Архітектура високопродуктивного нейрокомп'ютера для розв'язання задач комбінаторної оптимізації

Потебня А.В., Погорілий С.Д.

Київський національний університет імені Тараса Шевченка, просп. Глушкова, 4-г, Київ, Україна

poteba@yandex.ru, sdp@univ.net.ua

Анотація. *Запропоновано архітектуру новітнього нейрокомп'ютера для ефективного розв'язання задач комбінаторної оптимізації в режимі жорсткого реального часу. Висока продуктивність, компактність, надійність та енергоефективність нової платформи досягається за рахунок відтворення основних властивостей біологічних аналогів та врахування досвіду попередніх розробок. Наведено алгоритм адаптації запропонованої системи до розв'язання NP-складної задачі пошуку максимальної кліки. Сформовано підхід до масштабування нейрокомп'ютера шляхом його складання з блоків меншої розмірності. Показано, що продуктивність такої системи є сумірною з потужністю сучасних суперкомп'ютерів.*

Ключові слова

Задача комбінаторної оптимізації, визначення найбільшої кліки, нейрокомп'ютер, нейронна мережа Хопфілда, асоціативна пам'ять, NP-складні задачі, високопродуктивні обчислення, граф.

1 Вступ

Надзвичайно поширеними у багатьох сферах сучасних наукових досліджень є задачі комбінаторної оптимізації (ЗКО, від англ. *combinatorial optimization problem*), які потребують вибору оптимального розв'язку серед заданого набору варіантів [1]. До найважливіших з них належать задача визначення найбільшої кліки (*maximum clique problem*), задача комівояжера (*travelling salesman problem*), проблема пошуку мінімального дерева Штейнера (*minimum Steiner tree*), задача розбиття графів (*graph partition problem*) та багато інших. Побудова штучних молекулярних комплексів, розробка складних комп'ютерних мереж та, навіть, процес пошуку їжі тваринами, об'єднані потребою розв'язання задач комбінаторної оптимізації [2, 3].

Наприклад, при розробці надвеликих інтегральних схем (VLSI), виникає необхідність з'єднати їх компоненти у такий спосіб, щоб одержаний пристрій був компактним, споживав мало потужності та швидко проводив сигнали [3]. Подібні вимоги враховуються при проектуванні топології локальних мереж, визначенні розподілу задач між процесорними елементами, а також при синтезі ланцюжків ДНК та формуванні запитів до баз даних [4, 5].

Разом з тим, більшість задач комбінаторної оптимізації належать до класу NP-складних. Обчислювальні затрати на їх розв'язання експоненційно зростають зі збільшенням розмірності оброблюваних графів [6]. У зв'язку з цим виникає необхідність розробки новітніх методів ефективного розподілу ресурсів, що використовуються для пошуку оптимального розв'язку. Як наслідок, протягом останніх десятиріч було запропоновано низку програмних реалізацій штучних нейронних мереж, призначених для виконання на традиційній комп'ютерній архітектурі. Однак, застосування таких систем для розв'язання ЗКО пов'язано з втратою масового паралелізму, який є запорукою їх обчислювальної потужності. Крім того, при розробці медичних, авіаційних, військових та інших спеціалізованих систем, висуваються вимоги до розв'язання ЗКО в режимі жорсткого реального часу, порушення яких може призвести до катастрофічних наслідків [7].

Метою роботи є формування високопродуктивної апаратної платформи штучної нейронної мережі, придатної для розв'язання широкого класу задач оптимізації та розпізнавання образів. Запропонована модель нейрокомп'ютера, який містить 4 процесорні елементи, та сформовані рекомендації щодо його масштабування для розв'язання задач великої розмірності. До переваг розробленої системи належать значна швидкодія, компактність, енергоефективність та можливість роботи в режимі жорсткого реального часу.

2 Переваги апаратної реалізації штучних нейронних мереж

Останні десятиріччя пов'язані з бурхливими дослідженнями архітектур високопродуктивних комп'ютерних систем на основі штучних нейронних мереж. Придатні до застосування в умовах масового паралелізму, вони виявляються спроможними до ефективного розв'язання задач комбінаторної оптимізації, розпізнавання образів, стискання потокової відеоінформації та багатьох інших. Висока швидкодія, низьке споживання енергії та здатність до миттєвої обробки значних наборів даних стали запорукою широкого використання нейрокомп'ютерів (*neurocomputer*) для розв'язання цих задач [8, 9].

Однак, найбільш поширені програмні реалізації нейронних мереж позбавлені наведених переваг, що пов'язано з послідовною роботою класичної архітектури фон Неймана (*von Neumann*). Обчислювальна здатність таких систем суттєво знижується за рахунок можливості реалізації лише неявного (*implicit*) паралелізму [8]. У зв'язку з цим, виникає необхідність розробки їх апаратних реалізацій, придатних для впровадження явного (*explicit*) паралелізму та відповідного збільшення продуктивності. Наприклад, нині запропоновані реалізації клітинних нейронних мереж (*cellular neural network*) за надвисокого ступеня інтеграції (*VLSI*), які демонструють значення продуктивності на рівні декількох терафлопс. Крім того, такі системи мають високу «живучість», що пов'язано зі збереженням їх працездатності за пошкодження деякої частини нейронів.

Разом з тим, висока складність топології штучних нейронних мереж є значною завадою на шляху їх апаратної реалізації. Похибки, притаманні аналоговим та цифровим компонентам, можуть призвести до деградації процесу навчання, збільшення кількості необхідних циклів та формування помилкових результатів. Проектування системи додатково ускладнюється за рахунок нелінійності активаційної функції нейронів, відтворення якої є складною схемотехнічною задачею. Шляхи її розв'язання у сучасних нейрокомп'ютерах пов'язані з використанням спеціальних пошукових таблиць (*look-up table*) та кусково-лінійних апроксимованих функцій [9]. Тому застосування апаратних реалізацій штучних нейронних мереж для розв'язання прикладних задач залишається обмеженим.

3 Основні способи формування нейрокомп'ютерів

У наукових працях запропоновані найрізноманітніші підходи до класифікації апаратних моделей нейрокомп'ютерів. Ця задача суттєво ускладнюється за рахунок широкого різноманіття можливих топологій, активаційних функцій та алгоритмів навчання, що доповнюється параметрами технічної реалізації. У загальному випадку виділення окремих таксонів потребує врахування наступних параметрів:

- тип сигналів, що використовуються в мережі;
- метод реалізації та збереження вагових коефіцієнтів;
- вигляд активаційної функції нейронів.

Рис. 1. Класифікація апаратних реалізацій штучних нейронних мереж

Сигнали, що застосовуються для зв'язку між окремими нейронами системи, можуть бути представленими у вигляді величини струму або різниці потенціалів (аналоговий випадок). При цьому для реалізації вагових коефіцієнтів переважно застосовуються матриці резисторів або транзисторів [10]. Інші системи використовують цифровий підхід, за якого їх внутрішні сигнали можуть набувати лише деяких дискретних

значень. Реалізація вагових коефіцієнтів при цьому потребує застосування спеціальних компонентів цифрової схемотехніки (помножувачів, суматорів і т. п.). Для опису наведених систем використовуються загальні таксони аналогових та цифрових нейрокомп'ютерів. Окремо також слід виділити гібридні системи, утворені шляхом сполучення аналогових та цифрових схем (рис. 1).

До аналогового таксону можуть бути віднесені електричні нейросистеми та їх оптичні реалізації. Цифровий таксон охоплює цілу низку нейрокомп'ютерів з різноманітною внутрішньою архітектурою, зокрема, мультипроцесорні та векторні системи, оптимізовані для виконання матричних обчислень. Високу ефективність демонструють реалізації нейронних мереж у вигляді систолічних матриць (*systolic array*), утворені шляхом сполучення процесорних елементів з забезпеченням взаємодії між сусідніми вузлами системи. До цифрових нейрокомп'ютерів також належать системи суперскалярного (*superscalar*) типу, багатопроцесорні системи класу SIMD (*Single Instruction, Multiple Data*) та багато інших [11, 12].

У системах з імпульсним кодуванням (*pulse coded network*) застосовується підхід до представлення інформаційного сигналу у вигляді набору імпульсів з заданою частотою, коефіцієнтом заповнення та амплітудою [8]. При цьому можливість кодування аналогових даних цифровими імпульсами поєднується зі значною складністю представлення від'ємних чисел та проблемами реалізації основних арифметичних операцій.

Разом з тим, наведені таксони є надто загальними, що спричинює непридатність їх використання для характеристики апаратних реалізацій нейрокомп'ютерів. Тому, доцільним є введення складних таксонів вигляду $X_1X_2X_3$, де X_1 описує тип вхідних і вихідних сигналів, X_2 задає вигляд активаційної функції нейронів, а X_3 визначає представлення вагових коефіцієнтів. Такий підхід призводить до породження 8 класів нейрокомп'ютерів, які відрізняються способами реалізації їх складових частин. При цьому тип ААА відповідає аналоговим системам, ЦЦЦ – цифровим, а всі інші можливі варіанти є гібридними. Наприклад, таксон ААЦ об'єднує нейрокомп'ютери з аналоговими входами та виходами, неперервною активаційною функцією і цифровим форматом вагових коефіцієнтів.

Таблиця 1. Критерії продуктивності нейрокомп'ютерів

Параметр продуктивності	***Характеристики нейрокомп'ютера, що визначаються параметром***
Кількість оновлень зв'язків у секунду (*connection updates per second, CUPS*)	Швидкість розрахунку вагових коефіцієнтів та, відповідно, ефективність навчання нейронної мережі
Кількість зв'язків в секунду (*connections per second, CPS*)	Швидкість обчислення вхідних сигналів нейронів на основі розрахованих вагових коефіцієнтів
Енергія синапсів (*watt per connection per second, WCPS*)	Енергетичні витрати системи (вимірюються у ватах за секунду для кожного зв'язку)

У нейронних мережах застосовуються різні способи організації зв'язків між обчислювальними вузлами. Серед них виділяють повністю зв'язані топології (*fully connected*), системи з локальними зв'язками (*locally connected*), шаруваті (*layered*) структури і т. д. Крім того, поруч з фіксованими реалізаціями існують системи, які надають можливість зміни конфігурації міжнейронних зв'язків.

За ступенем паралелізму у наукових працях виділяють наступні класи нейрокомп'ютерів:

- системи з слабким паралелізмом (менше 10 – 15 процесорів);
- системи з середнім паралелізмом (10 – 100 процесорів);
- системи з сильним паралелізмом (100 – 1000 процесорів);
- системи з масовим паралелізмом (понад 1000 процесорів).

Найпоширеніший набір критеріїв, що використовуються для визначення ефективності нейросистем, міститься в таблиці 1. Відомо, що аналогові реалізації звичайно виявляються значно більш продуктивними, компактними та енергоефективними, оскільки сама природа обробки інформації в біологічних нейронних мережах є аналоговою [9]. Однак, цифрові системи надають більш високу точність обчислень та можливість використання віртуальних мереж (*virtual network*) з гнучким відображенням на апаратні ресурси. У зв'язку з цим, вибір цифрової або аналогової елементної бази для реалізації нейрокомп'ютерів визначається критичними параметрами кожної окремої задачі (таблиця 2).

Таблиця 2. Порівняння типів елементної бази для реалізації нейрокомп'ютерів

Тип елементної бази	*Переваги*	*Недоліки*
Аналогова	Висока швидкодія, енергоефективність та компактність	Високі технологічні вимоги, низька точність обчислень, чутливість до зовнішнього впливу, складність реалізації масових з'єднань, вузька спеціалізація
Цифрова	Висока точність, можливість програмування та стійкість до технологічних варіацій	Складність схемотехнічних рішень, менша швидкодія за рахунок багататактового виконання базових операцій

Регулярність структури штучних нейронних мереж надає широкі можливості для їх організації у вигляді спеціальних мікросхем штучного інтелекту (*neurochip*). Зокрема, одним з найпоширеніших є аналоговий модуль *Intel 8017NW ETANN* (*Electrically Trainable Analog Neural Network*), виконаний у вигляді великої інтегральної мікросхеми, що містить 64 повністю зв'язаних нейрони та 10240 синапсів. Мікросхема не містить вбудованих засобів для навчання нейронної мережі та працює на основі завантажених значень вагових коефіцієнтів, розрахованих віддаленою системою. Структура синапсу мікросхеми забезпечує множення вхідного сигналу з врахуванням знаку на значення вагового коефіцієнта, збережене в пам'яті. Значення добутків незалежно підсумовуються за допомогою аналогового суматора, вихід якого сполучений з входом пристрою, який реалізує активаційну функцію з програмованим нахилом. Обчислювальна потужність мікросхеми складає 2 GCPS, що є сумірним з потужністю сучасних суперкомп'ютерів [13].

Таблиця 3. Порівняння мікросхем штучних нейронних мереж

Назва схеми	*Виробник*	*Тип*	*Засоби для навчання*	*Блок синапсів*	*Кількість нейронів*	*GCPS*
ETANN	Intel	ААЦ	-	+	64	2,0
ANNA	AT&T Bell Labs	ЦАА	-	+	16 – 256	2,1
Neuro-Classifier	Mesa Research	ААЦ	-	+	6	21
N64000	Inova	ЦЦЦ	+	+	64	0,871
MA-16	Siemens	ЦЦЦ	-	-	16	0,4
RN-200	Ricoh	ЦЦЦ	+	+	7	3,0
NLX - 420	NeuraLogix	ЦЦЦ	-	-	16	3×10^{-7}
100 - NAP	HNC	ЦЦЦ	+	-	100	0,25
MT19003	Micro Circuit Engineering	ЦЦЦ	-	-	8	0,032
L-Neuro 1	Philips	ЦЦЦ	-	+	16	0,026

У таблиці 3 наведено порівняння розроблених мікросхем штучних нейронних мереж. Деякі з них не містять блоків для збереження вагових коефіцієнтів, що пов'язано з потребою використання вбудованих систем пам'яті та відповідним збільшенням енерговитрат [10]. Крім того, більшість мікросхем потребують залучення віддаленої системи для здійснення процедури навчання у зв'язку з відсутністю засобів для самостійного оновлення значень синапсів.

Складні нейрокомп'ютери утворюються шляхом сполучення базових мікросхем. Наприклад, система *Mod2* складається з 12 блоків *ETANN* та застосовується для обробки зображень в режимі реального часу. Цифровий нейрокомп'ютер *CNAPS* (*Connected Network of Adaptive Processors*) формується шляхом приєднання базових нейроблоків *N6400* (утворених з 64 нейронів) до загальних 8-бітних шин введення та виведення інформації [14].

Разом з тим, до складу людського мозку входить близько 10^{10} нейронів, а його продуктивність досягає 10^{16} CPS, що в мільйон разів перевищує можливості найефективніших нейрокомп'ютерів. Тому важливою є задача формування новітніх систем для досягнення вищих рівнів обчислювальної потужності. Найбільш перспективними, зокрема, виявляються швидкісні аналогові, оптичні та молекулярні реалізації. Тому в цій

роботі для формування архітектури нейрокомп'ютера використовується аналогова елементна база, а кодування інформаційних сигналів відбувається за допомогою значень напруги.

4 Розробка схеми аналогового нейрона

Відомо, що нейрон є найпростішим обчислювальним елементом, який здійснює формування вихідного сигналу на основі вхідних даних, отриманих від інших вузлів системи. При цьому біологічні нейрони мають типову структуру, яка складається з входів (дендритів), тіла клітини, необхідного для підтримки її життєдіяльності, та єдиного вихідного відростка – аксона. Сформований електричний імпульс, проходячи вздовж аксона, досягає міжнейронного синаптичного простору, в якому він спричинює вивільнення нейромедіаторів (*neurotransmitter*). Дифундуючи крізь проміжну щілину, молекули медіаторів реагують з рецепторними білками клітинної мембрани та призводять до зміни потенціалу нейрона шляхом активації механізмів йонного транспорту [15].

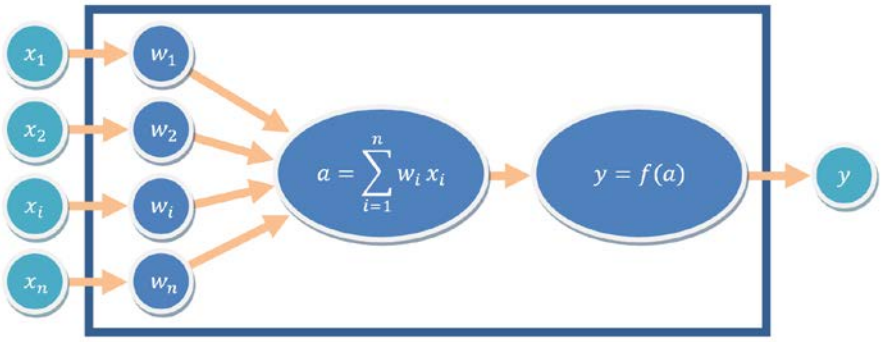

Рис. 2. Найпростіша модель штучного нейрона

Такий тип взаємодії може бути відтворений у електричній еквівалентній схемі з використанням джерел живлення та резисторів. При цьому джерела живлення відповідають збереженому потенціалу, а резистори визначають кількість йонів, що беруть участь у транспортних механізмах. Ємнісний ефект клітинної мембрани моделюється шляхом введення RC-комірок, які забезпечують усунення різких змін напруги та відтворення інерційних властивостей біологічних аналогів.

Структуру найпростішої нейронної моделі МакКаллока-Піттса (*McCulloch-Pitts*), що широко використовується при розробці систем штучного інтелекту, наведено на рис. 2. Формування вихідного імпульсу y при цьому відбувається шляхом застосування активаційної функції f до суми вхідних сигналів $x_1, x_2, ..., x_n$, помножених на відповідні вагові коефіцієнти $w_1, w_2, ..., w_n$. Вибір вихідної функції нейронів f має значний вплив на якість запропонованих розв'язків. Залежно від її вигляду часто виділяють дискретні (*discrete*) та неперервні (*continuous*) штучні нейронні мережі. Складові нейрони дискретних систем мають біполярні активаційні функції на відміну від неперервного випадку, в якому вихідні залежності є більш складними для реалізації [9].

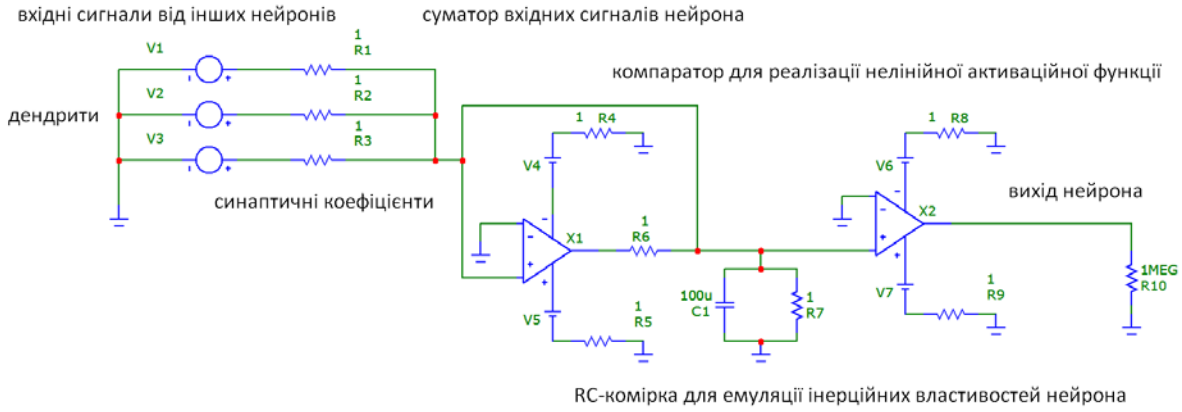

Рис. 3. Схема аналогового нейрона, розроблена в середовищі *Micro-Cap*

При розробці схеми аналогового нейрона виникає потреба реалізації компонентів наведеної формальної моделі. Для підсумовування вхідних сигналів використовується схема на основі операційного підсилювача.

Нульовий потенціал на його входах забезпечує усунення взаємних впливів та, як наслідок, надає широкі можливості для масштабування системи. Важливим є застосування RC-комірки, параметри якої визначають час реакції нейрона на вхідне збудження. Для реалізації біполярної активаційної функції застосовується схема аналогового компаратора. Розроблена на основі цих міркувань модель аналогового нейрона міститься на рис. 3.

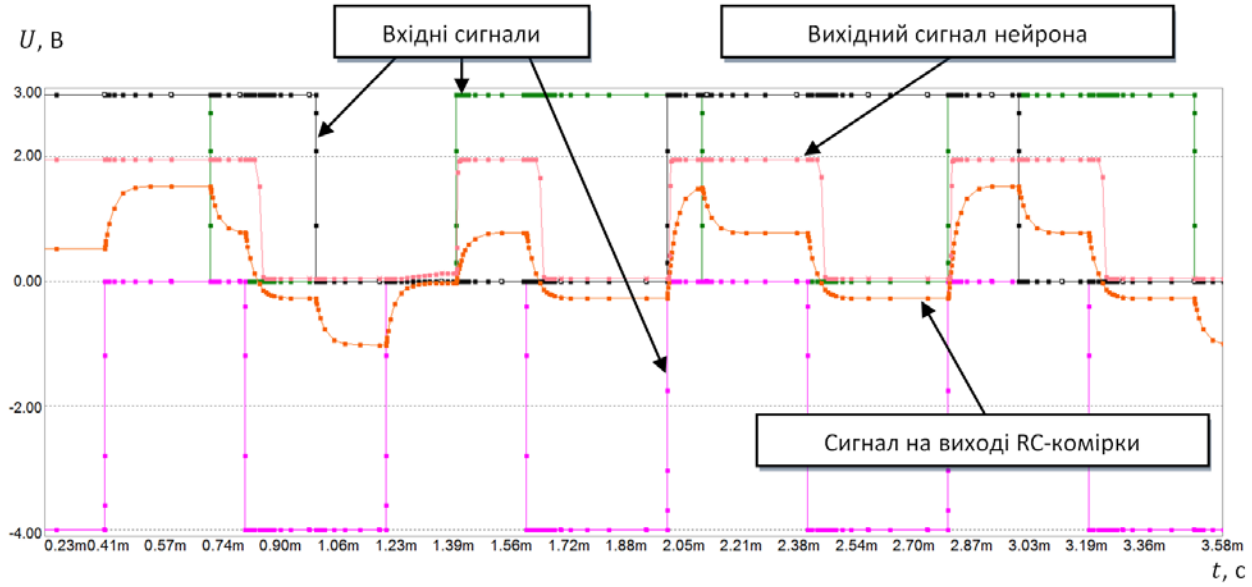

Рис. 4. Часові діаграми вхідних та вихідних сигналів розробленого нейрона

На рис. 4 наведено часові діаграми для демонстрації роботи запропонованого пристрою. Видно, що його вихідні імпульси формуються, якщо результат операції підсумовування вхідних сигналів є додатнім. При цьому процеси заряджання та розряджання конденсатора призводять до затримки реакції пристрою на зміну вхідних сигналів. Таким чином, розроблена схема відтворює основні властивості нейрона та може бути застосована для формування аналогових нейронних мереж.

5 Формування асоціативної пам'яті нейронної мережі

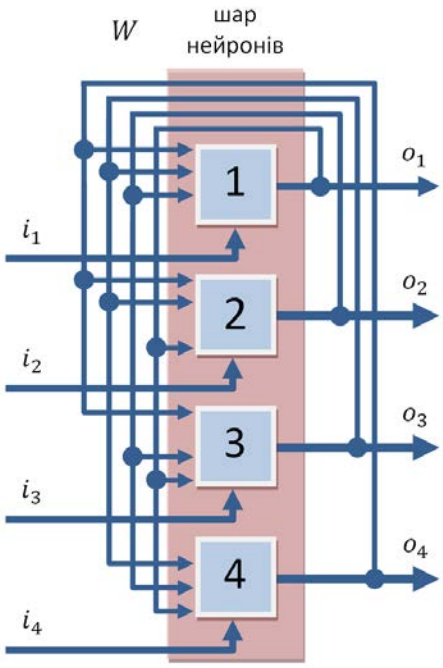

Рис. 5. Архітектура нейронної мережі Хопфілда

Надзвичайно потужним є механізм асоціативної (*associative*) пам'яті, який надає можливості для ефективного розв'язання багатьох задач за допомогою ланцюжків асоціацій, що є недосяжним для традиційних комп'ютерів. Такі системи реалізують відображення F множини вхідних сигналів X у множину результатів Y відповідно до правил, встановлених на етапі навчання. При цьому до матриці асоціативної пам'яті W записується набір шаблонів (*patterns*) у вигляді векторів $A_1, A_2, ..., A_p$, а робота нейронної мережі зводиться до вибору одного з них на основі заданої множини сигналів. Взаємозв'язки між векторами утворюються за рахунок збереження у пам'яті W пар асоціацій $(A_1, B_1), (A_2, B_2), ..., (A_p, B_p)$ [16].

У загальному випадку такі системи є гетероасоціативними (*hetero-associative*), тобто придатними до формування складних зв'язків між різними векторами. Натомість можливості автоасоціативної пам'яті (*auto-associative*) обмежені відтворенням окремих шаблонів за їх компонентами. У цьому випадку збережені пари мають вигляд $(A_1, A_1), (A_2, A_2), ..., (A_p, A_p)$.

На основі вхідного набору сигналів A шляхом множення AW та застосування активаційної функції здійснюється формування вихідного вектора A'. За допомогою рекурентних зв'язків отриманий результат може бути повторно надісланий до входів матриці W з утворенням наступного вектора A''. Стабільні системи мають закінчити цю послідовність перетворень формуванням

стійкого набору A_f, який є одним зі збережених шаблонів $A_1,...,A_p$. Таким чином, процес знаходження розв'язку A_f може бути описаний ланцюгом надходжень даних до пам'яті: $A \to W \to A' \to W \to A'' \to W \to ... \to A_f \to W \to A_f \to ...$.

Найпростішою є статична схема лінійної асоціативної пам'яті (*linear associative memory*), яка реалізує найкоротшу послідовність перетворень $A \to W \to B$. Архітектура системи складається з m входів та n виходів, сполучених матрицею вагових коефіцієнтів w_{ij}, які визначають пари асоціацій $\{(A_k, B_k) \mid k = 1,2,..., p\}$. Найвідомішою апаратною реалізацією такої пам'яті є система *Lernmatrix*, в якій значення w_{ij} задаються за допомогою резисторів. Однак, більш продуктивними виявляються динамічні структури, побудовані шляхом введення зворотних зв'язків з відповідним ускладненням еволюції векторів [8].

У цій роботі для реалізації асоціативної пам'яті використовується модель Хопфілда (*Hopfield model*), яка здійснює рекурсивне формування вихідних розв'язків до досягнення стабільного вектора A_f. Архітектура такої системи (рис. 5) має вигляд сильно зв'язаного набору процесорних елементів [15]. Для забезпечення стійкості до матриці з'єднань W висуваються вимоги симетрії $w_{ij} = w_{ji}$ та рівності нулю діагональних елементів $w_{ii} = 0$. Крім того, кожен вузол має додатковий вхід, призначений для надходження вхідних векторів $X(t)$.

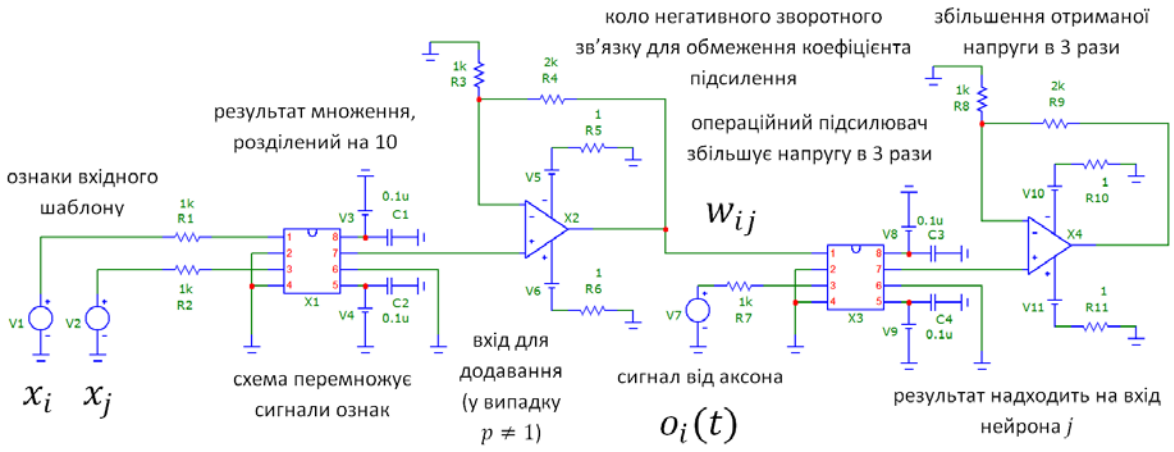

Рис. 6. Схема вузла асоціативної пам'яті

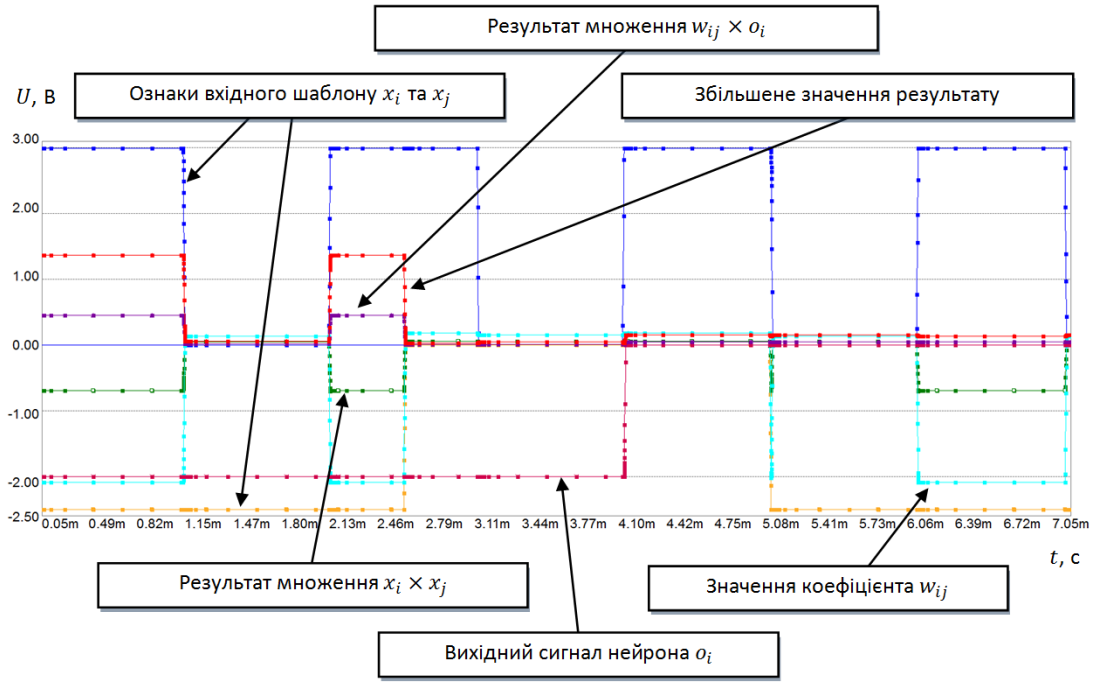

Рис. 7. Часові діаграми роботи вузла асоціативної пам'яті

Процедура навчання такої системи зводиться до обчислення матриці W_k для кожного записаного шаблону A_k. При цьому у вузлах $(w_{ij})_k$ зберігаються пари асоціацій у вигляді добутків компонентів вектора $(a_i)_k \times (a_j)_k$. Шляхом підсумовування матриць W_k формується вміст пам'яті для p векторів $W = \alpha \sum_{k=1}^{p} W_k$, де α – коефіцієнт, необхідний для захисту синапсів від перевантажень.

Слід зазначити, що організація вузлів пам'яті у вигляді резисторів значно обмежує можливості системи у зв'язку з відсутністю засобів для запису та оновлення шаблонів. Тому, в цій роботі значення асоціацій формуються зовнішнім хостом динамічно з використанням аналогових елементів множення AD633J. Схема розробленого у такий спосіб вузла пам'яті міститься на рис. 6. Вона містить частини для множення компонентів вектора $a_i \times a_j$, підсилення отриманого результату та врахування вихідного сигналу нейрона o_i. Часові діаграми для такого елемента пам'яті наведені на рис. 7.

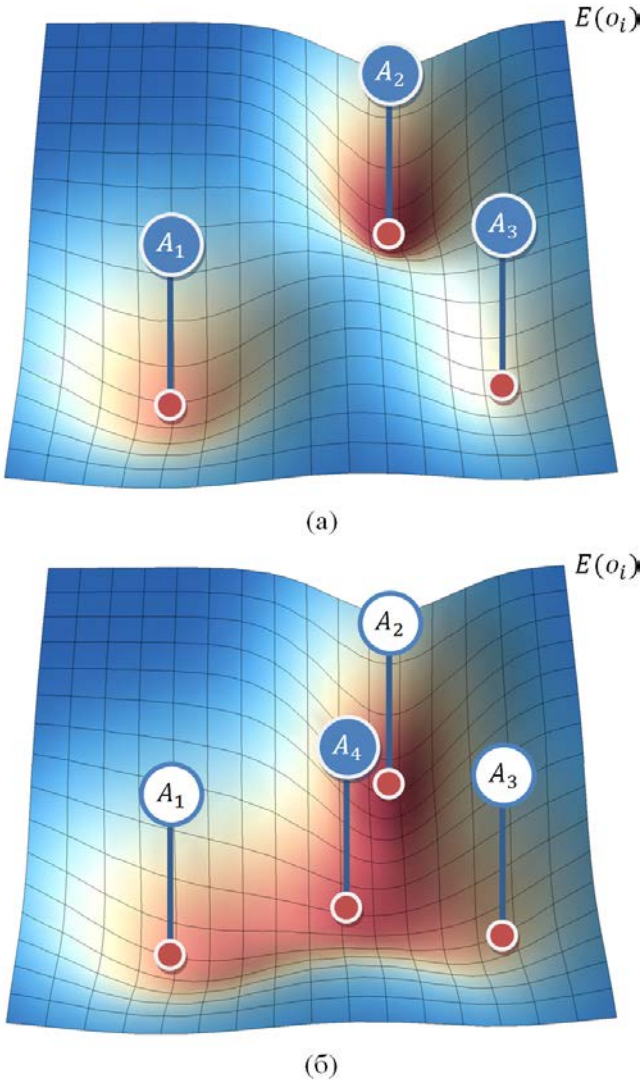

Рис. 8. Приклади енергетичного ландшафту системи

Процедура пошуку збереженого вектора може виконуватися з синхронним (паралельним) або асинхронним (послідовним) оновленням виходів o_i. Вибір схеми оновлення має значний вплив на продуктивність системи та її стійкість. У цій роботі застосовується синхронна схема, за якої дослідження всіх компонентів поточного вектора відбувається одночасно. При цьому ітераційна процедура формування розв'язків описується виразом

$$o_j(t+1) = f\left(\sum_{i=1}^{m} w_{ij} o_i(t)\right),$$

де $o_j(t)$ – стан відповідного нейрона в момент часу t, f – активаційна функція, m – розмірність шару процесорних елементів.

Для дослідження властивостей запропонованої системи застосовується спеціальна енергетична функція (*energy function*) [16], яка має вигляд:

$$E = -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} o_i w_{ij} o_j + \sum_{i=1}^{m} o_i \theta_i,$$

де θ_i – значення рівня активації i-ого нейрона. Локальні мінімуми цієї функції відповідають збереженим у пам'яті шаблонам $A_1, A_2, ..., A_p$, а процес роботи системи представлений їх пошуком. Запис додаткових векторів потребує введення відповідних мінімумів, що може призвести до спотворення ландшафту та формування паразитних станів системи.

На рис. 8а наведено приклад структури енергетичного ландшафту для асоціативної пам'яті, що містить три вектори A_1, A_2 та A_3. Кожному з них відповідає окремий локальний екстремум, що є необхідною умовою для утворення асоціацій. Однак, додавання шаблону A_4 (рис. 8б) призводить до порушення структури ландшафту та втрати екстремумів для деяких векторів. За таких умов застосування локального спуску може призвести до формування хибних розв'язків. Тому ємність асоціативної пам'яті обмежується розмірністю мереж і становить близько $0{,}15m$ [17].

Розглянемо приклад застосування розробленого нейрокомп'ютера для розпізнавання спотворених образів. Нехай, у пам'яті системи міститься шаблон $A = \{1,0,0,1\}$. Для кодування логічних нулів та одиниць

використовуються відповідні рівні напруг 0 та +5В. При цьому вміст асоціативної пам'яті W, утворений шляхом перемноження компонентів вектора A є наступним:

$$W = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

Відповідні значення напруги у вузлах матриці визначаються обраними коефіцієнтами підсилення. При цьому їх значення мають бути достатніми для розрізнення логічних нулів та одиниць. За обраних у цій роботі параметрів логічним нулям у вузлах матриці відповідає напруга близько 200 мВ, а логічним одиницям – 7 В.

Подамо на входи системи спотворений образ $X_0 = \{1,1,0,1\}$. При цьому до нейронів надходять сигнали $\{1,0,0,1\}$, а перемикання комутаторів спричинює миттєве формування на їх виходах відновленого вектора A. Таким чином, для усунення спотворень знадобилася лише одна ітерація. У зв'язку з переходом до стійкого стану отримані значення на виході нейрокомп'ютера залишаються незмінними.

Запропонована система містить додатковий шар комутаторів MAX4714, що забезпечують надходження вхідних сигналів до пам'яті та ініціюють процес утворення результату шляхом одночасного перемикання. Розроблена на основі цих міркувань функціональна схема нейрокомп'ютера наведена на рис. 9.

6 Розв'язання задач комбінаторної оптимізації за допомогою розробленого нейрокомп'ютера

Високий ступінь розпаралелювання при обробці інформації є запорукою ефективного використання нейрокомп'ютерів для розв'язання багатьох ЗКО. Крім того, властивість відшукання мінімумів енергетичної функції, притаманна матрицям асоціативної пам'яті, надає широкі можливості для їх адаптації до розв'язання задач оптимізації. Однак, запропоновані алгоритми повинні містити реалізації наступних етапів, специфічних для розробленої архітектури:

1. Кодування матриці асоціативної пам'яті на основі умов задачі для формування необхідних екстремумів енергетичної функції.
2. Вибір вхідного вектора X_0, який має значний вплив на процес пошуку.
3. Кодування отриманих результатів у вигляді векторів вихідних сигналів нейронів. У такий спосіб мають бути представлені цикли комівояжера, графи і т. д.

Розглянемо приклад адаптації розробленої системи до розв'язання NP-складної задачі знаходження найбільшої кліки (*clique*). Відомо, що клікою в неорієнтованому графі $G = (V, E)$ називається довільна підмножина $C \subseteq V$, в якій кожна пара вершин сполучена ребром графа [2]. Для відстеження найбільшої кліки завантажимо до пам'яті нейрокомп'ютера матрицю суміжності заданого графа. Оберемо кодування вихідних векторів, за якого вузли, присутні у складі найбільшої кліки $C \subseteq V$, позначені логічними одиницями. Наприклад, набір сигналів $\{1,0,1,0\}$ відповідає підграфу, який містить вершини $\{x_1, x_3\}$. Вважатимемо, що на початку всі вузли входять до складу найбільшої кліки, тобто, вхідний вектор є одиничним $\{1,1,1,1\}$. Наступні операції мають забезпечити відсіювання зайвих вершин та формування остаточного результату.

Однак, такий підхід дозволяє лише виявити у графі зв'язані компоненти та усунути ізольовані вузли. Всі зв'язані групи, незалежно від кількості ребер, будуть закодовані одиницями у вихідному наборі. Втім, ретельний аналіз показує, що значення сигналів на входах нейронів виявляються пропорційними до ступеня відповідних вершин графа. При цьому сильно зв'язані групи відрізняються найвищими значеннями вхідних сигналів.

Тому знаходження найбільшої кліки може бути реалізовано шляхом поступового підвищення порогу θ з плином ітерацій. У зв'язку з цим, нейрони, що відповідають слабко зв'язаним групам поступово втрачають здатність до активації та виключаються з розгляду. Найбільш стійкою до підвищення порогу θ виявляється максимальна кліка. За подальшого збільшення значення θ вихідний вектор стає нульовим.

Розглянемо приклад виконання цього алгоритму для графа, наведеного на рис. 10а. Обчислення розв'язку здійснюється на основі матриці пам'яті W (рис. 10б) та одиничного вхідного вектора X_0. Збільшення рівня активації до $\theta = 1$ спричинює виключення зайвого вузла x_1 з отриманого розв'язку X_1 (рис. 10в). За наступних

ітерацій отримані вектори залишаються стабільними. Лише встановлення порогу $\theta = 2$ призводить до формування нульового набору. Таким чином, до складу найбільшої кліки входять вузли $C = \{x_2, x_3, x_4\}$.

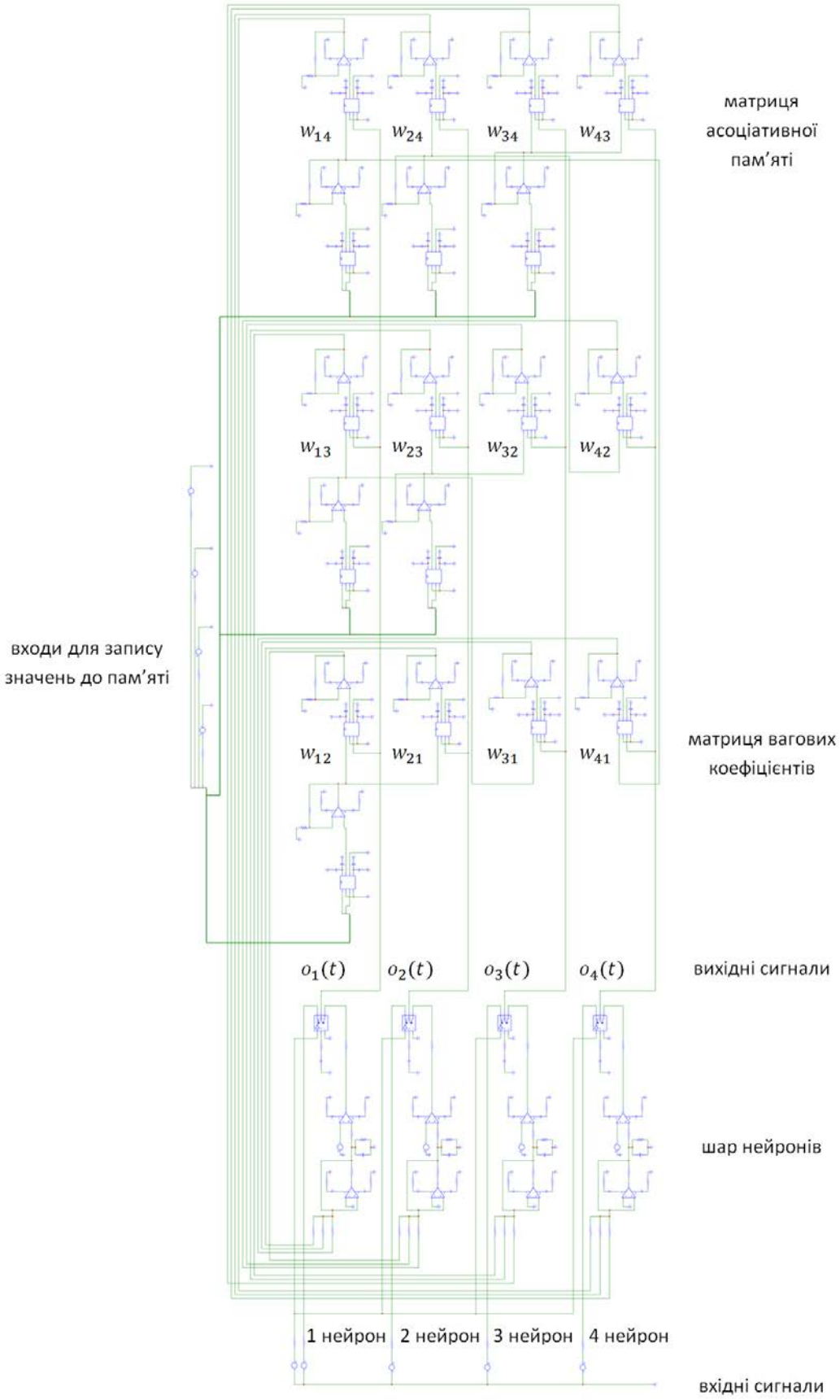

Рис. 9. Функціональна схема розробленого нейрокомп'ютера

Для розв'язання задач значної розмірності та збільшення обчислювальної потужності необхідним є масштабування розробленої системи. Зокрема, можливим є складання нейрокомп'ютерів з базових нейроблоків, які містять 4 процесори, та матриць асоціативної пам'яті з 4×4 вузлів. Наприклад, система з 8 нейронів, утворена у такий спосіб, міститиме 2 нейроблоки та 4 матриці вагових коефіцієнтів, необхідні для їх з'єднання.

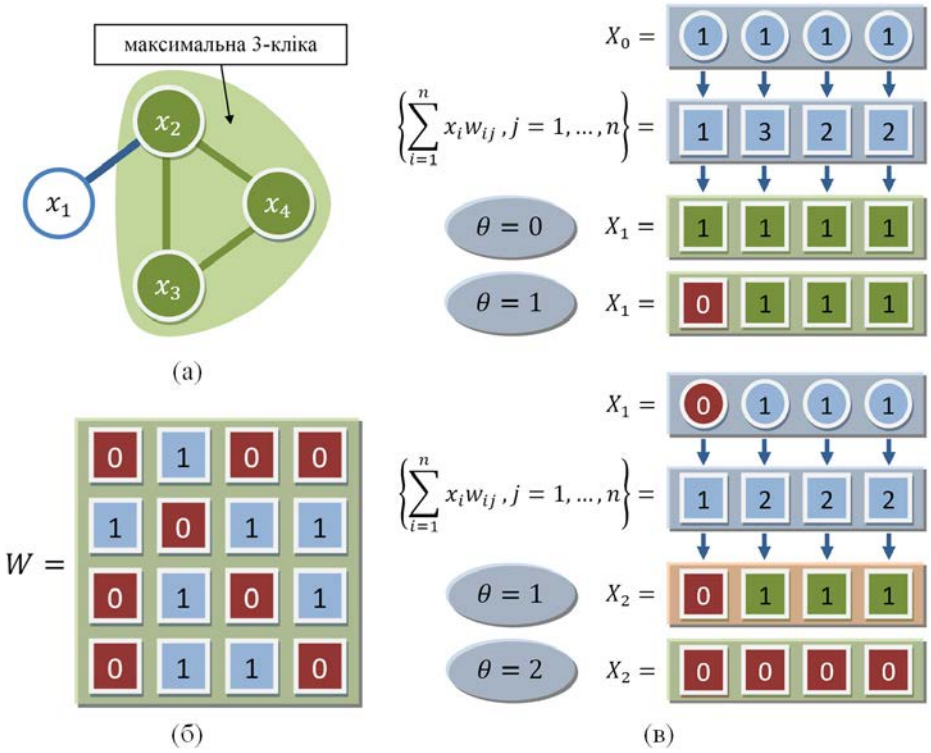

Рис. 10. Приклад знаходження найбільшої кліки за допомогою розробленого нейрокомп'ютера

Швидкодія запропонованої системи визначається затримками сигналів на операційних підсилювачах, елементах множення та часом заряджання RC-комірок нейронів. За наведених у цій статті параметрів обчислювальна потужність системи, яка містить 4 процесори, становить близько 7 MCPS. Продуктивність нейрокомп'ютера може бути додатково збільшена за рахунок зменшення сталої часу RC-комірок нейронів та зниження напруги у вузлах асоціативної пам'яті. При цьому необхідні значення параметрів визначаються критичними параметрами кожної окремої задачі та режимом енергоспоживання пристрою.

Разом з тим, продуктивність нейрокомп'ютера суттєво підвищується при здійсненні його масштабування. Зі збільшенням числа нейронів значно зростає кількість зв'язків між ними, оскільки топологія системи є сильно зв'язаною. Наприклад, пристрої, які містять 16 нейронів та 240 з'єднань між ними, мають обчислювальну потужність на рівні 140 MCPS. Продуктивність систем, складених з 64 процесорів та 4032 синапсів, досягає 2,352 GCPS. Приблизно таку обчислювальну потужність мають біологічні нейронні мережі, які забезпечують майстерність польоту та маневрування звичайних мух.

7 Висновки

У статті запропоновано архітектуру новітнього нейрокомп'ютера для ефективного розв'язання NP-складних задач комбінаторної оптимізації. На відміну від програмних реалізацій нейронних мереж, обчислювальний потенціал яких є суттєво обмеженим, розроблена система зберігає властивості явного паралелізму та має низку наступних переваг:

1. **Висока продуктивність, компактність та енергоефективність.** У запропонованій архітектурі відтворено основні властивості біологічних систем, до яких належать сильно зв'язана топологія та аналогова обробка інформації. Як наслідок, продуктивність системи з 64 нейронів досягає 2,352 GCPS, що є сумірним з потужністю сучасних суперкомп'ютерів.
2. **Масовий паралелізм**, реалізований на двох рівнях: процесорів та вузлів асоціативної пам'яті. У зв'язку з цим формування сигналів активації нейронів відбувається одночасно з обчисленням вагових коефіцієнтів.

3. **Висока відмовостійкість**, пов'язана зі збереженням працездатності системи в разі пошкодження частини нейронів та вузлів асоціативної пам'яті.
4. **Можливість масштабування системи для досягнення вищих рівнів продуктивності.** Шляхом поєднання базових нейроблоків та матриць асоціативної пам'яті можливим є формування нейрокомп'ютерів вищої розмірності. Обчислювальна потужність при цьому суттєво збільшується за рахунок посилення властивостей паралельного виконання.
5. **Можливість розв'язання широкого класу задач комбінаторної оптимізації.** При цьому шукані розв'язки мають бути представлені у вигляді мінімумів енергетичної функції системи. Для прикладу у статті наведено алгоритм адаптації розробленого нейрокомп'ютера до розв'язання задачі пошуку максимальної кліки.

Подальший напрямок досліджень пов'язаний з вдосконаленням запропонованої архітектури та здійсненням її адаптації до розв'язання інших задач комбінаторної оптимізації.

Перелік посилань

[1] *Ding-Zhu Du, Panos Pardalos*, Handbook of Combinatorial Optimization: Supplement Volume A. – Springer, 2010. – 648 p.

[2] Finding maximal cliques in massive networks / [*J. Cheng, Y. Ke, A. W.-C. Fu, J. X. Yu, L. Zhu*] // ACM Transactions on Database Systems, 36(4):21, 2011. – 34 p.

[3] *Погорілий С.Д., Потебня А.В.*, Дослідження та оптимізація архітектури еластичної нейронної мережі для розв'язання задачі комівояжера // Праці 9-ої міжнародної конференції «Теоретичні та прикладні аспекти побудови програмних систем (TAAPSD'2012)». (Україна, Київ, 4 – 7 грудня, 2012). – С. 236 – 244.

[4] *Погорілий С.Д., Потебня А.В.*, Розробка новітніх систем штучного інтелекту для розв'язання задачі комівояжера // Вісник Київського національного університету імені Тараса Шевченка. Серія фізико-математичні науки. – Київ, 2012 – №4. – С. 173 – 184.

[5] *Погорілий С.Д., Потебня А.В.*, Формування та дослідження паралельної схеми алгоритму Крускала для систем зі спільною пам'яттю // Наукові праці ДонНТУ. Серія «Інформатика, кібернетика та обчислювальна техніка». – Донецьк, 2012 – №16 (204). – С. 82 – 89.

[6] *Погорілий С.Д., Потебня А.В.*, Новітній швидкий алгоритм знаходження мінімальних опуклих оболонок // Праці міжнародної конференції «Високопродуктивні обчислення HPC-UA 2013». (Україна, Київ, 7 – 11 жовтня, 2013). – С. 322 – 329.

[7] *Погорілий С.Д., Потебня А.В.*, Новітній метод розв'язання задач комбінаторної оптимізації великої розмірності // Наукові праці ДонНТУ. Серія «Інформатика, кібернетика та обчислювальна техніка». – Донецьк, 2014 – №1 (19). – С. 114 – 125.

[8] Artificial neural networks: a review of commercial hardware / [*Dias F.M., Antunes A., Mota, A.M.*] // Engineering Applications of Artificial Intelligence. – December 2004 – Vol. 17, № 8. – P. 945 – 952.

[9] *Janardan Misra, Indranil Saha*, Artificial neural networks in hardware: A survey of two decades of progress // Neurocomputing. – December 2010 – Vol. 74, № 1 – 3. P. 239 – 255.

[10] *Maurizio Valle*, Analog VLSI Implementation of Artificial Neural Networks with Supervised On-Chip Learning // Analog Integrated Circuits and Signal Processing. –2002 – Vol. 33, № 3. – P. 263 – 287.

[11] DDCI: Simple Dynamic Semiautomatic Parallelizing for Heterogeneous Multicomputer Systems / [*Levchenko R.I., Sudakov O.O., Pogorilyy S.D.*] // Proceedings of the 5th IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, 21 – 23 September 2009, Rende (Cosenza), Italy.

[12] *Pogorilyy S.D., Gusarov A.D.*, Paralleling Of Edmonds-Karp Net Flow Algorithm // Applied and Computational Mathematics. – 2006 – Vol. 5, №2. – P. 121 – 130.

[13] An electrically trainable artificial neural network (ETANN) with 10240 «floating gate» synapses / [*Holler M. (Intel Corp., Santa Clara, CA, USA), Simon Tam, Castro H., Benson R.*] // Artificial neural networks. – 1990 – P. 50 – 55.

[14] The Mod 2 Neurocomputer system design / [*Mumford M.L., Andes D.K., Kern, L.R.*] // IEEE Transactions on Neural Networks. – May 1992. – Vol. 3, № 3. – P. 423 – 433.

[15] *Draghici S.*, Neural networks in analog hardware design and implementation issues // International Journal of Neural Systems. – 2000 – Vol. 10 (1). – P. 19 – 42.

[16] *Palm G.*, Neural associative memories and sparse coding // Neural Networks, 2013. – Vol. 37. – P. 165 – 171.

[17] Storage Capacity of the Hopfield Network Associative Memory / [*Yue Wu, Jianqing Hu, Wei Wu, Yong Zhou, Du K.L.*] // Fifth International Conference on Intelligent Computation Technology and Automation (ICICTA). – January 2012. – P. 330 – 336.