УДК 004.9

AI-ПІДХОДИ ДО ОПТИМІЗАЦІЇ РЕСУРСІВ У ДИНАМІЧНИХ ХМАРНИХ ІНФРМАСТРУКТУРАХ

Роман Дашківський, Андрій Копп

AI-APPROACHES TO RESOURCE OPTIMIZATION IN DYNAMIC CLOUD INFRASTRUCTURES

Roman Dashkivskyy, Andrii Kopp

Національний технічний університет «Харківський політехнічний інститут», вулиця Кирпичова, 2, Харків, 61000, Україна

***Abstract**. This paper explores modern artificial intelligence (AI)-based approaches to resource optimization in cloud computing environments. It focuses on the use of deep reinforcement learning, evolutionary algorithms, and federated learning for managing dynamic workloads, enhancing energy efficiency, and ensuring cybersecurity. Recent studies demonstrating the advantages of hybrid AI systems are reviewed and generalized. The paper also emphasizes the importance of integrating such approaches into multi-tenant cloud infrastructures, taking into account platform-specific constraints and emerging threat vectors. Special attention is paid to data confidentiality, fault tolerance, and compliance with international security standards.*

Хмарні обчислення сьогодні стали критично важливими для цифрової трансформації в бізнесі, науці, медицині, транспорті та державному управлінні. Їх гнучкість, масштабованість і економічність дозволяють розгортати високонавантажені сервіси майже в реальному часі. Проте збільшення обсягів оброблюваних даних і запитів призводить до нових викликів — від нестабільного часу відгуку до перевитрати енергії та ризиків безпеки. Класичні методи управління ресурсами часто не встигають адаптуватись до цих змін. Штучний інтелект (ШІ), зокрема його адаптивні, самонавчальні та прогнозні властивості, виявляється ефективним інструментом для вирішення цих проблем. Саме тому дослідження, спрямовані на впровадження AI-підходів у системи розподілу обчислювальних ресурсів, мають надзвичайну прикладну цінність.

Глибоке підкріплювальне навчання (DRL) формується на основі взаємодії з середовищем, де модель вчиться обирати оптимальні дії на основі зворотного зв’язку. У хмарних інфраструктурах це дозволяє оперативно приймати рішення щодо розподілу обчислювальних задач, зокрема вибирати між локальною обробкою чи передачею на сервер. Використання DRL у мобільних edge-системах дає змогу зменшити енергоспоживання до 40% та водночас покращити якість досвіду користувача на 35% [1].

Еволюційні алгоритми (ЕА) показують ефективність у глобальній оптимізації, оскільки дозволяють моделювати широкий простір рішень і поступово шукати найбільш ефективну конфігурацію параметрів. Їхня здатність адаптуватися до складної топології хмарної інфраструктури та до зміни запитів у часі робить їх корисним інструментом, зокрема у багатосерверних середовищах.

Федеративне навчання (FL), у свою чергу, дає можливість тренувати моделі на розподілених даних без необхідності їх централізованого збирання. Це не лише зменшує навантаження на мережу, але й забезпечує вищий рівень конфіденційності. Системи, що поєднують FL з постквантовими алгоритмами шифрування, як-от Kyber-1024, та з доказами з нульовим розголошенням (zk-SNARK), можуть досягати точності до 99,2% і знижувати ризик витоку інформації на понад 60% [2].

Окрему увагу заслуговують підходи, що поєднують методи скорочення порядку моделі з AI-модулями. Такі рішення дають змогу знизити обчислювальну складність системи без втрати продуктивності. Наприклад, застосування Model Order Reduction у розподілі задач забезпечує скорочення часу відгуку системи майже вдвічі, що важливо для критичних сценаріїв обробки даних у реальному часі [3].

Попри високий потенціал AI-алгоритмів у хмарних системах, існує низка викликів, які слід враховувати під час їхньої реалізації. Одним з ключових є динамічність навантаження. Запити користувачів, трафік і ресурсні потреби постійно змінюються, що вимагає від системи здатності швидко адаптуватися. DRL-алгоритми демонструють гнучкість у таких умовах, проте залежать від часу навчання та обчислювальних потужностей.

Ще однією важливою проблемою є безпека. Централізовані системи наражаються на ризики атак на модель (наприклад, через методи інверсії або отруєння даних), що може призвести до витоку конфіденційної інформації. Поєднання моделей без нагляду з зовнішніми джерелами загроз, зокрема індикаторами компрометації (IOC) та знанням про тактики й техніки зловмисників (TTP), дозволяє істотно підвищити точність виявлення атак та зменшити кількість хибних спрацювань [4].

Також слід враховувати архітектурні обмеження існуючих хмарних платформ. Наприклад, у середовищах на зразок Salesforce важливо адаптувати AI-моделі до політик API, багатокористувацької архітектури, розподілу пропускної здатності та затримок. Прогнозна аналітика в поєднанні з динамічним масштабуванням показує високу ефективність у таких умовах, забезпечуючи рівномірне навантаження і зменшення затримок [5].

Таким чином AI-підходи до оптимізації ресурсів у хмарних інфраструктурах є важливим напрямом сучасних досліджень. Вони відкривають нові можливості для адаптивного управління навантаженням, зниження енерговитрат та покращення стабільності обслуговування. Найбільший ефект спостерігається при використанні гібридних архітектур, де поєднано DRL, FL та EA, що забезпечує баланс між адаптивністю, безпекою та ефективністю. Однак широке впровадження таких підходів вимагає вирішення проблем конфіденційності, сумісності та забезпечення стабільної роботи в реальних умовах. Подальші дослідження мають зосередитися на масштабуванні таких систем у виробничих середовищах, оцінці їхньої відмовостійкості та відповідності стандартам інформаційної безпеки.

Бібліографія

1. Nishad D. K. et al. Adaptive AI-enhanced computation offloading with machine learning for QoE optimization and energy-efficient mobile edge systems //Scientific Reports. – 2025. – Т. 15. – №. 1. – С. 15263.
2. Mohammed A. et al. Advancing AI-Cloud Integration: Comparative Analysis of Algorithms and Novel Solutions //International Journal of Innovative Science and Research Technology. – 2025. – Т. 10. – №. 4. – С. 1555-1560.
3. Chaudhary H. et al. Advanced queueing and scheduling techniques in cloud computing using AI-based model order reduction //Discover Computing. – 2025. – Т. 28. – №. 1. – С. 1-40.
4. Whitman, J., El-Karim, A., Priya Nandakumar, Ortega, F., & William, E. Threat Intelligence Integration with Cloud-Based Anomaly Detectors. 2024. 14 November. URL: https://www.researchgate.net/publication/391497472_Threat_Intelligence_Integration_with_Cloud-Based_Anomaly_Detectors (date of access: 14.05.2025).
5. Hanan Lutfiyya, & Researcher Scholar V. A framework for optimizing resource allocation in AI-powered Cloud platforms under salesforce deployment... ResearchGate, P. 1–7. URL: https://www.researchgate.net/publication/391629185_a_framework_for_optimizing_resource_allocation_in_ai-powered_cloud_platforms_under_salesforce_deployment_constraints (date of access: 15.05.2025).