

**Андрій Пазинін**

аспірант

Інститут телекомунікацій і глобального інформаційного простору НАН України

ORCID: 0009-0002-9506-9539

pazinin@gmail.com

МЕТОД АДАПТИВНОГО МАСШТАБУВАННЯ ХМАРНИХ РЕСУРСІВ НА ОСНОВІ НАВЧАННЯ З ПІДКРІПЛЕННЯМ

Анотація. У статті розглянуто проблему адаптивного масштабування ресурсів хмарної інфраструктури в умовах динамічної зміни навантаження, нерівномірного використання обчислювальних ресурсів та необхідності забезпечення стабільності функціонування інформаційних сервісів. Встановлено, що традиційні підходи до горизонтального й вертикального масштабування переважно базуються на статичних порогових правилах або прогнозних моделях, які не забезпечують достатньої гнучкості в умовах змінного середовища та не враховують багатокритеріальний характер процесу керування ресурсами. Обґрунтовано доцільність використання методів навчання з підкріпленням для формування адаптивної стратегії керування хмарною інфраструктурою. У роботі проаналізовано сучасні методи автоматизованого масштабування ресурсів, моделі прийняття рішень у хмарних системах та підходи до використання методів навчання з підкріпленням у задачах оптимізації навантаження. Запропоновано метод адаптивного масштабування хмарних ресурсів, який базується на використанні інтелектуального агента, що здійснює оцінювання поточного стану інфраструктури, аналізує параметри навантаження, затримки обробки запитів, рівень використання процесорних ресурсів та оперативної пам'яті й формує оптимальну стратегію масштабування. Розроблено модель станів, систему винагороди та механізм вибору керуючих дій для автоматизованого розподілу ресурсів. У межах експериментального дослідження проведено порівняння запропонованого методу з класичними алгоритмами автоматизованого масштабування. Отримані результати підтвердили зниження затримок обробки запитів, підвищення ефективності використання ресурсів і покращення стабільності функціонування хмарної системи. Практичне значення роботи полягає у можливості застосування запропонованого методу для автоматизованого керування ресурсами у великих розподілених інформаційних системах та сучасних сервісах.

Ключові слова: хмарна інфраструктура; адаптивне масштабування; навчання з підкріпленням; хмарні ресурси; автоматизоване керування; оптимізація ресурсів; інтелектуальні системи; розподілені обчислення; масштабування навантаження; прийняття рішень.

ВСТУП

Сучасний розвиток інформаційних технологій, зростання обсягів цифрових даних та активне впровадження хмарних обчислень у різних сферах діяльності зумовлюють необхідність підвищення ефективності керування обчислювальними ресурсами. Особливої актуальності це набуває в умовах динамічної зміни навантаження на хмарну інфраструктуру, коли традиційні методи масштабування не забезпечують достатньої швидкості реагування та ефективного використання ресурсів [1, 4, 15]. Значне збільшення кількості користувачів інформаційних сервісів, розвиток розподілених систем, потокової обробки даних та інтелектуальних інформаційних платформ призводять до постійних коливань навантаження, що потребує використання адаптивних методів автоматизованого керування хмарною інфраструктурою [5, 13].

Актуальність дослідження зумовлена необхідністю створення інтелектуальних методів масштабування хмарних ресурсів, здатних забезпечувати стабільність функціонування системи, мінімізувати затримки обробки запитів і водночас оптимізувати використання обчислювальних ресурсів [2, 6, 27]. Традиційні підходи до автоматизованого масштабування переважно базуються на використанні фіксованих порогових значень навантаження або прогнозних моделей, що не дозволяє враховувати складну динаміку функціонування сучасних хмарних середовищ [3-4, 17]. У таких умовах перспективним напрямом є використання методів навчання з підкріпленням, які забезпечують формування адаптивної



стратегії прийняття рішень на основі аналізу поточного стану системи та накопиченого досвіду взаємодії із середовищем [1-2, 19, 21].

Наукова новизна роботи полягає у розробленні методу адаптивного масштабування хмарних ресурсів на основі навчання з підкріпленням, який забезпечує інтелектуальне керування процесом розподілу обчислювальних ресурсів залежно від зміни навантаження, стану системи та параметрів функціонування хмарної інфраструктури. На відміну від існуючих підходів, запропонований метод враховує багатокритеріальний характер процесу масштабування та забезпечує автоматизований вибір керуючих дій із урахуванням продуктивності системи, рівня використання ресурсів і стабільності функціонування сервісів.

Теоретичне значення дослідження полягає у розвитку методів інтелектуального керування хмарними ресурсами, удосконаленні моделей прийняття рішень у розподілених інформаційних системах та формуванні підходів до використання навчання з підкріпленням у задачах адаптивного масштабування [10, 12]. Використання імітаційного моделювання дозволяє досліджувати поведінку хмарної інфраструктури в умовах змінного навантаження та оцінювати ефективність механізмів масштабування.

Практичне значення роботи полягає у можливості застосування запропонованого методу в сучасних хмарних середовищах, центрах обробки даних, корпоративних інформаційних системах та сервісах із динамічним навантаженням для підвищення ефективності використання ресурсів і забезпечення стабільності функціонування інформаційних сервісів [5-6, 16]. Важливим аспектом функціонування хмарної інфраструктури є також продуктивність протоколів доступу та швидкість взаємодії із сервісами хмарного середовища.

Постановка проблеми. Ефективне керування ресурсами хмарної інфраструктури є одним із ключових завдань сучасних інформаційних систем, оскільки від правильності розподілу обчислювальних ресурсів залежить стабільність функціонування сервісів, швидкість обробки запитів, рівень доступності інформаційних ресурсів та економічна ефективність використання інфраструктури [11, 13]. Сучасні хмарні середовища функціонують в умовах постійної зміни навантаження, що зумовлено нерівномірною активністю користувачів, збільшенням обсягів даних, масштабуванням інформаційних сервісів та динамічним характером мережевої взаємодії [5, 23, 25]. Зростання масштабів використання хмарних технологій у різних галузях діяльності додатково підвищує вимоги до ефективності масштабування ресурсів та стабільності функціонування сервісів. У таких умовах виникає необхідність оперативного адаптивного керування ресурсами з метою запобігання перевантаженню системи або неефективному використанню обчислювальних потужностей.

Традиційні методи автоматизованого масштабування переважно базуються на використанні фіксованих порогових значень навантаження, часових правил або прогнозних моделей [15, 17]. Однак такі підходи мають обмежену адаптивність і не здатні повною мірою враховувати складну динаміку функціонування сучасної хмарної інфраструктури [11, 13]. Використання статичних механізмів масштабування часто призводить до надлишкового виділення ресурсів або, навпаки, до дефіциту обчислювальних потужностей у моменти пікового навантаження [2, 5, 27]. Це негативно впливає на продуктивність інформаційних сервісів, збільшує затримки обробки запитів та підвищує витрати на підтримку хмарної інфраструктури.

Особливої актуальності проблема адаптивного масштабування набуває в умовах функціонування розподілених інформаційних систем, сервісів потокової обробки даних, інтелектуальних платформ, систем електронної комерції та середовищ із високим рівнем змінності навантаження [18, 24]. Особливо це актуально для інтелектуальних систем, які інтегрують технології штучного інтелекту та інтернету речей і характеризуються нерівномірною інтенсивністю обміну даними.

У таких системах необхідно забезпечити автоматизоване прийняття рішень щодо масштабування ресурсів у режимі реального часу з урахуванням поточного стану інфраструктури, прогнозованого навантаження та ефективності використання обчислювальних ресурсів [9, 20, 25]. Адаптивне керування ресурсами є важливим також для забезпечення стабільності функціонування хмарно-орієнтованих інтелектуальних систем у динамічних умовах експлуатації.

Одним із перспективних напрямів вирішення зазначеної проблеми є використання методів навчання з підкріпленням, які дозволяють формувати адаптивну стратегію керування ресурсами на основі взаємодії інтелектуального агента із середовищем [1-2, 19, 21]. Використання такого підходу створює передумови для підвищення ефективності функціонування хмарної інфраструктури, зменшення витрат ресурсів та забезпечення стабільності роботи інформаційних сервісів у динамічних умовах функціонування системи.

Аналіз останніх досліджень і публікацій. У сучасних дослідженнях проблема автоматизованого масштабування хмарних ресурсів розглядається як важливий напрям оптимізації продуктивності, вартості та стабільності функціонування розподілених інформаційних систем. У роботі Н. Qiu та співавторів запропоновано підхід AWARE для автоматизації масштабування навантаження з використанням навчання



з підкріпленням у виробничих хмарних системах [1]. Автори наголошують, що складність практичного впровадження таких методів пов'язана з безпечним дослідженням середовища, адаптацією до різних типів навантаження та необхідністю швидкого переналаштування агента без повного повторного навчання.

У праці Р. Mishra та співавторів розглянуто масштабування хмарних навантажень із використанням навчання з підкріпленням [2]. Основна увага приділяється зниженню операційних витрат і підтриманню якості обслуговування під час зміни навантаження, що є важливим для постачальників хмарних сервісів та інфраструктурних платформ.

У дослідженні S. Alharthi та співавторів подано огляд сучасних методів автоматичного масштабування у хмарних обчисленнях [3]. Автори систематизують підходи до масштабування, виокремлюють переваги інтелектуальних методів та підкреслюють, що залишаються відкритими питання точності прогнозування, вибору метрик, адаптивності та узгодження масштабування з вимогами якості обслуговування.

У роботі I. Pintye та співавторів досліджено підвищення ефективності методів машинного навчання для автоматизованого масштабування хмарних застосунків [4]. Важливим результатом цієї праці є висновок про те, що різні застосунки потребують різних показників для коректного визначення потреби в ресурсах, тому універсальні статичні правила масштабування не забезпечують достатньої ефективності.

У статті X. Feng та співавторів запропоновано метод адаптивного масштабування контейнерів для змінних навантажень у хмарному середовищі [5]. Автори зазначають, що методи на основі навчання з підкріпленням можуть втрачати ефективність за умов різких коливань навантаження, якщо використовують фіксовану множину дій, тому актуальним є формування динамічної моделі керуючих дій.

У дослідженні R. Moreno-Vozmediano та співавторів представлено інтелектуальний підхід до розподілу ресурсів і автоматичного масштабування у середовищах периферійних обчислень, мереж п'ятого покоління та інтернету речей [6]. Автори підкреслюють проблему надлишкового й недостатнього виділення ресурсів, що безпосередньо впливає на якість обслуговування та ефективність функціонування сервісів.

У роботі M. N. Alatawi розглянуто адаптивний розподіл ресурсів у багатокористувацьких безсерверних хмарних середовищах із використанням навчання з підкріпленням [7]. Отримані результати підтверджують перспективність такого підходу для керування ресурсами хмарних систем, однак водночас актуальними залишаються питання узгодження продуктивності, енергоефективності, вартості та стабільності обслуговування.

Отже, аналіз останніх досліджень показує, що навчання з підкріпленням активно використовується для розв'язання задач автоматизованого масштабування хмарних ресурсів, однак раніше невирішеними залишаються питання побудови цілісного методу, який одночасно враховує поточний стан інфраструктури, зміну навантаження, затримки обробки запитів, рівень використання ресурсів і стабільність функціонування сервісів. Саме цим аспектам присвячене подальше дослідження.

Мета статті. Метою статті є розроблення методу адаптивного масштабування хмарних ресурсів на основі навчання з підкріпленням, який забезпечує автоматизоване прийняття рішень щодо розподілу обчислювальних ресурсів залежно від поточного стану хмарної інфраструктури, рівня навантаження та параметрів функціонування інформаційної системи.

Для досягнення поставленої мети у роботі необхідно вирішити такі завдання: провести аналіз сучасних методів автоматизованого масштабування ресурсів у хмарних середовищах; дослідити підходи до використання навчання з підкріпленням у задачах інтелектуального керування ресурсами; розробити модель станів, систему винагороди та механізм вибору керуючих дій для адаптивного масштабування; сформувати метод автоматизованого керування ресурсами хмарної інфраструктури; провести експериментальне дослідження ефективності запропонованого методу та виконати порівняння отриманих результатів із традиційними алгоритмами масштабування.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

У межах дослідження розглянуто проблему підвищення ефективності адаптивного масштабування хмарних ресурсів у сучасних розподілених інформаційних системах шляхом використання методів навчання з підкріпленням для автоматизованого прийняття рішень щодо керування обчислювальними ресурсами [17, 21]. Запропонований підхід орієнтований на створення інтелектуальної системи масштабування, здатної забезпечувати підвищення продуктивності хмарної інфраструктури, зменшення затримок обробки запитів та ефективне використання обчислювальних ресурсів в умовах динамічної зміни навантаження. У розділі наведено архітектуру системи адаптивного масштабування, математичну модель взаємодії інтелектуального агента із середовищем, функцію винагороди, алгоритм прийняття рішень та результати експериментального дослідження ефективності запропонованого методу.

Архітектура системи адаптивного масштабування хмарних ресурсів

Запропонований метод адаптивного масштабування хмарних ресурсів базується на використанні навчання з підкріпленням для автоматизованого прийняття рішень щодо зміни кількості обчислювальних ресурсів залежно від поточного стану хмарної інфраструктури [1-2, 19]. Основною метою функціонування системи є забезпечення стабільної роботи інформаційних сервісів за мінімально можливих витрат ресурсів та підтримання необхідного рівня продуктивності в умовах змінного навантаження.

Архітектура системи включає модуль моніторингу, модуль аналізу стану інфраструктури, інтелектуального агента навчання з підкріпленням, підсистему прийняття рішень та модуль керування масштабуванням ресурсів [4, 12, 18]. Подібні інтелектуальні підходи до керування ресурсами використовуються також у кіберфізичних і хмарних середовищах, де необхідне адаптивне прийняття рішень в умовах змінного навантаження.

Модуль моніторингу здійснює безперервний збір параметрів функціонування системи, зокрема рівня завантаження процесорних ресурсів, використання оперативної пам'яті, інтенсивності мережевого трафіку, кількості активних користувачів та затримки обробки запитів [16, 25]. Аналіз мережевого трафіку та параметрів функціонування інфраструктури є важливим елементом формування ефективної політики адаптивного масштабування ресурсів [8]. Отримані дані передаються до модуля аналізу, де формується поточний стан середовища для інтелектуального агента. Під час аналізу стану середовища важливим є врахування взаємозалежності між параметрами навантаження, станом ресурсів та керуючими діями інтелектуального агента, що узгоджується з підходами до оцінювання взаємодії агентів у складних інформаційних системах [26].

Інтелектуальний агент взаємодіє із хмарним середовищем та формує рішення щодо збільшення або зменшення кількості обчислювальних ресурсів [6-7, 21]. У процесі функціонування агент накопичує досвід, оцінює ефективність виконаних дій та поступово формує оптимальну стратегію керування ресурсами.

На рис. 1 наведено архітектуру системи адаптивного масштабування хмарних ресурсів на основі навчання з підкріпленням. Система включає модулі моніторингу, аналізу стану, інтелектуального агента, підсистему прийняття рішень і механізм масштабування ресурсів. На основі параметрів навантаження інтелектуальний агент формує рішення щодо зміни обсягу обчислювальних ресурсів, а механізм зворотного зв'язку та винагороди забезпечує навчання агента й поступове підвищення ефективності керування хмарною інфраструктурою.

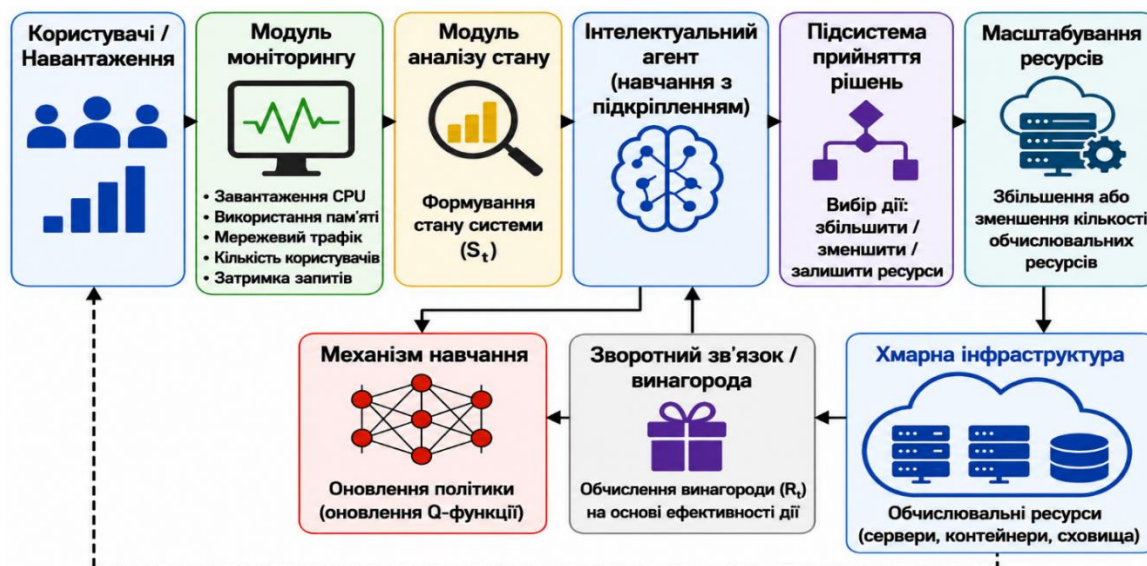


Рис. 1. Архітектура системи адаптивного масштабування хмарних ресурсів на основі навчання з підкріпленням

Архітектура, наведена на рис. 1, забезпечує безперервну взаємодію між модулями моніторингу, аналізу стану, прийняття рішень і механізмом навчання інтелектуального агента. Використання механізму зворотного зв'язку та функції винагороди дозволяє системі адаптувати процес масштабування ресурсів до змін навантаження та підвищувати ефективність функціонування хмарної інфраструктури.

**Модель зміни навантаження у хмарному середовищі**

Для забезпечення адаптивності запропонованого методу сформовано модель зміни навантаження у хмарному середовищі [9, 25]. Інтенсивність навантаження розглядається як змінна величина, що залежить від кількості активних користувачів, частоти надходження запитів, складності їх обробки та поточного стану обчислювальних ресурсів.

Математично навантаження хмарної інфраструктури можна подати у вигляді:

$$L(t) = \sum_{i=1}^n q_i(t) c_i, \quad (1)$$

де $L(t)$ – загальний рівень навантаження у момент часу t , $q_i(t)$ – інтенсивність запитів від i -го користувача або сервісу, c_i – обчислювальна складність обробки відповідного запиту, n – кількість активних джерел навантаження. Використання такої моделі дозволяє враховувати динамічний характер зміни навантаження та формувати більш точні рішення щодо збільшення, зменшення або збереження поточного обсягу хмарних ресурсів.

Формалізація моделі адаптивного масштабування

Процес адаптивного масштабування подається у вигляді моделі взаємодії агента із середовищем:

$$M = \langle S, A, R, P \rangle, \quad (2)$$

де S – множина станів системи, A – множина можливих керуючих дій, R – функція винагороди, P – ймовірність переходу системи між станами.

Стан системи формується як вектор параметрів:

$$S_t = \{u_{cpu}, u_{mem}, u_{net}, d, n\}, \quad (3)$$

де u_{cpu} – рівень використання процесорних ресурсів, u_{mem} – рівень використання оперативної пам'яті, u_{net} – інтенсивність мережевого навантаження, d – затримка обробки запитів, n – кількість активних вузлів хмарної інфраструктури.

Множина керуючих дій містить операції масштабування:

$$A = \{a_{up}, a_{down}, a_{keep}\}, \quad (4)$$

де a_{up} – збільшення ресурсів, a_{down} – зменшення ресурсів, a_{keep} – збереження поточного стану системи.

Формування функції винагороди

Для забезпечення ефективного керування ресурсами сформовано функцію винагороди, яка враховує рівень використання ресурсів, затримки обробки запитів та стабільність функціонування системи:

$$R_t = \alpha U_{eff} - \beta D - \gamma C, \quad (5)$$

де U_{eff} – ефективність використання ресурсів, D – затримка обробки запитів, C – вартість використання ресурсів, α, β, γ – вагові коефіцієнти.

Максимізація функції винагороди дозволяє агенту формувати стратегію масштабування, за якої забезпечується баланс між продуктивністю системи та витратами ресурсів.

Оновлення оцінки корисності виконаних дій здійснюється відповідно до рівняння:

$$Q(s_t, a_t) = Q(s_t, a_t) + \eta [r_t + \lambda \max_a Q(s_{t+1}, a) - Q(s_t, a_t)], \quad (6)$$

де η – коефіцієнт навчання, λ – коефіцієнт дисконтування майбутньої винагороди. Вагові коефіцієнти α, β та γ визначають пріоритети керування хмарною інфраструктурою. Якщо головною вимогою є мінімізація затримки, збільшується значення β ; якщо важливішою є економія ресурсів, підвищується значення γ ; якщо пріоритетом є продуктивність системи, більша вага надається коефіцієнту α . Це дозволяє адаптувати запропонований метод до різних типів хмарних сервісів та умов їх функціонування.

Алгоритм функціонування методу

Функціонування запропонованого методу складається з декількох етапів. На першому етапі здійснюється моніторинг параметрів хмарної інфраструктури та формування поточного стану системи. Далі інтелектуальний агент аналізує отримані параметри та обирає керуючу дію відповідно до сформованої політики прийняття рішень.

Після виконання дії система переходить у новий стан, а агент отримує значення винагороди, яке характеризує ефективність прийнятого рішення. На основі накопиченого досвіду виконується коригування параметрів політики керування [17, 27]. Таким чином забезпечується поступове покращення ефективності масштабування та адаптація системи до змін навантаження.

На рис. 2 наведено блок-схему алгоритму адаптивного масштабування хмарних ресурсів на основі навчання з підкріпленням. Алгоритм включає етапи моніторингу навантаження, формування стану системи, аналізу параметрів та вибору керуючої дії інтелектуальним агентом. Залежно від поточного стану інфраструктури система виконує масштабування ресурсів або збереження поточної конфігурації, після чого обчислюється винагорода, оновлюється Q -функція та формується новий стан системи.

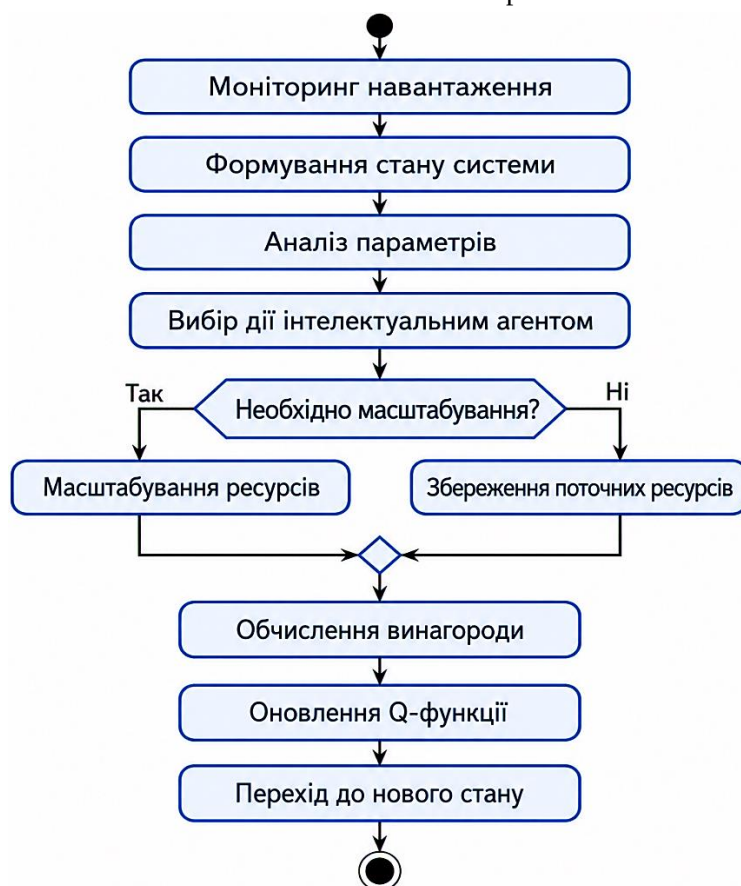


Рис. 2. Блок-схема алгоритму адаптивного масштабування хмарних ресурсів

Блок-схема, наведена на рис. 2, відображає послідовність функціонування запропонованого методу та взаємодію між етапами моніторингу, прийняття рішень і навчання інтелектуального агента [4, 12]. Використання такого підходу забезпечує адаптивне керування хмарними ресурсами та поступове підвищення ефективності масштабування в умовах змінного навантаження.

Механізм навчання інтелектуального агента

У запропонованому методі інтелектуальний агент навчається шляхом багаторазової взаємодії з хмарним середовищем [17, 21]. На кожному кроці агент отримує поточний стан системи, обирає керуючу дію, після чого середовище переходить у новий стан і формує значення винагороди. Якщо прийняте рішення сприяє зменшенню затримки, підвищенню ефективності використання ресурсів і зниженню витрат, агент отримує додатну винагороду [2, 6, 27]. Якщо дія призводить до перевантаження або надлишкового виділення ресурсів, значення винагороди зменшується.

Для вибору дій може використовуватися стратегія, яка поєднує дослідження нових рішень і використання вже накопиченого досвіду:

$$P(a) = 1 - \varepsilon + \frac{\varepsilon}{|A|}, \text{ якщо } a = a^* \quad (7)$$

$$P(a) = \frac{\varepsilon}{|A|}, \text{ якщо } a \neq a^*$$

де ε – коефіцієнт дослідження середовища, $|A|$ – кількість можливих керуючих дій, a^* – дія, яка має найбільшу очікувану корисність. Такий механізм забезпечує поступове вдосконалення політики масштабування та дає змогу агенту адаптуватися до різних сценаріїв зміни навантаження.

Для узагальненого оцінювання ефективності запропонованого методу доцільно використати інтегральний критерій, який поєднує продуктивність, використання ресурсів, затримку обробки запитів та вартість функціонування хмарної інфраструктури:

$$E = w_1 P + w_2 U - w_3 D - w_4 C, \quad (8)$$

де E – інтегральна ефективність масштабування, P – продуктивність хмарного сервісу, U – рівень корисного використання ресурсів, D – середня затримка обробки запитів, C – вартість використання ресурсів, w_1, w_2, w_3, w_4 – вагові коефіцієнти важливості відповідних показників. Використання такого

критерію дозволяє оцінювати не лише окремі технічні показники, а й загальну якість прийнятих рішень щодо масштабування ресурсів.

Для коректного об'єднання показників різної розмірності у межах інтегрального критерію доцільно виконувати їх нормалізацію. Нормалізоване значення показника визначається за виразом:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}, \quad (9)$$

де X_{norm} – нормалізоване значення показника, X – поточне значення показника, X_{min} та X_{max} – мінімальне й максимальне значення показника у межах експериментального дослідження. Використання нормалізації дозволяє коректно порівнювати продуктивність, затримку, використання ресурсів і вартість функціонування хмарної інфраструктури в єдиній системі оцінювання.

На рис. 3 наведено архітектуру нейромережевого агента навчання з підкріпленням для адаптивного масштабування хмарних ресурсів. Агент отримує входні параметри навантаження, зокрема рівень використання процесора, оперативної пам'яті, мережевого трафіку, затримку запитів та кількість активних користувачів [9, 13]. Після обробки даних у входньому та прихованих шарах нейромережі виконується оцінювання Q-значень і вибір оптимальної дії щодо масштабування ресурсів [17]. Механізм функції винагороди та зворотного зв'язку забезпечує навчання агента й адаптацію процесу прийняття рішень до змін стану хмарної інфраструктури.

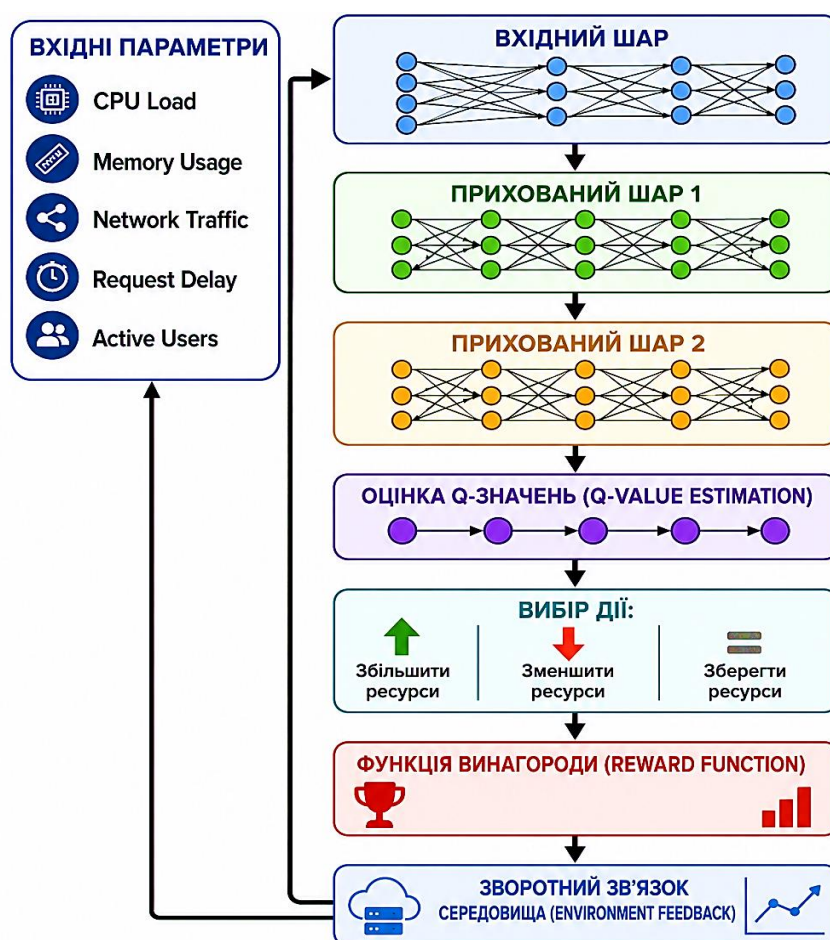


Рис. 3. Архітектура нейромережевого агента навчання з підкріпленням для адаптивного масштабування хмарних ресурсів

Архітектура, наведена на рис. 3, відображає послідовність обробки параметрів навантаження та механізм формування керуючих дій інтелектуальним агентом [18]. Використання нейромережевого підходу дозволяє агенту аналізувати поточний стан хмарної інфраструктури, оцінювати ефективність виконаних дій і поступово формувати оптимальну політику адаптивного масштабування ресурсів.

**Методика експериментального дослідження**

Експериментальне дослідження ефективності запропонованого методу проводилося шляхом моделювання роботи хмарної інфраструктури в умовах змінного навантаження [10, 16]. Для оцінювання використовувалися такі показники: середня затримка обробки запитів, рівень використання обчислювальних ресурсів, кількість випадків надлишкового масштабування, стабільність функціонування сервісів та кількість керуючих дій масштабування.

У межах експерименту розглядалися три сценарії: стабільне навантаження, поступове зростання кількості запитів і різкий стрибок навантаження [9, 25]. Запропонований метод порівнювався з класичним пороговим алгоритмом автоматизованого масштабування, у якому рішення про зміну кількості ресурсів приймається після перевищення заздалегідь установлених порогових значень.

Для оцінювання ефективності запропонованого методу було сформовано декілька сценаріїв функціонування хмарної інфраструктури, які відображають різні варіанти зміни навантаження та поведінки користувачів. Основні сценарії експериментального дослідження наведено у табл. 1.

Таблиця 1

Сценарії експериментального дослідження

Сценарій	Характер навантаження	Мета перевірки
Сценарій 1	Стабільне навантаження	Перевірка здатності системи підтримувати стабільний рівень продуктивності
Сценарій 2	Поступове зростання кількості запитів	Оцінювання здатності методу своєчасно збільшувати ресурси
Сценарій 3	Різкий стрибок навантаження	Перевірка швидкості реакції агента на пікові навантаження
Сценарій 4	Зменшення активності користувачів	Оцінювання здатності методу зменшувати надлишкові ресурси

Наведені сценарії дозволяють комплексно оцінити ефективність запропонованого методу в умовах стабільного, поступово зростаючого та різко змінного навантаження [11, 25]. Використання декількох сценаріїв забезпечує можливість перевірки адаптивності інтелектуального агента та стабільності прийняття рішень щодо масштабування ресурсів.

Результати експериментального дослідження

Для оцінювання ефективності запропонованого методу проведено експериментальне дослідження в умовах змінного навантаження хмарної інфраструктури [10, 16]. Під час дослідження виконано порівняння запропонованого підходу з класичними пороговими алгоритмами автоматизованого масштабування.

Результати дослідження показали, що використання навчання з підкріпленням дозволило зменшити середню затримку обробки запитів на 18–25 % порівняно з традиційними методами масштабування. Одночасно спостерігалось підвищення ефективності використання обчислювальних ресурсів та зменшення кількості випадків надлишкового виділення ресурсів.

Для кількісного оцінювання ефективності запропонованого методу проведено порівняння з класичним пороговим алгоритмом автоматизованого масштабування. Результати порівняння основних показників функціонування системи наведено у табл. 2.

Таблиця 2

Порівняння ефективності методів масштабування хмарних ресурсів

Показник	Пороговий алгоритм	Запропонований метод
Середня затримка обробки запитів, мс	145	112
Середнє використання ресурсів, %	68	82
Випадки надлишкового масштабування	14	6
Стабільність функціонування сервісу, %	91	97
Середній час реакції на зміну навантаження, с	42	27

Аналіз результатів, наведених у табл. 2, показує, що запропонований метод забезпечує кращий баланс між продуктивністю та витратами ресурсів порівняно з класичним пороговим алгоритмом [4, 27]. Найбільша перевага спостерігається за показниками середньої затримки обробки запитів, стабільності функціонування сервісу та кількості випадків надлишкового масштабування.



Для наочного порівняння ефективності методів масштабування доцільно проаналізувати зміну середньої затримки обробки запитів залежно від рівня навантаження. На рис. 4 наведено порівняння середньої затримки обробки запитів для класичного порогового алгоритму та запропонованого методу адаптивного масштабування хмарних ресурсів на основі навчання з підкріпленням. Результати показують, що зі збільшенням рівня навантаження середня затримка обробки запитів зростає для обох методів, однак запропонований метод забезпечує значно менші затримки порівняно з пороговим алгоритмом. Найбільша різниця спостерігається при високому рівні навантаження, що підтверджує вищу адаптивність та ефективність інтелектуального підходу до керування хмарними ресурсами.

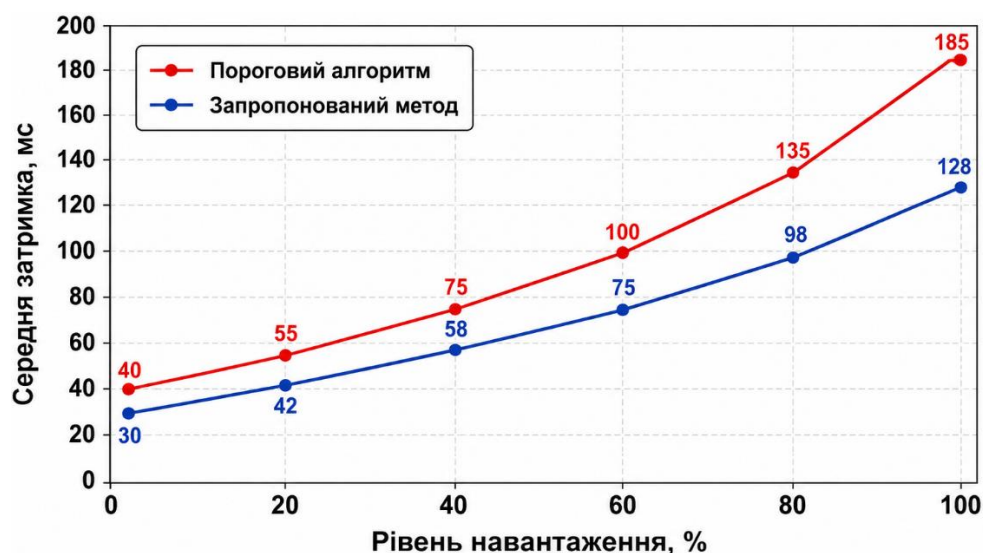


Рис. 4. Порівняння середньої затримки обробки запитів для різних методів масштабування

Отримані результати пояснюються тим, що інтелектуальний агент не лише реагує на перевищення порогових значень, а й враховує поточний стан хмарної інфраструктури, попередній досвід прийняття рішень та очікувану корисність кожної керуючої дії. Це дозволяє уникати запізненого масштабування, зменшувати надлишкове виділення ресурсів і підтримувати стабільну роботу сервісів за умов змінного навантаження.

Для узагальнення особливостей різних підходів до автоматизованого масштабування доцільно провести порівняння їх переваг і обмежень. Порівняльну характеристику методів масштабування наведено у табл. 3.

Таблиця 3

Порівняння переваг і обмежень методів масштабування

Метод масштабування	Переваги	Обмеження
Пороговий алгоритм	Простота реалізації; низькі обчислювальні витрати	Слабка адаптивність; запізнена реакція на зміну навантаження
Прогнозне масштабування	Можливість попереднього планування ресурсів	Залежність від точності прогнозної моделі
Запропонований метод на основі навчання з підкріпленням	Адаптивність; урахування попереднього досвіду; баланс між продуктивністю та витратами	Потребує початкового етапу навчання та налаштування функції винагороди

Проведене порівняння показує, що запропонований метод на основі навчання з підкріпленням забезпечує вищий рівень адаптивності та більш ефективне керування ресурсами порівняно з традиційними підходами [17]. Незважаючи на необхідність попереднього навчання інтелектуального агента, використання такого підходу дозволяє підвищити стабільність функціонування хмарної інфраструктури та зменшити витрати обчислювальних ресурсів у довгостроковій перспективі.

Водночас запропонований метод має певні обмеження, пов'язані з необхідністю початкового етапу навчання інтелектуального агента, вибором параметрів функції винагороди та залежністю результатів від якості вхідних даних моніторингу [11, 21]. У разі недостатньо повного або некоректного збору показників навантаження рішення агента можуть бути менш точними. Тому практичне впровадження методу



потребує налаштування параметрів навчання, перевірки стабільності роботи агента та адаптації функції винагороди до конкретного типу хмарного сервісу.

Додатково встановлено, що запропонований метод забезпечує більш стабільне функціонування системи в умовах різких змін навантаження та дозволяє швидше адаптуватися до зміни інтенсивності запитів користувачів. Отримані результати підтверджують доцільність використання навчання з підкріпленням для побудови інтелектуальних систем адаптивного масштабування хмарної інфраструктури.

Таким чином, отримані результати безпосередньо відповідають поставленій меті статті, оскільки підтверджують можливість побудови методу адаптивного масштабування хмарних ресурсів на основі навчання з підкріпленням. Розроблені модель станів, функція винагороди, механізм вибору керуючих дій та інтегральний критерій ефективності забезпечують цілісну основу для автоматизованого керування ресурсами в умовах змінного навантаження.

Отже, основним науковим результатом дослідження є розроблення методу адаптивного масштабування хмарних ресурсів, який поєднує модель зміни навантаження, формалізований опис станів системи, функцію винагороди та механізм навчання інтелектуального агента [17]. На відміну від класичних порогових алгоритмів, запропонований метод забезпечує не лише реакцію на поточне перевищення навантаження, а й поступове формування оптимальної політики керування ресурсами на основі накопиченого досвіду [27]. Це дозволяє підвищити стабільність функціонування хмарних сервісів, зменшити затримки обробки запитів і скоротити випадки надлишкового виділення обчислювальних ресурсів.

Обговорення. Отримані результати дослідження підтверджують доцільність використання методів навчання з підкріпленням для розв'язання задач адаптивного масштабування хмарних ресурсів у сучасних розподілених інформаційних системах. На відміну від традиційних порогових алгоритмів автоматизованого масштабування, запропонований метод забезпечує формування адаптивної стратегії прийняття рішень з урахуванням поточного стану хмарної інфраструктури, зміни навантаження та накопиченого досвіду функціонування системи. Запропонований підхід також узгоджується з сучасними методами адаптивного розподілу ресурсів у розподілених середовищах, де ефективність функціонування системи залежить від динамічного перерозподілу обчислювальних ресурсів відповідно до поточного рівня навантаження [14].

Результати експериментального дослідження показали, що використання інтелектуального агента дозволяє зменшити середню затримку обробки запитів, підвищити ефективність використання обчислювальних ресурсів та скоротити кількість випадків надлишкового масштабування. Найбільша ефективність запропонованого підходу спостерігалася в умовах різкої зміни навантаження, коли класичні порогові алгоритми не забезпечували достатньої швидкості адаптації до змін стану системи.

Важливою перевагою запропонованого методу є можливість формування політики масштабування на основі взаємодії інтелектуального агента із середовищем без необхідності жорсткого визначення правил керування ресурсами. Використання функції винагороди та механізму оцінювання корисності керуючих дій дозволяє забезпечити баланс між продуктивністю системи, стабільністю функціонування сервісів та витратами обчислювальних ресурсів. Додатково важливим аспектом є ефективність обробки та аналізу великих обсягів даних моніторингу, що впливає на швидкість формування керуючих дій і загальну продуктивність системи масштабування [22].

Разом із тим запропонований підхід має певні обмеження. Ефективність функціонування системи значною мірою залежить від коректності формування функції винагороди, вибору параметрів навчання та якості вхідних даних моніторингу. Крім того, на початкових етапах навчання агент може приймати недостатньо ефективні рішення, що потребує додаткового налаштування параметрів середовища та механізмів навчання. Незважаючи на це, результати дослідження підтверджують перспективність використання навчання з підкріпленням для побудови інтелектуальних систем автоматизованого керування хмарними ресурсами. Отримані результати також підтверджують актуальність використання методів оптимізації розподілу завдань і ресурсів у розподілених та периферійних обчислювальних середовищах, де адаптивне керування навантаженням є одним із ключових факторів забезпечення стабільності функціонування сервісів [28].

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

У результаті виконаного дослідження розв'язано актуальне науково-прикладне завдання розроблення методу адаптивного масштабування хмарних ресурсів на основі навчання з підкріпленням для автоматизованого керування обчислювальними ресурсами в умовах динамічної зміни навантаження.



Проведене дослідження дозволило обґрунтувати доцільність використання інтелектуальних методів прийняття рішень для підвищення ефективності функціонування хмарної інфраструктури.

У межах роботи проведено аналіз сучасних підходів до автоматизованого масштабування хмарних ресурсів та визначено основні недоліки традиційних порогових алгоритмів керування. Розроблено модель адаптивного масштабування, яка включає формалізований опис станів системи, множину керуючих дій, функцію винагороди та механізм навчання інтелектуального агента. Запропоновано інтегральний критерій оцінювання ефективності масштабування та модель зміни навантаження хмарної інфраструктури.

Проведене експериментальне дослідження підтвердило ефективність запропонованого методу. Отримані результати показали зниження затримки обробки запитів, підвищення ефективності використання обчислювальних ресурсів і покращення стабільності функціонування сервісів порівняно з класичними алгоритмами автоматизованого масштабування. Встановлено, що використання навчання з підкріпленням дозволяє забезпечити більш адаптивне та гнучке керування ресурсами хмарної інфраструктури в умовах змінного навантаження.

Перспективи подальших досліджень пов'язані з удосконаленням механізмів навчання інтелектуального агента, використанням глибоких нейронних мереж для прогнозування навантаження, інтеграцією методів багатокритеріальної оптимізації та дослідженням функціонування системи у багатохмарних і контейнеризованих середовищах. Перспективним напрямом є також підвищення енергоефективності адаптивного масштабування та розроблення механізмів автоматизованого керування ресурсами для великих розподілених інформаційних платформ.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Qiu, H., Mao, W., Wang, C., Franke, H., Youssef, A., Kalbarczyk, Z. T., Başar, T., & Iyer, R. K. (2023). AWARE: Automate workload autoscaling with reinforcement learning in production cloud systems. In 2023 USENIX Annual Technical Conference (USENIX ATC 23) (pp. 387–402).
2. Mishra, P., Hans, S., Saha, D., & Moogi, P. (2024). Optimizing cloud workloads: Autoscaling with reinforcement learning. In 2024 IEEE 17th International Conference on Cloud Computing (CLOUD) (pp. 217–222). IEEE. <https://doi.org/10.1109/CLOUD62652.2024.00033>
3. Alharthi, S., Alshamsi, A., Alseiyari, A., & Alwarafy, A. (2024). Auto-scaling techniques in cloud computing: Issues and research directions. *Sensors*, 24(17), 5551. <https://doi.org/10.3390/s24175551>
4. Pintye, I., Kovács, J., & Lovas, R. (2024). Enhancing machine learning-based autoscaling for cloud resource orchestration. *Journal of Grid Computing*, 22, 68. <https://doi.org/10.1007/s10723-024-09783-1>
5. Feng, X., Zhang, S., Jiao, T., Guo, C., & Song, J. (2025). Adaptive container auto-scaling for fluctuating workloads in cloud. *Future Generation Computer Systems*, 172, 107872. <https://doi.org/10.1016/j.future.2025.107872>
6. Moreno-Vozmediano, R., Huedo, E., Montero, R. S., & Llorente, I. M. (2025). AI-driven resource allocation and auto-scaling of VNFs in edge-5G-IoT ecosystems. *Electronics*, 14(9), 1808. <https://doi.org/10.3390/electronics14091808>
7. Alatawi, M. N. (2025). Optimizing multitierancy: Adaptive resource allocation in serverless cloud environments using reinforcement learning. *Electronics*, 14(15), 3004. <https://doi.org/10.3390/electronics14153004>
8. Skladannyi, P., Kostyuk, Y., Sokolov, V., & Khorolska, K. (2026). Network traffic analysis and management. In *Handbook of Cybersecurity Challenges and Solutions for Emerging Technologies* (pp. 148–172). <https://doi.org/10.1201/9781003640790-9>
9. Xu, M., Song, C., Wu, H., Gill, S. S., Ye, K., & Xu, C. (2022). esdnn: Deep neural network based multivariate workload prediction in cloud computing environments. *ACM Transactions on Internet Technology*, 22(3). <https://doi.org/10.1145/3524114>
10. Kostyuk, Y., Skladannyi, P., Sokolov, V., & Rzaieva, S. (2025). Intelligent system for simulation modeling and research of information objects. In *Proceedings of the 1st Workshop Software Engineering and Semantic Technologies (SEST 2025)* (Vol. 4053, pp. 237–251). CEUR Workshop Proceedings.
11. Khan, T., Tian, W., Zhou, G., Ilager, S., Gong, M., & Buyya, R. (2022). Machine learning-centric resource management in cloud computing: A review and future directions. *Journal of Network and Computer Applications*, 204, 103405.
12. Kostyuk, Y., Skladannyi, P., Sokolov, V., & Vorokhob, M. (2025). Models and technologies of cognitive agents for decision-making with integration of artificial intelligence. In *Proceedings of the Modern Data Science Technologies Doctoral Consortium (MoDaST 2025)* (Vol. 4005, pp. 82–96). CEUR Workshop Proceedings.



13. Zhong, Z., Xu, M., Rodríguez, M., Xu, C.-Z., & Buyya, R. (2022). Machine learning-based orchestration of containers: A taxonomy and future directions. *ACM Computing Surveys*, 54. <https://doi.org/10.1145/3510415>
14. Kostiuk, Y., Skladannyi, P., Korshun, N., Bebashko, B., & Khorolska, K. (2024). Integrated protection strategies and adaptive resource distribution for secure video streaming over a Bluetooth network. In *Workshop on Cybersecurity Providing in Information and Telecommunication Systems II (CPITS-II 2024)* (Vol. 3826, pp. 129–138). *CEUR Workshop Proceedings*.
15. Dixit, A., Gupta, R. K., Dubey, A., & Misra, R. (2021). Machine learning based adaptive auto-scaling policy for resource orchestration in Kubernetes clusters. In *International Conference on Internet of Things and Connected Technologies* (pp. 1–16). Springer.
16. Складанний, П. М., Костюк, Ю. В., Мазур, Н. П., & Пітайчук, М. А. (2025). Дослідження характеристик та продуктивності протоколів доступу до хмарних обчислювальних середовищ на основі універсального тестування. *Телекомунікаційні та інформаційні технології*, 1(86), 61–74. <https://doi.org/10.31673/2412-4338.2025.014649>
17. Garí, Y., Monge, D. A., Pacini, E., Mateos, C., & Garino, C. G. (2021). Reinforcement learning-based application autoscaling in the cloud: A survey. *Engineering Applications of Artificial Intelligence*, 102, 104288.
18. Костюк, Ю., Складанний, П., Рзаєва, С., Самойленко, Ю., & Коршун, Н. (2025). Інтелектуальні системи керування та захисту в кіберфізичних і хмарних середовищах Smart Grid. *Кібербезпека: освіта, наука, техніка*, 2(30), 125–156. <https://doi.org/10.28925/2663-4023.2025.30.956>
19. Benedetti, P., Femminella, M., Reali, G., & Steenhaut, K. (2022). Reinforcement learning applicability for resource-based auto-scaling in serverless edge applications. In *2022 IEEE International Conference on Pervasive Computing and Communications Workshops* (pp. 674–679). IEEE.
20. Rzaeva, S., Skladannyi, P., Kostiuk, Y., Abramov, V., & Kravchenko, V. (2025). Adaptive information security management in cloud-oriented intelligent transportation systems. *Ukrainian Scientific Journal of Information Security*, 31(1), 23–36. <https://doi.org/10.18372/2225-5036.31.20634>
21. Schuler, L., Jamil, S., & Kühl, N. (2021). AI-based resource allocation: Reinforcement learning for adaptive auto-scaling in serverless environments. In *2021 IEEE/ACM 21st International Symposium on Cluster, Cloud and Internet Computing (CCGrid)* (pp. 804–811). IEEE.
22. Складанний, П., Костюк, Ю., Рзаєва, С., & Мазур, Н. (2025). Паралельна обробка даних у розширюваних хеш-структурах та оцінка їх продуктивності. *Кібербезпека: освіта, наука, техніка*, 3(31), 242–269. <https://doi.org/10.28925/2663-4023.2025.31.1015>
23. Atadoga, A., Umoga, U. J., Lottu, O. A., & Sodiya, E. O. (2024). Evaluating the impact of cloud computing on accounting firms: A review of efficiency, scalability, and data security. *Global Journal of Engineering and Technology Advances*, 18(2), 065–074.
24. Довженко, Н., Мазур, Н., Костюк, Ю., & Рзаєва, С. (2024). Інтеграція ІОТ та штучного інтелекту в інтелектуальні транспортні системи. *Кібербезпека: освіта, наука, техніка*, 2(26), 430–444. <https://doi.org/10.28925/2663-4023.2024.26.708>
25. Patel, Y. S., & Bedi, J. (2023). Mag-d: A multivariate attention network based approach for cloud workload forecasting. *Future Generation Computer Systems*, 142, 376–392.
26. Kostiuk, Y., Skladannyi, P., Samoilenko, Y., Khorolska, K., Bebashko, B., & Sokolov, V. (2024). A system for assessing the interdependencies of information system agents in information security risk management using cognitive maps. In *Third International Conference on Cyber Hygiene & Conflict Management in Global Information Networks (CH&CMiGIN'24)* (Vol. 3925, pp. 249–264). *CEUR Workshop Proceedings*.
27. Chen, X., Zhu, F., Chen, Z., Min, G., Zheng, X., & Rong, C. (2022). Resource allocation for cloud-based software services using prediction-enabled feedback control with reinforcement learning. *IEEE Transactions on Cloud Computing*, 10(2), 1117–1129. <https://doi.org/10.1109/TCC.2020.2992537>
28. Patsias, V., Amanatidis, P., Karampatzakis, D., Lagkas, T., Michalakopoulou, K., & Nikitas, A. (2023). Task allocation methods and optimization techniques in edge computing: A systematic review of the literature. *Future Internet*, 15(8), 254.

**Andrii Pazynin**

Postgraduate student,

Institute of Telecommunications and Global Information

Space of the National Academy of Sciences of Ukraine, Kyiv, Ukraine

ORCID: 0009-0002-9506-9539

pazinin@gmail.com

METHOD OF ADAPTIVE SCALING OF CLOUD RESOURCES BASED ON REINFORCEMENT LEARNING

Abstract. The article considers the problem of adaptive scaling of cloud infrastructure resources under conditions of dynamic workload changes, uneven utilization of computational resources, and the necessity to ensure stable operation of information services. It has been determined that traditional approaches to horizontal and vertical scaling are mainly based on static threshold rules or predictive models that do not provide sufficient flexibility in a changing environment and do not take into account the multicriteria nature of the resource management process. The feasibility of using reinforcement learning methods for the formation of an adaptive cloud infrastructure management strategy is substantiated. The paper analyzes modern methods of automated resource scaling, decision-making models in cloud systems, and approaches to the use of reinforcement learning methods in workload optimization tasks. A method for adaptive scaling of cloud resources is proposed, based on the use of an intelligent agent that evaluates the current state of the infrastructure, analyzes workload parameters, request processing delays, processor resource utilization, and memory usage, and forms an optimal scaling strategy. A state model, reward system, and mechanism for selecting control actions for automated resource allocation have been developed. Within the experimental study, the proposed method was compared with classical automated scaling algorithms. The obtained results confirmed a reduction in request processing delays, improved resource utilization efficiency, and enhanced stability of cloud system operation. The practical significance of the work lies in the possibility of applying the proposed method for automated resource management in large distributed information systems and modern services.

Keywords: cloud infrastructure; adaptive scaling; reinforcement learning; cloud resources; automated management; resource optimization; intelligent systems; distributed computing; workload scaling; decision-making.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Qiu, H., Mao, W., Wang, C., Franke, H., Youssef, A., Kalbarczyk, Z. T., Başar, T., & Iyer, R. K. (2023). AWARE: Automate workload autoscaling with reinforcement learning in production cloud systems. In Proceedings of the 2023 USENIX Annual Technical Conference (USENIX ATC 23) (pp. 387–402).
2. Mishra, P., Hans, S., Saha, D., & Moogi, P. (2024). Optimizing cloud workloads: Autoscaling with reinforcement learning. In Proceedings of the 2024 IEEE 17th International Conference on Cloud Computing (CLOUD) (pp. 217–222). IEEE. <https://doi.org/10.1109/CLOUD62652.2024.00033>
3. Alharthi, S., Alshamsi, A., Alseiri, A., & Alwarafy, A. (2024). Auto-scaling techniques in cloud computing: Issues and research directions. *Sensors*, 24(17), 5551. <https://doi.org/10.3390/s24175551>
4. Pintye, I., Kovács, J., & Lovas, R. (2024). Enhancing machine learning-based autoscaling for cloud resource orchestration. *Journal of Grid Computing*, 22, 68. <https://doi.org/10.1007/s10723-024-09783-1>
5. Feng, X., Zhang, S., Jiao, T., Guo, C., & Song, J. (2025). Adaptive container auto-scaling for fluctuating workloads in cloud environments. *Future Generation Computer Systems*, 172, 107872. <https://doi.org/10.1016/j.future.2025.107872>
6. Moreno-Vozmediano, R., Huedo, E., Montero, R. S., & Llorente, I. M. (2025). AI-driven resource allocation and auto-scaling of VNFs in edge-5G-IoT ecosystems. *Electronics*, 14(9), 1808. <https://doi.org/10.3390/electronics14091808>
7. Alatawi, M. N. (2025). Optimizing multitierancy: Adaptive resource allocation in serverless cloud environments using reinforcement learning. *Electronics*, 14(15), 3004. <https://doi.org/10.3390/electronics14153004>
8. Skladannyi, P., Kostyuk, Y., Sokolov, V., & Khorolska, K. (2026). Network traffic analysis and management. In *Handbook of Cybersecurity Challenges and Solutions for Emerging Technologies* (pp. 148–172). CRC Press. <https://doi.org/10.1201/9781003640790-9>



9. Xu, M., Song, C., Wu, H., Gill, S. S., Ye, K., & Xu, C. (2022). esdnn: Deep neural network based multivariate workload prediction in cloud computing environments. *ACM Transactions on Internet Technology*, 22(3). <https://doi.org/10.1145/3524114>
10. Kostiuk, Y., Skladannyi, P., Sokolov, V., & Rzaieva, S. (2025). Intelligent system for simulation modeling and research of information objects. In *Proceedings of the 1st Workshop Software Engineering and Semantic Technologies (SEST 2025)* (Vol. 4053, pp. 237–251). CEUR Workshop Proceedings.
11. Khan, T., Tian, W., Zhou, G., Ilager, S., Gong, M., & Buyya, R. (2022). Machine learning-centric resource management in cloud computing: A review and future directions. *Journal of Network and Computer Applications*, 204, 103405.
12. Kostiuk, Y., Skladannyi, P., Sokolov, V., & Vorokhob, M. (2025). Models and technologies of cognitive agents for decision-making with integration of artificial intelligence. In *Proceedings of the Modern Data Science Technologies Doctoral Consortium (MoDaST 2025)* (Vol. 4005, pp. 82–96). CEUR Workshop Proceedings.
13. Zhong, Z., Xu, M., Rodríguez, M., Xu, C.-Z., & Buyya, R. (2022). Machine learning-based orchestration of containers: A taxonomy and future directions. *ACM Computing Surveys*, 54. <https://doi.org/10.1145/3510415>
14. Kostiuk, Y., Skladannyi, P., Korshun, N., Bebesko, B., & Khorolska, K. (2024). Integrated protection strategies and adaptive resource distribution for secure video streaming over a Bluetooth network. In *Proceedings of the Workshop on Cybersecurity Providing in Information and Telecommunication Systems II (CPITS-II 2024)* (Vol. 3826, pp. 129–138). CEUR Workshop Proceedings.
15. Dixit, A., Gupta, R. K., Dubey, A., & Misra, R. (2021). Machine learning based adaptive auto-scaling policy for resource orchestration in Kubernetes clusters. In *Proceedings of the International Conference on Internet of Things and Connected Technologies* (pp. 1–16). Springer.
16. Skladannyi, P. M., Kostiuk, Y. V., Mazur, N. P., & Pitaichuk, M. A. (2025). Study of characteristics and performance of access protocols to cloud computing environments based on universal testing. *Telecommunication and Information Technologies*, 1(86), 61–74. <https://doi.org/10.31673/2412-4338.2025.014649>
17. Garí, Y., Monge, D. A., Pacini, E., Mateos, C., & Garino, C. G. (2021). Reinforcement learning-based application autoscaling in the cloud: A survey. *Engineering Applications of Artificial Intelligence*, 102, 104288.
18. Kostiuk, Y., Skladannyi, P., Rzaieva, S., Samoilenko, Y., & Korshun, N. (2025). Intelligent control and protection systems in cyber-physical and cloud Smart Grid environments. *Cybersecurity: Education, Science, Technique*, 2(30), 125–156. <https://doi.org/10.28925/2663-4023.2025.30.956>
19. Benedetti, P., Femminella, M., Reali, G., & Steenhaut, K. (2022). Reinforcement learning applicability for resource-based auto-scaling in serverless edge applications. In *Proceedings of the 2022 IEEE International Conference on Pervasive Computing and Communications Workshops* (pp. 674–679). IEEE.
20. Rzaeva, S., Skladannyi, P., Kostiuk, Y., Abramov, V., & Kravchenko, V. (2025). Adaptive information security management in cloud-oriented intelligent transportation systems. *Ukrainian Scientific Journal of Information Security*, 31(1), 23–36. <https://doi.org/10.18372/2225-5036.31.20634>
21. Schuler, L., Jamil, S., & Kühl, N. (2021). AI-based resource allocation: Reinforcement learning for adaptive auto-scaling in serverless environments. In *Proceedings of the 2021 IEEE/ACM 21st International Symposium on Cluster, Cloud and Internet Computing (CCGrid)* (pp. 804–811). IEEE.
22. Skladannyi, P., Kostiuk, Y., Rzaieva, S., & Mazur, N. (2025). Parallel data processing in extensible hash structures and performance evaluation. *Cybersecurity: Education, Science, Technique*, 3(31), 242–269. <https://doi.org/10.28925/2663-4023.2025.31.1015>
23. Atadoga, A., Umoga, U. J., Lottu, O. A., & Sodiya, E. O. (2024). Evaluating the impact of cloud computing on accounting firms: A review of efficiency, scalability, and data security. *Global Journal of Engineering and Technology Advances*, 18(2), 065–074.
24. Dovzhenko, N., Mazur, N., Kostiuk, Y., & Rzaieva, S. (2024). Integration of IoT and artificial intelligence into intelligent transportation systems. *Cybersecurity: Education, Science, Technique*, 2(26), 430–444. <https://doi.org/10.28925/2663-4023.2024.26.708>
25. Patel, Y. S., & Bedi, J. (2023). Mag-d: A multivariate attention network based approach for cloud workload forecasting. *Future Generation Computer Systems*, 142, 376–392.
26. Kostiuk, Y., Skladannyi, P., Samoilenko, Y., Khorolska, K., Bebesko, B., & Sokolov, V. (2024). A system for assessing the interdependencies of information system agents in information security risk management using cognitive maps. In *Proceedings of the Third International Conference on Cyber Hygiene & Conflict Management in Global Information Networks (CH&CMiGIN'24)* (Vol. 3925, pp. 249–264). CEUR Workshop Proceedings.



27. Chen, X., Zhu, F., Chen, Z., Min, G., Zheng, X., & Rong, C. (2022). Resource allocation for cloud-based software services using prediction-enabled feedback control with reinforcement learning. *IEEE Transactions on Cloud Computing*, 10(2), 1117–1129. <https://doi.org/10.1109/TCC.2020.2992537>
28. Patsias, V., Amanatidis, P., Karampatzakis, D., Lagkas, T., Michalakopoulou, K., & Nikitas, A. (2023). Task allocation methods and optimization techniques in edge computing: A systematic review of the literature. *Future Internet*, 15(8), 254.

Отримано редакцією журналу / Received: 28.05.26

Прорецензовано / Revised: 16.06.26

Схвалено до друку / Accepted: 26.10.26



This work is licensed under Creative Commons Attribution-noncommercial-sharealike 4.0 International License.