УДК 004.75:004.02
DOI https://doi.org/10.32782/2663-5941/2026.2.1/05

Волк М.О.
https://orcid.org/0000-0003-4229-9904
Харківський національний університет радіоелектроніки

Ковтун Є.І.
https://orcid.org/0009-0006-4775-1292
Харківський національний університет радіоелектроніки

МОДЕЛІ ДИНАМІЧНОГО РОЗПОДІЛЕННЯ ОБЧИСЛЮВАЛЬНИХ ЗАДАЧ У ГЕТЕРОГЕННИХ КОМП'ЮТЕРНИХ СИСТЕМАХ

Стаття присвячена проблемі динамічного розподілення обчислювальних задач у гетерогенних комп'ютерних системах. Такі системи мають ресурси з відмінними характеристиками продуктивності та енергоспоживання. Актуальність цього дослідження зумовлена зростанням навантажень у хмарних і розподілених середовищах, а також необхідністю забезпечувати баланс між енергоефективністю та дотриманням вимог до якості обслуговування (QoS). Існуючі підходи до планування часто орієнтуються на окремі критерії, що знижує їх ефективність в умовах високої динаміки. Тому перспективними є дослідження, пов'язані з врахуванням значної кількості характеристик таких систем. В роботі запропоновано математичну модель багатокритеріальної оптимізації, у якій функція мети формулюється як зважена сума трьох складових: сумарного енергоспоживання обчислювальних вузлів, штрафу за перевантаження ресурсів та вартості міграції віртуальних машин. Енергоспоживання описується лінійною залежністю. Для мінімізації ризику деградації продуктивності застосовано квадратичну штрафну функцію за перевищення допустимого порогу завантаження. Це сприяє рівномірному розподілу навантаження. Додатково враховано параметр вартості міграції, що дозволяє обмежувати частоту перенесення задач і підвищувати стабільність системи. Експериментальна перевірка показала зниження середнього енергоспоживання до 18% порівняно з базовими моделями та суттєве скорочення часу перебування вузлів у перевантаженому стані. Найкращі інтегральні показники досягнуто у збалансованому режимі налаштування вагових коефіцієнтів. Запропоновані моделі демонструють адаптивність до змін навантаження та придатність для застосування у хмарних і високопродуктивних обчислювальних системах. Перспективи подальших досліджень пов'язані з розширенням моделі на багаторесурсні середовища та інтеграцією гібридних метаевристичних алгоритмів глобальної оптимізації.

Ключові слова: *гетерогенні комп'ютерні системи, динамічне розподілення задач, планування ресурсів, енергоефективність, QoS, міграція віртуальних машин, багатокритеріальна оптимізація, функція мети, хмарні обчислення.*

Постановка проблеми. Швидке зростання обчислювальних навантажень разом із різнорідністю апаратного забезпечення істотно ускладнює задачі планування та розподілення ресурсів у сучасних гетерогенних комп'ютерних системах. У практичних конфігураціях такі системи, як правило, поєднують високопродуктивні сервери, звичайні CPU-вузли, графічні прискорювачі, а також спеціалізовані обчислювальні модулі, які функціонують у спільному інформаційному середовищі. Водночас, попри наявність значної кількості підходів до балансування навантаження, забезпечення стабільної продуктивності й досі залишається проблемним, особливо за динамічних змін вхідних потоків, енергетичних обмежень та змінної доступності обчислювальних ресурсів[1].

Надійність та ефективність функціонування гетерогенних обчислювальних платформ значною мірою визначаються здатністю системи оперативно адаптуватися до змін характеристик навантаження та доступності ресурсів. Традиційні підходи до розподілу обчислювальних задач, як правило, ґрунтуються на наперед заданих правилах або евристичних алгоритмах, що обмежує їхню гнучкість в умовах високої динаміки та варіативності робочих сценаріїв.

Але значна частина існуючих оптимізаційних моделей динамічного планування не враховує

повністю складність взаємодії між компонентами, що мають різнорідний характер. Особливо це виникає за наявності вимог з декількома критеріями, які пов'язані із якістю обслуговування, часовими обмеженнями та необхідністю мінімізації енергоспоживання [2].

Тому вважаємо актуальним обговорення математичних моделей і методів, які здатні забезпечювати збалансоване керування обчислювальними задачами та ресурсами у гетерогенних середовищах. При цьому важливо одночасно враховувати продуктивність, затримки в передачі даних та обробки, енергетичні параметри і вартість балансування та перерозподілу. Додаткової складності додає багатопараметричність самих цільових функцій, необхідність врахування ймовірностного характеру навантажень та особливості роботи окремих модулів.

У науковій літературі багато уваги приділяється еволюційним та метаевристичним алгоритмам, які надають можливість підвищити ефективність процесів динамічного розподілення задач. Досягається це за рахунок адаптивного пошуку рішень у просторах з декількаами вимірами. Перспективними напрямками є алгоритми, які моделюють біологічні або фізичні процеси та демонструють здатність до швидкої самоадаптації.

У статті запропоновано математичні моделі динамічного розподілення задач у гетерогенних комп'ютерних системах. Запропонований підхід розрахован на методи оптимізації з адаптивними евристиками, що дозволяє врахувати як апаратні особливості ресурсів, так і мінливий характер обчислювальних потоків. Розроблені рішення спрямовані на підвищення стабільності системи за динамічних умов і можуть бути застосовані у хмарних середовищах, на HPC-платформах та в інфраструктурах обробки великих даних.

Аналіз останніх досліджень і публікацій. Проблема динамічного розподілення обчислювальних задач у гетерогенних системах привертає значну увагу дослідників, насамперед через збільшення варіативності апаратної платформи (CPU, GPU, FPGA) та непередбачуваність навантажень у реальному часі. У класичному підході до віртуалізації і балансування навантаження застосовуються прості евристики та правила (наприклад, round-robin, least-loaded), що, однак, часто не дозволяють ефективно поєднати енергетичну ефективність і суворі SLA у сучасних датацентрах. Комплексні багатокритеріальні підходи, які враховують одночасно енергію, час відгуку та вартість міграцій, все активніше розглядаються як більш привабливі для практичних систем, але вимагають високоефективних моделей для методів пошуку [1, с.638].

За останні роки помітно зросла кількість робіт, що застосовують сучасні метаевристичні методи до задач розміщення віртуальних машин та консолідації ресурсів. Зокрема, поліпшені версії Slime Mould Algorithm (SMA) демонструють високі результати в оптимізаційних задачах, а в роботах 2023–2024 років запропоновано низку модифікацій, які підвищують збіжність та зменшують ризик застрягання у локальних мінімумах [1, с.641; 2, с.3]. Ці вдосконалення роблять SMA перспективною базою для задач управління ресурсами.

Окремий напрям становлять метаевристики, що використовують чисельні схеми на основі методів Рунге–Кутти. У роботах 2023 року показано ефективність таких підходів у задачах з високою розмірністю параметричного простору, де класичні еволюційні алгоритми втрачають стабільність [3,с.124]. Це робить RUN-підходи привабливими як компонент локального вдосконалення всередині комбінованих оптимізаторів.

У практичних системах хмарних обчислень 2023–2024 років спостерігається тенденція до одночасного врахування енергоспоживання, часу відповіді та вартості переміщення віртуальних машин. Автори робіт [4, с.276; 5, с. 76] підкреслюють, що оптимальні стратегії балансування навантаження базуються на гібридних методах: швидкі евристики забезпечують оперативність рішень, а метаевристики – точність на довших часових горизонтах.

Ще одним активно досліджуваним напрямом є використання адаптивних популяційних алгоритмів, зокрема модифікованих Harris Hawks Optimization (HHO), для задач розподілу навантаження у хмарних та контейнерних середовищах [6, с. 3097; 7, с. 5]. Результати цих робіт демонструють, що такі методи забезпечують високий рівень стабільності навіть при обмежених обчислювальних ресурсах. Водночас автори наголошують на потребі точнішого визначення штрафних функцій і динамічних критеріїв міграції, що і визначає наукові розриви у сучасних дослідженнях [8, с. 14].

Підсумовуючи**,** у наявній літературі бракує узгоджених моделей, які одночасно:

– формалізують багатокритеріальну ціль із урахуванням енергоспоживання, QoS та вартості міграцій;

– забезпечують локальний критерій прийняття рішень, придатний для онлайн-орієнтованих оркестраторів;

– інтегрують комбіновані метаевристики у єдину узагальнену процедуру пошуку.

Запропоновані у статті моделі спрямовані саме на усунення цих прогалин.

Постановка завдання. У гетерогенних обчислювальних системах, що поєднують різнорідні апаратні ресурси (CPU, GPU, високошвидкісні прискорювачі), виникає потреба у розподіленні задач таким чином, щоб одночасно досягати високої продуктивності та енергетичної ефективності. Динамічне коливання навантажень, відмінності в архітектурі вузлів і різна інтенсивність обчислень у задачах створюють багатокритеріальну проблему оптимізації, яка ускладнюється необхідністю реагувати в реальному часі. Існуючі методи балансування навантаження часто орієнтуються на один критерій (наприклад, мінімізацію вартості обчислень, часу виконання або енергії), а тому не забезпечують збалансованості рішень у сучасних середовищах із високою динамікою стану.

У цьому дослідженні ставиться задача формалізувати модель, що дозволяє описати процес динамічного розподілення обчислювальних задач між вузлами гетерогенної системи, де кожний вузол характеризується обмеженою ємністю, а задачі – змінними у часі вимогами. Ключовою особливістю моделі є одночасне врахування трьох компонентів:

– енергоспоживання обчислювальних вузлів, яке залежить від їх активного стану і рівня завантаження;

– відхилення від допустимого рівня продуктивності (QoS), що відображає ризик перевантаження та затримок;

– вартості міграції між вузлами, яка включає часові та енергетичні втрати, пов'язані з перенесенням задач.

Модель передбачає, що система керування отримує дані моніторингу в дискретні моменти часу з інтервалом T, у межах якого навантаження задач вважається кусково-сталим. Кожному вузлу приписується індивідуальна характеристика енергоспоживання, яка дає змогу описати як базові витрати, так і динамічну складову, пропорційну його завантаженню. Це дозволяє враховувати ключові відмінності між типами апаратури, що є критичним для гетерогенних систем.

Проблема розподілення формулюється як оптимізація комбінованої цільової функції, яка включає енергетичний складник, штраф за перевантаження та вартість міграцій. Таким чином, модель забезпечує рівновагу між мінімізацією енергоспоживання та дотриманням вимог щодо продуктивності.

Структура цільової функції забезпечує гладкість штрафів за QoS-порушення та дозволяє використовувати як аналітичні методи, так і метаевристичні процедури пошуку.

Особливістю запропонованої постановки є її придатність до інтеграції з онлайн-орієнтованими алгоритмами управління: у межах кожного інтервалу T контролер може оцінити стан вузлів, спрогнозувати короткострокові зміни навантаження та ініціювати локальні або глобальні перенесення задач. Обмеження міграцій та можливість переведення малозавантажених вузлів у режим standby дозволяють моделі відображати реальні процеси в сучасних датацентрах.

Таким чином, сформульована модель задає формальний апарат для подальшого розроблення методів оптимального або наближеного керування ресурсами, зокрема гібридних алгоритмів, що поєднують швидкі локальні рішення та глобальну оптимізацію на основі сучасних метаевристик.

Нехай існують N віртуальних машин VM_i $(i = 1..N)$ та M фізичних вузлів PM_j $(j = 1..M)$. Нехай для кожної VM відомі її поточні потреби в ресурсах (агрегований показник) $r_i \geq 0$. Ємність кожного фізичного вузла позначимо $C_j > 0$. Нехай бінарна змінна x_{ij} означає, що VM_i розміщена на вузлі j або ні (xij=0). Обмеження цілісності розміщення: для кожної VM має виконуватись $\sum_{j=1}^{M} x_{ij} = 1$.

Завантаження вузла j (у відносних одиницях) позначимо Uj, його розрахунок:

$$U_j = \frac{1}{C_j} \sum_{i=1}^{N} r_i x_{ij}.$$

Енергоспоживання вузла j нехай залежить від його завантаження й складатиметься з базової витрати P_j^{idle} (споживання у режимі простою) та динамічної частини, пропорційної завантаженню: для простоти приймемо модель

$$P_j = \begin{cases} 0, & \text{якщо вузел вимкнено} \\ P_j^{idle} + \gamma_j U_j, & \text{якщо вузол увімкнено} \end{cases}$$

де γ_j – коефіцієнт, що відображає додаткове споживання при завантаженні. У моделі враховуємо, що при низькому сумарному навантаженні деякі вузли можуть бути поставлені у стан standby (вимкнені) з нульовим активним споживанням (спрощено).

Функція мети. Оскільки потрібно врахувати дві цілі – мінімізацію енергоспоживання та підтримку продуктивності – формулюємо комбіновану функцію мети у вигляді зваженої суми:

$$\min(F) = \alpha \cdot E + \beta \cdot Q,$$

де $E = \sum_{j=1}^{M} P_j$ – сумарне енергоспоживання кластера, Q – скоригований показник «неякісності» (quality penalty), що вимірює відхилення від бажаних QoS-

параметрів (наприклад, перевантаження вузлів або перевищення допустимого часу відповіді), а α, β – невід'ємні вагові коефіцієнти, які задає оператор або система управління відповідно до пріоритетів (баланс між енергією та продуктивністю). Важливо: вибір α та β дає змогу налаштувати політику – від енергоорієнтованої до performance-oriented.

Пропонуємо просту інтерпретацію Q: нехай перевантаження вузла j вважається критичним, якщо $U_j > U^{thr}$ (наприклад, $U^{thr} = 0.8$). Тоді

$$Q = \sum_{j=1}^{M} \max\left(0, U_j - U^{thr}\right)^2 .$$

Це означає, що невелике перевищення порогу штрафується слабо, а значні перевантаження отримують квадратичне покарання – це спонукає до більш рівномірного розподілу навантаження. Квадратична форма дає гладкість і ускладнює концентрацію багатьох VM на одному вузлі.

Обмеження. Модель доповнюється стандартними обмеженнями:

– ємнісні обмеження: для кожного j виконується $\sum_{i=1}^{N} r_i x_{ij} \le C_j$;

– у кожної VM – одне розміщення: $\sum_{j=1}^{M} x_{ij} = 1$;

– бінарність: $x_{ij} \in \{0,1\}$.

Вартість міграції. Окремо вводимо штраф за міграцію VM, оскільки часті переміщення знижують QoS і витрачають ресурси. Нехай m_i – вартість переміщення VM_i (у енергетичних або часових одиницях), а y_i – індикатор того, чи була VM переміщена у поточному кроці (1 – так, 0 – ні). Тоді додаємо до функції мети терм $\delta \sum_{j=1}^{M} m_i y_i$, де δ – ваговий коефіцієнт для зменшення частоти міграцій. У підсумку повна функція мети:

$$F = \alpha \sum_{j=1}^{M} P_j + \beta \sum_{j=1}^{M} \max\left(0, U_j - U^{thr}\right)^2 + \delta \sum_{i=1}^{N} m_i y_i .$$

Зауваження: в умовах реального датацентру параметри $P_j^{idle}, \gamma_j, C_j, m_i$ вимірюються або оцінюються на підставі моніторингу. Для цілей роботи їх можна задати експериментально у моделюванні (симуляції).

Формальна модель локальної оптимізації

Нехай поточна конфігурація системи описується матрицею $X=[x_{ij}]$, а зміна стану після переміщення однієї віртуальної машини VM_i з вузла a на вузол b позначається $X'=X+\Delta_{i,a,b}$. Визначимо градієнтну апроксимацію зміни функції мети (Ідейна формула локальної вигоди переносу i:$a \to b$):

$$\Delta F_{i,a,b} = F(X') - F(X) \approx \alpha \cdot \Delta E_{a,b} + \beta \cdot \Delta Q_{a,b} + \delta \cdot m_i,$$

$$\Delta E_{a,b} = \gamma_a (U_a' - U_a) + \gamma_b (U_b' - U_b),$$

$$\Delta Q_{a,b} = \max(0, (U_a' - U^{thr})^2 - \max(0, (U_a - U^{thr})^2 + \max(0, (U_b' - U^{thr})^2 - \max(0, (U_b - U^{thr})^2,$$

де U_a', U_b' – нові завантаження після переносу обчислюються:

$$U_a' = \frac{\sum_i (r - r_i)}{c_a}, \; U_b' = \frac{\sum_i (r - r_i)}{c_b}.$$

Рішення вважається доцільним, якщо $\Delta F_{i,a,b} < 0$ та після переносу не порушуються ємнісні обмеження: $\sum r_i x_{ib} \le C_b$.

Пояснення вибору компонентів моделі та інтерпретація параметрів

Математична структура моделі (див. рис. 1) обрана так, щоб відобразити два взаємопротилежні прагнення: мінімізувати енергоспоживання (через консолідацію та вимикання вузлів) та уникати перевантаження ресурсів (через квадратичну штрафну функцію).

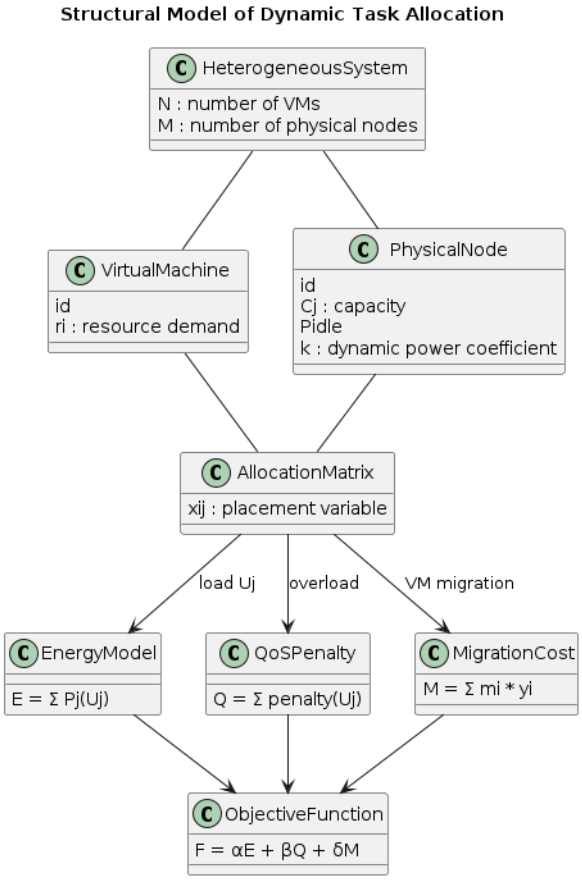

Рис. 1. Структурна UML діаграма взаємозв'язків моделі

Квадратичний штраф за перевантаження стимулює більш рівномірний розподіл навантаження, ніж лінійна форма, і робить стратегію більш «консервативною» щодо локальних перевищень порогу. Коефіцієнти α і β – інструмент налаштування політики. Якщо оператор ставить пріоритет на енергію (економія витрат), α збільшують відносно β; якщо ж на першому місці QoS – навпаки.

Параметр δ контролює агресивність міграцій: велике δ зменшує кількість переміщень і робить систему більш «стабільною» у плані QoS. Поріг U^{thr} та допоміжні U^{upper}, U^{target} задаються експериментально: наприклад, $U^{thr} = 0.8, U^{upper} = 0.95$, U^{target} =0.65. Така схема дозволяє уникати частих осциляцій (hysteresis) – вузол вважається перевантаженим тільки при досягненні вищого порогу, і після перерозподілу його завантаження приводиться до цільового значення нижчого рівня. Модель міграції спрощена: вартість m_i може залежати від об'єму пам'яті VM та доступної пропускної здатності мережі. У реальних умовах можна використовувати більш складні оцінки (включаючи час зупинки сервісу), але спрощена модель достатня для демонстрації поведінки алгоритму.

Експериментальна перевірка пропонуємих моделей динамічного розподілення задач виконувалася з використанням симуляційного середовища GridSim з відкритим кодом [9, с. 1190]. Воно дозволяє моделювати розподілені обчислювальні системи різнорідної природи, які можуть включати ресурси різної продуктивності, мережевої затримки та врахувати динамічні потоки задач. Використовувалась Java-платформа із розширенням стандартного планувальника шляхом інтеграції запропонованої багатокритеріальної функції мети. Протягом експерименту було змодельовано кластер із 10 гетерогенних фізичних вузлів трьох типів із різними характеристиками продуктивності та енергоспоживання. У систему надходило 60 віртуальних машин із змінними вимогами до ресурсів, що формували квазіперіодичні пікові навантаження. Тривалість кожного експериментального прогону становила одну годину модельного часу, а результати усереднювалися за десятьма незалежними ітераціями для зменшення впливу стохастичних факторів. Для порівняння було реалізовано три режими керування: базовий евристичний алгоритм типу round-robin, запропоновану модель у збалансованому режимі вагових коефіцієнтів та енергоорієнтований режим із підвищеним значенням коефіцієнта при енергетичній складовій функції мети. Оцінювання проводилося за такими показниками: середнє енергоспоживання кластера, частка часу перевантаження вузлів, середня кількість міграцій за годину та інтегральне значення функції мети. Узагальнені кількісні результати наведено в таблиці 1.

Отримані результати свідчать про те, що застосування запропонованої моделі дозволяє зменшити середнє енергоспоживання на 13–18 % порівняно з базовим алгоритмом. При цьому скорочується частка часу перебування вузлів у перевантаженому стані. У збалансованому режимі досягнуто найменшого інтегрального значення функції мети. Це підтверджує доцільність використання багатокритеріального підходу. З іншого боку, енергоорієнтований режим забезпечує мінімальне енергоспоживання, однак супроводжується більшим рівнем перевантаження та відбувається зростання кількості міграцій. Збільшення кількості міграцій у запропонованій моделі можно пояснити активною реакцією алгоритму на зміни навантаження. Введення вагового коефіцієнта при вартості міграцій дозволяє контролювати їх інтенсивність та уникати надмірних переміщень у короткочасних флуктуаціях.

Висновки. Проведене експериментальне дослідження підтвердило ефективність запропонованої моделі динамічного розподілення обчислювальних задач у гетерогенних комп'ютерних системах. Багатокритеріальна функції мети, яка поєднує енергетичну складову, штраф за перевантаження та вартість міграцій дозволила забезпечити збалансоване керування ресурсами за умов динамічної зміни навантаження. Результати моделювання демонструють можливість зниження енергоспоживання кластера до 18 % порівняно з традиційними евристичними підходами. Одночасно зменшеншення рівня порушень QoS сталося більш ніж у два рази. На нашу думку, найкращий інтегральний результат було отримано у режимі збалансованих вагових коефіцієнтів. Це підтверджує доцільність адаптивного налаштування параметрів моделі залежно від

Таблиця 1

Порівняння результатів моделювання

Режим керування	Середнє енергоспоживання, Вт	Частка перевантаження, %	Кількість міграцій (за годину)	Значення функції мети F
Round-robin	1245	12.8	6	1.00 (нормоване)
Запропонована модель (збаланс.)	1082	4.6	14	0.71
Запропонована модель (енергоорієнт.)	1015	7.9	18	0.76

стратегії управління. Отримані результати свідчать про практичну придатність запропонованих моделей для застосування у хмарних та гетерогенних обчислювальних середовищах. Подальші дослідження доцільно спрямувати на розширення моделі до багаторесурсного випадку з окремим урахуванням CPU, пам'яті та мережевих параметрів, а також на інтеграцію гібридних метаевристичних алгоритмів для глобальної оптимізації розміщення задач.

Список літератури:

1. Verma P., Maurya A. K., Yadav R. A survey on energy-efficient workflow scheduling algorithms in cloud computing. *Software: Practice and Experience.* 2023. vol. 54. pp. 637–682. DOI: 10.1002/spe.3292

2. Bano F, Ahmad F, Shahid M, Alam M, Hasan F, Sajid M. A Levelized Multiple Workflow Heterogeneous Earliest Finish Time Allocation Model for Infrastructure as a Service (IaaS) Cloud Environment. *Algorithms.* 2025. 18(2):99. https://doi.org/10.3390/a18020099

3. Mamchych O., Volk M. A unified model and method for forecasting energy consumption in distributed computing systems based on stationary and mobile devices. *Radioelectronic and Computer Systems.* [S.l.], v. 2024, n. 2. p. 120-135. DOI: https://doi.org/10.32620/ reks.2024.2.10.

4. Senthilkumar G., Anandamurugan S. Energy and time-aware scheduling in diverse virtualized cloud computing environments using optimized self-attention progressive generative adversarial network. *Network: Computation in Neural Systems.* 2025. vol. 36. no. 2. p. 274–293. DOI: 10.1080/0954898X.2024.2391401

5. Волк М.О., Курочкін В. С., Запорожченко А.П., Паронікян П.А. Гібридний метод розподілу ресурсів в хмарних системах. *Системи управління, навігації та зв'язку.* 2025. випуск 2(76). с. 70-83. doi:10.26906/ SUNZ.2024.2.070.

6. Ahmed W., Gautam G., Alam B. et al. An analytical review and performance measures of state-of-art scheduling algorithms in heterogeneous computing environment. *Archives of Computational Methods in Engineering.* 2024. vol. 31. pp. 3091–3113. DOI: 10.1007/s11831-024-10069-8

7. Chandrasiri S., Meedeniya D. Energy-efficient dynamic workflow scheduling in cloud environments using deep learning. *Sensors.* 2025, vol. 25, art. 1428. DOI: 10.3390/s25051428

8. Miao Z., Shao C., Li H., Tang Z. Review of task-scheduling methods for heterogeneous chips. *Electronics.* 2025. vol. 14, art. 1191. DOI: 10.3390/electronics14061191

9. Buyya R., Murshed M. GridSim: a toolkit for the modeling and simulation of distributed resource management and scheduling for Grid computing. Concurrency and Computation: Practice and Experience, 2002, vol. 14, no. 13–15, pp. 1175–1220. DOI: 10.1002/cpe.710

Volk M.O., Kovtun Ye.I. MODELS OF DYNAMIC DISTRIBUTION OF COMPUTING TASKS IN HETEROGENEOUS COMPUTER SYSTEMS

The article is devoted to the dynamic distribution of computational tasks in heterogeneous computer systems. Such systems have resources with different performance and energy consumption characteristics. The relevance of this study is due to the growth of loads in cloud and distributed environments, as well as the need to ensure a balance between energy efficiency and compliance with quality of service (QoS) requirements. Existing planning approaches are often focused on individual criteria, which reduces their effectiveness in conditions of high dynamics. Therefore, research related to taking into account a significant number of characteristics of such systems is promising. The paper proposes a mathematical model of multi-criteria optimisation, in which the objective function is formulated as a weighted sum of three components: the total energy consumption of computing nodes, the penalty for resource overload, and the cost of virtual machine migration. Energy consumption is described by a linear dependence. To minimise the risk of performance degradation, a quadratic penalty function is applied for exceeding the permissible load threshold. This contributes to uniform load distribution. Additionally, the migration cost parameter is taken into account, which allows limiting the frequency of task transfer and increasing system stability. Experimental verification showed a decrease in average energy consumption of up to 18% compared to the baseline models and a significant reduction in the time nodes spend in an overloaded state. The best integral indicators were achieved in the balanced mode of setting weight coefficients. The proposed models demonstrate adaptability to load changes and suitability for use in cloud and high-performance computing systems. Prospects for further research are related to the expansion of the model to multi-resource environments and the integration of hybrid metaheuristic algorithms for global optimisation.

Keywords: *heterogeneous computer systems, dynamic task scheduling, resource scheduling, energy efficiency, QoS, virtual machine migration, multi-criteria optimization, objective function, cloud computing.*

Дата першого надходження статті до видання: 06.02.2026
Дата прийняття статті до друку після рецензування: 04.03.2026
Дата публікації (оприлюднення) статті 11.05.2026