 виходів суматора См першої сходінки конвеєра $СК_1$ переписуються в регістри $РгW_1, \dots, РгW_N$, $РгX_1, \dots, РгX_N$ та $РгZ_1$ другої сходінки конвеєра $СК_2$.

Перший результат обчислення отримуємо на виході Z після N -о конвеєрних тактових імпульсу TI_k . У кожному наступному конвеєрному такті на виході Z будемо отримувати наступні результати обчислення скалярного добутку.

Особливістю розробленої структури паралельно-потоків пристрою вертикального обчислення скалярного добутку з прямим формуванням часткових добутків на основі аналізу одного розряду множників є те, що кількість сходінок конвеєра дорівнює розрядності множників X_j . Швидкодія такого пристрою визначається тривалістю такту роботи конвеєра, який обчислюється за такою формулою:

$$T_{k1} = t_{P_r} + t_I + t_{NC_{см}} + t_{C_{см}}, \quad (23)$$

де t_{P_r} – час запису у регістр; t_I – час затримки на логічному елементі I ; $t_{NC_{см}}$ – час підсумовування N макрочасткових добутків, $t_{C_{см}}$ – час додавання двох чисел.

Структури багатовходових суматорів з вертикальним підсумовуванням

Для групового підсумовування багаторозрядних чисел використовуються багатовходові однорозрядні суматори. Такі суматори об'єднуються за принципом дерева Уоллеса та перетворюють багаторядний код в дворядний, який перетворюється в однорядний за допомогою паралельного суматора. Для реалізації таких операцій використовуються 3-, 7- і 15-входові однорозрядні суматори.

Для синтезу 7-входного однорозрядного суматора розроблено наступні аналітичні вирази:

$$\begin{aligned} S_2(2^0) &= Y_0 L_1 \vee Y_1 L_0 \vee Y_0 L_3 \vee Y_1 L_2 \vee Y_2 L_1 \vee Y_3 L_0 \vee Y_1 L_4 \vee Y_2 L_3 \vee Y_3 L_2 \vee Y_3 L_4, \\ S_1(2^1) &= Y_0 L_2 \vee Y_1 L_1 \vee Y_2 L_0 \vee Y_2 L_1 \vee Y_3 L_0 \vee Y_2 L_4 \vee Y_3 L_3 \vee Y_3 L_4 \vee Y_0 L_3 \vee Y_1 L_2, \\ P(2^2) &= Y_0 L_4 \vee Y_1 L_3 \vee Y_2 L_2 \vee Y_3 L_1 \vee Y_1 L_4 \vee Y_2 L_3 \vee Y_3 L_2 \vee Y_2 L_4 \vee Y_3 L_3 \vee Y_3 L_4, \end{aligned} \quad (24)$$

де $Y_0 = \underline{C_7 C_6 C_5}$; $Y_1 = \underline{C_7 C_6 C_5} \vee \underline{C_7 C_6 C_5} \vee \underline{C_7 C_6 C_5}$; $Y_2 = \underline{C_7 C_6 C_5} \vee \underline{C_7 C_6 C_5} \vee \underline{C_7 C_6 C_5}$; $Y_3 = \underline{C_7 C_6 C_5}$; $L_0 = \underline{C_4 C_3 C_2 C_1}$; $L_1 = \underline{C_4 C_3 C_2 C_1} \vee \underline{C_4 C_3 C_2 C_1} \vee \underline{C_4 C_3 C_2 C_1} \vee \underline{C_4 C_3 C_2 C_1}$; $L_2 = \underline{C_4 C_3 C_2 C_1} \vee \underline{C_4 C_3 C_2 C_1} \vee \underline{C_4 C_3 C_2 C_1} \vee \underline{C_4 C_3 C_2 C_1}$; $L_3 = \underline{C_4 C_3 C_2 C_1} \vee \underline{C_4 C_3 C_2 C_1} \vee \underline{C_4 C_3 C_2 C_1} \vee \underline{C_4 C_3 C_2 C_1}$; $L_4 = \underline{C_4 C_3 C_2 C_1}$

На основі аналітичних виразів (10) синтезуємо комбінаційний 7-входовий однорозрядний суматор, схема якого подана на рис.4. Швидкодія комбінаційного 7-входового однорозрядного суматора визначається часом затримки проходження даних з входу на вихід і обчислюється так:

$$t_{\text{СМ7-3}} = 5t_{\text{логI}}, \quad (25)$$

Для підсумовування восьми чисел розрядністю вісім на основі 7-входового однорозрядного суматора синтезовано структура багатовходового суматора з вертикальним підсумовуванням, яка наведена на рис.5.

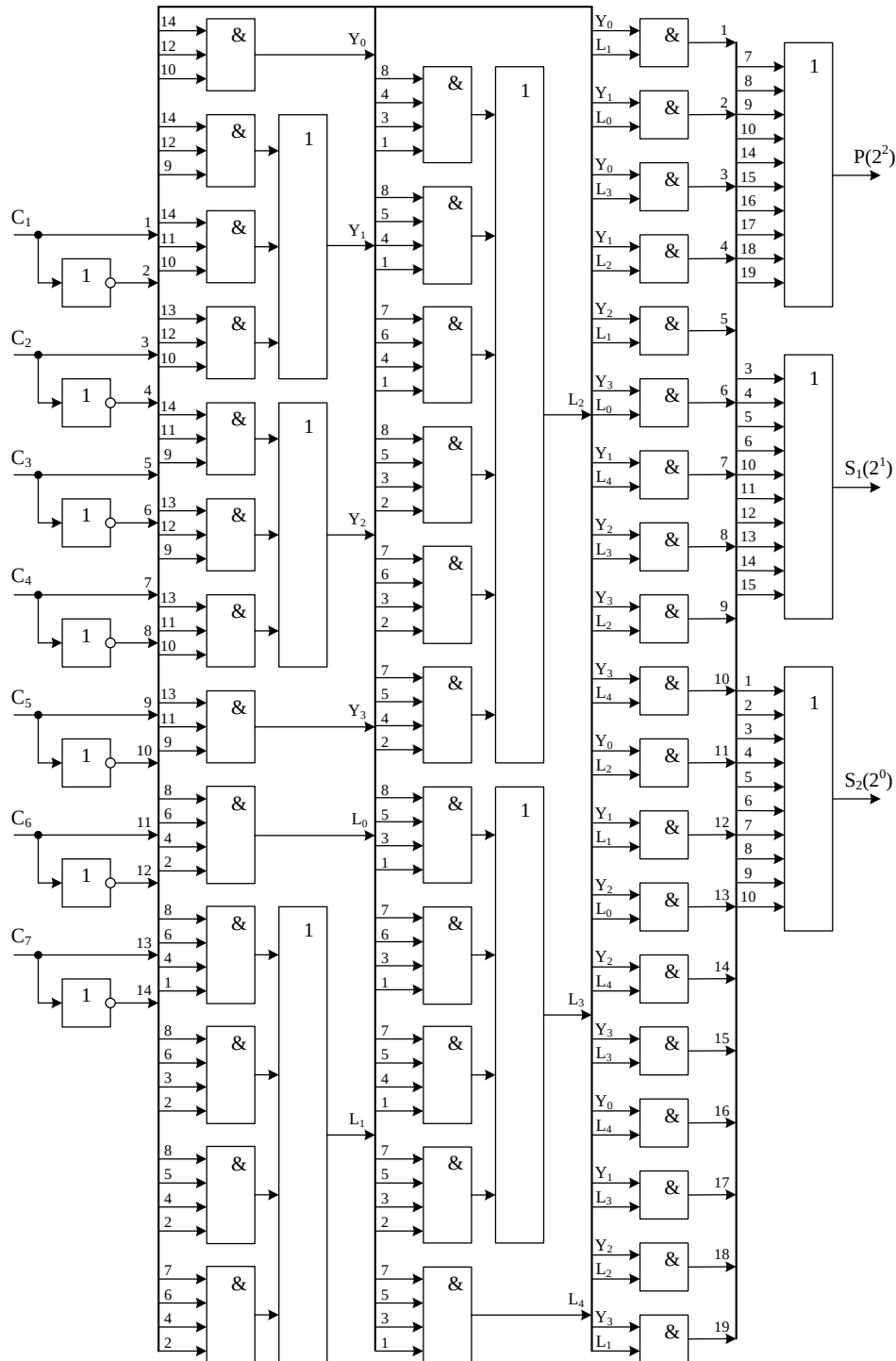


Рис. 4. Схема комбінаційного 7-входового однорозрядного суматора

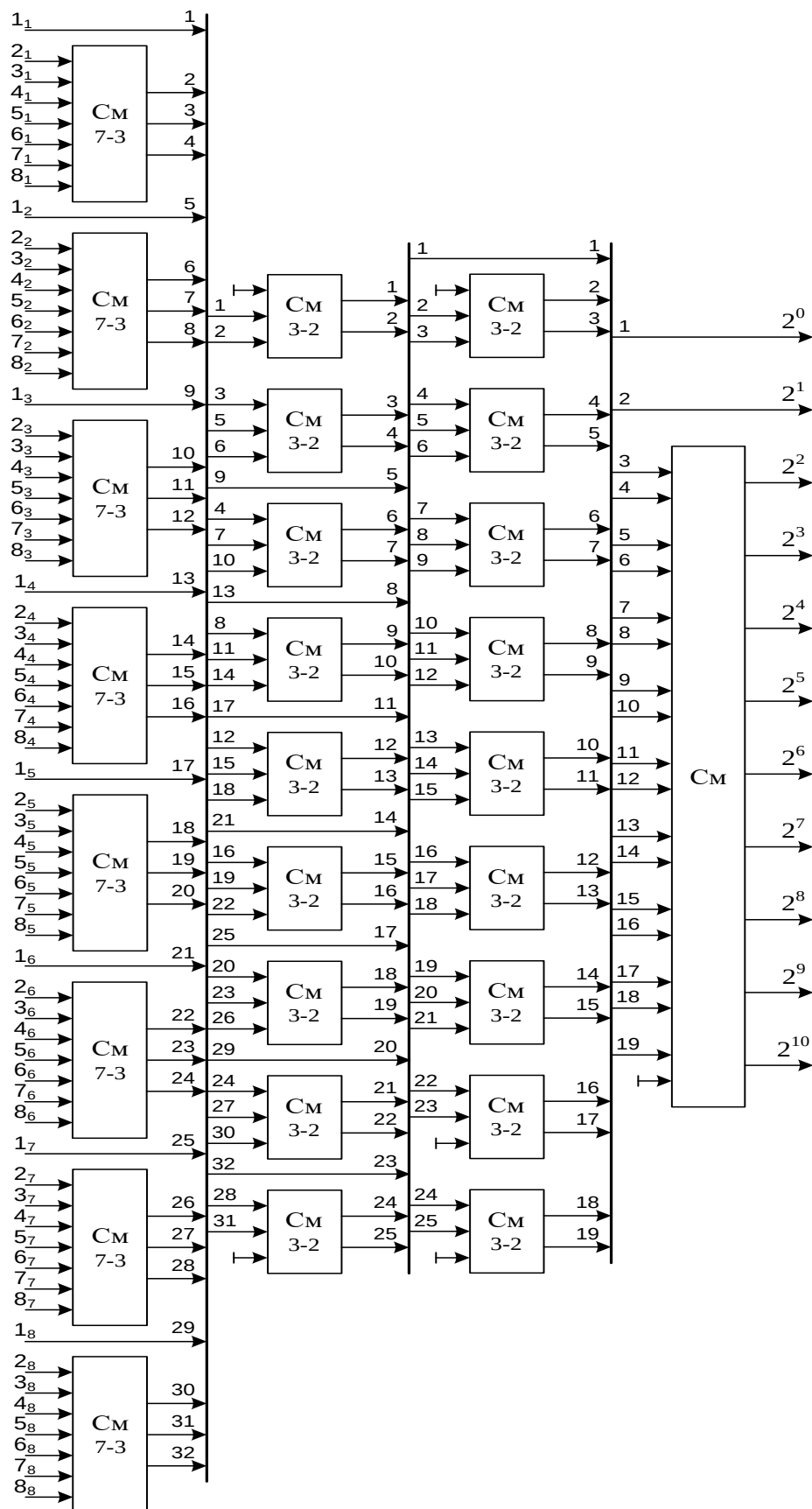


Рис.5. Структура багатовходового суматора
з вертикальним підсумовування восьми чисел розрядністю вісім

Підсумовування восьми чисел розрядністю вісім у багатовходовому виконується за чотири етапи шляхом використання таких перетворень: 7-рядного коду в 3-рядний, 3-рядного коду в 2-рядний та 2-рядного коду в 1-рядний код. На кожному етапі виконується такі операції:

- перший етап – виконується перетворення 7-рядного коду в 3-рядний за допомогою восьми 7-входових однорозрядних суматорів;
- другий і третій етапи – виконується перетворення з 3-рядного коду в 2-рядний з допомогою двох шарів дев'яти 3-входових однорозрядних суматорів;
- четвертий етап – виконується перетворення 2-рядного коду в 1-рядний за допомогою 9-розрядного паралельного суматора.

Час підсумовування восьми чисел розрядністю вісім у багатовходовому суматорі з вертикальним підсумовуванням визначається так:

$$t_{БСМ} = t_{СМ7-3} + 2t_{СМ3-2} + t_{СМ2-1}, \quad (26)$$

де $t_{СМ7-3}$ – час перетворення 7-рядного коду в 3-рядний; $t_{СМ3-2}$ – час перетворення 3-рядного коду в 2-рядний; $t_{СМ2-1}$ – час перетворення 2-рядного коду в 1-рядний код.

Вдосконалення методу синтезу паралельно-потоків пристроїв вертикального обчислення базових багатооперандних нейрооперацій

Паралельно-потоків пристроїв вертикального обчислення базових багатооперандних нейрооперацій повинні забезпечувати:

- обробку інтенсивних потоків даних у реальному часі;
- реалізацію на однотипних конвеєрних сходинах;
- високу ефективність використання обладнання та узгодження часу надходження даних з часом обчислення;
- адаптацію до конкретних застосувань і орієнтацію на НВІС-реалізацію;
- низьку вартість, малий час розробки та мінімальну кількість інтерфейсних виводів.

Ефективність таких пристроїв оцінюється критерієм $E_{ББН}$, який враховує складність алгоритму, кількість інтерфейсних виводів, однорідність структури та співвідношення часу виконання операції з витратами обладнання, що дозволяє кількісно оцінити продуктивність елементів пристрою. Кількісно величина ефективності використання обладнання для пристроїв обчислення базових багатооперандних нейрооперацій визначається так:

$$E_{ПББН} = \frac{R_{БНО}}{t_{БНО}(k_1 W_{ПББН} + k_2 Q)}, \quad (24)$$

де $E_{ПББН}$ – ефективність використання обладнання пристроєм обчислення базової багатооперандної нейрооперації; $R_{БНО}$ – складність алгоритму реалізації базової нейрооперації; $t_{БНО}$ – час виконання базової нейрооперації; $W_{ПББН}$ – витрати обладнання на реалізацію пристрою обчислення базових багатооперандних нейрооперацій; k_1 – коефіцієнт врахування однорідності структури; k_2 – коефіцієнт врахування кількості виводів зовнішнього інтерфейсу; Q – кількість виводів зовнішнього інтерфейсу.

Задача синтезу паралельно-потоків пристроїв вертикального обчислення базових багатооперандних нейрооперацій зводиться до забезпечення роботи у реальному часі при мінімальних апаратних затратах на їх реалізацію. Вихідними даними для синтезу паралельно-потоків пристроїв вертикального обчислення базових багатооперандних нейрооперацій є час t_d надходження масивів із N вхідних даних, який визначається так:

$$t_d = N t_{ТІ1}, \quad (25)$$

де $t_{ТІ1} F_d$ – тривалість періоду надходження вхідних даних. Для забезпечення роботи паралельно-потоків пристроїв вертикального обчислення базових багатооперандних нейрооперацій у реальному часі вимагається виконання наступної умови:

$$t_{ББН} > t_{ТІК}, \quad (26)$$

де $t_{ББН}$ – час обчислення багатооперандної базової нейрооперації. У конвеєрних паралельно-потоків пристроях, які опрацьовують потоки даних, час обчислення $t_{ББН}$ визначається тривалістю

конвеєрного такту, тобто $t_{\text{Б БН}} > t_{\text{Т Ік}}$, де $t_{\text{Т Ік}}$ - тривалість конвеєрного такту. Тривалість конвеєрного такту роботи $t_{\text{Т Ік}}$ залежить від швидкодії елементної бази, часу запису(читання) даних у буферні регістри та часу спрацювання операційного блоку в сходінці конвеєра. Збільшення кількості сходінок конвеєра веде до зменшення тривалості такту $t_{\text{Т Ік}}$ внаслідок зменшення складності операцій, які виконуються на сходінці конвеєра. При зменшенні кількості сходінок збільшується складність операцій, які виконуються на сходінці, а відповідно і збільшується тривалість такту. Апаратні витрати на реалізацію компонент зростають із збільшенням кількості сходінок конвеєра. Для вибору варіанту реалізації паралельно-потоків пристрою вертикального обчислення базових багатооперандних нейрооперацій у реальному часі будемо використовувати критерій ефективності використання обладнання $E_{\text{ПББН}}$. Висока ефективність використання обладнання $E_{\text{ПББН}}$ при реалізації паралельно-потоків пристрою обчислення базової багатооперандної нейрооперації у реальному часі досягається шляхом узгодження часу надходження даних t_d з тривалістю конвеєрного такту $t_{\text{Т Ік}}$. Таке узгодження може вимагати як збільшення, так і зменшення $t_{\text{Т Ік}}$.

Основними шляхами зменшення $t_{\text{Т Ік}}$ є:

- конвеєризація N -входового суматора шляхом розбиття його на сходінки;
- паралельне включення двох і більше пристроїв обчислення базових багатооперандних нейрооперацій, кількість яких визначається в основному часом t_d надходження вхідних даних.

У випадку коли час надходження даних t_d є значно більшим від часу $t_{\text{Т Ік}}$ то для забезпечення високої ефективності використання обладнання необхідно їх збільшити $t_{\text{Т Ік}}$. Основними шляхами збільшення $t_{\text{Т Ік}}$ є:

- використання алгоритмів, які зменшують кількість сходінок конвеєра (алгоритмів з формуванням групових часткових результатів);
- реалізація однією сходінкою конвеєра двох і більше ітерацій алгоритму.

Синтез паралельно-потоків пристроїв з вертикальним обчисленням базових багатооперандних нейрооперацій у реальному часі з заданим часом t_d надходження даних будемо виконувати з використанням розроблених базових структур пристроїв (рис.1, рис.2, рис.3 і рис.5). Синтез таких паралельно-потоків пристроїв вимагає виконання наступних етапів:

- 1) вибір базової паралельно-потоків структури для реалізації конкретного пристрою обчислення базової багатооперандної нейрооперації у реальному часі;
- 2) оцінювання тривалості такту $t_{\text{Т Ік}}$ конвеєра, затрат обладнання та ефективності використання обладнання для вибраної базової структури пристрою обчислення багатооперандної нейрооперації;
- 3) порівняння часу t_d з тривалістю такту $t_{\text{Т Ік}}$ конвеєра;
- 4) вибір шляхів узгодження часу t_d надходження вхідних даних з тривалістю такту $t_{\text{Т Ік}}$ конвеєра пристрою;
- 5) розроблення різних варіантів структур паралельно-потоків пристроїв з узгодженим часом обчислення ($t_d \geq t_{\text{Т Ік}}$) багатооперандної нейрооперації;
- 6) оцінювання ефективності використання обладнання для розроблених варіантів паралельно-потоків структур пристроїв з узгодженим часом обчислення багатооперандної нейрооперації;
- 7) вибір варіанту структури для реалізації паралельно-потоків пристрою обчислення багатооперандної нейрооперації шляхом порівняння значень ефективності використання обладнання;

Особливістю методів і паралельно-потоків структур пристроїв вертикального обчислення базових багатооперандних нейрооперацій (пошуку максимальних і мінімальних значень у одновимірному масиві даних, обчислення суми квадратів різниць, обчислення скалярного добутку, групового підсумовування), описаних у (Tsmots et al., 2017; Wu et al., 2022) є складність узгодження часу надходження даних з часом обчислення. Розроблені базові структури паралельно-потоків пристроїв з вертикальною обробкою даних забезпечують узгодження часу надходження вхідних даних з часом обчислення. Це може досягатися шляхом вибору алгоритмів, які зменшують кількість

сходинок конвеєра (для випадку коли $t_d > t_{Tlk}$), або конвеєризацією багатовходових суматорів (для випадку коли $t_d < t_{Tlk}$). Завдяки використанню послідовно-паралельних перетворювачів форматів даних забезпечується зменшення кількості виводів інтерфейсу зв'язку.

Недоліком дослідження можна вважати відсутність аналізу варіантів реалізації паралельно-потоківих пристроїв з вертикальним обчисленням базових багатооперандних нейрооперацій з прив'язкою до конкретних архітектури програмованих логічних інтегральних схем, наприклад, CPLD. Подальші дослідження буде спрямовано на розроблення алгоритмів і паралельно-потоківих структур для вертикального обчислення базових багатооперандних нейрооперацій з операндами з плаваючою комою. Отже, за результатами виконаної роботи можна сформулювати таку наукову новизну та практичну значущість результатів дослідження.

Наукова новизна: вдосконалено методи вертикального обчислення базових багатооперандних нейрооперацій (пошук екстремумів, сума квадратів різниць, скалярний добуток, групове підсумовування), що завдяки паралельній обробці розрядних зрізів та узгодженню часу надходження даних підвищують ефективність апаратної реалізації; удосконалено метод синтезу паралельно-потоківих пристроїв, який через узгодження тривалості конвеєрного такту з потоком даних забезпечує роботу в реальному часі.

Практична значущість: використання вдосконалених методів, базових структур (пошук екстремумів, сума квадратів різниць, скалярний добуток, багатовходові суматори) та методу синтезу забезпечує режим реального часу й високу ефективність апаратної реалізації багатооперандних нейрооперацій.

Висновки

Виділено операційний базис ШНМ, що включає групи нейрооперацій: попередньої обробки, процесорних та обчислення передатних функцій. Для апаратної реалізації вибрано базові багатооперандні нейрооперації: пошук максимуму й мінімуму в одновимірному масиві, обчислення суми квадратів різниць, скалярного добутку та групового підсумовування. Вдосконалено методи вертикального обчислення цих операцій, що за рахунок одночасної обробки розрядних зрізів та узгодження часу надходження даних з часом обчислення забезпечують високу ефективність обладнання. Запропоновано інтегрований підхід до розробки паралельно-потоківих пристроїв вертикального обчислення, який враховує можливості сучасної елементної бази, вертикальні методи та алгоритми, а також вимоги конкретних застосувань. Вибрано принципи побудови таких пристроїв: багатооперандність, модульність, конвеєризація, просторовий паралелізм, регулярність структури, узгодження часу обчислень з потоком даних, а також спеціалізація під конкретні задачі. Розроблено послідовно-паралельний перетворювач форматів даних та базові паралельно-потоківі структури, що апаратно реалізують вертикальні алгоритми. Вдосконалено метод синтезу пристроїв, який через узгодження тривалості конвеєрного такту з часом надходження даних забезпечує обробку в реальному часі при високій ефективності використання обладнання. Дослідження виконано в Національному університеті «Львівська політехніка» у рамках НДР «Методи та засоби інтелектуального вимірювання параметрів руху та визначення просторової орієнтації наземних мобільних робототехнічних платформ».

СПИСОК ЛІТЕРАТУРИ

- Albert Chun-Chen Liu, & Ming, O. (2021). Artificial Intelligence Hardware Design. John Wiley & Sons.
- Baranovsky S, Bilokobylskyi O, Ye, B., Bomba A, Dovbysh A, Zhokhin A, Yeroshenko T, Kazymyr V, Klymenko M, Kovalevskyy S, Kozlov O, Yu, K., Kupin A, Lande D, Malyarets L, Mincer O, Pankratova N, Pysarenko V, Ramazanov S, & Roskladka A. (2023). Strategy for Artificial Intelligence Development in Ukraine. https://doi.org/10.15407/development_strategy_2023
- Das, A., & Serrano-Gotarredona, T. (2024). Spike-based learning application for neuromorphic engineering. Frontiers Media SA.

- Davies, M., Srinivasa, N., Lin, T.-H., China, G., Cao, Y., Choday, S. H., Dimou, G., Joshi, P., Imam, N., Jain, S., Liao, Y., Lin, C.-K., Lines, A., Liu, R., Mathaikutty, D., McCoy, S., Paul, A., Tse, J., Venkataramanan, G., & Weng, Y.-H. (2018). Loihi: A Neuromorphic Manycore Processor with On-Chip Learning. *IEEE Micro*, 38(1), 82–99. <https://doi.org/10.1109/mm.2018.112130359>
- Gjorgjievska Perusheska, M., Dimitrova, V., Popovska-Mitrovikj, A., & Andonov, S. (2021). Application of Machine Learning in Cryptanalysis Concerning Algorithms from Symmetric Cryptography. *Lecture Notes in Networks and Systems*, 885–903. https://doi.org/10.1007/978-3-030-80129-8_59
- Guo, K., Zeng, S., Yu, J., Wang, Y., & Yang, H. (2018, December 6). A Survey of FPGA-Based Neural Network Accelerator. *ArXiv.org*. <https://doi.org/10.48550/arXiv.1712.08934>
- Hager, G., & Wellein, G. (2010). *Introduction to high performance computing for scientists and engineers*. CRC Press.
- Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, 234, 11–26. <https://doi.org/10.1016/j.neucom.2016.12.038>
- Muhammad Sarg, Khalil, A. H., & Mostafa, H. (2021). Efficient HLS Implementation for Convolutional Neural Networks Accelerator on an SoC. 1–4. <https://doi.org/10.1109/icm52667.2021.9664920>
- Nouacer, R., Hussein, M., Espinoza, H., Ouhammou, Y., Ladeira, M., & Castiñeira, R. (2020). Towards a framework of key technologies for drones. *Microprocessors and Microsystems*, 77, 103142. <https://doi.org/10.1016/j.micpro.2020.103142>
- Olaoye, G. (2025). Self-Learning Neural Networks in the Cloud: Towards Autonomous AI Systems. <https://doi.org/10.2139/ssrn.5129553>
- Panda, A. K., Palisetty, R., & Ray, K. C. (2020). High-Speed Area-Efficient VLSI Architecture of Three-Operand Binary Adder. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 67(11), 3944–3953. <https://doi.org/10.1109/tcsi.2020.3016275>
- Preethi, A. P. (2021). Secure Data Communication Using Diffie Rivest Cryptography Algorithm. *Neural, Parallel, and Scientific Computations*, 29(2). <https://doi.org/10.46719/npsc20212923>
- Saini, S., Lata, K., & Sinha, G. R. (2021). *VLSI and Hardware Implementations using Modern Machine Learning Methods*. CRC Press.
- Sze, V., Chen, Y.-H., Emer, J., Suleiman, A., & Zhang, Z. (2017, April 1). Hardware for machine learning: Challenges and opportunities. *IEEE Xplore*. <https://doi.org/10.1109/CICC.2017.7993626>
- Tsmots, I. G., Tesliuk, V. M., & Opotyak, Y. V. (2024). Device for determining maximum and minimum numbers in a two-dimensional array of numbers: Patent for invention No. 128150 Ukraine. Published 14.04.2024, Bulletin No. 16.
- Tsmots, I. G., Tesliuk, V. M., Lukashchuk, Yu. A., & Kazymyra, I. Ya. (2023). Device for calculating the scalar product: Patent for invention No. 127774 Ukraine. IPC G06G 6/33, Application No. u202010852; Filed 19.05.2020; Published 27.12.2023, Bulletin No. 52.
- Tsmots, I. H., & Antoniv, V. A. (2022). METHODS AND TOOLS FOR VERTICAL-PARALLEL SEARCHING OF MAXIMUM AND MINIMUM NUMBERS IN ARRAYS. *Ukrainian Journal of Information Technology*, 4(1), 68–77. <https://doi.org/10.23939/ujit2022.01.068>
- Tsmots, I. H., Lukashchuk, Yu. A., Ihnatyev, I. V., & Kazymyra, I. Ya. (2021). COMPONENTS OF HARDWARE NEURAL NETWORKS FOR COORDINATED PARALLEL-VERTICAL DATA PROCESSING IN REAL TIME. *Ukrainian Journal of Information Technology*, 3(1), 63–72. <https://doi.org/10.23939/ujit2021.03.063>
- 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>
- Tsmots, I., Skorokhoda, O., Ignatyev, I., & Rabyk, V. (2017). Basic vertical-parallel real time neural network components. 2017 12th International Scientific and Technical Conference on Computer Sciences and Information Technologies (CSIT), 344–347. <https://doi.org/10.1109/stc-csit.2017.8098801>
- Usha, S., & Kanthimathi, M. (2024). Design and Comparison of 24-bit Three Operand Adders using Parallel Prefix method for Efficient Computations. *ICST Transactions on Scalable Information Systems*. <https://doi.org/10.4108/eetsis.5004>
- Wang, G., & Fu, D. (2024). Spike Neural Network with Delayed Propagation Characteristics and Hardware Implementation. 1181–1185. <https://doi.org/10.1109/eei63073.2024.10696338>

- Wu, J., Zhao, B., Wen, H., & Zhao, Q. (2022). Design of Neural Network Accelerator Based on In-Memory Computing Theory. 2022 4th International Conference on Natural Language Processing (ICNLP), 547–551. <https://doi.org/10.1109/icnlp55136.2022.00100>
- Xian, A., Martini, B., & Culurciello, E. (2015). Recurrent Neural Networks Hardware Implementation on FPGA. ArXiv (Cornell University). <https://doi.org/10.48550/arxiv.1511.05552>
- Zhang, C., Patras, P., & Haddadi, H. (2019). Deep Learning in Mobile and Wireless Networking: A Survey. IEEE Communications Surveys & Tutorials, 21(3), 2224–2287. <https://doi.org/10.1109/comst.2019.2904897>

SYNTHESIS OF PARALLEL-PIPELINE DEVICES FOR VERTICAL COMPUTATION OF BASIC MULTI-OPERAND NEURAL OPERATIONS

Ivan Tsmots¹, Bohdan Shtohrinets²

^{1,2} Lviv Polytechnic National University,
Department of Automated Control Systems, Lviv, Ukraine.

¹ E-mail: ivan.h.tsmots@lpnu.ua, ORCID: 0000-0002-4033-8618

² E-mail: bohdan.v.shtohrinets@lpnu.ua, ORCID: 0009-0001-4956-3862

© Tsmots I., Shtohrinets B., 2025

This paper presents a study of hardware implementation of basic multi-operand neural operations for artificial neural networks. The operational basis of neural networks is identified, which includes groups of preprocessing operations, processor operations, and transfer function computations. The selection of basic multi-operand neural operations is justified: finding extreme values in one-dimensional arrays, calculating the sum of squared differences, scalar product computation, and group summation. Vertical computation methods for the specified operations have been improved through simultaneous processing of bit slices of all operands and adaptive complexity changes of operations in pipeline stages. This ensures synchronization of data arrival time with computation time and high equipment utilization efficiency. An integrated approach to developing parallel-pipeline devices is proposed, which is based on the capabilities of modern element base and considers the requirements of specific applications. The principles of developing parallel-pipeline devices for vertical-group computation have been defined: using the basis of elementary arithmetic operations, modularity, pipelining, spatial parallelism, structural uniformity, timing parameter coordination, and specialization for specific tasks. A serial-parallel data format converter has been developed for transforming the stream of word-sequential input data into a parallel one-dimensional data array. Basic parallel-pipeline structures have been created that implement vertical computation algorithms in hardware. The synthesis method for parallel-pipeline devices has been improved using mechanisms for coordinating pipeline cycle duration with data arrival time. It is shown that the application of the developed methods and structures ensures real-time data processing with high equipment utilization efficiency.

Keywords: multi-operand operations, parallel-pipeline devices, vertical data processing, real-time, equipment utilization efficiency, artificial neural networks, pipeline processing, hardware implementation.