

Волк М.О.

Харківський національний університет радіоелектроніки

Бугрій А.М.

Харківський національний університет радіоелектроніки

Ковтун Є.І.

Харківський національний університет радіоелектроніки

Самойлов І.А.

Харківський національний університет радіоелектроніки

Брестовицький Р.М.

Харківський національний університет радіоелектроніки

Соробей Б.В.

Харківський національний університет радіоелектроніки

МЕТОДИ РОЗПОДІЛЕННЯ ОБЧИСЛЮВАЛЬНИХ ЗАДАЧ У ГЕТЕРОГЕННИХ КОМП'ЮТЕРНИХ СИСТЕМАХ

Гетерогенні обчислювальні системи, які поєднують у собі обчислювальні елементи різної архітектури – центральні процесори (CPU), графічні процесори (GPU), матриці FPGA, системи на кристалі (SoC), спеціалізовані інтегральні схеми (ASIC), а також периферійні edge пристрої – стають основою сучасної інфраструктури розподілених обчислень. Така інфраструктура потребує нових підходів до ефективного розподілу обчислювальних задач між обчислювальними елементами з різними обчислювальними характеристиками, енергоспоживанням, архітектурною сумісністю та пропускну здатністю мережевого з'єднання.

У статті здійснено ґрунтовний огляд сучасних методів розподілення обчислювальних задач у гетерогенних комп'ютерних системах, що поєднують ресурси з різними архітектурами, продуктивністю та енергетичними характеристиками. Зважаючи на швидке зростання обсягів даних, різноманітність обчислювальних завдань і необхідність їх паралельного виконання, задача ефективного планування в подібних середовищах набуває критичного значення. Розглянуто класичні алгоритми, евристичні та метаевристичні підходи, включаючи HEFT, Min-Min, генетичні алгоритми (GA), рої частинок (PSO), мурашині алгоритми (ACO), а також сучасні методи глибокого навчання, зокрема методи з підкріпленням (Deep RL), що демонструють значну адаптивність до змін середовища.

Особлива увага приділяється energy-aware методам, які дозволяють досягати балансу між продуктивністю та енергоспоживанням, що є особливо актуальним для IoT-пристроїв, мобільних обчислень і edge-середовищ. У статті запропоновано класифікацію задач за архітектурними (тип зв'язності, залежності) та прикладними (реальний час, HPC, AI) ознаками, що дозволяє встановити доцільність застосування конкретних моделей планування. Наведені у статті порівняльні таблиці методів дозволяють фахівцям обрати найдоцільніший підхід до планування. Отримані результати можуть бути корисними при побудові систем автономного управління завданнями у хмарних, fog- або edge-середовищах.

Ключові слова: гетерогенні обчислювальні системи, розподілення задач, машинне навчання, методи розподілу ресурсів, хмарні системи, енергоефективність, евристичні методи.

Постановка проблеми. З ростом обсягу даних, розвитку штучного інтелекту та переходу до edge-обчислень, з'явилася потреба обробляти величезні обсяги інформації в розподіленому, але енергоефективному середовищі. Гетерогенні комп'ютерні системи дозволяють масштабувати обчислення та оптимізувати продуктивність за рахунок використання спеціалізованих пристроїв. Наприклад: GPU є ефективними для паралельних задач (графіка, матричні обчислення, нейронні мережі); FPGA забезпечують низьку затримку й гнучке налаштування; CPU обробляє послідовні, менш структуровані задачі; ASIC/TPU – вузькоспеціалізовані та високоефективні; edge-пристрої – критичні для low-latency задач у smart city (розумний будинок) та IoT (Інтернет речей).

Проблема полягає у виборі ефективної політики розподілу задач, яка враховує множину факторів: час виконання задачі на конкретному пристрої; архітектурна сумісність і вартість переносу даних; енергоспоживання; QoS-вимоги (затримка, доступність); динамічність навантаження.

Задача розподілу є NP-складною [1], а отже вимагає застосування наближених, адаптивних або навчальних методів. Зокрема, у середовищах, де задачі можуть динамічно надходити, а стан ресурсів – змінюватися в реальному часі, – необхідні онлайн-алгоритми з можливістю самонавчання.

Крім того, у контексті глобального тренду на декарбонізацію IT-сектору, енергоефективність розподілу обчислень також стає критично важливою метою.

Аналіз останніх досліджень і публікацій. У літературі представлено широкий спектр підходів до задачі планування (scheduling) в гетерогенних середовищах. Їх умовно можна поділити на кілька напрямів.

Серед класичних напрямків одним із базових алгоритмів є HEFT (Heterogeneous Earliest Finish Time) [1], який забезпечує ефективне планування DAG-завдань з урахуванням часу передачі та обчислення. HEFT продовжує залишатись популярним через простоту й добру якість рішень у статичних системах.

Гарну рекомендацію серед класичних напрямків мають метаевристичні алгоритми. Genetic Algorithm (GA) дозволяє шукати розподіл, мінімізуючи глобальну цільову функцію (makespan, енергія, QoS). Ant Colony Optimization (ACO) має користь при великій кількості зв'язаних задач. Particle Swarm Optimization (PSO) є методом,

адаптованим для багатовимірних оптимізацій із динамічним середовищем. Сьогодні дослідники використовують комбінації відомих метаевристичних методів для підвищення ефективності. Наприклад, автори статті [2] застосовували гібрид GA + PSO для гетерогенних середовищ із динамічним надходженням (arrival rate) задач. Результати показали покращення на 12–18% щодо стандартного HEFT. Гібридний метод на основі використання GA та методу випадкового лісу (Random forest) запропоновано у роботі [3]. Проведений аналіз показав, що комбінований метод показує зменшення часу виконання на 10–12%.

Останні роки активно розвиваються моделі та методи на базі машинного навчання. До найвідоміших методів відносяться наступні: Reinforcement Learning (RL) – дозволяє приймати рішення на основі стану середовища; Imitation Learning дозволяє агенту, який приймає рішення, навчатись від експертної системи (наприклад, Oracle scheduler); Graph Neural Networks (GNN) – дозволяє представляти DAG та відносини між задачами у вигляді графа. Останні публікації у цьому напрямку також використовують гібридні підходи. Наприклад, для QoS-aware сценаріїв в [4] запропоновано метод HRLHS (Hybrid Reinforcement Learning-based Heuristic Scheduling), заснований на комбінації методів RL та PSO.

Ще один з сучасних напрямків досліджень стосується енергозатратності методів управління ресурсами. Так в [5] запропоновано метод DVFS (Dynamic Voltage and Frequency Scaling) з інтегрованим плануванням, що дозволяє зменшити споживання енергії при незначному падінні продуктивності. Важливо також планувати не лише самі задачі, але й консолідувати ресурси. Проблемам дослідження енергоспоживання в системах розподілених ресурсів присвячено роботи [6, 7].

Стандартні методи багатоцільової оптимізації також застосовуються до розподілу обчислювальних ресурсів. Основними критеріями виступають, як правило, мінімізація часу (makespan), енергоспоживання, забезпечення QoS (latency, availability), балансування навантаження. Прикладами виступають методи NSGA-II або Pareto front [8].

У 2025 році у роботі [9] було представлено найбільш повний систематичний огляд моделей розподілу задач у гетерогенних чипах, класифікуючи їх за критеріями централізації, часового horizon (offline / online), використання машинного навчання, оптимізаційних цілей (QoS, енергія, latency).

Постановка завдання. Метою цієї статті є систематизація та узагальнення існуючих моделей і методів розподілення обчислювальних задач у гетерогенних комп'ютерних системах, з урахуванням їх ефективності, сфери застосування та обмежень. Основна увага приділяється аналізу підходів до планування задач у середовищах із неоднорідними ресурсами, зокрема в контексті задач реального часу, обчислень у хмарі, високопродуктивних обчислень та енергоефективних систем. Стаття також має на меті визначити перспективні напрямки досліджень і сформулювати рекомендації щодо вибору відповідних методів для різних класів обчислювальних задач.

Виклад основного матеріалу. У сфері гетерогенних обчислювальних систем існує широкий спектр підходів до розподілення задач, які розрізняються за своєю складністю, адаптивністю та ефективністю. Одними з найпоширеніших є евристичні алгоритми, зокрема HEFT та Min-Min. Метод HEFT ґрунтується на побудові розкладу задач, орієнтуючись на оцінку найранішого можливого часу завершення завдання на кожному обчислювальному ресурсі. Цей підхід добре працює у випадках, коли задачі мають залежності, представлені у вигляді орієнтованих ациклічних графів (DAG). Алгоритм демонструє високу продуктивність за помірної обчислювальної складності, що робить його зручним для застосування в реальному часі та наукових симуляціях. Метод Min-Min, у свою чергу, фокусується на першочерговому виконанні задач з найменшим прогнозованим часом виконання, що дозволяє досягти високої швидкодії, хоча й часто призводить до менш ефективного завантаження системи.

Значного поширення також набули метаевристичні методи, серед яких провідне місце займають генетичні алгоритми (GA), мурашині алгоритми (ACO) та метод рою частинок (PSO). Генетичні алгоритми базуються на принципах природного добору і використовують операції схрещування, мутації та селекції для пошуку оптимальних або близьких до оптимальних рішень. Їхня гнучкість дозволяє ефективно враховувати різноманітні обмеження задачі, проте вони можуть вимагати значних обчислювальних ресурсів для досягнення збіжності. Мурашині алгоритми моделюють процеси колективного навчання колоній комах і демонструють гарні результати при вирішенні задач з багатьма альтернативними маршрутами, що актуально для планування з урахуванням залежностей. PSO, натхненний поведінкою зграї птахів, швидко сходиться до рішень за умов низь-

кої складності проблеми, проте вимагає ретельного налаштування параметрів і може застрягати в локальних мінімумах.

У сучасних дослідженнях дедалі частіше використовуються методи машинного навчання, особливо навчання з підкріпленням (reinforcement learning, RL), зокрема його глибокі варіанти (deep RL). Однією з відомих реалізацій такого підходу є система Decima, яка дозволяє здійснювати динамічне планування DAG-завдань у розподілених середовищах шляхом навчання агента на основі досвіду з попередніх розкладів. Перевагою цих методів є здатність до самонавчання та адаптації до змін у системі, включаючи змінну доступність ресурсів або флуктуації у навантаженні. Проте застосування глибоких нейронних мереж потребує значної кількості даних для тренування та обчислювальних ресурсів, що обмежує їх використання в деяких практичних сценаріях.

Окремий напрямок досліджень становлять energy-aware методи, які враховують енергоспоживання при розподілі задач. Такі підходи часто реалізуються через застосування технологій динамічного масштабування напруги та частоти (DVFS), що дозволяє знижувати енергоспоживання процесорів без істотної втрати продуктивності. Наприклад, під час виконання задач, не критичних до затримки, система може зменшити частоту процесора, що істотно знижує загальне споживання енергії. Енергетично-ефективне планування є особливо важливим у контексті IoT, edge computing та мобільних пристроїв, де обмеження акумуляторів чи теплові ліміти суттєво впливають на працездатність системи.

У багатьох сучасних розробках спостерігається тенденція до інтеграції гібридних підходів. Зокрема, поєднання RL з евристичними методами, як-от HEFT, дозволяє досягти як високої адаптивності, так і гарантованого початкового рівня продуктивності. Такі гібридні рішення можуть динамічно змінювати свою стратегію в залежності від змін в середовищі виконання або типу задачі. У перспективі саме гібридні моделі, що поєднують інтелектуальне планування, енергетичні обмеження та евристичну стабільність, можуть стати основою для повністю автономних систем розподілу задач.

Проведемо аналіз застосування моделей та методів для різних класів задач. Застосування моделей і методів розподілення задач у гетерогенних системах значною мірою залежить від природи самих задач, архітектури обчислювального середовища та пріоритетів розробників –

зокрема, чи важливіше досягти мінімального часу виконання, забезпечити енергоефективність, гарантувати дотримання дедлайнів чи адаптивно реагувати на змінне навантаження. Залежно від цих параметрів формуються характерні переваги та недоліки різних груп методів, що відображено в типових сценаріях їхнього використання.

Для задач без залежностей між компонентами – так званих незалежних задач – найчастіше застосовуються прості евристичні алгоритми, такі як HEFT чи Min-Min. Їхньою перевагою є швидкість розрахунку та достатньо висока ефективність при статичному розподілі. Проте при збільшенні кількості задач ці методи втрачають гнучкість, оскільки не враховують динамічні зміни у стані системи. У таких випадках доповнення класичних методів навчальними підходами або гібридними схемами може дати кращі результати.

Для задач, що мають структуру DAG (орієнтований ациклічний граф), характерна складніша топологія виконання, тому ефективне планування вимагає врахування як порядку виконання, так і обчислювальних витрат. Саме тут алгоритм HEFT виявляється найбільш популярним, оскільки він добре справляється з такими обмеженнями. Водночас зростає інтерес до застосування навчання з підкріпленням, зокрема у випадках, коли задача виконується в динамічному середовищі або ресурсна доступність змінюється в реальному часі.

Для систем реального часу ключовим є дотримання часових обмежень, тому найкраще зарекомендували себе методи з високою передбачуваністю поведінки. Тут RL демонструє цікаві результати завдяки своїй здатності навчатися оптимальним реакціям на зміну навантаження, проте він потребує попереднього тренування. Евристичні підходи, такі як EDF (earliest deadline first), у поєднанні з HEFT також залишаються конкурентоспроможними за рахунок своєї простоти.

У контексті енергообмежених середовищ, таких як IoT або edge-пристрої, пріоритет надається методам, орієнтованим на енергоспоживання. Energy-aware підходи, включаючи DVFS та планування із врахуванням електроспоживання, значно переважають інші методи. Також певне поширення мають варіанти RL, які включають енергетичні фактори до функції винагороди, однак їх використання вимагає тонкого налаштування та профілювання пристроїв.

Щодо задач машинного навчання та штучного інтелекту, які часто характеризуються складною структурою, високим навантаженням та змінними

параметрами, найбільш ефективними виявилися моделі глибокого навчання, зокрема Deep RL. Вони дозволяють знаходити ефективні стратегії планування навіть у випадках, коли залежності між підзадачами є складними або змінюються в процесі виконання. Однак ці методи також мають високу вартість у плані обчислень і часу тренування.

Нарешті, в задачах високопродуктивних обчислень (HPC), де зазвичай використовуються великі обсяги даних і обчислювальні навантаження, важливо забезпечити максимальне завантаження всіх доступних ресурсів. Тут переважають класичні HEFT-подібні підходи, які демонструють хорошу продуктивність у стабільних середовищах. Проте з ростом складності завдань поступово зростає роль метасевристичних алгоритмів і навіть навчальних моделей, які забезпечують адаптивність і масштабованість.

Результати дослідження областей застосування різних підходів відображено у таблиці 1.

Наприклад, у системах автономного водіння inference-моделі розподіляються між on-board GPU та edge сервером: задачі з дедлайном <20 мс залишаються локально, решта – вивантажуються.

У таблиці 2 представлено приблизний розподіл поширеності застосування методів планування у гетерогенних системах для різних класів задач з урахуванням гібридних підходів, тобто коли для даного типу задач використовується комбінація з методів, які відносяться до різних груп.

На основі аналізу публікацій оцінено частоту використання основних підходів до планування задач у гетерогенних системах (рис. 1).

З огляду на стрімкий розвиток гетерогенних обчислювальних систем, зростаючі вимоги до адаптивності, енергоефективності та інтелектуального управління ресурсами, подальші дослідження у цій сфері повинні орієнтуватися не лише на покращення існуючих алгоритмів, але й на формування нових міждисциплінарних рішень. Серед ключових напрямків, що мають потенціал значного впливу, варто передусім виділити поглиблену інтеграцію методів глибокого навчання у процес планування обчислень. Особливо перспективними виглядають моделі навчання з підкріпленням у поєднанні з контекстною інформацією про задачі та системні обмеження, що дозволяє формувати розподіли з врахуванням реальних умов виконання.

Однією з важливих тенденцій майбутнього є створення самонавчальних систем планування, які здатні не лише адаптуватися до змін у струк-

Таблиця 1

Розподіл методів планування за типами задач

Тип задачі	Характеристика	Оптимальні методи
Незалежні задачі (bag-of-tasks)	Немає залежностей між задачами	Min-Min, GA, RL
DAG-завдання (workflow)	Завдання із залежностями	HEFT, ACO, GNN+RL
Реального часу (real-time)	Час виконання обмежений дедлайном	EDF + RL, A3C
Edge/IoT задачі	Обмеження по енергії, латентності, зв'язку	Energy-aware PSO, DVFS, кластеризація задач
ML/AI inference	GPU-heavy, багатопоточні	Static + profile-based mapping, reinforcement scheduling
High-performance computing (HPC)	Великі симуляції, CPU-GPU співпраця	Graph partitioning, Task co-location

Таблиця 2

Розподіл методів планування за типами задач

Тип задачі	HEFT/Min-Min	GA/PSO/ACO	Deep RL	Energy-aware
Незалежні задачі	70%	50%	20%	10%
DAG-завдання	80%	40%	30%	15%
Реального часу	30%	20%	50%	10%
IoT/Енергообмежені	10%	25%	30%	80%
ML/AI	20%	10%	70%	40%
HPC	60%	30%	40%	20%

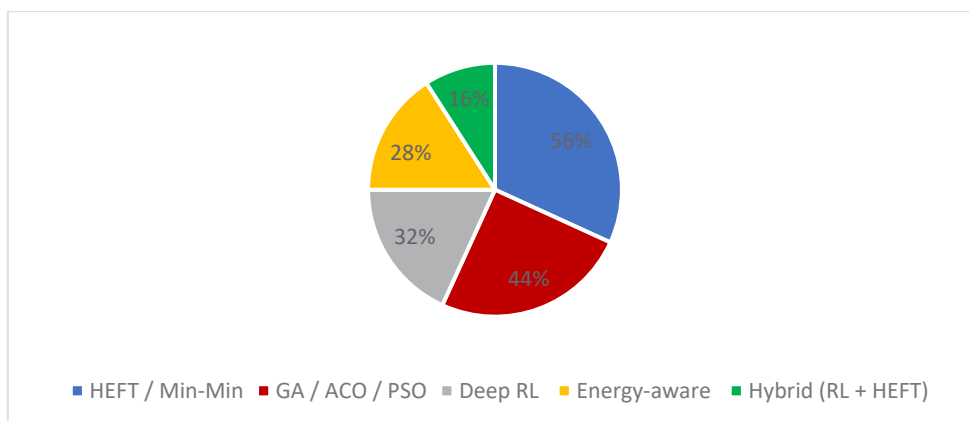


Рис. 1. Використання методів у публікаціях (відсоток від загальної кількості)

турі задач, але й прогнозувати поведінку ресурсів на основі історичних даних. Зокрема, розвиток моделей із короткотерміною пам'яттю (LSTM) або трансформерів для передбачення часу виконання та енергоспоживання задач відкриває нові горизонти для динамічного планування в режимі онлайн. Такі системи потенційно здатні працювати ефективно навіть в умовах відмов або непередбачених змін у мережевій структурі.

Окрему увагу заслуговує розробка гібридних моделей, які поєднують класичні евристичні з сучасними навчальними підходами. Наприклад, використання HEFT чи Min-Min на першому етапі

для грубого попереднього планування може доповнюватися більш тонким коригуванням через RL-агента або енергетичний модулятор. Такий багаторівневий підхід дозволяє досягти балансу між швидкістю та адаптивністю системи.

Перспективним вбачається і впровадження онтологій ресурсів та семантичного опису задач. Ще один напрямок, який варто досліджувати глибше – це федеративне навчання у плануванні, що дозволяє уникнути централізації при збиранні даних. Не менш актуальним є питання надійності й безпеки при розподіленні задач у системах із низьким рівнем довіри до окремих вузлів (напри-

клад, у відкритих edge-мережах або volunteer computing).

Підсумовуючи, можна стверджувати, що подальший розвиток у цій галузі буде все більше орієнтований на адаптивність, контекстну обізнаність, енергоефективність і безпеку. Це вимагає міждисциплінарного підходу, що об'єднує теорію розподілених систем, штучний інтелект, семантичне моделювання та прикладні задачі з реального світу.

Висновки. У статті було проведено системний огляд сучасних моделей та методів розподілення обчислювальних задач у гетерогенних комп'ютерних системах. З'ясовано, що з розвитком гібридних систем, включаючи CPU, GPU, FPGA, ASIC, edge вузли, значно ускладнюється задача ефективного планування обчислень.

Розглянуті підходи демонструють наступне. Класичні евристики (HEFT, Min-Min) залишаються важливими базовими методами, особливо в статичних задачах із відомими графами

виконання. Metaheuristic-алгоритми (GA, PSO, ACO) забезпечують високу адаптивність, особливо в задачах з кількома обмеженнями або метриками. Методи глибокого навчання (RL, GNN-RL) відкривають нові можливості для гнучкого онлайн-розподілу задач в умовах високої варіативності й динаміки. Edge/IoT-сценарії вимагають врахування нових параметрів – енергоспоживання, латентності, надійності зв'язку.

Зазначено, що вибір конкретної моделі повинен враховувати тип задачі (незалежна, реального часу), архітектурну топологію системи, доступні метрики, а також можливість часткового або повного навчання на історичних даних.

До перспективних напрямків досліджень належать: гібридизація RL та евристичних scheduler-ів, врахування семантичного профілю ресурсів, безпечне планування у відкритих гетерогенних середовищах, побудова стандартних benchmark-середовищ для відтворення результатів.

Список літератури:

1. Topcuoglu H., Hariri S., Wu M. Performance-effective and low-complexity task scheduling for heterogeneous computing. *IEEE Transactions on Parallel and Distributed Systems*. 2002. № 13(3). P. 260–274. <https://doi.org/10.1109/71.993206>
2. Wang Y., Li K., Xu Y. Hybrid Genetic and Particle Swarm Algorithm for Energy-aware Task Scheduling. *Future Generation Computer Systems*. 2022. № 108. P. 480–492. <https://doi.org/10.1016/j.future.2020.03.003>
3. Волк М.О., Курочкін В. С., Запороженко А.П., Паронікян П.А. Гібридний метод розподілу ресурсів в хмарних системах. *Системи управління, навігації та зв'язку*. 2024. № 2(76). С. 70–83. doi:10.26906/SUNZ.2024.2.070
4. Shao Q., Zhang Y., Wang H. HRLHS: Hybrid Reinforcement Learning-based Heuristic Scheduling in Edge Computing. *Scientific Reports*. 2024. № 14(1), 8461. URL: <https://www.nature.com/articles/s41598-025-94068-0> (дата звернення: 08.07.2025)
5. Kazemi F., Barzegar B., Motameni H., Yadollahzadeh-Tabari M. An energy-aware scheduling in DVFS-enabled heterogeneous edge computing environments. *The Journal of Supercomputing*. 2025. № 8, 1078. DOI: 10.1007/s11227-025-07432-2
6. Mamchych O., Volk M. A. A unified model and method for forecasting energy consumption in distributed computing systems based on stationary and mobile devices. *Radioelectronic and Computer Systems*. 2024. № 2. P. 120–135. DOI: <https://doi.org/10.32620/reks.2024.2.10>.
7. Mamchych O., Volk M. Smartphone Based Computing Cloud and Energy Efficiency. *12th International Conference on Dependable Systems, Services and Technologies (DESSERT)*. Athens. Greece. 2022. P. 1–5. DOI: 10.1109/DESSERT58054.2022.10018740
8. Zhou M., Li J., Chen Y. Multi-objective scheduling using NSGA-II in heterogeneous IoT systems. *Expert Systems with Applications*. 2024. № 239, 120184. <https://doi.org/10.1016/j.eswa.2024.120184>
9. Miao C., Ren Z., Qian Z. Scheduling in Heterogeneous Computing Systems: A Survey. *Electronics*. 2025. № 14(6), 1191. <https://doi.org/10.3390/electronics14061191>

Volk M.O., Buhrii A.M., Kovtun Ye.I., Samoilov I.A., Brestovytskyi R.M., Sorobey B.V. METHODS OF DISTRIBUTION OF COMPUTATIONAL TASKS IN HETEROGENEOUS COMPUTER SYSTEMS

Heterogeneous computing systems that combine computing elements of different architectures – central processing units (CPUs), graphics processing units (GPUs), FPGAs, systems on a chip (SoCs), application-specific integrated circuits (ASICs), and peripheral edge devices – are becoming the basis of modern distributed computing infrastructure. Such infrastructure requires new approaches to effectively distribute computing

tasks between computing elements with different computing characteristics, power consumption, architectural compatibility, and network connection bandwidth.

The article provides a thorough review of modern models and methods for distributing computing tasks in heterogeneous computer systems that combine resources with different architectures, performance, and power characteristics. Given the rapid growth of data volumes, the variety of computing tasks, and the need for their parallel execution, the problem of effective scheduling in such environments is becoming critical. Classical algorithms, heuristic and metaheuristic approaches, including HEFT, Min-Min, genetic algorithms (GA), particle swarm optimization (PSO), ant algorithms (ACO), as well as modern deep learning methods, in particular methods with reinforcement (Deep RL), which demonstrate significant adaptability to environmental changes, are considered.

Special attention is paid to energy-aware methods that allow achieving a balance between performance and energy consumption, which is especially relevant for IoT devices, mobile computing and edge environments. The article proposes a classification of tasks by architectural (type of connectivity, dependencies) and application (real-time, HPC, AI) features, which allows establishing the feasibility of using specific planning models. The comparative tables of methods presented in the article allow specialists to choose the most appropriate approach to planning. The results obtained can be useful when building autonomous task management systems in cloud, fog or edge environments.

Key words: heterogeneous computing systems, task distribution, machine learning, resource allocation methods, cloud systems, energy efficiency, heuristic methods.

Дата надходження статті: 05.07.2025

Дата прийняття статті: 11.07.2025

Опубліковано: 27.10.2025