

*the risk of over-reliance on model outputs without critical thinking, and challenges regarding data security and confidentiality when using external AI services. Overall, the findings suggest that AI integration in Scrum can significantly boost team productivity and project outcomes, serving as a catalyst for an **AI-Augmented Scrum** approach where AI not only supports but also informs management decisions.*

Keywords: Scrum, Agile, artificial intelligence, automation, project efficiency, AI integration, project management.

УДК 004.023

DOI: 10.18372/2073-4751.83.20546

Воронцов В.І., аспірант.

Лукашенко В.В., д.т.н.
orcid.org/0000-0002-8898-2269

ОГЛЯД АПАРАТНОЇ РЕАЛІЗАЦІЇ ПРИСТРОЇВ З ВИКОРИСТАННЯМ POSIT ЧИСЛЕННЯ

Національний авіаційний університет

e-mail: vitalik1998v@gmail.com

e-mail: viktoriiia.lukashenko@npp.nau.edu.ua

Вступ

Сучасний розвиток комп'ютерних обчислень супроводжується зростанням вимог до точності числових представлень, енергоефективності та апаратної оптимізації. Традиційний формат чисел з плаваючою крапкою IEEE 754, який протягом десятиліть слугував базовим стандартом у чисельних обчисленнях, виявляє низку обмежень, що стають критичними в умовах нових технологічних викликів. Зокрема, проблемними залишаються питання надмірної складності апаратних реалізацій, обмеженої точності та неефективного використання обчислювальних ресурсів.

Арифметика Posit, запропонована у 2017 році, розглядається як перспективна альтернатива класичним підходам. Її концептуальні особливості дозволяють підвищити точність подання чисел, зменшити апаратні витрати та досягти кращого балансу між продуктивністю та енергоефективністю. Це робить систему Posit актуальним об'єктом дослідження у таких сферах, як глибинне навчання, вбудовані системи та високопродуктивні обчислення.

Однак практична реалізація арифметики Posit на апаратному рівні потребує поглибленого аналізу. Життєздатність та

масштабованість нової системи числового представлення визначається ефективністю апаратних архітектур, здатних забезпечити необхідні характеристики точності, продуктивності та стійкості до обчислювальних похибок.

В основній частині даної статті, підготовленої в рамках здобуття наукового ступеня доктора філософії, описана основа арифметики Posit, її структуру, переваги та виклики порівняно з IEEE 754. Після цього будуть ретельно досліджені апаратні реалізації одиниць арифметики Posit. Буде досліджено застосування в глибокому навчанні, аналізи точності та ефективності, дослідження стійкості та дослідження помилок. Огляд завершується підсумком ключових висновків та наслідками для майбутніх досліджень.

Мета

Метою цієї статті є надання всеосяжного огляду недавніх наукових статей, присвячених апаратній реалізації пристроїв побудованих на арифметиці Posit та дослідження наявної сучасної літератури з метою виявлення поточного стану дослідженості Posit арифметики. Огляд включає аналіз, включаючи апаратні реалізації, застосування в глибокому навчанні, оцінки точності та ефективності, дос-

лідження стійкості та аналізу помилок. Синтезуючи висновки з восьми різних статей, ми маємо на меті надати цілісне розуміння поточного стану та майбутнього потенціалу Posit арифметики.

Основна частина

Система чисел Posit, створена Джоном Л. Густафсоном у 2017 році, представляє революційну альтернативу традиційним представленням плаваючої точки, особливо стандарту IEEE 754. Posits розроблені для вирішення обмежень систем фіксованої точки та плаваючої точки, пропонуючи універсальний формат, який динамічно коригує свою точність на основі величини числа.

На відміну від IEEE 754, арифметика Posit відмовляється від експоненти фіксованої ширини бітів, дозволяючи більшу гнучкість у представленні широкого діапазону величин. Режим, експонента та компоненти дробу в Posit є взаємозалежними, що дозволяє більш ефективне використання бітів. Це динамічне представлення на основі режиму відрізняє Posit від форматів плаваючої точки з фіксованою шириною, пропонуючи переваги в точності та діапазоні для як малих, так і великих чисел.

Переваги арифметики Posit:

1. Арифметика Posit досягає балансу між точністю та діапазоном, налаштовуючи використання бітів для різних величин. Ця гнучкість надає їй помітну перевагу над форматами з фіксованою шириною, де розподіл бітів залишається постійним, незалежно від масштабу даних.
2. На відміну від представлень з фіксованою точкою, які схильні до помилок квантування через однорідний розподіл бітів, динамічна структура числа Posit мінімізує такі помилки, особливо для чисел, близьких до одиниці.
3. Структура числа Posit дає змогу представляти дійсні числа з вищою точністю, зменшуючи помилки округлення, пов'язані з традиційною арифметикою плаваючої точки, має єдину систему правил округлення та підтримує доповняльний код, що дозволяє виконувати порів-

няння чисел формату posit як знакові цілі числа.

4. Менше виключень, відповідно більше комбінацій біт для чисел, має лише 1 значення NaN, коли у IEEE 754 їх кількість може досягати 9 квадрильйонів, 4, має малу кількість виключень, а саме NaN (не є числом), нескінченність, underflow.

Виклики та критика нового формату:

1. Динамічна структура числа Posit ставить виклики перед апаратною реалізацією, з потребою в спеціалізованих обчислювальних блоках для режимів змінної довжини. З іншого боку, при реалізації фіксованого числа Posit, апаратна реалізація простіше ніж реалізація IEEE 754.
2. На сьогоднішній день немає апаратної реалізації у комерційних продуктах.
3. Перехід від традиційної арифметики плаваючої точки до Posit вимагає зміни апаратної та програмної частини, що вимагає корекції алгоритмів та обчислювальних методологій.

Актуальна наукова література свідчить про зростання досліджень, присвячених арифметиці Posit.[1-8] Вчені досліджували її апаратні реалізації, порівняльні аналізи з традиційними форматами та застосування в нових галузях, таких як глибоке навчання та вбудовані системи. Цей контекст створює основу для всеосяжного огляду різних аспектів арифметики Posit, які досліджуються в статтях, розглянутих у цьому огляді.

Розглянемо вісім статей присвячених дослідженню апаратних реалізацій пристроїв з використанням Posit арифметики для оцінки доцільності використання арифметики Posit, висвітлюючи її поточний стан та майбутні перспективи.

1. PLAM: A Posit Logarithm-Approximate Multiplier.

Стаття, запропонована Мурильо та ін.[7], представляє новий апроксимований множник PLAM для формату Posit, що використовує логарифмічну арифметику. Дослідження пропонує енергоефективне рішення для прискорення обчислень у пристроях з обмеженими ресурсами, зокрема в архітектурах

глибокого навчання нейронних мереж (DNN). Метою роботи є зменшення апаратних витрат, споживання енергії та затримок при збереженні прийнятної точності обчислень. Методологія базується на перетворенні операндів у логарифмічну систему числення з наступним виконанням операцій додавання замість множення, що значно зменшує складність апаратної реалізації. Результати демонструють значне покращення енергоефективності, продуктивності та апаратних показників порівняно з точними множниками Posit при мінімальній втраті точності для широкого кола застосувань, що робить цей підхід перспективним для підвищення ефективності архітектур DNN.

Хоча підхід ефективно зменшує апаратні витрати, його головним обмеженням є втрата точності, властива для апроксимаційних обчислень. Таким чином, його застосування найбільш доцільне в задачах, де енергоефективність і швидкість пріоритетніші за абсолютну точність, наприклад в нейронних мережах.

2. Unified Posit/IEEE-754 Vector MAC Unit.

Стаття, запропонована Креспо та ін.[2], обговорює концепцію трансточних обчислень, яка має на меті покращити енергоефективність за допомогою використання кількох модулів з плаваючою точкою з різними точностями для задоволення вимог застосувань. Автори висвітлюють обмеження архітектур змінної точності, які часто базуються на стандарті IEEE-754 без підтримки арифметики низької точності. Вони пропонують формат Posit як підходящу альтернативу для арифметики низької точності.

Однак вони зазначають, що для вищих точностей апаратні вимоги до формату Posit стають неприйнятними. Для вирішення цього питання автори пропонують новий об'єднаний блок Vector Multiply-Accumulate (VMAC) Posit/IEEE-754. Цей блок складається з векторизованого шляху передачі даних змінної точно-

сті з спільною підтримкою форматів Posit та IEEE-754.

Реалізація пропозиції авторів у вигляді ASIC за 28-нм техпроцесом продемонструвала скорочення займаної площі на 50% і зменшення енергоспоживання у 2.9 рази в порівнянні з традиційними архітектурами, розрахованими на роботу з даними змінної точності.

Реалізація єдиного блоку є перспективним кроком для поступового впровадження Posit без повного відмови від існуючої інфраструктури. Однак складність такого об'єднання, особливо для форматів Posit з високою точністю, залишається серйозною проблемою для масового впровадження.

3. Posit Processing Element for Energy-Efficient DNN Accelerators.

Це дослідження, запропоноване Зольфагарінеджадом та ін.[3], представляє енергоефективний елемент обробки (ЕО) Posit для використання в прискорювачах глибокого навчання нейронних мереж на основі масивів. Автори пропонують метод наближення для подальшого зниження споживання енергії блоку. Арифметика Posit, що використовується в запропонованому ЕО, забезпечує високу точність для розглянутих ширин даних, навіть коли для операцій використовується наближення. Автори пропонують модуль додавання Posit та використовують деякі підходи до модифікації/спрощення для зменшення складності використовуваного множника-акумулятора в запропонованому ЕО.

Ефективність запропонованого ЕО досліджується за допомогою технології CMOS 45 нм. Результати показують поліпшення затримки в 3,5 рази та споживання енергії на 92% в порівнянні з найкращими ЕО Posit.

Впровадження апроксимаційних методів є логічним кроком для ще більшого зниження апаратних витрат. Однак ключовим завданням є кількісна оцінка впливу цих наближень на кінцевий результат після впровадження таких оптимізацій.

4. A Lightweight Posit Processing Unit for RISC-V Processors.

Стаття, запропонована Кокоціоні та ін.[4], обговорює два нових фактори в нейронних мережах. Перший - це відкрита архітектура набору інструкцій RISC-V, яка дозволяє безшовну реалізацію користувачьких наборів інструкцій. Другий - це декілька нових форматів для арифметики дійсних чисел. Автори поєднали ці два аспекти, використовуючи перспективний формат Posit, розробивши легкий обчислювальний блок Posit (PPU-light). Вони представляють розширення базового ISA RISC-V, яке дозволяє конвертацію між 8- або 16-бітними Posit та 32-бітними плаваючими числами IEEE або форматами з фіксованою точкою. Це пропонує стисле представлення дійсних чисел з мінімальним або відсутнім зниженням точності. Автори розглядають інтеграцію апаратного та програмного забезпечення їх PPU-light всередині ядра Ariane RISC-V та його інструментарію. Вони показують, як мало він впливає на складність схеми та споживання енергії. Дійсно, лише 0,36% схеми призначено для PPU-light, тоді як повне ядро RISC-V займає 33% загальної складності схеми. Нарешті, вони представляють вплив їх PPU-light на завдання глибокої нейронної мережі, повідомляючи про прискорення до 10 разів часу обробки вибіркового виведення. Для оцінки ефективності запропонованого PE автори моделюють 8-бітний прискорювач DNN та використовують його для реалізації деяких архітектур DNN.

Результати вказують, що запропонований PE та його наближений варіант забезпечують в середньому на 19,3% та 29,6% нижче споживання енергії порівняно з останньою попередньою роботою, надаючи при цьому на 10,6% та 5,8% вищу точність відповідно.

Мінімальне збільшення апаратних витрат робить це рішення дуже привабливим. Головне питання для подальшого розвитку — це масштабування підтримки для більших форматів Posit (наприклад, 32-бітних), які можуть знадобитися для складніших обчислень.

5. Posit8 Decompression Operator for DNN Inference.

Стаття, запропонована Десрентесом та ін.[5], представляє оператор апаратного забезпечення, розроблений для декомпресії представлень Posit8 з розмірами експоненти 0, 1, 2, 3 до представлення IEEE 754 binary 16 (FP16). Мотивацією цього є використання тензорних блоків багатоядерного процесора, який вже підтримує операції множення-акумуляції матриць FP16.32 для виведення глибокого навчання. Автори пропонують, що додавання інструкцій для декомпресії чисел Posit8 у числа FP16 може подальше зменшити відбиток параметрів глибокої нейронної мережі з прийнятним втратам точності або прецизійності. Презентується дизайн оператора декомпресії та порівнюється з реалізаціями таблиць пошуку для технологічного вузла цільового процесора.

Автори також оцінюють ефекти стиснення параметрів глибокого навчання IEEE 754 binary 32 (FP32) до представлень Posit8 на класичних мережах класифікації та виявлення. Вони обговорюють виклики декомпресії представлень Posit8 до FP16.

Робота представляє розумний компроміс для поступового переходу на Posit8. Проте втрати точності при декомпресії можуть бути критичними для деяких застосувань, тому цей метод потребує ретельного тестування для кожного конкретного випадку використання.

6. Customized Posit Adders and Multipliers using the FloPoCo Core Generator.

Стаття, запропонована Мурильо та ін.[6], представляє метод автоматизованого синтезу апаратних прискорювачів для обчислень у форматі Posit з використанням генератора апаратних ядер FloPoCo. Дослідження пропонує рішення для створення модулів додавання та множення з підтримкою формату Posit різної розрядності (16 та 32 біт). Метою роботи є спрощення та прискорення розробки енергоефективних сопроцесорів шляхом запропонованих готових рішень замість ручного проектування. Методологія базується на модифікації

фреймворку FloPoCo для генерації високо-продуктивних конвеєрних архітектур, оптимізованих під специфіку формату Posit. Результати демонструють переваги запропонованих блоків у покращенні продуктивності (тактової частоти) порівняно з існуючими аналогами при збереженні конкурентних показників займаної площі кристалу.

Автоматизація проектування за допомогою інструментів на кшталт FloPoCo може значно прискорити дослідження та впровадження арифметики Posit. Однак ефективність цього підходу повністю залежить від якості та гнучкості базового фреймворку, який може не оптимально враховувати всі особливості формату Posit.

7. Lossless FFTs Using Posit Arithmetic.

Стаття, запропонована Густавсоном та Леонгом[7], обговорює використання швидкого перетворення Фур'є (FFT), яке є важливим інструментом в різних галузях, таких як хімія, метеорологія, оборона та обробка сигналів для сейсмічного дослідження та радіоастрономії. Автори підкреслюють, що FFT є зв'язаним з комунікацією, що робить суперкомп'ютери значно повільнішими при виконанні FFT, ніж при виконанні лінійної алгебри. За словами авторів, ключем до прискорення FFT є мінімізація бітів на дані без жертвування точністю. Вони стверджують, що 16-бітний формат фіксованої точки та тип плаваючої точки IEEE не мають достатньої точності для FFT з 1024 та 4096 входами даних з аналого-цифрових перетворювачів.

Автори пропонують використання 16-бітного Posit, який має вищу точність та більший динамічний діапазон. Вони демон-

струють, що він може виконувати FFT настільки точно, що пряме, а потім зворотне FFT відновлює оригінальний сигнал ідеально. Це призводить до "обернених" FFT з Posit, які є безвтратними, що виключає потребу в точності 32 біт або вище.

Аналогічно, вони пропонують, що 32-бітні FFT Posit можуть замінити 64-бітні FFT з плаваючою точкою для багатьох завдань високопродуктивних обчислень (HPC). В результаті швидкість, енергоефективність та вартість зберігання можуть бути покращені в 2 рази для широкого діапазону навантажень HPC.

Здатність 16-бітного Posit забезпечувати безвтратні обчислення FFT є вагомим аргументом на користь його переваг у точності порівняно з IEEE-754, що підтверджується сучасними дослідженнями. Це відкриває значні перспективи для економії пам'яті та енергії в високопродуктивних обчисленнях. Однак універсальність цієї переваги для інших, складніших алгоритмів, окрім FFT, потребує подальшого вивчення.

8. Design of Energy Efficient and Low Delay Posit Multiplier.

Стаття, запропонована Haripriya та ін.[8], представляє комплексне дослідження апаратної реалізації блоку множення Posit. Автори пропонують альтернативу блоку з плаваючою точкою IEEE-754 - систему чисел Posit, яка надає більш точні результати при тій же ширині слова. Ключовою інновацією є використання суматора на основі Leading One Detector (LOD) замість традиційної логіки на основі мультиплексера.

Таблиця 1. Порівняльний аналіз досліджень по апаратній реалізації Posit

	Предмет дослідження	Ключові переваги порівняно з IEEE 754 або аналогами	Виявлені обмеження
[1]	Використання логарифмічної арифметики для заміни множення додаванням	Зменшення площі, споживання енергії та затримки при мінімальній втраті точності	Втрата точності, обмеженість множенням
[2]	Єдиний Векторний блок MAC з підтримкою як Posit, так і IEEE 754)	Скорочення площі на 50% та енергоспоживання у 2.9 разів порівняно з роздільними блоками	Складність реалізації для високоточних Posit

[3]	Енергоефективний обчислювальний блок	Зменшення затримки в 3.5 рази, споживання енергії на 92%	Апроксимація вимагає додаткової перевірки якості результатів
[4]	Легкий блок обчислень як розширення RISC-V	Прискорення обчислень до 10 разів, мінімальний вплив на площу кристалу (+0.36%)	Обмеження 8-16 бітними форматами
[5]	Апаратний оператор декомпресії Posit8 в FP16	Можливість використання існуючих тензорних ядер (FP16)	Втрати точності при конвертації форматів
[6]	Автоматизована генерація ядер з використанням FloPoCo	Прискорення розробки енергоефективних сопроцесорів	Ефективність залежить від якості фреймворку FloPoCo
[7]	Використання 16-бітних Posit для безвтратних перетворень Фур'є	Безвтратне обчислення для 16-бітних даних posit при перетворенні Фур'є	Спеціалізація на конкретному алгоритмі
[8]	Оптимізація блоку множення використанням суматора на основі LOD	Зменшення затримки на 17.55%, енергоефективність вище на 28.35% у порівнянні з існуючими рішеннями	Тестування проведене тільки у симуляторі

Ця зміна зменшує як затримку, так і площу блоку множення Posit. LOD використовується для витягування бітів даних з даних Posit для будь-якої заданої ширини даних Posit (N), розміру експоненти (ES) та розміру режиму (RS), що призводить до меншої площі та затримки шляху передачі даних. Крім того, автори пропонують окремий розрахунок кінцевої експоненти та кінцевого режиму, що призводить до подальших поліпшень в області через зменшення розміру суматора.

Блоки множення Posit були синтезовані для 8-біт, 16-біт та 32-біт та порівнювалися з існуючою роботою. Результати синтезу показали зменшення затримки на 17,55% та збільшення енергоефективності на 28,35%.

Запропоновані оптимізації, зокрема використання LOD-суматора, є важливим внеском у розвиток ефективних апаратних реалізацій Posit. Проте заявлені переваги потребують верифікації шляхом стендових випробувань на ASIC або FPGA. Симуляція на рівні RTL-моделей, хоч і є важливим етапом проектування, лише обмежено відображає енергоефективність та швидкодію рішення в реальних умовах експлуатації.

Зведений аналіз досліджень представлено у таблиці 1.

Висновки

Аналіз сучасних досліджень засвідчив, що апаратні реалізації арифметики Posit здатні забезпечити істотні переваги у швидкодії та енергоефективності порівняно зі стандартом IEEE-754. Зокрема було продемонстровано, що апроксимований множник PLAM на основі логарифмічної арифметики знижує апаратні витрати та затримки, водночас суттєво підвищуючи енергоефективність у задачах глибокого навчання [1]. Був запропонований об'єднаний блок VMAC Posit/IEEE-754 показав скорочення площі кристалу на 50% та зменшення енергоспоживання у 2,9 рази порівняно з класичними рішеннями [2]. Досліджений елемент обробки Posit для прискорювачів нейронних мереж забезпечує зменшення затримки у 3,5 разів та зниження споживання енергії на 92%, що робить його особливо ефективним для архітектур DNN [3]. У свою чергу, інтеграція обчислювального блоку PPU-light у RISC-V процесор потребувала лише 0,36% додаткових ресурсів, але дозволила прискорити час виведення нейронної мережі до 10 разів. При цьому середнє зниження енергоспоживання складало 19,3% у звичайному та 29,6% у наближеному варіантах [4]. Було показано, що

використання суматора на основі Leading One Detector у множнику Posit дозволяє знизити затримку на 17,55% і підвищити енергоефективність на 28,35% порівняно з традиційними реалізаціями [8]. Дана робота при інтеграції у інші реалізації може суттєво збільшити якісні характеристики рішень на Posit.

У роботі [5] представлено оператор декомпресії Posit8 до FP16, що дозволяє скоротити обсяг пам'яті, необхідний для зберігання параметрів нейронних мереж при збереженні прийнятної точності, хоча втрати на етапі декомпресії можуть бути критичними для деяких застосувань. У роботі [6] було показано, що використання автоматизованого генератора FloPoCo для синтезу додатків і множників Posit значно пришвидшує проектування та забезпечує вищі тактові частоти при конкурентних апаратних витратах, проте якість синтезованих ядер буде залежати від використаного фреймворку. У роботі [7] 16-бітний Posit продемонстрував здатність виконувати безвтратні FFT, перетворення відновлюють вихідний сигнал ідеально.

Таким чином, найбільш відчутні переваги Posit проявляються у випадках роботи з малими розрядностями (8–16 біт), де досягається значна економія енергії та апаратних ресурсів при прийнятній точності. Це робить Posit особливо перспективним для прискорювачів глибокого навчання та обчислень FFT у високопродуктивних системах. Натомість для задач, що вимагають високої точності при великих розрядностях, формат IEEE-754 залишається більш доцільним через надмірну складність апаратних реалізацій Posit. Подальші дослідження мають бути спрямовані на масштабування архітектур Posit до старших розрядностей та експериментальну верифікацію результатів на ASIC і FPGA.

Перспективними напрямками майбутніх досліджень є подальша оптимізація апаратних реалізацій для форматів високої точності, розширення інструментарію автоматизованого синтезу та дослідження гібридних підходів, що поєднують пере-

ваги Posit та IEEE-754 у гетерогенних обчислювальних системах.

Література

1. Murillo R., Del Barrio A. A., Botella G., Kim M. S., Kim H., Bagherzadeh N. PLAM: A Posit Logarithm-Approximate Multiplier. *IEEE Transactions on Emerging Topics in Computing*. 2022. Vol. 10, № 4. P. 2079–2085. DOI: 10.1109/TETC.2021.3109127.
2. Crespo L., Tomás P., Roma N., Neves N. Unified Posit/IEEE-754 Vector MAC Unit for Transprecision Computing. *IEEE Transactions on Circuits and Systems II: Express Briefs*. 2022. Vol. 69, No. 5. P. 2478–2482. DOI: 10.1109/TCSII.2022.3160191.
3. Zolfagharinejad M., Kamal M., Afzali-Khusha A., Pedram M. Posit Process Element for Using in Energy-Efficient DNN Accelerators. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*. 2022. Vol. 30, № 6. P. 844–848. DOI: 10.1109/TVLSI.2022.3165510.
4. Cococcioni M., Rossi F., Ruffaldi E., Saponara S. A Lightweight Posit Processing Unit for RISC-V Processors in Deep Neural Network Applications. *IEEE Transactions on Emerging Topics in Computing*. 2022. Vol. 10, № 4. P. 1898–1908. DOI: 10.1109/TETC.2021.3120538.
5. Desrentes O., Resmerita D., Dupont de Dinechin B. A Posit8 Decompression Operator for Deep Neural Network Inference. In: Gustafson J., Dimitrov V. (eds) *Next Generation Arithmetic*. CoNGA 2022. Lecture Notes in Computer Science, vol 13253. Springer, Cham, 2022. P. 19–33. DOI: 10.1007/978-3-031-09779-9_2.
6. Murillo R., Del Barrio A. A., Botella G. Customized Posit Adders and Multipliers using the FloPoCo Core Generator. In: 2020 *IEEE International Symposium on Circuits and Systems (ISCAS)*. Seville, Spain, 2020. P. 1–5. DOI: 10.1109/ISCAS45731.2020.9180771.
7. Leong S.H., Gustafson J.L. Lossless FFTs Using Posit Arithmetic. In: Gustafson J., Leong S.H., Michalewicz M. (eds) *Next Generation Arithmetic*. CoNGA 2023. Lec-

ture Notes in Computer Science, vol 13851. Springer, Cham. 2023. P. 3–15. DOI: 10.1007/978-3-031-32180-1_1.

8. L.B.R.K., H.R.S., Puli K., Annapalli S.R.R., Pudi V. Design of Energy Efficient and Low Delay Posit Multiplier. *In: 2023*

36th International Conference on VLSI Design and 2023 22nd International Conference on Embedded Systems (VLSID). Hyderabad, India, 2023. P. 1–6. DOI: 10.1109/VLSID57277.2023.00042

Воронцов В.І., Лукашенко В.В.

ОГЛЯД АПАРАТНОЇ РЕАЛІЗАЦІЇ ПРИСТРОЇВ З ВИКОРИСТАННЯМ POSIT ЧИСЛЕННЯ

Сучасний розвиток комп'ютерних обчислень супроводжується зростанням вимог до точності числових представлень, енергоефективності та апаратної оптимізації. Традиційний формат чисел з плаваючою крапкою IEEE 754, який протягом десятиліть слугував базовим стандартом у чисельних обчисленнях, виявляє низку обмежень, що стають критичними в умовах нових технологічних викликів. Зокрема, проблемними залишаються питання надмірної складності апаратних реалізацій, обмеженої точності та неефективного використання обчислювальних ресурсів.

Арифметика Posit розглядається як перспективна альтернатива класичним підходам. Її концептуальні особливості дозволяють підвищити точність подання чисел, зменшити апаратні витрати та досягти кращого балансу між продуктивністю та енергоефективністю. Це робить систему Posit актуальним об'єктом дослідження у таких сферах, як глибоке навчання, вбудовані системи та високопродуктивні обчислення.

Однак практична реалізація арифметики Posit на апаратному рівні потребує поглибленого аналізу. Життєздатність та масштабованість нової системи числового представлення визначається ефективністю апаратних архітектур, здатних забезпечити необхідні характеристики точності, продуктивності та стійкості до обчислювальних похибок.

В основній частині даної статті, підготовленої в рамках здобуття наукового ступеня доктора філософії, описана основа арифметики Posit, її структуру, переваги та виклики порівняно з IEEE 754. Після цього будуть ретельно досліджені апаратні реалізації одиниць арифметики Posit. Буде досліджено застосування в глибокому навчанні, аналізи точності та ефективності, дослідження стійкості та дослідження помилок. Огляд завершується підсумком ключових висновків та наслідками для майбутніх досліджень.

Ключові слова: Система числення Posit, числа з плаваючою комою, *embedded systems*, апаратне забезпечення, глибоке навчання нейронних мереж.

Vorontsov V.I., Lukashenko V.V.

OVERVIEW OF THE HARDWARE IMPLEMENTATION OF DEVICES USING POSIT NUMBER SYSTEM

The modern development of computer computing is accompanied by increasing demands for numerical representation accuracy, energy efficiency, and hardware optimization. The traditional floating-point format IEEE 754, which has served as the fundamental standard in numerical computations for decades, reveals a number of limitations that become critical under new technological challenges. In particular, issues such as excessive hardware complexity, limited accuracy, and inefficient utilization of computational resources remain problematic.

The Posit arithmetic is considered a promising alternative to classical approaches. Its conceptual features make it possible to enhance numerical accuracy, reduce hardware costs, and achieve a better balance between performance and energy efficiency. This makes the Posit system a relevant subject of investigation in such fields as deep learning, embedded systems, and high-performance computing.

However, the practical implementation of Posit arithmetic at the hardware level requires in-depth analysis. The viability and scalability of this new number representation system are determined by the efficiency of hardware architectures capable of providing the required levels of accuracy, performance, and robustness against computational errors.

In the main part of this article, prepared within the framework of obtaining the degree of Doctor of Philosophy, the fundamentals of Posit arithmetic, its structure, advantages, and challenges in comparison with IEEE 754 are described. This is followed by a detailed examination of hardware implementations of Posit arithmetic units. Applications in deep learning, accuracy and efficiency analyses, robustness studies, and error investigations will be considered. The review concludes with a summary of the key findings and implications for future research.

Keywords: Posit Number System, IEEE floating point unit, embedded systems, hardware, deep neural network.

УДК 519.725:629.7.052

DOI: 10.18372/2073-4751.83.20525

Гільгурт С.Я., д.т.н.,
orcid.org/0000-0003-1647-1790

Давиденко А.М., д.т.н.,
orcid.org/0000-0001-6466-1690

ОЦІНКА МОЖЛИВОСТЕЙ СТВОРЕННЯ ЗАВАДОСТІЙКИХ КОДІВ ШЛЯХОМ ПЕРЕБОРУ КОМБІНАЦІЙ КОДОВИХ СЛІВ¹

Інститут проблем моделювання в енергетиці ім. Г.Є. Пухова НАН України

e-mail: hilgurt@ukr.net
e-mail: davidenkoan@gmail.com

Вступ

Під час пересилання даних комунікаційними засобами, що піддаються впливу перешкод, виникає проблема запобігання втраті інформації. Наприклад, передача радіоканалом відеозображення з бортової камери БПЛА, які останнім часом все частіше використовуються не тільки для цивільних, але й для військових застосувань, ускладнюється внаслідок впливу різних факторів. Основним дже-

релом перешкод є засоби радіоелектронної боротьби, причому, не тільки ворожі, але також дружні [1, 2]. Заміна радіоканалу на оптоволоконну лінію зв'язку обмежує радіус дії, ускладнює маневреність та вносить ризик втрати зв'язку внаслідок пошкодження волоконного кабелю [3]. Тому радіозв'язок досі залишається основним рішенням для забезпечення керування та передачі відеосигналу. Тому проблема підвищення завадостійкості залишається гострою. Відомі суто радіотехнічні підходи щодо боротьби з перешкодами. Наприклад, розробляються спеціалізовані радіомодеми для БПЛА, які здатні динамічно змінювати частоту та знаходити "вікна" в спектрі ворожих глуш-

¹ Дослідження виконане за фінансової підтримки з боку Національного фонду досліджень України, проєкт 2025.6/0109 "Захищена завадостійка система передачі відеоінформації з безпілотного літального апарата", номер державної реєстрації 0125U003164.