ISSN 1998-7927(print) ISSN 2664-6498 (online)

DOI: https://doi.org/10.33216/1998-7927-2026-304-6-35-43

УДК 004.032.26:004.272:004.8

МАТЕМАТИЧНА МОДЕЛЬ УЗГОДЖЕНОЇ БАГАТОРІВНЕВОЇ ПАРАЛЕЛЬНОЇ ОБРОБКИ ПОТОКОВИХ ДАНИХ У СИСТЕМАХ ШТУЧНОГО ІНТЕЛЕКТУ РЕАЛЬНОГО ЧАСУ

Гентош Л.І.

MATHEMATICAL MODEL OF COORDINATED MULTILEVEL PARALLEL PROCESSING OF STREAMING DATA IN REAL-TIME ARTIFICIAL INTELLIGENCE SYSTEMS

Hentosh L.I.

Запропоновано математичну модель узгодженої багаторівневої паралельної обробки потокових даних у системах штучного інтелекту реального часу, орієнтовану на зменшення наскрізної затримки за збереження заданих показників результативності інтелектуального опрацювання. Актуальність дослідження зумовлена тим, що сучасні системи штучного інтелекту функціонують як багатостадійні потокові обчислювальні контури, у яких часові обмеження формуються не лише тривалістю виконання окремої моделі, а сукупним впливом локальних обчислень і передавань, структури алгоритмічних залежностей, комунікацій, синхронізації, диспетчеризації та черг. У моделі потік подано послідовністю дискретних порцій зі сталою або заданою інтенсивністю надходження, а обробку порції — множиною взаємопов'язаних етапів із визначеними часовими характеристиками. Формалізовано стаціонарний інтервал формування результатів і пропускну здатність системи та встановлено умову потокової працездатності, за якої досяжна пропускна здатність не є меншою за інтенсивність надходження порцій. Наскрізну затримку декомпоновано на локальну, алгоритмічну та системну складові. Локальна складова характеризує внутрішньовузлові обчислення, доступ до пам'яті, передавання між пристроями та локальну синхронізацію; алгоритмічна визначається структурою графа залежностей і довжиною критичного шляху; системна охоплює комунікації, глобальну синхронізацію, диспетчеризацію та очікування в чергах. На основі декомпозиції введено домінуючу складову наскрізної затримки та встановлено відповідність між її типом і локальним, алгоритмічним або системним рівнем паралельної організації. Наукова новизна полягає у формалізації в межах єдиної математичної моделі взаємозв'язку між характеристиками потоку, складовими наскрізної затримки, пропускною здатністю та рівнями паралельної організації, що дає змогу обґрунтовано визначати напрям паралельного перетворення і повторно оцінювати структуру затримки після його застосування. Практичну придатність моделі проілюстровано на потоковій системі аналізу МРТ-зображень: визначення локальної складової як домінуючої обґрунтувало застосування багатопотокової попередньої обробки, що зменшило час проходу з 9,96 с до 1,09 с і підвищило пропускну здатність з 705,12 до 6443,12 зображення/с без погіршення точності класифікації. Запропонована модель не прив'язана до конкретного типу потокових даних або апаратної платформи та може використовуватися як формальна основа вибору рівня паралельної обробки в багатостадійних системах штучного інтелекту реального часу.

Ключові слова: потокова обробка даних, системи штучного інтелекту, наскрізна затримка, багаторівнева паралельність, пропускна здатність, математична модель.

Вступ. Інтелектуальні системи реального часу формують окремий клас обчислювальних систем, у яких результат повинен бути не лише коректним з погляду прикладної задачі, а й отриманим у допустимому часовому інтервалі [1]. Така вимога є принциповою для систем комп'ютерного зору, медичної діагностики, автономного керування, робототехніки, технічного контролю, сенсорної аналітики, аудіообробки та підтримки прийняття рішень. У цих застосуваннях дані надходять у вигляді потоку порцій, а система має підтримувати сталий темп їх опрацювання без неконтрольованого накопичення черги [2], [3].

Сучасна інтелектуальна система дедалі рідше зводиться до однієї моделі штучного інтелекту. На практиці результат формується багатостадійною схемою, яка включає надходження даних, попередню підготовку, нормалізацію, виділення ознак,

застосування моделі, постобробку, агрегацію та передавання результату. Через це повна затримка визначається не тільки часом виконання нейромережевої або іншої моделі, а взаємодією локальних обчислень, алгоритмічних залежностей і системної організації виконання [4], [5].

У сучасних потокових системах значну увагу приділено питанням пропускної здатності, затримки, масштабованості, відмовостійкості та керування навантаженням [6], [7], [8]. Водночас у задачах штучного інтелекту додатково виникає вимога збереження цільових показників результату: точності класифікації, точності сегментації, стійкості оцінювання або коректності керувального рішення. Це означає, що оптимізація часу не може розглядатися ізольовано від функціональної коректності обробки.

Розвинено локальні методи прискорення, зокрема багатопотокове виконання, використання графічних процесорів, оптимізація доступу до пам'яті та асинхронні передавання між центральним і графічним процесорами [9], [10]. Окремий напрям становлять методи алгоритмічної перебудови, у яких зменшення часу досягається не лише за рахунок швидшого виконання операцій, а через зміну структури обчислювальної схеми, скорочення критичного шляху або виділення незалежних гілок. На системному рівні досліджуються розподілене виконання, керування чергами, синхронізація, диспетчеризація, балансування навантаження і масштабування [11], [12].

Проте для потокових систем штучного інтелекту реального часу принциповим є не сам факт використання певного виду паралельності, а відповідність обраного механізму природі тієї складової наскрізної затримки, яка фактично обмежує систему. Прискорення локального оператора може не зменшити повну затримку, якщо обмеження визначається критичним шляхом алгоритму. Аналогічно, алгоритмічна перебудова може не забезпечити зменшення наскрізної затримки, якщо система обмежена комунікаціями, синхронізацією або чергами. Отже, необхідна математична модель, яка формалізує взаємозв'язок між характеристиками потоку, складовими наскрізної затримки, пропускною здатністю та рівнями паралельної організації обчислень.

Відомі підходи до організації потокових і паралельних обчислень розв'язують окремі складові цієї задачі. Модель Dataflow [6] зосереджена насамперед на організації опрацювання необмежених потоків даних з урахуванням часової семантики, коректності, затримки та вартості; системи на основі дискретизованих потоків [7] – на масштабованому та відмовостійкому потоковому виконанні; InferLine [5] – на виборі ресурсів, реплікації, пакетуванні та масштабуванні етапів конвеєра прогнозування для виконання заданих обмежень на затримку; підходи до керування ресурсами в розподілених потокових системах [13] охоплюють виділення ресурсів, паралелізацію операторів і планування завдань. На відміну від зазначених підходів, у запропонованій моделі вихідною ознакою вибору механізму паралельного перетворення є не наперед заданий тип ресурсу, спосіб масштабування або організації потоку, а природа домінуючої складової наскрізної затримки. Для цього в єдиній математичній моделі локальну, алгоритмічну та системну складові зіставлено з відповідними рівнями паралельної організації, а після застосування перетворення передбачено повторне визначення домінуючої складової, що дає змогу виявляти перенесення часового обмеження на інший рівень.

Метою роботи є розроблення математичної моделі узгодженої багаторівневої паралельної обробки потокових даних у системах штучного інтелекту реального часу, яка шляхом декомпозиції наскрізної затримки на локальну, алгоритмічну та системну складові забезпечує визначення домінуючої складової, вибір відповідного рівня паралельної організації та перевірку часово-пропускних обмежень за умови збереження заданих показників результативності інтелектуального опрацювання.

Виклад основного матеріалу дослідження. Нехай задано потік порцій даних $X = \{X_k\}_{k=1}^{\infty}$, де X_k – k-та порція потоку. Порцією може бути кадр зображення, фрагмент аудіосигналу, блок сенсорних вимірювань, телеметричний пакет або інший дискретний елемент, що надходить до інтелектуальної системи для розпізнавання, класифікації, оцінювання стану або формування керувального рішення.

Інтенсивність надходження порцій позначимо через λ. Тоді середній інтервал між надходженнями визначається як $\Delta_{in} = \frac{1}{\lambda}$.

Обробку порції подамо як багатоетапну схему

$$X_k \to F_1(X_k) \to F_2(X_k) \to \ldots \to F_M(X_k), \tag{1}$$

де F_j – j-й етап обробки, а T_j – час його виконання. Якщо всі етапи виконуються послідовно без перекриття, стаціонарний інтервал формування результатів дорівнює сумі тривалостей етапів:

$$\Delta_{seq} = \sum_{j=1}^{M} T_j. \tag{2}$$

Якщо етапи організовані конвеєрно або асинхронно, то після фази заповнення конвеєра стаціонарний інтервал у граничному випадку визначається найдовшою стадією:

$$\Delta_{pipe} = \max_{1 \le j \le M} T_j. \quad (3)$$

Пропускна здатність системи становить

$$\mu = \frac{1}{\Delta}, \quad (4)$$

де Δ – стаціонарний інтервал формування результатів. Умова працездатності потокової системи в режимі, наближеному до реального часу, має вигляд

$$\lambda \le \mu. \quad (5)$$

Якщо ця умова не виконується, порції накопичуються в черзі, а затримка зростає навіть тоді, коли окремі оператори виконуються швидко. Це узгоджується з базовими положеннями теорії черг і законом Літтла, який пов'язує середню довжину черги, інтенсивність надходження та час очікування [13].

Для подальшого аналізу наскрізну затримку формування результату подамо як

$$T_{end} = T_{loc} + T_{alg} + T_{sys}, \quad (6)$$

де T_{loc} – локальна складова, T_{alg} – алгоритмічна складова, T_{sys} – системна складова. Задача полягає у побудові математичної моделі узгодженої багаторівневої паралельної обробки, яка забезпечує зменшення T_{end}, виконання умови (5) та збереження заданих цільових метрик інтелектуальної обробки.

Сукупність співвідношень (1)–(6) задає базову частину математичної моделі, у якій стан потокового виконання описується інтенсивністю надходження порцій, стаціонарним інтервалом формування результатів, пропускною здатністю та складовими наскрізної затримки. Співвідношення (7)–(11) встановлюють правило визначення домінуючої часової складової та формалізують локальний, алгоритмічний і системний рівні її утворення. Таким чином, модель пов'язує вхідні характеристики потоку, структуру обчислювального процесу та часові обмеження з рівнем паралельної організації, на якому доцільно виконувати перетворення.

Ключовим положенням запропонованої математичної моделі є принцип домінуючої затримки. Він полягає в тому, що зменшення повної затримки системи досягається не довільним прискоренням окремих операцій, а цілеспрямованим впливом на ту складову, яка в поточній конфігурації обмежує пропускну здатність або формує найбільший внесок у час отримання результату.

Домінуючу складову визначимо як

$$T_{dom} = \max\{T_{loc}, T_{alg}, T_{sys}\}. \quad (7)$$

Якщо $T_{dom} = T_{loc}$, то обмеження має локальну природу і пов'язане з внутрішньовузловими обчисленнями, пам'яттю, передаваннями між пристроями або локальною синхронізацією. Якщо $T_{dom} = T_{alg}$, то обмеження визначається структурою залежностей і критичним шляхом обчислювальної схеми. Якщо $T_{dom} = T_{sys}$, то повний час визначається системними чинниками: комунікаціями, чергами, диспетчеризацією, глобальною синхронізацією або нерівномірністю навантаження.

Звідси випливає, що кожен тип домінуючої затримки потребує власного рівня паралельної організації. Локальне прискорення не усуває довгий критичний шлях, а алгоритмічна перебудова не усуває системних втрат, якщо головним обмеженням є синхронізація або комунікації. Тому вибір паралельної організації має ґрунтуватися на часовій природі обмеження. Класичні закони Амдала і Густафсона [11], [12] також показують, що фактичний результат масштабування визначається не лише швидкістю виконання окремої паралельної частини, а співвідношенням між паралельними, послідовними та накладними складовими часу.

У запропонованій математичній моделі виділено три взаємопов'язані рівні: локальний, алгоритмічний і системний. Кожен рівень визначає власний об'єкт впливу, одиницю паралельності, механізм реалізації та критерій оцінювання часових характеристик.

Локальний рівень спрямований на зменшення внутрішньовузлової затримки. Його об'єктом є операції над порцією даних або її елементами в межах одного обчислювального вузла. До таких операцій належать фільтрація, нормалізація, локальні перетворення, виділення ознак, обчислювальні ядра, доступ до пам'яті та передавання між центральним і графічним процесорами. Одиницею паралельності є елемент, блок, фрагмент або порція потоку. Основними механізмами є паралельність за даними, багатопотокове виконання, векторизація, графічне прискорення, асинхронні передавання, буферизація та внутрішньовузлова конвеєризація.

$$T_{loc} = \max\left\{T_{comp}, T_{mem}, T_{transfer}, T_{sync}\right\}, \quad (8)$$

де T_{comp} – час локальних обчислень, T_{mem} – витрати доступу до пам'яті, $T_{transfer}$ – час передавання між пристроями, T_{sync} – витрати локальної синхронізації.

Алгоритмічний рівень спрямований на зменшення затримки, що виникає через структуру обчислювального процесу. Його об'єктом є граф залежностей між операціями, модулями, шарами, гілками або етапами формування результату. Якщо обчислювальний процес подано графом

$$G = (V, E), \tag{9}$$

де V – множина операцій, а E – множина залежностей, то алгоритмічна затримка визначається критичним шляхом

$$T_{alg} = \max_{\pi \in \mathrm{P}(G)} \sum_{v \in \pi} T(v), \tag{10}$$

де $\mathrm{P}(G)$ – множина шляхів у графі залежностей. Одиницею паралельності на цьому рівні є модуль, гілка, шар, підзадача або макрокрок. Основними механізмами є декомпозиція, виділення незалежних або слабкозалежних гілок, перебудова графа залежностей, пакетна організація та скорочення критичного шляху.

Системний рівень спрямований на зменшення затримки, що виникає внаслідок взаємодії компонентів системи. Його об'єктом є процеси, вузли, пристрої, черги, репліки моделей, канали передавання, диспетчери виконання та механізми синхронізації. Системну складову можна подати як

$$T_{sys} = \max_{q} T_{comp}^{(q)} + T_{comm} + T_{sync} + T_{sched} + T_{queue}, \tag{11}$$

де $T_{comp}^{(q)}$ – час локального обчислення на q-му ресурсі, T_{comm} – час комунікацій, T_{sync} – час синхронізації, T_{sched} – витрати диспетчеризації, T_{queue} – час очікування в чергах. Системний рівень є визначальним тоді, коли локальні обчислення й алгоритмічна структура вже не формують основного обмеження, а повна затримка визначається взаємодією компонентів.

На основі співвідношень математичної моделі формується послідовна процедура її застосування.

1. Формалізується потік порцій X_k, інтенсивність надходження λ, допустимий інтервал обробки та цільові метрики результату.

2. Багатостадійний процес подається у вигляді сукупності етапів $F_1, ..., F_M$ для яких визначаються тривалості $T_1, ..., T_M$ і залежності між ними.

3. Наскрізна затримка декомпонується на складові T_{loc}, T_{alg}, T_{sys}.

4. Визначається домінуюча складова T_{dom} за формулою (7).

5. Обирається відповідний рівень паралельної організації. За домінування локальної складової застосовуються локальні механізми паралельності; за домінування алгоритмічної – перебудова графа залежностей; за домінування системної – зміна організації взаємодії компонентів.

6. Після оптимізації виконується повторна оцінка $T'_{end} = T'_{loc} + T'_{alg} + T'_{sys}$.

7. Перевіряється, чи не перенесено вузьке місце на інший рівень. Якщо новою домінуючою стає інша складова, процедура повторюється.

8. Остаточно перевіряється умова потокової працездатності $\lambda \le \mu'$, де μ' – досягнута пропускна здатність після оптимізації. Оптимізація вважається коректною лише тоді, коли поряд зі зменшенням часових характеристик зберігаються або не знижуються цільові метрики, визначені постановкою задачі.

Ключовим результатом запропонованої математичної моделі є класифікаційно-проєктна відповідність між типом домінуючої затримки, рівнем паралельної організації та прогнозованою зміною часових характеристик системи. Така відповідність наведена в таблиці 1.

Таблиця 1

Відповідність між домінуючою затримкою та рівнем паралельної організації

Тип домінуючої затримки	Що обмежує систему	Рівень паралельної організації	Що змінюється	Досяжна зміна часових характеристик
Локальна	Локальні обчислення, доступ до пам'яті, передавання між пристроями, локальна синхронізація	Локальний	Спосіб виконання операцій у межах вузла	Зменшення внутрішньовузлової затримки та стабілізація стаціонарного кроку
Алгоритмічна	Критичний шлях, залежності між етапами, структура моделі або процедури	Алгоритмічний	Граф залежностей, порядок виконання, декомпозиція	Скорочення критичного шляху формування результату
Системна	Комунікації, синхронізація, черги, диспетчеризація, глобальна агрегація	Системний	Організація взаємодії компонентів	Збереження масштабованості та запобігання накопиченню черг

Таблиця 1 має не лише класифікаційне, а й проєктне значення. Вона дозволяє встановити, який рівень паралельності доцільно застосовувати залежно від того, яка складова часу фактично обмежує роботу системи. Це дає змогу уникати неефективної оптимізації, коли прискорюється стадія, яка не визначає повну затримку.

Практичну інтерпретацію математичної моделі доцільно розглядати через типові режими домінування затримки. Такий підхід робить модель незалежною від конкретного типу даних, моделі або предметної області.

Перший режим відповідає домінуванню локальної затримки. Він виникає тоді, коли повний час роботи системи обмежений локальними операціями в межах одного вузла: обчисленнями над елементами порції, доступом до пам'яті, передаваннями між пристроями або простоєм між локальними стадіями. У цьому випадку доцільними є паралельність за даними, багатопотокове виконання, графічне прискорення, буферизація та асинхронна конвеєризація. Досягнення цільового часово-обчислювального результату підтверджується зменшенням локальної затримки та стабілізацією темпу обробки порцій.

Другий режим відповідає домінуванню алгоритмічної затримки. Він характерний для систем, у яких основне обмеження визначається не швидкістю окремого оператора, а структурою залежностей між етапами. У такому випадку доцільним є перехід до алгоритмічного рівня: перебудова графа залежностей, виділення незалежних гілок, декомпозиція моделі, пакетна організація або скорочення кількості послідовно залежних етапів. Досягнення цільового часово-обчислювального результату підтверджується скороченням критичного шляху без порушення коректності обчислювальної процедури.

Третій режим відповідає домінуванню системної затримки. Він виникає тоді, коли локальні обчислення й алгоритмічна структура не є головним обмеженням, а повна затримка визначається взаємодією компонентів. Основними обмеженнями стають синхронізація, комунікації, диспетчеризація, нерівномірний розподіл навантаження, колективні операції або накопичення черг. У цьому випадку необхідними є системне узгодження темпу роботи компонентів, балансування навантаження, мінімізація накладних витрат і контроль пропускної здатності.

Таким чином, запропонована математична модель не прив'язана до конкретного класу даних або прикладної галузі. Вона задає універсальний порядок дій для багатостадійних потокових систем: визначити природу домінуючої затримки, обрати відповідний рівень паралельності, застосувати механізм зменшення саме цієї складової та перевірити, чи не виникло нове обмеження на іншому рівні.

Запропонована математична модель відрізняється від наявних підходів до потокових і паралельних обчислень тим, що об'єднує в межах однієї формалізації три різні джерела наскрізної затримки та встановлює правило вибору рівня паралельного перетворення за її домінуючою складовою. На відміну від підходів, орієнтованих переважно на організацію потокового виконання, виділення та масштабування ресурсів, планування завдань або прискорення окремих обчислювальних операцій, центральним поняттям запропонованої моделі є домінуюча складова наскрізної затримки. При цьому локальна, алгоритмічна та системна складові розглядаються не як незалежні напрями оптимізації, а як взаємопов'язані складові єдиного часового обмеження системи. Повторне оцінювання їх співвідношення після кожного перетворення дає змогу виявити перенесення обмеження на інший рівень і відповідно змінити напрям подальшої паралельної організації.

Якщо обмеження визначається критичним шляхом, то прискорення окремого ядра матиме обмежений внесок у зменшення наскрізної затримки. Якщо система обмежена синхронізацією або чергами, то навіть істотне прискорення локальних обчислень може не покращити стаціонарну пропускну здатність. Тому математична модель передбачає повторну оцінку після кожного етапу оптимізації. Це дає змогу виявити перенесення лімітувального компонента на інший рівень і перейти до відповідного механізму паралельної організації.

Практична цінність математичної моделі полягає в тому, що вона задає єдиний підхід до проєктування потокових інтелектуальних систем: спочатку аналізується структура затримки, далі визначається рівень паралельної організації, після чого оцінюється виконання умови потокової працездатності та збереження заданих цільових метрик інтелектуальної обробки. Прикладні реалізації окремих локальних механізмів для обробки зображень і контурно-орієнтованого підсилення наведено в [14], [15], [16], однак запропонована математична модель не обмежується цими прикладами й описує ширший принцип вибору рівня паралельної організації.

Для ілюстрації практичного застосування запропонованої математичної моделі розглянуто потокову інтелектуальну систему аналізу МРТ-зображень головного мозку. Як вхідні дані використано відкритий набір даних Brain Tumor MRI Dataset, розміщений на платформі Kaggle [17]. Набір містить 7023 МРТ-зображення, розподілені на чотири діагностичні класи: гліома, менінгіома, пухлина гіпофіза та відсутність пухлини. Дані поділено на навчальну і тестову підмножини: 5712 та 1311 зображень відповідно. Кожна порція потоку в розглянутій системі відповідає одному МРТ-зображенню, а результатом обробки є клас діагностичної категорії, сформований нейромережевою моделлю.

Обчислювальний контур системи включав надходження зображення, приведення його до узгодженого формату, фільтраційну підготовку, нормалізацію, формування тензорного подання, застосування класифікатора на основі згорткової нейронної мережі та формування результату класифікації. Як цільову модель використано згорткову нейронну мережу з чотирма згортковими шарами, шарами максимального агрегування, повнозв'язним шаром, регуляризацією шляхом випадкового вимкнення частини нейронів та вихідним шаром на чотири класи. Навчання виконувалося з використанням оптимізатора Adam, функції втрат на основі розрідженої категоріальної перехресної ентропії та показника точності класифікації.

Відповідно до запропонованої математичної моделі спочатку було визначено домінуючу складову затримки. У розглянутому прикладі нею була локальна складова, пов'язана з підготовкою зображень, фільтраційними перетвореннями, нормалізацією та формуванням порцій даних для подальшого застосування нейромережевої моделі. Тому для зменшення затримки було обрано локальний рівень паралельної організації, а одиницею паралельності визначено окреме зображення або групу зображень потоку. Механізмом реалізації стала параметрично керована багатопотокова попередня обробка на центральному процесорі з узгодженням темпу подачі даних у нейромережевий модуль.

Обчислювальні експерименти виконано на робочій станції з процесором Intel Core i7-7700, 32 ГБ оперативної пам'яті та графічним процесором NVIDIA Tesla T4. Паралельність попередньої обробки задавалася кількістю робочих потоків. Для порівняння оцінювалися час проходу, еквівалентного одній епосі навчання, пропускна здатність, прискорення, відносне зменшення затримки, точність класифікації та значення функції втрат.

Отримані результати, наведені в [16], підтвердили доцільність вибору саме локального рівня паралельної організації. Перехід від базового режиму без паралельності до режиму з чотирма робочими потоками зменшив час проходу, еквівалентного одній епосі навчання, з 9,96 с до 1,09 с, що відповідає скороченню затримки на 89,06 % та зростанню пропускної здатності з 705,12 до 6443,12 зображення/с. Для фільтраційної стадії перехід від одного до чотирьох потоків забезпечив прискорення у 2,60 раза, а до восьми потоків – у 2,99 раза, після чого спостерігалося насичення прискорення.

Важливо, що зменшення локальної затримки не супроводжувалося погіршенням результативності інтелектуального аналізу. Точність класифікації у базовому режимі становила 92,45 %, тоді як після застосування обраного режиму підготовки даних вона зросла до 94,58 %, а фінальна конфігурація досягла точності 95,42 % за значення функції втрат 0,1283. Отже, наведений приклад показує, що запропонована математична модель дає змогу визначити домінуючу складову затримки, обрати відповідний рівень паралельної організації та перевірити, що зменшення часу обробки не призводить до погіршення цільових показників інтелектуальної системи.

Висновки. Розроблено математичну модель узгодженої багаторівневої паралельної обробки потокових даних у системах штучного інтелекту реального часу. Модель ґрунтується на формалізації потоку порцій, стаціонарного інтервалу формування результатів і пропускної здатності, а також на декомпозиції наскрізної затримки на локальну, алгоритмічну та системну складові. Введене правило визначення домінуючої складової встановлює формальний зв'язок між джерелом часових обмежень і відповідним рівнем паралельної організації обчислень.

Наукова новизна одержаного результату полягає у формалізації в межах єдиної математичної моделі взаємозв'язку між інтенсивністю надходження порцій, пропускною здатністю, локальною, алгоритмічною та системною складовими наскрізної затримки і відповідними рівнями паралельної організації. На відміну від підходів, у яких окремо розглядаються прискорення обчислень, перебудова обчислювальної схеми, масштабування ресурсів або організація потокового виконання, у запропонованій моделі введено правило вибору рівня паралельного перетворення за домінуючою складовою наскрізної затримки та передбачено її повторне визначення після внесення змін. Це дає змогу виявляти перенесення часового обмеження між рівнями та послідовно узгоджувати застосовані механізми паралельної організації з фактичним джерелом затримки.

Показано, що локальний рівень відповідає домінуванню внутрішньовузлових обчислень, доступу до пам'яті, передавань і локальної синхронізації; алгоритмічний – домінуванню критичного шляху та структурних залежностей; системний – домінуванню комунікацій, глобальної синхронізації, диспетчеризації та черг. Умова потокової працездатності в моделі визначається співвідношенням між інтенсивністю надходження порцій і досяжною пропускною здатністю системи.

Практична перевірка на потоковій системі аналізу МРТ-зображень підтвердила коректність використання моделі для вибору рівня паралельного перетворення. Визначення локальної складової як домінуючої обґрунтувало багатопотокову попередню обробку, що зменшила час проходу з 9,96 с до 1,09 с та підвищила пропускну здатність з 705,12 до 6443,12 зображення/с без погіршення точності класифікації.

Перспективи подальших досліджень пов'язані з розширенням математичної моделі на нестаціонарні та адаптивні потокові режими, у яких інтенсивність надходження даних, доступні обчислювальні ресурси й домінуюча складова наскрізної затримки можуть змінюватися в часі.

Література

1. Саламун К., Павіч І., Дзапо Х., Чуляк І. Слабко жорстка модель реального часу для систем керування: огляд. Sensors. 2023. Т. 23, № 10. С. 4652. DOI: https://doi.org/10.3390/s23104652.
2. Гомес Е., Коста Ф., Де Ролт К., Пленц П., Дантас М. Огляд застосувань від реального до майже реального часу в середовищах туманних обчислень. Telecom. 2021. Т. 2, № 4. С. 489–517. DOI: https://doi.org/10.3390/telecom2040028.
3. Сінгх Р., Гілл С. С. Edge AI: огляд. Internet of Things and Cyber-Physical Systems. 2023. Т. 3. С. 71–92. DOI: https://doi.org/10.1016/j.iotcps.2023.02.004.
4. Сзе В., Чен Ю.-Х., Янг Т.-Дж., Емер Дж. С. Ефективне опрацювання глибоких нейронних мереж: навчальний матеріал та огляд. Proceedings of the IEEE. 2017. Т. 105, № 12. С. 2295–2329. DOI: https://doi.org/10.1109/JPROC.2017.2761740.
5. Кренкшоу Д. та ін. InferLine: урахування затримки під час виділення ресурсів і масштабування конвеєрів прогнозування. Proceedings of the 11th ACM Symposium on Cloud Computing. 2020. С. 477–491. DOI: https://doi.org/10.1145/3419111.3421285.
6. Акідау Т. та ін. Модель Dataflow: практичний підхід до узгодження коректності, затримки та вартості в масштабній необмеженій потоковій обробці з невпорядкованими даними. Proceedings of the VLDB Endowment. 2015. Т. 8, № 12. С. 1792–1803. DOI: https://doi.org/10.14778/2824032.2824076.
7. Захарія М., Дас Т., Лі Х., Хантер Т., Шенкер С., Стойка І. Дискретизовані потоки: відмовостійкі потокові обчислення у масштабі. Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles. 2013. С. 423–438. DOI: https://doi.org/10.1145/2517349.2522737.
8. Квінтареллі Е., Рабосіо Е., Танка Л. Ефективне використання контекстного впливу для рекомендації нових елементів тимчасовим групам. Information Systems. 2019. Т. 84. С. 197–213. DOI: https://doi.org/10.1016/j.is.2019.05.003.
9. Міттал С., Веттер Дж. С. Огляд методів гетерогенних обчислень CPU–GPU. ACM Computing Surveys. 2015. Т. 47, № 4. С. 1–35. DOI: https://doi.org/10.1145/2788396.
10. Хестнесс Дж., Кеклер С. В., Вуд Д. А. Неефективність конвеєрів GPU-обчислень і можливості оптимізації в гетерогенних процесорах CPU–GPU. 2015 IEEE International Symposium on Workload Characterization. 2015. С. 87–97. DOI: https://doi.org/10.1109/IISWC.2015.15.
11. Шраєн Г. Прискорення та ефективність обчислювального розпаралелювання: узагальнювальний підхід і асимптотичний аналіз. Journal of Parallel and Distributed Computing. 2024. Т. 187. С. 104835. DOI: https://doi.org/10.1016/j.jpdc.2023.104835.
12. Кемерон К. В. За межами закону Амдала: як вимірювання й моделювання співвідношення енергоспоживання та продуктивності в масштабі визначають проєктування серверів і суперкомп'ютерів. Journal of Computer Science and Technology. 2023. Т. 38, № 1. С. 80–86. DOI: https://doi.org/10.1007/s11390-022-2950-7.
13. Лю С., Буйя Р. Керування ресурсами та планування в розподілених системах потокової обробки: таксономія, огляд і перспективні напрями. ACM Computing Surveys. 2021. Т. 53, № 3. С. 1–41. DOI: https://doi.org/10.1145/3355399.
14. Мочурад Л. Підхід до підвищення точності семантичної сегментації рентгенівських зображень грудної клітки шляхом інтеграції детектування контурів і глибокого навчання. Frontiers in Artificial Intelligence. 2025. Т. 8. С. 1522730. DOI: https://doi.org/10.3389/frai.2025.1522730.
15. Мочурад Л. І. Аналіз детектування контурів Canny на основі паралельного алгоритму, побудованої шкали складності та CUDA. Computing and Informatics. 2022. Т. 41, № 4. С. 957–980. DOI: https://doi.org/10.31577/cai_2022_4_957.
16. Мочурад Л., Мельничук К., Мочурад Ю. Оптимізація попередньої обробки МРТ головного мозку з використанням паралельних обчислень для ефективної класифікації пухлин мозку. The Journal of Supercomputing. 2026. Т. 82, № 3. С. 123. DOI: https://doi.org/10.1007/s11227-026-08282-2.
17. Нікпарвар М. Brain Tumor MRI Dataset. Kaggle. 2025. URL: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset.

References

1. K. Salamun, I. Pavić, H. Džapo, and I. Čuljak, "Weakly Hard Real-Time Model for Control Systems: A Survey," Sensors, vol. 23, no. 10, p. 4652, May 2023, doi: https://doi.org/10.3390/s23104652.
2. E. Gomes, F. Costa, C. De Rolt, P. Plentz, and M. Dantas, "A Survey from Real-Time to Near Real-Time Applications in Fog Computing Environments," Telecom, vol. 2, no. 4, pp. 489–517, Dec. 2021, doi: https://doi.org/10.3390/telecom2040028.
3. R. Singh and S. S. Gill, "Edge AI: A survey," Internet of Things and Cyber-Physical Systems, vol. 3, pp. 71–92, 2023, doi: https://doi.org/10.1016/j.iotcps.2023.02.004.
4. V. Sze, Y.-H. Chen, T.-J. Yang, and J. S. Emer, "Efficient Processing of Deep Neural Networks: A Tutorial and Survey," Proc. IEEE, vol. 105, no. 12, pp. 2295–2329, Dec. 2017, doi: https://doi.org/10.1109/JPROC.2017.2761740.
5. D. Crankshaw et al., "InferLine: latency-aware provisioning and scaling for prediction serving pipelines," in Proceedings of the 11th ACM Symposium on Cloud Computing, Virtual Event USA: ACM, Oct. 2020, pp. 477–491. doi: https://doi.org/10.1145/3419111.3421285.
6. T. Akidau et al., "The dataflow model: a practical approach to balancing correctness, latency, and cost in massive-scale, unbounded, out-of-order data processing," Proc. VLDB Endow., vol. 8, no. 12, pp. 1792–1803, Aug. 2015, doi: https://doi.org/10.14778/2824032.2824076.
7. M. Zaharia, T. Das, H. Li, T. Hunter, S. Shenker, and I. Stoica, "Discretized streams: fault-tolerant streaming computation at scale," in Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles, Farminton Pennsylvania: ACM, Nov. 2013, pp. 423–438. doi: https://doi.org/10.1145/2517349.2522737.
8. E. Quintarelli, E. Rabosio, and L. Tanca, "Efficiently using contextual influence to recommend new items to ephemeral groups," Information Systems, vol. 84, pp. 197–213, Sep. 2019, doi: https://doi.org/10.1016/j.is.2019.05.003.
9. S. Mittal and J. S. Vetter, "A Survey of CPU-GPU Heterogeneous Computing Techniques," ACM Comput.

Surv., vol. 47, no. 4, pp. 1–35, Jul. 2015, doi: https://doi.org/10.1145/2788396.
10. J. Hestness, S. W. Keckler, and D. A. Wood, “GPU Computing Pipeline Inefficiencies and Optimization Opportunities in Heterogeneous CPU-GPU Processors,” in 2015 IEEE International Symposium on Workload Characterization, Atlanta, GA, USA: IEEE, Oct. 2015, pp. 87–97. doi: https://doi.org/10.1109/IISWC.2015.15.
11. G. Schryen, “Speedup and efficiency of computational parallelization: A unifying approach and asymptotic analysis,” Journal of Parallel and Distributed Computing, vol. 187, p. 104835, May 2024, doi: https://doi.org/10.1016/j.jpdc.2023.104835.
12. K. W. Cameron, “Adventures Beyond Amdahl’s Law: How Power-Performance Measurement and Modeling at Scale Drive Server and Supercomputer Design,” J. Comput. Sci. Technol., vol. 38, no. 1, pp. 80–86, Feb. 2023, doi: https://doi.org/10.1007/s11390-022-2950-7.
13. X. Liu and R. Buyya, “Resource Management and Scheduling in Distributed Stream Processing Systems: A Taxonomy, Review, and Future Directions,” ACM Comput. Surv., vol. 53, no. 3, pp. 1–41, May 2021, doi: https://doi.org/10.1145/3355399.
14. L. Mochurad, “Approach for enhancing the accuracy of semantic segmentation of chest X-ray images by edge detection and deep learning integration,” Front. Artif. Intell., vol. 8, p. 1522730, Apr. 2025, doi: https://doi.org/10.3389/frai.2025.1522730.
15. L. I. Mochurad, “Canny Edge Detection Analysis Based on Parallel Algorithm, Constructed Complexity Scale and CUDA,” cai, vol. 41, no. 4, pp. 957–980, 2022, doi: https://doi.org/10.31577/cai_2022_4_957.
16. L. Mochurad, K. Melnychuk, and Y. Mochurad, “Optimization of brain MRI preprocessing using parallel computing for efficient brain tumour classification,” J Supercomput, vol. 82, no. 3, p. 123, Feb. 2026, doi: https://doi.org/10.1007/s11227-026-08282-2.
17. M. Nickparvar, “Brain Tumor MRI Dataset.” Jan. 10, 2025. [Online]. Available: https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset

Hentosh L.I. Mathematical model of coordinated multilevel parallel processing of streaming data in real-time artificial intelligence systems

A mathematical model of coordinated multilevel parallel processing of streaming data in real-time artificial intelligence systems is proposed to reduce end-to-end latency while preserving specified performance indicators of intelligent processing. The relevance of the study follows from the fact that modern artificial intelligence systems operate as multistage streaming computational pipelines in which timing constraints are determined not only by the execution time of an individual model but also by the combined influence of local computations and data transfers, algorithmic dependencies, communications, synchronization, scheduling, and queues. In the proposed model, the input is represented as a sequence of discrete data portions with a specified arrival rate, whereas the processing of each portion is represented as a set of interrelated stages with defined timing characteristics. The steady-state interval between consecutive outputs and the system throughput are formalized, and the streaming operability condition is established, requiring the achievable throughput to be not lower than the incoming portion rate. End-to-end latency is decomposed into local, algorithmic, and system components. The local component characterizes intra-node computations, memory access, device-to-device transfers, and local synchronization; the algorithmic component is determined by the dependency graph and its critical path; the system component covers communication, global synchronization, scheduling, and queueing delays. Based on this decomposition, the dominant latency component is introduced and a correspondence is established between its type and the local, algorithmic, or system level of parallel organization. The scientific novelty consists in formalizing, within a single mathematical model, the relationship among stream characteristics, end-to-end latency components, throughput, and levels of parallel organization. This relationship enables a justified selection of the target level for parallel transformation and repeated evaluation of the latency structure after the transformation has been applied. The practical applicability of the model is illustrated using a streaming brain MRI analysis system. Identification of the local component as dominant justified multithreaded preprocessing, which reduced the processing pass time from 9.96 s to 1.09 s and increased throughput from 705.12 to 6443.12 images/s without degrading classification accuracy. The proposed model is independent of a specific streaming-data type or hardware platform and can be used as a formal basis for selecting a parallel-processing level in multistage real-time artificial intelligence systems.

Keywords: *streaming data processing, artificial intelligence systems, end-to-end latency, multilevel parallelism, throughput, mathematical model.*

Гентош Леся Ігорівна – кандидат технічних наук, доцент, доцент кафедри систем штучного інтелекту, Національний університет «Львівська політехніка».
ORCID: https://orcid.org/0000-0002-4957-1512
e-mail: lesia.i.mochurad@lpnu.ua

Дата першого надходження статті 28.07.2026.
Дата прийняття статті до друку після рецензування 26.08.2026.
Дата публікації 30.09.2026.

Стаття з відкритим доступом, відповідно до умов ліцензії **Creative Commons (CC BY 4.0)**